Model-based human gait tracking, 3D reconstruction and recognition in uncalibrated monocular video

E Adeli-Mosabbeb*[a], M Fathy[a] and **F Zargari[b]**
[a]Department of Computer Engineering, Iran University of Science and Technology, Tehran 1684613114, Iran

[b]Multimedia Systems Group, Department of Information Technology, Iran Telecommunication Research Center, Tehran 1439955471, Iran

Abstract: Automatic analysis of human motion includes initialisation, tracking, pose recovery and activity recognition. In this paper, a computing framework is developed to automatically analyse human motions through uncalibrated monocular video sequences. A model-based kinematics approach is proposed for human gait tracking. Based on the tracking results, 3D human poses and gait features are recovered and extracted. The recognition performance is evaluated by using different classifiers. The proposed method is advantageous in its capability of recognising human subjects walking non-parallel to the image plane.

Keywords: computer vision, gait recognition, human motion analysis, human tracking, support vector machines

1 INTRODUCTION

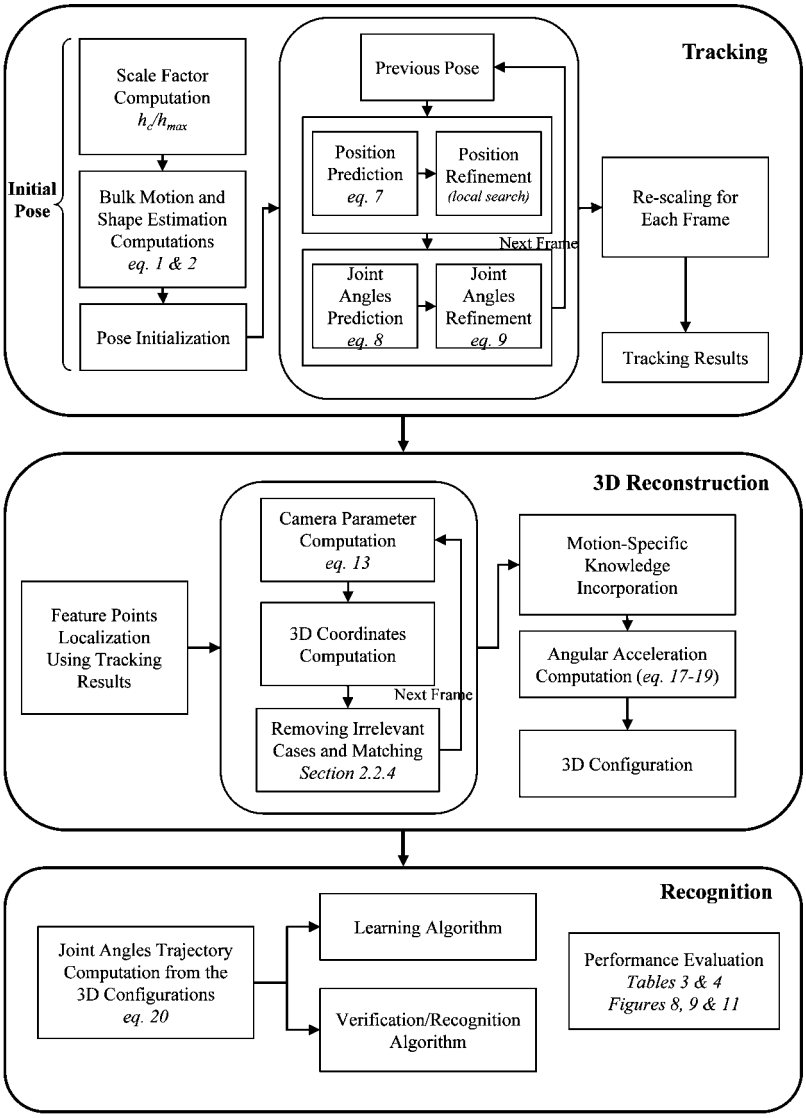
Many people know their relatives form their gait. In this paper, the possibility of recognising people by the means of their gait using computer vision techniques is studies. Human model-based tracking gives a suitable human gait model in a sequence of frames, after which the human 3D configuration could be extracted to be used in a robust recognition process, free from many constraints and limitations considered in previous works.

Object tracking in video sequences is one of the most basic video processing steps in the process of extracting objects' dynamic behaviours. Tracking results can be used to interpret object dynamics via mathematical analysis. Two approaches for tracking have been generally used: model-based and motion-based. Model-based methods provide an insightful interpretation of images at the cost of computational complexity out of geometrical transforms. Motion-based methods, by contrast, can reduce the computational cost without understanding tracked regions, and hence cannot deal with complex or nuanced tracking scenarios.

The complicacy of human tracking depends on different sources, some of which are due to the non-rigid structure of the body. Human body includes many joints, and each body part can move in a wide range around its corresponding joint. As a result, human motion involves a large number of degrees of freedom and frequent self-occlusions of body parts. Another source of complication is derived from the sensors, clothes and background. Sensors of low quality, cluttered background, changing brightness and monotone clothes often add inaccuracy to the motion segmentation.[1,2] Considering all these issues in a real-world system using current technologies is likely impossible. This is why in most of the previous works, some assumptions on walking direction and image acquisition are made to cope with a simplifier problem.[3]

In many researches in the filed of computer vision working on model-based gait tracking, recognition

The MS was accepted for publication on 24 February 2011.
* *Corresponding author: Ehsan Adeli-Mosabbeb, Department of Computer Engineering, Iran University of Science and Technology, Tehran 1684613114, Iran; email: eadeli@iust.ac.ir*



1 The work flow of the proposed algorithm

and human motion analysis (HMA), it is assumed that the subject walks fronto-parallel relative to the image plane. In this paper, simple considerations such as scaling and shape estimation via bulk motion computation are used to remove these restrictions. Figure 1 illustrates the flow diagram of the proposed approach. As could be seen, the proposed algorithm includes three main stages, human tracking, 3D reconstruction of the subject's walking figure and gait recognition. The tracking process uses a model-based approach to track the individual body parts and recover the joint angles in the video sequence. The tracking results, joint angles, could be utilised for various analyses on the human movement patterns. But in this work, the model is projected onto the image plane and positions of the joints are determined for each subject's figure in any frame. In order to design a robust, reliable method to confront the problems of human walking direction and occlusion, a suitable tracking procedure is required. The tracking phase uses both regional and boundary information to match a model on the human body in each frame. This algorithm employs scaling and shape estimation procedures to make the tracking independent from the walking direction. After the tracking phase, using the extracted 2D models and the joint angle positions for each frame along with the orthographic projection, together with the knowledge of human gait, the configuration of the body can be deduced in a 3D space. The algorithm uses the monocular video sequence and the angular velocity of the joints to reconstruct the joints positions in the 3D space. These

3D body configurations could be incorporated in the gait recognition feature extraction phase. The features are fed into some machine learning algorithms to form a complete gait recognition system.

There are two types of gait features, static and dynamic.[15] Static features[48] are those which could describe the appearance of the human body and the dynamic features[2] summarise the dynamics of the gait motion. In general, the results of recognition are more robust when using the dynamic features, because they show the process of walking in more details. In this work, a 3D reconstruction algorithm is used to recover the dynamic gait features of a subject not walking parallel to the image plane.

Generally, it is very hard to recover gait features from video sequences of a walking human. Therefore, in the previous several restrictions are imposed on the imaging technologies and/or the subject's way of walking. From a technological point of view for 3D reconstruction, some works assume that they have the camera parameters or they use multiple cameras to be able to calculate the 3D coordinates.[3,8,15,35] And in order to be able to calculate the gait features, most of the previous works on gait recognition assume that the subject is walking fronto-parallel to the image plane.[2,3,25,48] In this work, a 3D reconstruction algorithm is proposed to extract dynamic gait features out of monocular uncalibrated video sequences and the main contribution is that the proposed gait recognition algorithm is not sensitive to the subject's walking direction and is path-independent.

The rest of this paper is organised as what follows: next section reviews some of the previous works on human tracking and recognition. Section 3 introduces the proposed tracking, 3D reconstruction and recognition algorithms. Section 4 describes the experimental results followed by the conclusions and directions for future works in Section 5.

2 RELATED WORKS

According to great number of applications and inherent computational complexity, HMA has been of great interest. In order to analyse the behaviour of moving objects in video sequences, we need to first track them throughout the sequence.

2.1 Tracking

Tracking is the process of repeatedly predicting and refining the sequence of frames. Tracking approaches could be divided to four classes:[3] (1) region-based tracking; (2) feature-based tracking; (3) model-based tracking; and (4) active-contour-based tracking.

In region-based tracking, the region resulting from the motion segmentation process is used as the key feature to be tracked.[4,5] Region-based tracking approaches are likely fast, but in comparisons with the other approaches, they do not provide good means of dealing with occlusion. On the other hand, using these approaches does not offer well-defined situations for HMA, as far as they do not give meaningful information about the within region elements.

In feature-based tracking approaches, some features of the moving object are chosen as the key feature to be tracked, such as straight lines, angles, scale invariant feature transforms, etc.[6] After extracting the features, the extracted features are matched frame by frame by minimising a cost function. An advantage of these methods is that when partial occlusion occurs, some of the features are still present and this gives a great opportunity to deal with the occlusion. Mehmood *et al.*[7] proposed a fast compressed domain motion analysis for an object tracking in the surveillance videos. Motion vectors are counted as features to retrieve the object's trajectory from standard H.264/MPEG-4 videos. Feature-based tracking could simply solve the problem of occlusion and retains a good run time. Selecting a set of features, not sensitive to rotation and magnification, for tracking is a challenge. Since this approach does not give enough information about the moving object's geometry and region, it is not useful for human motion and behaviour analysis.

In model-based tracking, a 2D or 3D model is assumed priority. In each frame, regarding the tracking history, an initial pose is estimated. After that a pose evaluation function (PEF) calculates the measure of similarity of the model projected on the image plane and the original subject. A search strategy recursively finds an optimal state for the model.[8] In a similarly hybrid approach, Thome *et al.*[9] proposed a method to construct a 2D human appearance model. They used the structural properties of the articulated human body in a general region-based tracking approach to track people in real-time in complex situations.

In active-contour-based tracking, each object is represented by its surrounding contours. These contours update in each frame dynamically. The

aim to this approach is to obtain the shape of the moving objects, and as a result to provide better descriptions for that object. In active-contour-based tracking, the goal is to minimise an energy function. For each frame, the contour from previous frame is projected onto the current frame image plane and then refined to minimise the energy function. As an example, Paragios and Deriche[10] proposed a technique to extract the active contour by finding the minimum distance between pairs of points to detect the moving object in the sequence of frames. They have used a density function of the object boundaries in two consecutive frames to refine the contour. Bertalmío *et al.*[11] proposed a partial differential equation to reform the contour in the first image to the one in the second. In another work, Erdem[12] proposed an active-contour-based algorithm which uses region data. He used the colour histogram difference of the tracked object between the reference and the current frames. For the scenes that there is an out-of-focus blur difference between the object and the background, defocus energy is also used. As far as active-contour-based approaches provide enough information about the person's body pose and geometrical state, it would be a good approach for problems of HMA and tracking. These approaches also solve the occlusion problem simply. The only drawback could be their high computational complexity, although in comparison with the model-based tracking approaches they perform better.

2.2 Recognition

Human activity representation and recognition has not been a long-time research issue. Its main goal is to recognise human motion using scene description and computer vision techniques. These methods generally seek the extraction and recognition of some information such as gender, identity and activity in a sequence of frames. Such processes can work as an important part in any surveillance system.

Cunado *et al.*[13] gave an overview of different gait recognition approaches and proposed a new approach to gait recognition which automatically extracts and describes human gait for recognition. It is done via the use of Fourier series to describe the motion of the upper leg and apply temporal evidence gathering techniques to extract the moving model from a sequence of images. As one of the initial works on gait recognition, Niyogi and Adelson[46] extracted a signature from the space–time pattern of a person's gait. In XT dimensions (translation–time), head and legs have different patterns. These patterns are processed to extract boundary contour motions. Little and Boyd[14] aimed to determine which part of motion is more significant in the process of recognition. While this work presents that a model-based approach for recognising human from their gait is not compulsory, it states that classification using simple gait and human motion features fails when the database grows.

One of the best works in this field is done by Green and Guan.[15,16] In Ref. 15, they have presented a continuous human movement recognition framework. In this framework, the human motion edges and texture regions were tracked using a 3D colour clone-body-model. The tracking process also included an estimation of the joint angles, using the well-known particle filter. The human movement recognition was implemented by the use of multiple hidden Markov models. They have used this continuous human movement recognition in Ref. 16 to demonstrate applications of HMA systems to the biometric authentication of gait, anthropometric data, human activities and movement disorders. This is done by the use of human body propositions and dimensions and by utilising the concept of EigenGait.

Gait recognition methods could also be divided into two major classes: model-based and motion-based. The former extracts 2D or 3D models of the human body and uses the model to extract human gait dynamic features for recognition. Of these approaches one can refer to Ref. 17 as one of the most perfect ones. It uses an anatomical data to generate shape models consistent with normal human body proportions. To extract gait signature, it estimates the articulated motion using mean joint rotation patterns, then adapts a model by Gaussian addition. The authors have tested the algorithm on a large gait database and have acquired an 84% correct classification rate. References 2 and 18 are other samples of model-based gait recognition, in which approaches to recover joint angles trajectory during fronto-parallel walking of the subjects are proposed.

The second class of gait recognition approaches uses statistical features extracted from the silhouette and the gait motion. References 19–22 are examples of these approaches. BenAbdelkader *et al.*[20] used image self-similarity plots to propose an EigenGait method, Collins *et al.*[21] described a method for

template matching of body silhouettes in key frames during a walk cycle for human identification and Phillips *et al.*[22] considered spatial–temporal silhouette correlation to propose a baseline algorithm for gait identification.

Generally, motion-based approaches result in simpler algorithms, easier to implement with lower computational complexity, whereas model-based approaches give more robust systems in exchange for the high computational complexity.

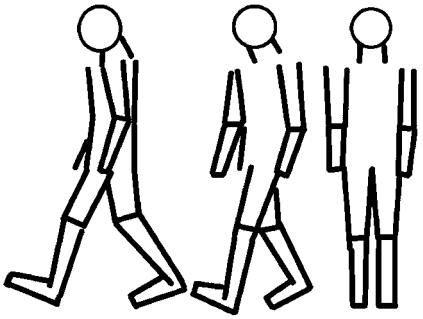
3 PROPOSED APPROACH

The proposed approach, as illustrated in Fig. 1, includes a model-based tracking process of the detected human in the video sequence, a 3D reconstruction algorithm performed on the results of the tracking process and finally a recognition strategy. This section describes these processes one after the other.

Path-invariant gait recognition is a challenging problem due to the nature of the 2D video images. To overcome this issue, only a simple robust 3D configuration of the gait and its trajectory could help. The feature extraction for gait recognition plays an important role in the quality of the classification sub-systems and as a result in the final correct recognition rate. In this paper, dynamic features of the human gait are used, which have also shown good results in previous works. Joint angles and their changes over time could properly model the human gait's dynamicity. That is why some lower body limbs' joint angles are used for our purpose, which is explained in details in the Section 4.

3.1 Tracking

A process of tracking is composed of two basic stages: pose prediction and refinement. Prediction is usually done via a dynamic model, which models the human body in the image and is obtained using the previous information. In the refinement stage, the acquired model is modified such that it fits the human body. As far as the search and the optimisation in the human body configurations space are very complex and time consuming, this stage has always been greatly important. In this paper, a kinematics method[8,23] is utilised for the refinement stage, which uses the physical forces for each body element. This method does not necessarily need differentiable PEFs unlike Taylor models[24] and has less computational complexity compared with the stochastic sampling method.



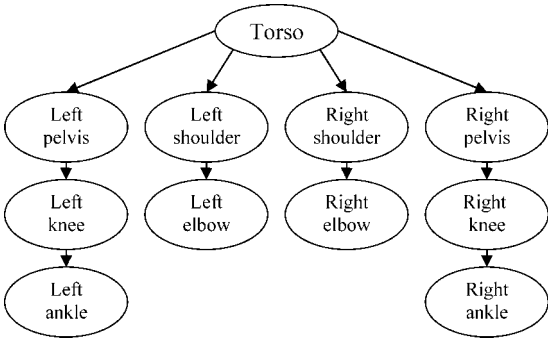
2 The human body model, projected on the image plane

For the cases that the person is not walking parallel to the image plane, the silhouette size in the video sequence goes smaller or bigger during the video sequence, depending on the distance that the subject has from the camera in each frame. To execute the tracking algorithm on the walking human, simply we scale the person to a same size in each frame. In order to acquire a robust initial model of the walking human, the person's shape is estimated using bulk motion. The tracking process is started as follows: in each frame, using the information and the model from the previous frame, the position is first estimated and then refined. The refinement stage is composed of searching the space of different body poses and minimising a PEF. According to the body kinematics structure, it could be modelled as a tree and a hierarchical search strategy simply could be developed. A dynamic model is introduced here for pose and joint angle trajectory refinement using physical forces, like in Refs. 2, 8 and 23.

3.1.1 Human body model

The human body model used includes 14 rigid parts: upper torso, lower torso, neck, two arms, two forearms, two thighs, two legs, two foots and a head. Each body part is represented by two semi-parallel cone-shaped lines, except the head which is modelled by a circle. These parts are connected to one another using the joints and as a result they form angles. These angles are modelled by utilising Euler angles. Figure 2 shows the sample model used in this paper.

For each joint one degree of freedom is assumed, which reduces the problem dimensionality to 12: 10 for the joints and two for the position. These 10 joint angles include the joint angles of the two shoulders,

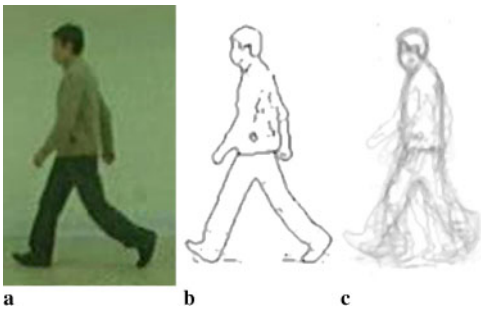


3 A schematic of the kinematic tree used for modelling the human model

two elbows, two pelvises, two knees and two ankles. Therefore, a vector $P=\{x, y, \theta_1, \theta_2, \ldots, \theta_{10}\}$, could simply represent a pose, where (x,y) is the body position and θ_i is the ith joint angle. The aim of the tracking process is to recover this vector, frame by frame. The model is projected onto the image plane, and the coordinates of each point on the model are calculated and then translated to the camera coordinates. The human model is treated as a kinematic tree, the torso is its root and each body part has its own local coordinates (Fig. 3). The (x,y) coordinates in the pose vector P are the base coordinates of the torso and are almost located at the centre of the body. The kinematical modelling helps find a path $s_i s_{i-1} \ldots s_0$ for each body part s_i in the kinematics tree, which starts with s_i and ends to the root s_0.

3.1.2 *Model initialisation*

Throughout the video sequence, the tracking process predicts the model for the current frame regarding the existing model from the previous frame, and then refines it. This requires the tracking results from previous frames. At the beginning of the tracking process, such a previous model does not exist and an initialisation is needed. In many previous works, the initialisation is done manually.[25,26] In another work, Ning *et al.*[2] have clustered the human pose to six different poses to reduce the computational complexity of the initialisation stage. In the initial frames, they evaluate the six poses and search for a frame to minimise the PEF with one of these six poses. After finding the frame, they refine the projected model for that frame and start the tracking from there. Here, for initialisation the bulk motion is calculated. This bulk motion can be used to estimate the walking person's shape and body size. This approach is driven similarly to what was used in Refs. 17 and 27.



4 Applying the bulk motion algorithm on a sample: (a) frame 65 of sample 006 of database B from the CASIA gait database; (b) extracted edges using Sobel operator; (c) the bulk motion image

After a background subtraction, an edge detection algorithm extracts the subject's silhouette edges, and here canny is used. Let A be the velocity accumulator, v (pixels per frame) be the velocity, I_n be the edge image intensity function in frame n, i and j bethe image coordinates, dy_n be the displacement in the direction of y axis and N be the number of frames in the sequence, we have:

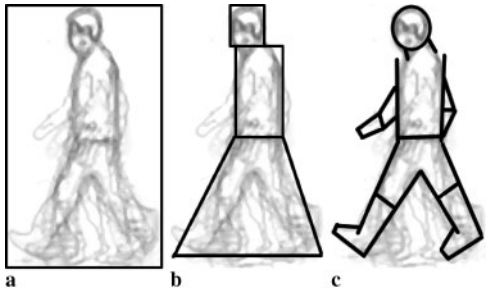
$$A(i,j)=\sum_{n=0}^{N-1} I_n\left[i+v\left(\frac{N}{2}-n\right),\ j-dy_n\right] \tag{1}$$

Applying equation (1) on a video sequence sorts the moving objects with regard to their velocity and start point and generates an accumulator of their motion. The level of interference of each edge in the bulk motion depends on the intensity of that edge, number of frames that this edge is present in and its velocity. This is the same as a global averaging and is not influenced from the partial occlusion.

If this accumulation performs properly, in other words if the velocity is considered correctly, the edges would be accumulated and would output the shape average (Fig. 4). To acquire this bulk motion, we can test different velocities in a range and choose the bulk motion that has the maximum value in. The computational complexity of this approach would be $O(V \times E \times N)$, where V is the number of possible velocities in the range, E is the average number of edge points and N is the number of frames. This computational complexity even could simply be enhanced by estimating the person's velocity via averaging the difference of centroids frame by frame, current C_c and previous C_p frames:

$$v=\frac{1}{N}\sum_{n=0}^{N}\left\|C_p-C_c\right\| \tag{2}$$

A region growing algorithm could be used to simply bind a rectangle on the bulk motion image extracted in the



5 Applying the shape estimation algorithm on the bulk motion image: (a) the bounding region; (b) initial estimation; (c) the final estimation

previous section (Fig. 5). After extracting this model, the initial frames of the sequence are searched. The frame with the smallest value of PEF would be chosen and the model would be refined for that frame.

3.1.3 Scaling

The proposed tracking algorithm is supposed to work for both cases in which the human walks fronto-parallel the camera or walks on a straight line with an angle relative to the image plane. What is important is that when the subject walks not fronto-parallel relative to the camera, body size in terms of pixels is different in the initial and final frames. Therefore, to solve this problem the person's silhouette is scaled to a same size throughout the sequence. The tracking is performed in these images. After the tracking results are acquired, model is scaled back to be projected on the image plane.

At the beginning, before initialisation, all the frames are scanned in one pass to find the maximum height of the person $h_{\max}$. For each frame $h_c/h_{\max}$ is calculated as the scale factor, where h_c is the silhouette height in the current frame. In the tracking process, when matching the model to the subject in the image, the scaled silhouette is used. In this paper, the scaling is implemented via the bilinear interpolation.

Adding this phase, at the beginning and at the end of the model-based tracking algorithm guarantees that the same algorithm described works properly for both cases where the subject is walking fronto-parallel or with an angle relative to the camera.

3.1.4 PEF

The most important part of a model-based tracking algorithm is the PEF, which relates the pose vector $P=\{x, y, \theta_1, \theta_2, \ldots, \theta_{10}\}$ to the data present in the image. The most probable body pose is predicted using the previous frames tracking results, and then the predicted model is projected on the image plane. The projected model is matched to the new frame human pose and the match error forms the PEF output.

To reduce the PEF, recursively the pose is predicted and refined. Here, to compute the matching error, both boundary matching and region matching are incorporated. Then the physical forces are computed regarding this matching error, which are used to modify the model and its joint angles to decrease matching error.

Chamfer distance has always been a good measure to match the model edges to the boundary of the detected object.[2,18,28,29] Suppose that we have the body model $P=\{x, y, \theta_1, \theta_2, \ldots, \theta_{10}\}$, let $r(s)$ be the model curve ($0 \leqslant s < 1$) and $z(s)$ be the equivalent curve of the detected human in the image. The function $g(s)$ maps each point in $z(s)$ to the boundaries of the detected human in the image, $r[g(s)]$. Therefore the boundary matching error would be:

$$E_b = \frac{1}{C}\int_0^1 \min[\|z_1(s) - r(s)\|, u]\,ds \tag{3}$$

where C is the normalisation constant and usually is set to the length of $r(s)$. $z_1(s)$ is the nearest point on $z(s)$, relative to each point on $r(s)$:

$$z_1(s) = Z(s'), \quad s' = \underset{s' \in g^{-1}(s)}{\operatorname{argmin}} \|r(s) - z(s')\| \tag{4}$$

To refine the joint angles using this matching error, spring forces could be used.[8,23] Each F_i, calculated above, is considered as a spring which corresponds to the physical force $\|F_i\|$. The combination F_b pulls the model towards the corresponding points on the image.

$$F_b = \frac{1}{C}\int_0^1 f\left[F(s), \rho\frac{F(s)}{\|F(s)\|}\right] ds \tag{5}$$

where $F(s) = \overrightarrow{r(s)z_1(s)}$ and ρ is the spatial scale constant and

$$f(F_1, F_2) = \begin{cases} F_1 & \|F_1\| \leq \|F_2\| \\ F_2 & \|F_1\| > \|F_2\| \end{cases} \tag{6}$$

In order to avoid errors when the model lies in the space between two body parts, region matching is utilised. The model region is divided into two parts: p_1 is the overlapping area and p_2 is the rest. As a

result the region matching error would be:

$$E_r = \frac{|p_2|}{|p_1| + |p_2|} \tag{7}$$

where $|p_i|$ is the number of pixels in region p_i.

Similar to the previous case, a physical force could be defined. Let c_1 and c_2 be the centroids of the regions p_1 and p_2. The vector $F_r = \overrightarrow{c_1 c_2}$ is defined as the resultant physical force. This force would pull the model towards the equivalent region, such that more overlapping appears.

The boundary information enhances the process of locating different parts and region information makes it more robust. To achieve better accuracy and more robustness, the two matching errors are combined to form the PEF.[2] This is done for the two physical forces F_b and F_r.

$$E(P) = (1-\alpha)E_b + \alpha E_r \tag{8}$$

$$F = (1-\alpha)F_b + \alpha F_r \tag{9}$$

where P is the pose vector and α is the balancing factor. Lower values of α are used for the upper parts of the body to reduce the region matching error effects. This is because there often is a same texture all over the cloths and region information is less important.

3.1.5 PEF minimisation

The goal to tracking is to find a pose $P=\{x, y, \theta_1, \theta_2, \ldots, \theta_{10}\}$ in the 12-dimension space of the body configuration defined above. As stated before, the kinematics model could be replaced with a tree-like structure. Therefore, a hierarchical search strategy could be incorporated. The body position first is found and each part is tracked independently. As far as the body position has a direct effect on the PEF value, the position (x,y), the tree root, is first determined. And then for each joint angle, a path from the root is extracted and a kinematics-based method estimates that angle. In this method, first the tree root (torso) and then the sub-trees are estimated one by one recursively.

The position estimation is composed of two stages: prediction and refinement. In the prediction phase, it is assumed that the body displacement equals the position displacement:

$$\begin{aligned} x_c = C_c x - C_p x + x_p &\Rightarrow x_c - x_p = C_c x - C_p x \\ y_c = C_c y - C_p y + y_p &\Rightarrow y_c - y_p = C_c y - C_p y \end{aligned} \tag{10}$$

where c and p are current and previous frames indicators and (x,y) and C represent the position and centroid of the person. $C_i x$ and $C_i y$ indicate the x and y coordinates of C_i respectively.

The refinement stage involves a simple search of the pixels around the predicted point to decrease the PEF. As far as the body centroid is a good measure for predicting position, the real position is near the predicted one.

The estimation of the joint angles also includes two stages: prediction and refinement. For prediction, the angular velocity is used which is drawn by the tracking results from previous frames:

$$\theta_{ic} = \theta_{ip} + \dot{\theta}_{ip}\Delta t \tag{11}$$

where θ_i is the ith joint angle and $\dot{\theta}_{ip}$ is the angular velocity for the previous frame. Therefore, θ_{ic} is the predicted joint angle value. The angular acceleration is updated frame by frame, which leads to non-linear characteristics in θ_{ic} and $\dot{\theta}_{ip}$. Some examples of updating and refining joint angles could be found in Refs. 2, 18 and 30.

For refinement, a force F is applied on the model centroid and rotates the body part around the joint.[2] With L as the part length and $F^{\perp}$ the orthogonal component of the force, the angle difference to be added to the joint angle is calculated as:

$$\Delta\theta = \beta F^{\perp}/L \tag{12}$$

where β is a constant factor independent of F and L. In the refinement stage, the joint angles are recursively updated. Bigger values of β cause the sudden increase in the angle and lower values of β prevent the oscillation around the optimal solution. It is obvious that β being so small results in big number of iterations for achieving a suboptimal solution. Here, the same as in some previous works, the value of β is decreased as the number of iterations increases. As a result, in the kth iteration, $\beta=\gamma/k$ where γ is a constant. The stopping criteria could be: (1) the matching error drawn from the PEF is less than a threshold δ; or (2) the number of iterations exceeds a constant K; or (3) $\Delta\theta$ gets less than a threshold ε.

An iterative algorithm can now be introduced that minimises the PEF. First, we need to initialise $k=0$ and calculate θ^0 and the physical force F. Then by calculating $\Delta\theta$, whether it is less than ε or not is checked. If so, the algorithm exits, otherwise set $\theta^{k+1}=\theta^k+[\beta/(k+1)]\Delta\theta$. Update P, project the new pose on the image plane and calculate $E(P)$. Continue this until the termination criteria are met.

The output of this phase is the vector $P=\{x, y, \theta_1, \theta_2, \ldots, \theta_{10}\}$ for each frame, which includes the

position of the subject and the joint angles. Having this, the exact positions of the joints on the image plane could be exactly calculated. These positions for each frame are values used as inputs for the next phase, the 3D reconstruction algorithm.

3.1.6 *Occlusion handling*

The most probable reason for tracking failure is the occlusion between the moving objects in the image.[31] To handle the problem of occlusion, after the occlusion the sequence of each of the walking subjects needs to be reinitialised independently, after that the algorithm would be able to track the subjects again. In this process only the frames, in which the subjects are occluded by one another are missed. This occlusion handling algorithm can perform well in partial occlusions.

3.2 Gait 3D reconstruction

To estimate human pose and model, many 2D and 3D approaches have been proposed in the literature. But still an automatic deterministic approach to reliable human gait reconstruction is a challenge. In this section, regarding the knowledge of human walking, a simple but automatic approach is presented.

Human pose estimation is the same as the process of estimating the kinematics structure of the human skeleton. This can be a non-separable part of a tracking or an HMA system. In this area, vision-based techniques are of much interest and are more efficient than 3D scanners. Cheung *et al.*[32] used the segmented body parts acquired from multiple cameras to extract the kinematics structure. Menier *et al.*[33] also proposed an algorithm to initialise this structure from multiple cameras. They reinitialise the kinematics structure in each frame. A category of approaches use more than a single image from a same scene for 3D reconstruction. As an example, a two-stage algorithm for 3D reconstruction in stereo pair is proposed in Ref. 34. The two stages in this algorithm include a hierarchical disparity estimation and an edge-preserving disparity field regularisation. On the other hand, some of the approaches[35] extract the 3D pose from monocular video sequences. In these approaches, authors assume that the positions of joints are set manually. Taylor in Ref. 35 uses the relative distance of manually marked joints in a single image to reconstruct the articulated object. In this approach, if any of the joints is not visible in the image, the method fails significantly. Some other approaches assume that camera calibration parameters are available.[36] This assumption makes these techniques not useful in real-world applications.

Remondino and Roditakis,[37] the same as Taylor,[35] used the relative distance between joints. But they have refined the method by relaxing some constraints. Chen and Lee used some knowledge of human gait and converted the 3D reconstruction problem to a graph problem.[38] They have used the A* search strategy to solve the problem and extract the 3D coordinates. Ren *et al.*[39] also have illustrated a method for body configuration retrieval in static images. In this method, they solve a quadratic programming problem to extract the most probable body configuration.

After the tracking process is performed on a video sequence, described in the previous section, the positions of the joints in each frame are known. Incorporating some prior knowledge about human together with these 2D coordinates could help estimating the camera calibration parameter (or the scale factor), after which 3D coordinates are extracted. Under all these constraints, still remain a big number of possible configurations for the joints. Again incorporating the human gait knowledge, a method is proposed to extract the correct 3D pose. This is the case that in previous works a user used to decide between all possible configurations.

The output to this 3D reconstruction algorithm is meant to be used for gait recognition. Here a general human 3D pose reconstruction is presented, but only the lower body joints are reconstructed for the recognition purpose.

3.2.1 *Problem formulation*

In the ideal situation, the camera is assumed as a point in the space. The (x,y,z) coordinate system is the world and the (u,v) system is considered as the camera coordinate system. We can model the relation between these two coordinate systems using orthographic projection. With this assumption, the coordination of a point in (x,y,z) space would be related to a point in (u,v) as shown in equation (13):

$$\begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{13}$$

As a result, the effect of orthographic projection would be reduced to a simple scaling of x and y

coordinates.[35] Assuming a similarly constant z (depth), this model can be simply used. Under this circumstance, one can find an appropriate value for s:

$$\begin{aligned} u &= sx \\ v &= sy \end{aligned} \tag{14}$$

This model simplifies the reconstruction process and provides suitable results for visualisation. In the proposed algorithm, s will be estimated using the information in the image and the human gait knowledge.

3.2.2 3D reconstruction algorithm

In previous works on 3D reconstruction of body parts, in order to extract the camera parameters, a function is defined that relates the measurements in the image to model parameters. They next invert the function and extract the parameters.[40] This is very time-consuming and very difficult. The proposed approach in this paper avoids these problems and chooses a direct method.

In the case of orthographic projection, the two end points (x_1,y_1,z_1) and (x_2,y_2,z_2) are projected on (u_1,v_1) and (u_2,v_2). Considering s as the scale factor, we can use these data to retrieve the relative depth of the two end points.

In the 3D space the distance is:

$$l^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \tag{15}$$

Taking into account both equations (14) and (15), we can say:

$$\begin{aligned} (u_1 - u_2) &= s(x_1 - x_2) \\ (v_1 - v_2) &= s(y_1 - y_2) \\ \mathrm{d}z &= (z_1 - z_2) \end{aligned} \tag{16}$$

As a result

$$\mathrm{d}z^2 = l^2 - \left[(u_1 - u_2)^2 + (v_1 - v_2)^2\right]/s^2 \tag{17}$$

In other words, 3D coordinates of the end points are calculated as a function of s. As far as equation (17) is a quadratic equation, for each value of s in this case two different values for the coordinations could be extrapolated. We can decide whether point 1 will have a smaller z or point 2. In this condition, both of the lines will have a same projection on the image plane. This ambiguity is like the ones brought up in Refs. 36 and 40.

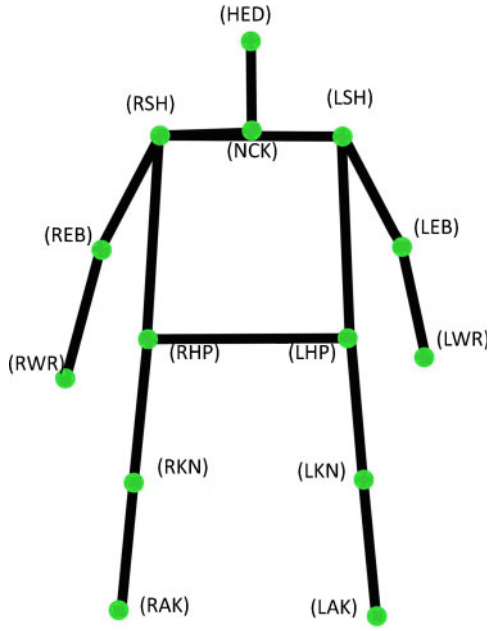
Let's assume that we have the scale factor s, considering this equation (17) says that we can express relative depth of the two points as a function of l. Therefore, the 3D reconstruction problem reduces to a simple problem of finding a good scale factor. In order to find an appropriate value for s, equation (17) could be reformed to determine a lower bound for s. As far as $\mathrm{d}z$ could not be a complex number, we have:

$$s \geq \frac{\left[(u_1 - u_2)^2 + (v_1 - v_2)^2\right]^{1/2}}{l} \tag{18}$$

This could be generalised for the articulated case.[35] For instance, consider a kinematics chain including three segments which we previously know the relative lengths. As mentioned above, the relative depth of the two ends could be calculated as a function of s. Then one of the end points in the image could be chosen as the reference point, after which the z coordinate of the other points could be calculated relative to that reference point. For each segment, two different configurations are plausible. This means that for each s, eight different cases for points coordination could be found to satisfy the measurements and constraints acquired from the image. Applying the inequality 18 on each line segment a lower bound for s could be acquired. According to Refs. 35 and 37, the maximum value of all the lower bounds of s for all line segments in the image could be considered as the scale factor. As mentioned above, one of the end points could be assumed to be closer to the camera. Considering an arbitrary value for the coordinates of that end point, one can calculate the coordination of the other point. As far as the model has a kinematics structure, this procedure could be performed for every segment and finally all the points will have a coordination regarding the first point coordinates. This leads to a big number of different configurations for the points. In general, for a kinematics structure with n joints, there will be 2^n different configurations. This is considered as an ambiguity and an approach is proposed to solve this problem.

3.2.3 Human body model

The human body behaves as a series of joined segments. In order to reconstruct the 3D body model, a number of feature points and their linkages need to be defined. The model used in this work is shown in Fig. 6. It consists of 14 feature points, 13 of which are body joints (right shoulder, left shoulder, neck, right elbow, left elbow, right wrist, left wrist, right hip, left hip, right knee, left knee, right ankle and left ankle) and the highest point is the representative of the head. This model is relative to what was presented in Section 3.1.1.



6 Human body model

As mentioned before, the proposed algorithm needs the relative distances of the feature points. In order to extract these relative distances in a human body, we refer to clinical studies, HMA databases and human body proportions studies.[16,37] Generally the human body could be modelled with a height of eight heads, and the distance of the joints could be expressed as proportions of the height (Table 1).

3.2.4 Dealing with the ambiguity

According to Section 3.2.2, for such a model with 14 feature points, 214 configurations for each frame are possible. If there are N frames in the video sequence with no constraints on the human gait, there will be $(2^{14})^{N}=2^{14\times N}$ different configurations for the human gait throughout this video sequence. A search on this space to find the most plausible configuration needs a very high computational complexity. To this end, we use the physiological knowledge of human body.

Table 1 Relative length of different human body parts

Segment	Relative length
Height	8
Forearm	2
Upper arm	1.5
Neck and head	1.25
Torso	2.5
Shoulder girdle	2
Pelvic girdle	2
Upper leg	2
Lower leg	2
Foot	1

3.2.4.1 Human gait rules: the single-frame case

The following three rules are used to reduce the number of body configurations in each single frame:

Rule 1: The two hands cannot lie on the same side of the torso (either front or behind). The same constraint could be taken into account for the legs.

Rule 2: The leg and the hand on the same side of the body could not together be front or behind the torso.

Rule 3: In each scene, the right hip (RHP), left hip (LHP), right shoulder (RSH), left shoulder (LSH) and neck (NCK) feature points almost lie on a same plane. To validate this rule, the plane composed of the first three points is created. Then the distance of the other points from this plane is calculated. If the distance is less than a predefined threshold that configuration is valid from this rule's point of view.

3.2.4.2 Motion analysis using the sequence of frames

The probable configurations for each frame should be integrated to examine if they form a logical motion. As stated in Ref. 41, because of passive viscoelastic and active chemicomechanical properties of muscles in the process of walking, the form of human gait in the sequence of frames which are captured in close time intervals is smooth. In the process of motion analysis, this smooth walking form is used for extraction of a reliable and suitable model. Assume that the variable x_i is one of the probable body configurations in frame i then $\overrightarrow{A}_j(x_i)$ will be a vector from feature point A to feature point B on $\overline{AB}$ in x_i. The gap between two configurations in two consecutive frames i and $i+1$ is defined as:

$$D(x_i,x_{i+1})=\sum_{j=1}^{14} d\left[\overrightarrow{A}_j(x_i),\overrightarrow{A}_j(x_{i+1})\right] \tag{19}$$

$d\left[\overrightarrow{A}_j(x_i),\overrightarrow{A}_j(x_{i+1})\right]$ is the Euclidean distance of feature point A_j in two configurations of x_i and x_{i+1} in 3D space. Suppose that the time interval between two consecutive frames is Δt, relative velocity of transition of $\overline{AB}$ from configuration x_i in ith frame to the configuration x_{i+1} in frame $i+1$ is defined as

$$\begin{aligned}\overrightarrow{V_{AB}}(x_i,x_{i+1}) &= \frac{\left[\overrightarrow{B}_{(x_{i+1})}-\overrightarrow{B}_{(x_i)}\right]-\left[\overrightarrow{A}_{(x_{i+1})}-\overrightarrow{A}_{(x_i)}\right]}{\Delta t}\\ &= \overrightarrow{AB}(x_{i+1})-\overrightarrow{AB}(x_i)/\Delta t\end{aligned} \tag{20}$$

The relative angular velocity of $\overrightarrow{AB}$ is defined as vector product of $\overrightarrow{AB}$ and $\overrightarrow{V}_{\overline{AB}}(x_i + x_{i+1})$.

$$\overrightarrow{w}_{\overline{AB}}(x_i, x_{i+1}) = \overrightarrow{AB}\,\overrightarrow{V}_{\overline{AB}}(x_i + x_{i+1}) \tag{21}$$

The relative angular acceleration is also defined as in equation (22):

$$\alpha_{\overline{AB}}(x_i, x_{i+1}, x_{i+2}) = \frac{\left|\overrightarrow{w}_{\overline{AB}}(x_{i+1}, x_{i+2}) - \overrightarrow{w}_{\overline{AB}}(x_i, x_{i+1})\right|}{\Delta t} \tag{22}$$

The human body motion is a complicated process and the external and internal forces act on body skeleton.[38] The smooth and continuous body kinematics motion can be formed as a set of smooth and continuous motion of all the body parts. The velocity of smooth motion of a part of body is constant angular velocity or its average angular acceleration is almost zero. Therefore, in order to estimate the smoothness of walking in the frames i, $i+1$ and $i+2$, the angular acceleration function is defined as equation (23):

$$f_i(x_i, x_{i+1}, x_{i+2}) = \sum_{\overline{AB}} \left|\overrightarrow{\alpha}_{\overline{AB}}(x_i, x_{i+1}, x_{i+2})\right| \tag{23}$$

In equation (23), the angular velocities of all body parts are aggregated. For simplicity the scale of angular velocity vector is used. Finally, for calculating the total angular acceleration of N consecutive frames, equation (24) is utilised.

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f_1(x_1, x_2, x_3) + f_2(x_2, x_3, x_4) + \\ &\quad \ldots + f_{N-2}(x_{N-2}, x_{N-1}, x_N) \\ &= \sum_{i=1}^{N-2} f_i(x_i, x_{i+1}, x_{i+2}) \end{aligned} \tag{24}$$

Therefore, finding a sequence with smooth motion in N frames of human motion while walking is the same as finding the set $x = \{x_1, x_2, \ldots, x_n\}$ that produce the minimum value of $f(x_1, x_2, \ldots, x_N)$.

3.2.4.3 Applying restrictions using the human gait knowledge

Since in the process of finding the smoothest set of motions, it is possible to find sets with a lot of similarities, some special rules for human gait should be checked. This set of rules are:

Rule 4: When shoulder and elbow joints are moving forward and backward, both are moving in the same direction. This rule also holds for joints of thigh and knee.

Rule 5: The motions of hands and legs are in the plane which is parallel to the direction of the body motion.

Rule 6: In each instance of walking, only one of the knees is bent.

In the 3D reconstruction algorithm, using all mentioned constraints a unique 3D configuration is extracted for each subject throughout the video sequence. These 3D coordinates for each frame could be used to extract the joint angles in that frame and for the sequence as a whole they could be used to calculate the joint angle changes. The behaviour and the changes for the lower limbs form the dynamic features for the subject's gait. Next section describes how to calculate the gait features from the 3D coordinates of the joints.

3.3 Gait recognition

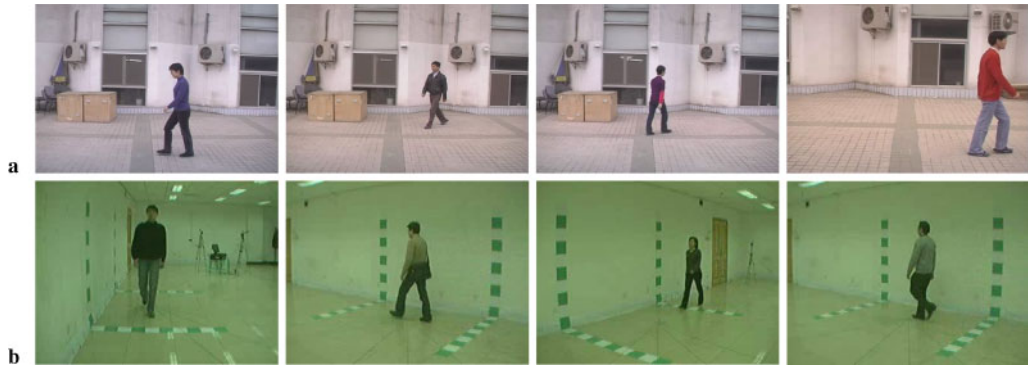
Automatic human identification in surveillance systems has always been an important issue. Previously, a model-based tracking algorithm was proposed in Section 3.1. Section 3.2 introduced a 3D reconstruction algorithm which used the human gait knowledge and tracking results from the previous section to reconstruct the human body in 3D space. Now, these 3D pose data are utilised to extract the gait dynamic features. As far as the arms, hands and torso movements are not consistent and reliable for identification only the lower body parts movements are used for gait dynamic feature extraction. So, the 3D reconstruction algorithm was only executed on the lower parts of the body, the two legs, which makes the reconstruction much easier and also more robust.

3.3.1 Gait dynamic feature extraction

For recognition, we only need to extract the human walking dynamics, which can be done via geometrical relations. A vector in 3D space is denoted using a triple (x,y,z), with (x_1,y_1,z_1) and (x_2,y_2,z_2), its two end points we have $x = x_2 - x_1$, $y = y_2 - y_1$ and $z = z_2 - z_1$. Each of the line segments of the 3D model could be treated as a vector. To calculate the angle between two vectors, dot multiplication can be used, so the angle between the two vectors θ could be calculated as:

$$\theta = \cos^{-1} \frac{\overrightarrow{a}\,\overrightarrow{b}}{|\overrightarrow{a}||\overrightarrow{b}|} \tag{25}$$

As the feature for recognition, the two hip and two knee joint angles are used. These angles are extracted from the 3D stick figure model shown in Fig. 6 using equation (25). The signals of these joint angles are normalised in a cycle, and the two end points of the signal are aligned.



7 Samples of (a) the outdoor data and (b) the indoor data. In these databases, the sequences from subjects walking outdoor are captured in angles 0, 45, 90 and 135°, and the indoor sequences are captured in 0, 18, 36, 54, 72, 90, 108, 126, 144, 162 and 180°

The following algorithm gives the whole work's flow step by step for better understanding:

Step 1. Input the video sequence and determine the background image.

Step 2. Find the walking subject using a background subtraction technique (*see* Section 4).

Step 3. Initialise the video sequence and project the initial model on the image plane (Section 3.1.2).

Step 4. For each frame:

Step 4.1. According to equation (11) predict the joint angle θ^0, set $k=0$ and perform the joint angle refinement as follows:

Step 4.1.1. According to equation (9) calculate the physical force F.

Step 4.1.2. Use equation (12) to calculate $\Delta\theta$.

Step 4.1.3. If $|\Delta\theta|<\varepsilon$, exit the loop.

Step 4.1.4 $\theta^{k+1}=\theta^k+[\beta/(k+1)]\Delta\theta$

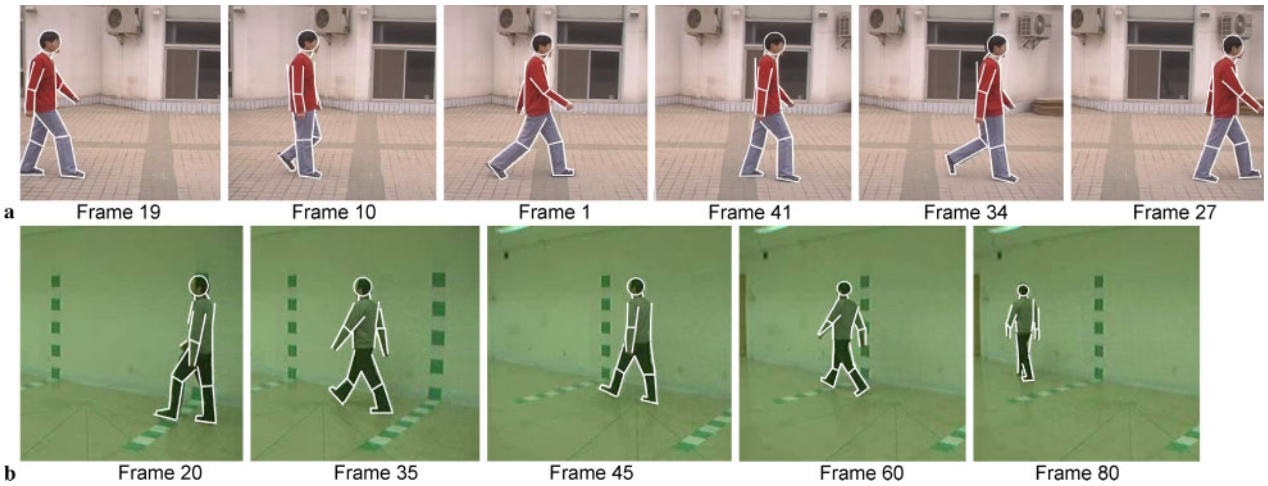
Step 4.1.5. After updating P, project the new pose on the image and calculate $E(P)$, (equation (8)).

Step 4.1.6. If $E(P)<\delta$ or $k\geqslant K$, exit the loop, otherwise go back to Step 4.1.1.

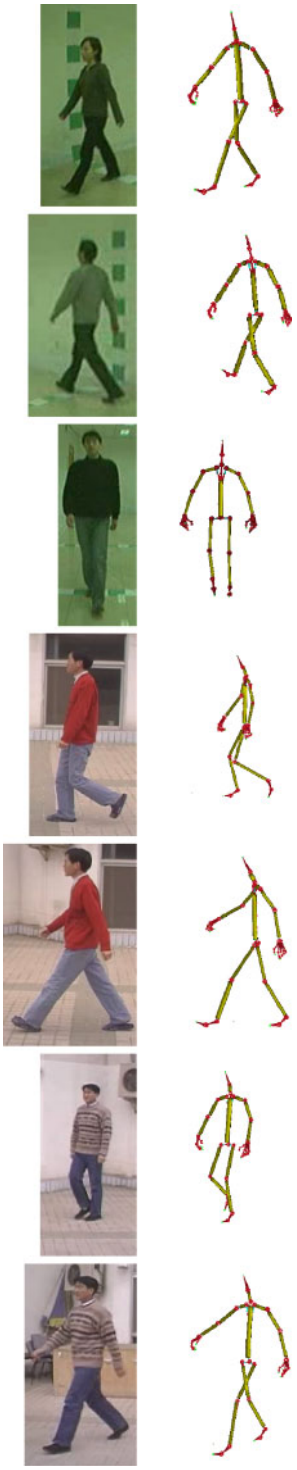
Step 5. After the tracking process, the vector $P=\{x, y, \theta_1, \theta_2, \ldots, \theta_{10}\}$ is available for each frame. Extract 2D coordinates of the body's joint angles as the feature points of the proposed model in Fig. 6.

Step 6. Estimate scale coefficients using equation (18), relative distances in Table 1 and position of feature points (joints).

Step 7. Initialise the frame counter to 1 (*cur_frame*=1) and for each frame, perform the followings:



8 Tracking results: (a) outdoor; (b) indoor



9 Results of 3D reconstruction on some samples of image sequences on test datasets

Step 7.1. Produce all 2^{14} body configurations for current frame.

Step 7.2. Enforce the rules in Section 3.2.4.1 and eliminate the invalid body configurations.

Step 7.3. If ($cur\_frame \geqslant 2$), match all the generated configurations of x_i with x_{i-1} in last frame ($cur\_frame-1$), choose one whose value is less than distance $D(x_i,x_{i-1})$ defined in equation (19). Add new configuration x_i to the end of picked sequence x_1, x_2, …, x_{i-1}, so the motions with suitable velocity are picked.

Step 7.3. If multiple x_i are picked for one x_{i-1}, only keep the one that produce the least $D(x_i,x_{i-1})$.

Step 7.4. If for one configuration of x_i the value of $D(x_i,x_{i-1})$ is not less than threshold, ignore that configuration of x_i.

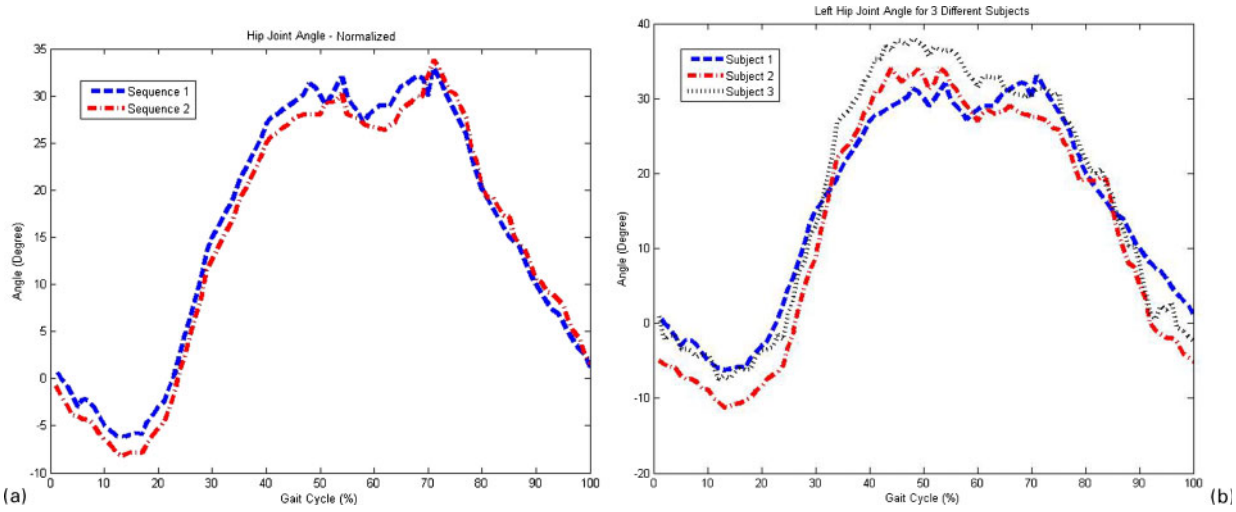
Step 7.5. At the end of the loop, save all the generated sequences $\{x_1, x_2, \ldots, x_{cur\_frame}\}$ to be used for configuration matching in next frames.

Step 7.6. Increment frame counter ($cur\_frame = cur\_frame + 1$).

Step 7.7 If there are no frames left, exit the loop. Otherwise return back to Step 7.1.

Step 7.8. Enforce the rules of Section 3.2.4.3 on generated sequences $\{x_1, x_2, \ldots, x_N\}$ and eliminate the invalid configurations.

Step 8. Now some sequences $\{x_1, x_2, \ldots, x_N\}$ are still remained. The maximum number of these sequences according to the numbers theory is 2^{14}. The maximum number of sequences are produced if each configuration in the *i*th frame is matched only with one configuration in frame $i-1$. Therefore by enforcing the rules in Step 7.2 and high possibility of mismatch in Step 7.3 and again enforcing the rules in Step 7.8, the number of configurations is much less than 2^{14}. Using equation (24) calculate the total angular acceleration of each configuration sequence $\{x_1, x_2, \ldots, x_N\}$. The sequence with minimum angular acceleration is chose as body configuration sequence in the sequence of video images.



10 (a) Normalised signal of left hip of a walking person in two different sequences; (b) the normalised signal of the left hip joint angle of three different subjects

Step 9. After choosing a configuration sequence for the whole sequence, the 3D coordinates of the body joints for each frame are available. Use the lower body limbs and their joints coordinates and the information in Section 3.3.1 to calculate the joint angles trajectory throughout the video sequence.

Step 10. Use a previously trained classifier (as explained in the Section 4), input these calculated joint angles for recognition purposes.

4 EXPERIMENTS

To show the efficiency of the proposed algorithms, different experiments on both indoor and outdoor data are conducted. For the indoor test, the CASIA NLPR[42] gait database B is used. This database includes 126 subjects, with 10 different walking sequences for each, and at the same time 11 cameras in different angles capture images of the person's gait. In fact, this database is a multi-camera dataset, but here data from each camera are used independently. For the outdoor test database A is chosen from the same set of data, CASIA NLPR. In this dataset, there are 20 different subjects with 12 sequences for each. These data, contrary to the ones of database B, are captured with a single camera. The subjects walk in the sequence with different angles relative to the camera. Figure 7 shows samples from both these datasets. In order to detect moving human in the scene, a simple background subtraction technique is used: a temporal averaging of the pixels to model the background and then a subtraction to reveal the moving objects.

As mentioned before, choosing factor α, for balancing the boundary matching and the region matching errors in PEF, is very important and affects the output. Here, α for upper body parts is set to 0.6 and for lower body parts to 0.8. As also stated before, when a part is missed or for a pixel p_i no match q_i could be found, the error value is set to μ. Under such a circumstance that part receives $E_b \simeq \mu$, so displacement and the physical forces to that part are set to be the same as its previous frame and that part is not projected on the image. Only those pixels are projected that have found an equivalent in the boundary matching stage (have matching error less than μ). Figure 8 shows the tracking results optimised for proper view.

Table 2 Inter-class and intra-class correlation coefficient for the four selected features for the recognition purpose

	Left hip joint angle	Right hip joint angle	Left knee joint angle	Right knee joint angle
Inter-subject correlation	0.956	0.96	0.973	0.945
Intra-subject correlation	0.89	0.881	0.868	0.904

After the tracking results are available, the joint positions are extracted. These positions could be incorporated to reconstruct the 3D model. Some of the results are shown in Fig. 9. Using the motion builder software[43] the 3D images of the subjects are drawn from the 3D reconstructed joint coordinate. In order to compare this 3D reconstruction approach, one can see the results of Refs. 37 and 49, which are similar works to this approach. Chen and Chai[49] constructed generative models for both human motion and skeleton from pre-recorded data and then deformed the models to best match a set of 2D joint trajectories derived from monocular video sequences and then developed a multi-resolution optimisation procedure to reconstruct motion in a coarse-to-fine manner. Remondino and Roditakis[37] also used a same approach described throughout this paper, but do not give a general solution for dealing with the ambiguities. The results from these approaches are visually the same as what presented here.

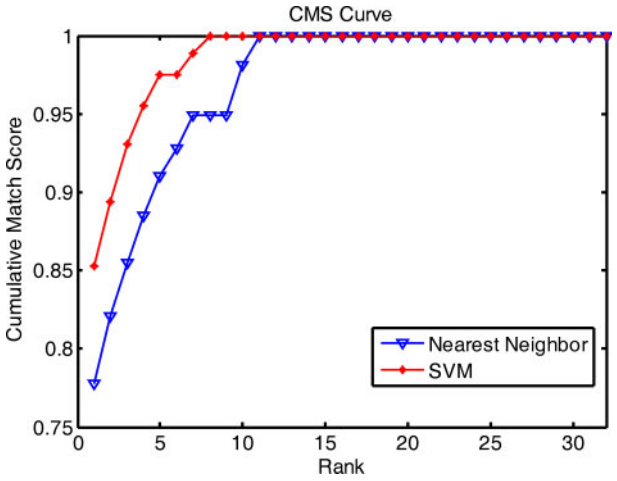
After the 3D reconstruction phase, as also stated before, the lower body parts are used to extract these dynamic features. Figure 10a shows a sample of the normalised signal of hip joint angle from two different walking sequences of a same person. It is obvious that not too much difference could be seen in the signals. Figure 10b shows the same signal for three different people; in this case the difference between signals is distinguishable. This illustrates that a good classifier could recognise different subjects using this feature.

In order to discuss the desirability of the four selected features, correlation coefficient could be utilised. Correlation shows the measure of the statistical relationships between two or more data values. If we have a series of n measurements of X and Y (as x_i and y_i where $i=1, 2, \ldots, n$), then the sample correlation coefficient can be used as a single real number to measure the level of similarity of the two datasets. It is written as:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - x)(y_i - y)}{(n-1)s_x s_y} \tag{26}$$

where x and y are the means of X and Y, and s_x and s_y are their standard deviations.

If we consider normalised signals of the joint angles as X_s and Y_s, the correlation coefficient could be simply calculated between them. Table 2 shows the averaged correlation coefficient between all the normalised signals used for training, for each



11 The cumulative match score curve for the 32 subjects

Table 3 Confusion matrix of the nearest neighbour classifier for 16 subjects with eight sequences from each

Subject ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
P1	6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
P2	0	7	0	0	0	0	0	0	0	0	0	0	1	0	0	0
P3	1	0	5	0	0	1	0	0	0	0	0	0	0	0	1	0
P4	0	0	0	6	0	0	1	0	0	0	0	0	1	0	0	0
P5	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
P6	0	0	0	0	0	6	0	0	0	1	0	0	0	0	0	1
P7	0	0	0	0	1	0	7	0	0	0	0	0	0	0	0	0
P8	0	0	0	0	0	0	0	6	0	1	0	0	1	0	0	0
P9	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0
P10	0	1	0	0	1	0	0	0	0	5	0	0	0	1	0	0
P11	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0
P12	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0
P13	0	0	0	0	1	0	0	0	0	0	0	0	7	0	0	0
P14	0	1	0	0	0	0	0	0	0	0	0	0	0	7	0	0
P15	0	0	0	0	1	0	0	0	0	0	0	0	1	0	6	0
P16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8

Table 4 The SVM classifier confusion matrix for 16 subjects of outdoor test with eight sequences for each

Subject ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
P1	7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
P2	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P3	1	0	6	0	0	0	0	1	0	0	0	0	0	0	0	0
P4	0	0	0	7	0	0	0	0	0	0	0	0	1	0	0	0
P5	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
P6	0	0	0	0	0	7	0	0	0	0	0	0	0	0	1	0
P7	0	0	1	0	0	0	7	0	0	0	0	0	0	0	0	0
P8	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0
P9	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	1
P10	0	1	0	0	1	0	0	0	0	6	0	0	0	0	0	0
P11	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0
P12	0	0	0	0	0	0	0	0	0	0	0	7	0	0	1	0
P13	0	0	0	0	1	0	0	0	0	0	0	0	7	0	0	0
P14	0	1	0	0	0	0	0	0	0	0	1	0	0	6	0	0
P15	0	0	0	0	0	1	0	0	0	0	0	0	0	0	7	0
P16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8

individual subject's joint angles. The second row of the tables gives the same averaged correlation coefficient value over training joint angles normalised signals of two different subjects. Let's call the first row, inter-subject correlation and the second intra-subject correlation. As could be seen, the inter-subject correlations are far more than the inter-subject correlations. This could simply show that the selected features are suitable enough for our recognition purpose.

The gait recognition algorithm was tested on the same dataset described before. Totally, 384 different

Table 5 Comparison of the proposed approach with some others

Correct recognition rate	Used approach	Used dataset	Method
85.3%	Kinematics model-based tracking, 3D reconstruction and recognition based on gait dynamic features	32 subjects, 12 sequences from each	Proposed approach
Variable between 60 and 80%	Extracting the gait characteristic from the time–space pattern of the silhouette of a fronto-parallel walking person	26 sequences from five subjects	Ref. 46
88–100%	Extracting the eigenvector from the moving subjects' silhouette, parallel to the image plane (a stochastic approach)	Seven subjects, 10 sequences from each	Ref. 47
84%	Kinematics tracking and using dynamic gait features of a walking person parallel to the image plane	26 subjects, 26 sequences from each	Ref. 2
Dependent on the combination method from 87.5 to 97.5%	Static (Procrustes shape analysis) and dynamic (joint angles trajectory) features fusion, subjects walking parallel to the image plane	20 subjects, four indoor sequences for each	Ref. 19
84% (indoor data), 64% (outdoor data)	Frequency and period estimation of the kinematics motion of the fronto-parallel walking person	115 subjects, indoor and outdoor data	Ref. 17
1. 59.6% 2. 83.8%	Person-dependent dynamic shape model extraction using kinematics tracking of the fronto-parallel walking subjects	Two experiments: 1. 10 subjects, 30 sequences from each 2. 37 subjects, four sequences from each	Ref. 30
86%	K-means clustering of the silhouettes and dynamic time warping for recognition of subjects walking parallel to the image plane	20 subjects	Ref. 48

sequences are incorporated in the training and testing process from the CASIA gait database (32 different people with four sequences for training and eight for testing from each person).

Here, for the recognition purpose, two different classifiers are utilised, nearest neighbour and support vector machines (SVMs). Four sequences from each subject are used for training these systems. From each of these four sequences, four signals are extracted, as discussed before two hip and two knee joint angles. As a result, for each subject, 16 different normalised joint angle signals are present. Then, the four signals of each joint are averaged to form the stored pattern for this subject. So, four different normalised averaged signals are stored for each person. A simple Euclidean distance is used as the distance metric for the nearest neighbour classifier to recognise subjects. When testing, the normalised signals of the subject's four feature joint angles are extracted and then compared with the pattern stored in the database for each person. The one with the least distance is chosen. Table 3 shows the confusion matrix for the 16 subjects of outdoor test (each tested with eight sequences). This matrix shows that the classification result for the outdoor test is 84.3%.

The SVM classifier[45] is also used to classify the data for the recognition purpose. The signals are now represented with 50 discrete points for this classifier. The four normalised joint angles for each sequence are given to the SVM as an input independently. As a result, 50×4 inputs (four signals, each 50 samples) are used as the inputs of the SVM classifier. There are 32 discrete output states, as there are 32 subjects in the database. They are coded with error correcting output codes (ECOCs) for the outputs of the SVM classifier.

ECOC is a method for combining outputs of an SVM classifier.f SVM is a two-class classifier, to create a multi-class classifier; we need to combine the outputs. In a simple method, if we have n classes, we can use n different binary SVM classifiers. Or we can encode the classes with $N = \lceil \log n \rceil$ bits and use N different SVM classifiers. But here, as proposed by Passerini *et al.*,[44] for encoding and decoding the classifier outputs, ECOCs are used. In this case, 31 bits are chosen as the code length. Each class is encoded using 1-bit string. When testing the SVM, the output bit string specifies the corresponding class. If the bit string does not match to a class code, ECOC selects the nearest class in terms of error.

Therefore, an SVM with 200 input nodes, 31 output nodes and a radial basis function kernel function are utilised, the used parameters are $\gamma = 0.5$ and $C = 10$. For implementing SVM, the MATLAB library[45] is used. In the training phase, four discretised signals of the knee and hip joints are fed into the SVM as the inputs. Table 4 shows the confusion matrix for the 16 subjects of the outdoor test with SVM classifier (eight sequences for each). This matrix shows that the classification rate for the outdoor data is 89.03%.

Figure 11 illustrates the cumulative match score curve for the nearest neighbour and the SVM classifiers on the test data for all the 32 subjects. Using this diagram, one can compare the classification rate for these two classifiers. The value at rank = 1 shows the classification rate which is 77.8% for the nearest neighbour and 85.3% for the SVM classifier.

To compare the proposed approach with the ones of other works, Table 5 is presented. In this table, the used dataset and a brief overview of the approach are given and the correct classification rate shows the performance of the method. The results are the same as what was reported in their original works. However, these works are not individually implemented and tested on the used dataset in this paper, but by looking at this table we can figure out that although the proposed approach in this paper removes a number of constraints and restrictions of this class of algorithms, the results are still acceptable. This shows the robustness and efficiency of the proposed algorithm.

5 CONCLUSION AND FUTURE WORKS

In this paper, an approach for gait tracking and recognition is proposed. It uses the model-based tracking results to recover the gait dynamic features for the recognition phase. To this end, results of the kinematics-based tracking are taken into use by a 3D reconstruction algorithm to reconstruct the lower body parts in a 3D space. This 3D reconstruction uses the human gait knowledge. Then after the 3D coordinates of the joints are available, the feature extraction phase extracts the dynamic features from the sequence. The advantage to this method is that it can track and recognise people not walking parallel to the image plane, which is the constraint and the assumption made in lots of previous works in this

area. To test the proposed algorithm, CASIA gait database with the nearest neighbour and SVM classifiers is used. The results show similarly same efficiency of the proposed approach in comparison with previous works, while here the subjects are not forced to be walking parallel to the image plane.

Some directions for future works could be listed as follows: the proposed algorithm can be generalised for the case that the subject walks with no restriction in any direction, not just on a straight line. Static and dynamic gait features could simply be fused to achieve better recognition results. Since this approach provides the 3D model, it could be used for gait static feature extraction. Also the shape estimation and the initialisation processes used in the tracking algorithm can be used as a gait static feature extraction. For the rotation and translation steps, quaternion could be effectively utilised, which could significantly reduce the computational complexity. And last but not least, running the algorithm on more realistic, real-world data is also still a challenge.

REFERENCES

1 Cheng, J. C. and Moura, J. M. F. Capture and representation of human walking in live video sequence. *IEEE Trans. Multimed.*, 1999, **1**, 144–156.

2 Ning, H., Tan, T., Wang, L. and Hu, W. Kinematics-based tracking of human walking in monocular video sequences. *Image Vis. Comput.*, 2004, **22**, 429–441.

3 Moeslund, T. B., Hilton, A. and Kruger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 2006, **104**, 90–126.

4 McKenna, S. J., Jabri, S., Duric, Z., Rosenfeld, A. and Wechsler, H. Tracking groups of people. *Comput. Vis. Image Underst.*, 2000, **80**, 42–56.

5 Collins, R. T., Lipton, A., Fujiyoshi, H. and Kanade, T. Algorithms for cooperative multi-sensor surveillance. *Proc. IEEE*, 2001, **89**, 1456–1477.

6 Li, L. 'On-line visual tracking with feature-based adaptive models', MSc thesis, University of British Columbia, Vancouver, BC, Australia, 2004.

7 Mehmood, K., Mrak, M., Calic, J. and Kondoz, A. Object tracking in surveillance videos using compressed domain features from scalable bit-streams. *Signal Process. Image Commun.*, 2009, **24**, 814–824.

8 Delamarre, Q. and Faugeras, O. 3D articulated models and multi-view tracking with physical forces. *Comput. Vis. Image Underst.*, 2001, **81**, 328–357.

9 Thome, N., Merad, D. and Miguet, S. Learning articulated appearance models for tracking humans: a spectral graph matching approach. *Signal Process. Image Commun.*, 2008, **23**, 769–787.

10 Paragios, N. and Deriche, R. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. Patt. Anal. Mach. Intell.*, 2000, **22**, 266–280.

11 Bertalmío, M., Sapiro, G. and Randall, G. Morphing active contours. *IEEE Trans. Patt. Anal. Mach. Intell.*, 2000, **22**, 733–737.

12 Erdem, Ç. E. Video object segmentation and tracking using region-based statistics. *Signal Process. Image Commun.*, 2007, **22**, (10), 891–905.

13 Cunado, D., Nixon, M. S. and Carter, J. N. Automatic extraction and description of human gait models for recognition purposes. *Comput. Vis. Image Underst.*, 2003, **90**, 1–41.

14 Little, J. J. and Boyd, J. E. Recognizing people by their gait: the shape of motion. *J. Comput. Vis. Res.*, 1998, **1**, 1–32.

15 Green, R. D. and Guan, L. Quantifying and recognizing human movement patterns from monocular video images — Part I: A new framework for modeling human motion. *IEEE Trans. Circuits Syst. Video Technol.*, 2004, **14**, 179–190.

16 Green, R. D. and Guan, L. Quantifying and recognizing human movement patterns form monocular video images — Part II: Applications to biometrics. *IEEE Trans. Circuits Syst. Video Technol.*, 2004, **14**, 191–198.

17 Wagg, D. K. and Nixon, M. S. On automated model-based extraction and analysis of gait, Proc. 6th IEEE Int. Conf. on Automatic face and gesture recognition*: FGR '04*, Seoul, Korea, May 2004, IEEE Computer Society, pp. 11–16.

18 Lok, W. W. and Chan, K. L. Model-based human motion analysis in monocular video, Proc. IEEE Int. Conf. on Acoustic speech and signal processing*: ICASSP '05*, Philadelphia, PA, USA, March 2005, IEEE, Vol. 2, pp. 697–700.

19 Wang, L., Ning, H., Tan, T. and Hu, W. Fusion of static and dynamic body biometrics for gait recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 2004, **14**, (2), 149–158.

20 BenAbdelkader, C., Culter, R., Nanda, H. and Davis, L. EigenGait: motion-based recognition of people using image self-similarity, Proc. Int. Conf. on Audio- and video-based biometric person authentication*: AVBPA '01*, Halmstad, Sweden, June 2001, pp. 284–294 (Springer).

21 Collins, R., Gross, R. and Shi, J. Silhouette-based human identification from body shape and gait, Proc. 5th IEEE Int. Conf. on Automatic face and gesture recognition*: FGR '02*, Washington DC, USA, May 2002, IEEE Computer Society, pp. 366–371.

22 Phillips, P., Sarkar, S., Robledo, I., Grother, P. and Bowyer, K. The gait identification challenge problem: data sets and baseline algorithm, Proc. 16th Int. Conf. on

Pattern recognition: *ICPR '02*, Quebec, Canada, August 2002, IEEE Computer Society, Vol. I, pp. 385–388.

23 Delamarre, Q. and Faugeras, O. 3D articulated models and multi-view tracking with silhouettes, Proc. 7th Int. Conf. on Computer vision: *ICCV '99*, Kerkyra, Greece, September 1999, IEEE Computer Society, pp. 716–721.

24 Lowe, D. Fitting parameterized 3-D models to images. *IEEE Trans. Patt. Anal. Mach. Intell.*, 1991, **13**, 441–450.

25 Wachter, S. and Nagel, H. H. Tracking persons in monocular image sequences. *Comput. Vis. Image Underst.*, 1999, **74**, 174–192.

26 Sidenbladh, H., Black, M. and Fleet, D. Stochastic tracking of 3D human figures using 2D image motion, Proc. 6th Eur. Conf. on Computer vision: *ECCV 2000*, Dublin, Ireland, June–July 2000, IEEE, Vol. 2, pp. 702–718.

27 Wagg, D. K. and Nixon, M. S. Model-based gait enrolment in real-world imagery, Proc. Workshop on *Multimodal user authentication*, Santa Barbara, CA, USA, December 2003, ISCA, EURASIP and IEEE, pp. 189–195.

28 Borgefors, G. Hierarchical chamfer matching: a parametric edge matching algorithm. *IEEE Trans. Patt. Anal. Mach. Intell.*, 1988, **10**, 849–865.

29 Isard, M. and Blake, A. CONDENSATION — conditional density propagation for visual tracking. *Int. J. Comput. Vis.*, 1998, **29**, 5–28.

30 Lee, C. S. and Elgammal, A. Gait tracking and recognition using person-dependent dynamic shape model, Proc. 7th Int. Conf. on Automatic face and gesture recognition: *FGR '06*, Southampton, UK, April 2006, IEEE Computer Society, pp. 553–559.

31 Dockstader, S. L. and Imennov, N. S. Prediction for human motion tracking failures. *IEEE Trans. Image Process.*, 2006, **15**, 411–421.

32 Cheung, G., Baker, S. and Kanade, T. Shape-from-silhouette for articulated objects and its use for human body kinematics estimation and motion capture, Proc. IEEE Computer Society Conf. on Computer vision and pattern recognition: *CVPR '03*, Madison, WI, USA, June 2003, IEEE Computer Society, pp. 77–84.

33 Menier, C., Boyer, E. and Raffin, B. 3D skeleton-based body pose recovery, Proc. 3rd Int. Symp. on 3D data processing, visualisation and transmission: *3DPVT '06*, Chapel Hill, NC, USA, June 2006, IEEE Computer Society, pp. 389–396.

34 Kim, H. and Sohn, K. 3D reconstruction from stereo images for interactions between real and virtual objects. *Signal Process. Image Commun.*, 2005, **20**, 61–75.

35 Taylor, C. J. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Comput. Vis. Image Underst.*, 2000, **80**, 349–363.

36 Lee, H. J. and Chen, Z. Determination of human body posture from a single view. *Comput. Vis. Graph. Image Process.*, 1985, **30**, 148–168.

37 Remondino, F. and Roditakis, A. 3D reconstruction of human skeleton from single images or monocular video sequences, In Pattern Recognition (Ed. B. Michaelis and G. Krell), 2003, pp. 100–107 (Springer, Berlin).

38 Chen, Z. and Lee, H. J. Knowledge-guided visual perception of 3-D human gait from a single image sequence. *IEEE Trans. Syst. Man Cybern.*, 1992, **22**, 336–342.

39 Ren, X., Berg, A. C. and Malik, J. Recovering human body configurations using pairwise constraints between parts, Proc. 10th IEEE Int. Conf. on Computer vision: *ICCV '05*, Beijing, China, October 2005, IEEE Computer Society, Vol. 1, pp. 824–831.

40 Webb, J. A. Static analysis of moving jointed objects, Proc. 1st Natl. Conf. of American Association of Artificial Intelligence, Stanford, CA, USA, August 1980, AAAI, pp. 35–37.

41 Jain, R. and Sethi, I. K. Establishing correspondence of nonrigid objects using smoothness of motions, Proc. 2nd IEEE Workshop on Computer vision: *representation and control*, Annapolis, MD, USA, April 1984, IEEE Computer Society, pp. 83–87.

42 Chinese Academy of Science, Institute of Automation. Available at: <http://www.sinobiometrics.com>

43 Available at: <http://www.autodesk.com>

44 Passerini, A., Pontil, M. and Frasconi, P. New results on error correcting output codes of kernel machines. *IEEE Trans. Neural Networks*, 2004, **15**, 45–54.

45 Schwaighofer, A. Support vector machine toolbox for MATLAB. Version 2.51, 2002.

46 Niyogi, S. A. and Adelson, E. H. Analyzing and recognizing walking figures in XYT, Proc. IEEE Computer Society Conf. on Computer vision and pattern recognition: *CVPR '94*, Seattle, WA, USA, June 1994, IEEE Computer Society, pp. 469–474.

47 Murase, H. and Sakai, R. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Patt. Recogn. Lett.*, 1996, **17**, 155–162.

48 Amiri, A., Fathy, M. and Tahery, R. Human identification by gait using k-mean clustering algorithm and dynamic time warping, Proc. 12th Int. CSI Computer Conf.: *CSICC'07*, Tehran, Iran, February 2007, CSI, pp. 2423–2427.

49 Chen, Y.-L. and Chai, J. 3D reconstruction of human motion and skeleton from uncalibrated monocular video, Proc. 9th Asian Conf. on Computer vision: *ACCV '09*, Xi'an, China, September 2009, Asian Federation of Computer Vision Societies, pp. 71–82.