People Detection and Recognition using Gait for Automated Visual Surveillance

Imed BOUCHRIKA, Mark S. NIXON

ISIS, Department of Electronics and Computer Science
University of Southampton, SO17 1BJ, UK
{ib04r, msn}@ecs.soton.ac.uk

Keywords: Visual surveillance, people tracking, gait Recognition, biometrics.

Abstract

In this paper, a computer vision system for automated visual surveillance in an unconstrained outdoor environment is described. We propose a method for tracking multiple moving objects based on shape-based feature correspondence between consecutive frames. We have explored a new approach for walking people detection and recognition based on their gait motion. The novelty of our approach is motivated by the latest research for people identification using gait. The gait signature is derived using a model-based method. The experimental results confirmed the robustness of our method to discriminate between single walking people, groups of people and vehicles with a detection rate of %100. Furthermore, the system is able to recognize walking people with a CCR of %92.

1 Introduction

In recent years, automatic visual surveillance has received considerable interest in the computer vision community. This is due to the incapability of human operators to monitor the large growing numbers of cameras deployed in sensitive areas. The main aim of a surveillance system is to detect and track people in the scene, to understand their behaviour and to report any suspicious activities to a control centre. Surveillance systems could recognize people by their gait. Gait is a new biometric aimed at recognizing people by the way they walk. Because gait is hard to conceal and does not require user cooperation, it has received significant interest due to its potential in numerous applications.

Detecting, tracking and recognizing people using a single camera is a challenging problem due to occlusion, shadows, entry and exit of objects into the scene, and natural background clutter. Furthermore, the flexible structure of the human body, which encompasses a wide range of possible motion transformations, exacerbates difficulties for developing a vision-based surveillance system to recognize people.

Existing surveillance systems are classified into several categories [6] according to their type (single or multiple camera) and their functionality (tracking single or multiple people, etc.). Wren *et al* [16] proposed the PFinder system which uses a uni-modal background model to locate moving objects. The drawback of this system is its constraint to analyse just single people in the scene. The W^4 [6] surveillance system employs an appearance model to track people whereby single or group are distinguished using a projection histogram. Each person in the group is located through the tracking of his/her head. Lipton *et al* [11] proposed a real time vision-based system to classify moving objects into either human or vehicle based on the ”dispersedness”. Two types of features are used for people detection in surveillance systems: shape-based or motion-based features. The first type of cue relies on the shape of human silhouettes such as dispersedenss [11], aspect ratio of bounding box, or just simple shape parameters. For the motion-based features, the periodicity of human motion is considered as a strong cue for people detection. Cutler [4] described a real time method for measuring periodicity based on self-similarity.

Two main approaches are being used for gait recognition [1, 14]: model-based and non-model based methods. For the first one, a priori shape model is established to match real images to this predefined model, and thereby extracting the corresponding features once the best match is obtained. Stick models and volumetric models are the most commonly used methods. Karaulova *et al.* [9] have used the stick figure model to build a novel hierarchical model of human dynamics represented using hidden Markov models. Recently, Wang *et al* [8] modelled the human body by articulated cones and a sphere whereby body and motion models are used for tracking. For the non-model based methods, feature correspondence between successive frames is based upon prediction, velocity, shape, texture and colour. Shio *et al.* [12] proposed a method to describe the human body using moving blobs or 2D ribbons. Kurakake and Nevatia [10] extracted the joint locations by establishing correspondence between extracted blobs.

In this paper, we propose a multi-object tracking method based on feature correspondence between consecutive frames. Moving objects are assigned to different layers whereby blobs corresponding to the same object are assigned to the same layer. The criteria for allocating objects to layers is based on the Mahalanobis distance measure of shape-based features. Because of the dearth of visual surveillance systems that exploit human gait for object classification and their limited aim to detect people only using simple shape-based features extracted from silhouettes, we have explored an alternative technique for walking people detection and recognition based their gait motion. The novelty of our approach is motivated by the latest research for people identification using gait. In our method, gait periodicity is considered a strong cue for people detection which is estimated via the extraction of the heel strikes. For people recognition, a new model-based method is described to extract the positions of the joints of walking people. Evidence gathering is used for the extraction process whereby spatial model templates for human motion are described in a parametrized form using Fourier descriptors. The proposed solution has capability to extract moving joints of human body with high accuracy. Further, we assess recognition capability to demonstrate the potency of the approach described.

2 Moving Regions Correspondence

The first problem for automated surveillance is the detection of moving objects in the scene. This is often performed via background subtraction. Moving objects are detected by taking the difference between the current image and background image in pixel by pixel fashion. The approach we used for the segmentation of moving objects, is the adaptive background subtraction proposed by Stauffer and Grimson [13]. A mixture of K (from 3 to 5) Gaussian distributions is used to model the RGB color changes. Since adaptive background subtraction lacks capability to remove shadows, we used the approach described in [7] to evaluate whether a foreground pixel corresponds to shadow based on brightness and color distortion.

The next step is to track detected moving objects over the sequence of frames. Tracking multiple objects is a challenging task and requires a robust region correspondence algorithm to handle occlusion, entry and exit of objects. Our approach models moving objects as temporal templates characterized with three extracted features: size, centroid position, and aspect ratio of height to width of the bounding box.

Moving objects are assigned to different layers, such that moving regions which correspond to the same object are allocated to the same layer. Each layer is defined by four parameters $L_i < s_i, a_i, x_i, y_i >$ where i is the layer index. s_i and a_i are the mean values for the sizes and aspect ratios of objects belonging to the i^{th} layer respectively. x_i and y_i are the predicted centroid position of the object in the next frame. The centroid position is estimated linearly via computing the velocity V_i as the spatial difference of the last two previous positions.

First, every moving object is allocated to a new layer L_i whereby, we update the layer parameters. Velocity is assumed zero at startup. In the next frame, we create a list containing the existing layers. Newly detected blobs are ordered according to their size. Starting from larger blobs, we take every object and search for the ideal layer in the list to assign this object to, if an allocation is made, the chosen layer is removed from the list. Because the Euclidean distance metric's allows dimensions with larger scales and variances to dominate the feature space, the cost function for allocating moving objects to their corresponding layers is based on the Mahalanobis metric measure using equation (1). The use of the Mahalanobis metric alleviates most of the Euclidean metric limitations, as it accounts automatically for the scaling of coordinate axes in the feature space.

$$C = \sqrt{(f_l - f_c)^T \, \Sigma^{-1} \, (f_l - f_c)} \tag{1}$$

where f_l and f_c are the features vectors for the layer and candidate object respectively. Σ is the covariance of the training set. We have defined a number of constraints to the allocation criteria to handle occlusion and entry and exit of moving objects into the monitored scene. A candidate will be allocated to layer L_i only if:

- The layer L_i has the smallest cost value C.
- $|s_i - S| < 3\sigma_i$ where S is the size of the candidate object, σ_i and s_i are the standard deviation and mean values of objects' sizes belonging to the i^{th} layer.
- If a candidate object does not have a corresponding layer, it will be allocated to layer L_i, if the object is mostly contained within the bounding box of L_i.

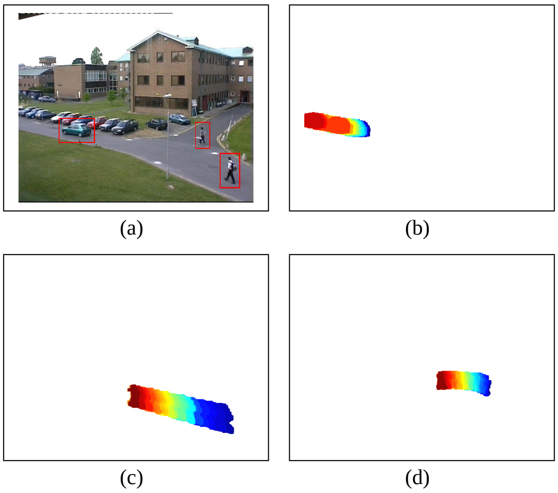


Figure 1: Tracking Multiple Objects (a) Frame from a Video Sequence. (b) Layer 1: Moving Vehicle (c) Layer 2: Walking Person (d) Layer 3: Walking Person

If an object is not assigned to one of the existing layers, a new layer is created for this new object. To cope with the appearance of uninteresting regions such as background clutter, we define a threshold $T = 5$, if a layer has a life span of T frames or less, then this layer is ignored and deleted. Figure (1) shows the results of tracking multiple moving objects

3 Object Classifications

Our method classifies moving objects into either: i) person, ii) group of people or iii) undefined objects (such as vehicles). The classification procedure is based on the analysis of gait motion. During the strike phase, the foot of the striking leg stays at the same position for half a gait cycle, whilst the rest of the human body moves forward. To avoid the drawbacks of approaches that employ color intensity as the low-level features for the purpose of classification, we used corners instead and apply the Harris corner detector on every frame t. For every moving object belonging to the i_{th} layer, we take the corresponding corners of this object, Afterwards, corner images corresponding to the same object are accumulated together using (2)

$$C_i = \sum_{t=1}^{N} (\ H(I_t)\ \wedge\ L_{i,t}\ ) \tag{2}$$

Where H is the output of the Harris corner detector, I_t is original image at frame t, $L_{i,t}$ is i^{th} layer. $\wedge$ is the logical conjunction operator. Because the striking foot is stabilized for half a gait cycle, as result, a dense area of corners is detected in the region where the leg strikes the ground. In order to locate these areas, we have estimated a measure for density of proximity. The value of proximity at point p is dependent on the number of corners within the region R_p and their corresponding distances from p. R_p is assumed to be a square area with centre p, and radius of r that is determined as the ratio of total image points to the total of corners in C_i which is about 10. We have first computed proximity value d_p of corners for all regions R_p in C_i using equation (3). This is an iterative process starting from a radius r. The process then iterates to accumulate proximity values of corners for point p.

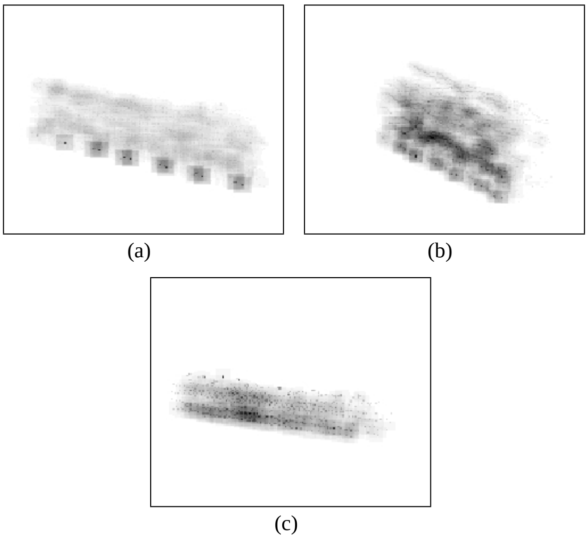
$$\begin{cases} d_p^r = \frac{N_r}{r} \\ d_p^i = d_p^{i+1} + \frac{N_i}{i} \end{cases} \tag{3}$$

where d_p^i is the proximity value for rings of radius i away from the centre p, and N_i is the number of corners which are of distance i from the centre, rings are single pixel wide. Afterwards, we accumulate all the densities for the subregions R_p for all points p into one image to produce the corners proximity image using (4).

$$D = \sum_{p \in Corners} shift(d_p) \tag{4}$$

where d_p is the corner proximity value for region R_p. The $shift$ function places the proximity value d_p on a blank image at the position p. An output of the corner proximity image for different moving objects is shown in Figure (2). The corner proximity image shows brighter peaks at the heel strikes areas. A similar algorithm to [5] is used to derive the positions of the peaks as local maxima.

Clearly, the corner proximity image for walking subjects has larger peaks at the bottom as legs have static periods. Furthermore, since gait is periodic, the stride length should be the same for different gait cycles, therefore the standard deviation of distances between two close strikes should tend to zero. For the classification of moving objects, we define a feature vector $<\ b, \sigma, \alpha\ >$ where b is the proportion of the lower part of the proximity image, and σ is the standard deviation value of distances between two successive peaks. α is the aspect ration of height to width of the bounding box.


(a)

(b)

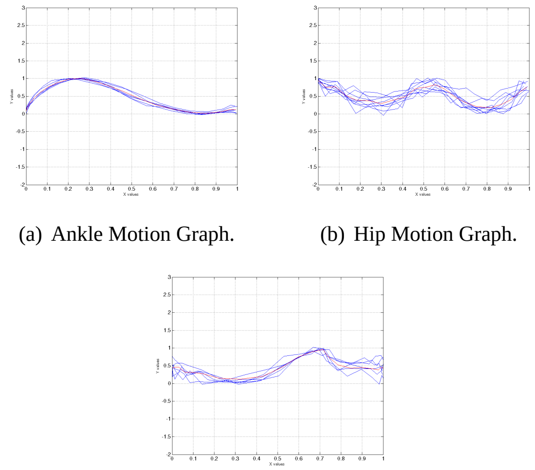
(c)

Figure 2: The Corners Proximity Images for : (a) Single Walking Person, (b) Group of People, (c) Moving Vehicle

4 Gait Recognition

In order to recognize people by their gait, a model-based approach is proposed to extract the joints trajectories of walking people. Although, the Fourier series is the most accurate way for modelling gait motion, most previous methods adopted simple models [3, 17] to extract gait angular motion via evidence gathering with a few parameters. This is mainly due to complexity and computational cost. In our method, human gait is modelled using the Fourier series. The heel strike data were used to reduce the number of parameters and therefore reduce significantly the computational cost. Model templates which describe joints' motions are constructed from the analysis of manually labelled data of 30 video sequences. Figure (3) shows the joint motion path between two successive heel strikes for the ankle, hip and knee. The mean patterns for gait motion are represented using elliptic Fourier Descriptors [2], where the Fourier series is based on a curve f expressed by a complex parametric form as shown in equation (5):

$$f(t) = x(t) + jy(t) \tag{5}$$



(a) Ankle Motion Graph.

(b) Hip Motion Graph.

(c) Knee Motion Graph.

Figure 3: Motion Analysis of the Joints.

where $x(t)$ and $y(t)$ are approximated via the Fourier summation by n terms as shown in equation (6) :

$$\begin{aligned} x(t) &= \sum_{k=1}^{n} a_{x_k} cos(kt) + b_{x_k} sin(kt) \\ y(t) &= \sum_{k=1}^{n} a_{y_k} cos(kt) + b_{y_k} sin(kt) \end{aligned} \tag{6}$$

where a_{x_k},a_{y_k}, b_{x_k} and b_{y_k} are the set of the elliptic phasors which can be computed by a Riemann summation [2]. In order to obtain a flexible motion model sufficient to describe gait motion, spatial model templates are created via representing f in a parametrized form by applying appearance transformations (rotation, scaling and translation). Spatial model template M describing gait motion is described in equation (7):

$$\left\{ \begin{array}{l} M = T + R_\alpha \left(s_x x(t) + s_y y(t) i \right) \\ T = a_0 + b_0 i \\ R_\alpha = \cos(\alpha) + \sin(\alpha) i \end{array} \right. \tag{7}$$

where T is the translation transform whose vector is (a_0, b_0). R is the rotation transform of angle α. s_x and s_y are the scaling factors across the horizontal and vertical axes respectively. Evidence gathering [3] i.e. Hough Transform is used as a first stage to extract the spatial motion path of the joints using model templates. A 5-dimensional accumulator is needed to to store the votes for the set of parameters $< a_0, b_0, \alpha, s_x, s_y >$. The matching process is carried out across the frames sequence for one gait cycle using the detected corners. However, due to the large number of parameters, the Hough Transform can be computationally complex and intensive. Using the heel strikes data extracted earlier, the search for the ankle motion model is reduced to only one parameter s_y, while it is reduced to two parameters b_0 and s_y for the case of the hip and knee motion models.

Afterwards, having extracted the spatial motion models for the joints, local search is run within every frame to estimate the positions of the joints using the temporal models as described in equation (8). Temporal models describe the horizontal displacement of moving joints with respect to time.

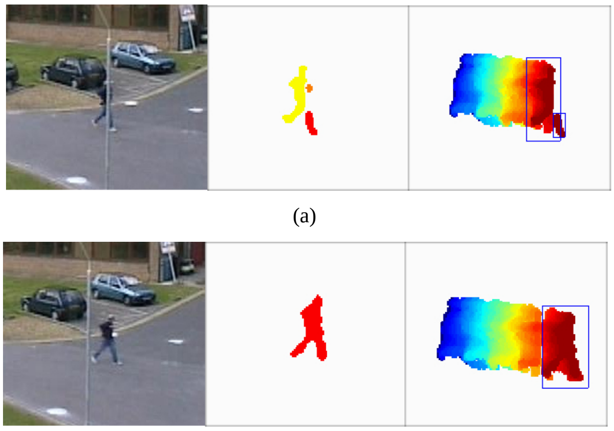
$$\left\{ \begin{array}{l} x_t = s_x \times V \left(\frac{t}{N} \right) + x_{si} \\ y_t = S(x_t) \end{array} \right. \tag{8}$$

where t is the frame index, N is the number of frames. x_{si} is the x coordinate of the heel strike s_i and V is the horizontal displacement model function for the joint. The function S is the joint spatial motion model extracted during the first stage. The joint position is determined as the centroid of the corners within the small region whose centre is (x_t, y_t) .

In order to derive the gait signature to uniquely identify people, we have fused both static and dynamic gait features, as combining both features has been proved to be more efficient. Dynamic features are acquired as the phase-weighted magnitude of the Fourier frequencies from the angular motion of the hip and thigh.

5 Experimental Results

To demonstrate the efficacy of our method for automated visual surveillance, the system has been extensively evaluated on a variety of scenarios and conditions. The proposed algorithm for people detection is applied on a set of videos provided by PETS 2001 compressed in JPEG format. Videos are filmed in an unconstrained outdoor environment with walking people and moving vehicles. The size of video frames is reduced to 384x288.



(a)

(b)

Figure 4: Experimental Results for Handling Occlusion : (a) Tracking During Occlusion. (b) Tracking recovery results of the walking subject after occlusion.

Moving objects are tracked successfully during their life span in the monitored scene. Furthermore, the system can handle occlusion efficiently, and reallocate the occluded

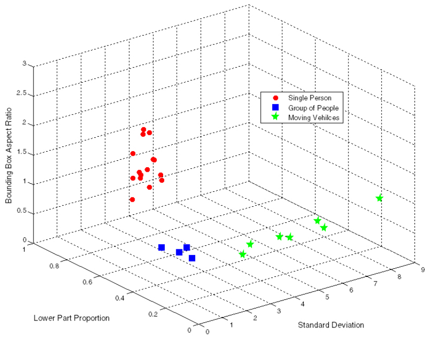
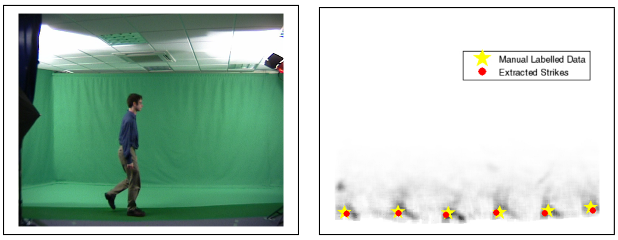


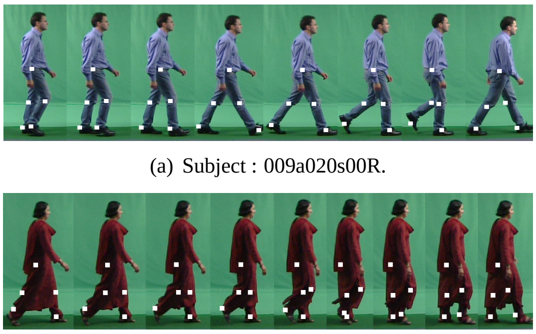
Figure 5: Feature Space

object to the correct layer when occlusion vanishes. The appearance of uninteresting regions such background clutter are ignored by the system. Figure (4) shows a walking person who is partially occluded by a lamppost. The moving subject is detected as multiple separate moving regions by the foreground segmentation process as shown in Figure 4(a). The tracking algorithm successfully allocates the detected blobs to the layer corresponding to the walking subject, since they are not allocated to existing layers and are mostly contained within the predicted bounding box of the walking subject. After occlusion, tracking is carried out successfully as shown in Figure 4(b). To verify the effectiveness of our approach to classify moving objects by their gait pattern, we have carried out a number of experiments on the whole PETS video data containing a total of 26 moving objects consisting of: 15 single walking subjects, 4 groups of people and 7 moving vehicles. The leave-one-out validation rule is used to evaluate the classification performance using the k-nearest neighbor classifier. The system was able to discriminate between single walking people, a group of people and vehicles efficiently using the proposed features and achieved a detection rate of %100. The feature vectors for the moving objects are projected into the feature space shown in Figure (5). Although, the classification results were promising, we have conducted further experiments to confirm the robustness of the proposed method for extracting the heel strikes, we have run the algorithm on a set of 120 video sequences from the SOTON database. The mean error for the positions of the strikes extracted by the algorithm compared to manually labelled strikes is %0.52 of the person height. The error is measured by Euclidean distance normalized to a percentage of the person's height. Figure (6) shows the results of heel strike extraction by the described method compared with the data labelled manually for one video sequence and it can be observed that the match is indeed close.



(a) (b)

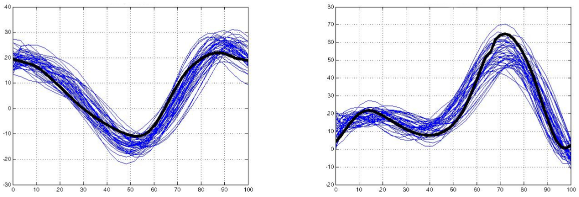
Figure 6: Experimental Results for Heel Strikes Extraction: (a) Walking subject. (b) Extracted strikes compared with data manually labelled



(a) Subject : 009a020s00R.

(b) Subject : 012a031s00R.

Figure 7: Joints Extraction for Indoor Data.



(a) (b)

Figure 8: Gait Angular Motion during one Gait Cycle: (a) Hip, (b) Knee

To evaluate the recognition performance of our algorithm, we have extracted the positions of for the ankle, knee and hip joints as shown in Figure (7) for a set of 120 video sequences containing 20 different subjects with 6 sequences for every subject. Furthermore, The algorithm is tested on a subject wearing Indian clothes which covered the legs, the joints positions are extracted successfully as shown in Figure 7(b) which reveals the potential of this approach to handle self-occlusion. The angular motions for the hip and thigh are determined from the joints' trajectories as shown in Figure (8) and it can observed that the results obtained via this approach are consistent with the biomechanical data by Winter [15] shown in bold. We have fused both dynamic and static gait features to yield a feature vector consisting of 48 features. Static features include the body height, stride and heights of different body parts whilst dynamic features are mainly the phase-weighted magnitudes of the Fourier frequencies for the hip and knee angular motions. The gait signature is derived using the adaptive forward floating search algorithm via selecting the features with higher discriminability values.

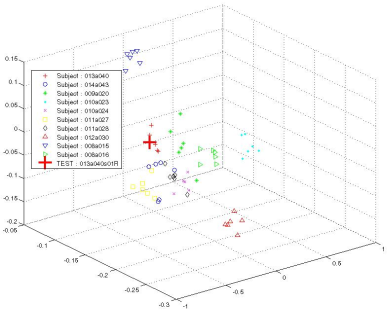


Figure 9: Canonical Space Projection

The gait signatures for every sequence are projected into the canonical space whereby the k-nearest neighbor rule is used for classification. The system achieved a correct classification rate of %92 using the leave-one-out validation rule. This is shown in Figure (9) with only 10 of the 20 subjects being projected.

6 Conclusions

We have proposed a new method to classify moving objects and recognize people for automated visual surveillance by their gait. Multiple objects are tracked successfully through the use of shape-based parameters to allocate them to different layers. Problems encountered during tracking such as background clutter, appearance of uninteresting objects and entry and exit of objects are handled efficiently. Finally moving regions are classified into either a single walking person, group of people or an undefined object such as vehicle. We have explored an alternative technique for walking people detection based on their gait motion. The experimental results confirm the robustness of our method to discriminate between moving objects with a detection rate of %100. For people recognition, a new model-based method is described to extract the joints positions via an evidence gathering technique. Spatial model templates for human motion are described in a parametrized form using the Fourier descriptor. The proposed solution has achieved a classification rate of %92 for people recognition. The model-based is suited to more generalized deplyment and this will be the focus for future work.

References

[1] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Nonrigid motion analysis: Articulated and elastic motion. *CVIU*, 70(2):142–156, 1998.

[2] A. S. Aguado, M. S. Nixon, and M. E. Montiel. Parameterising arbitrary shapes via fourier descriptors for evidence-gathering extraction. *CVGIP: Image Understanding*, 2:547–51, 1998.

[3] D. Cunado, M. S. Nixon, and J. N. Carter. Automatic extraction and description of human gait models for recognition purposes. *CVIU*, 90(1):1–41, 2003.

[4] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE TPAMI*, 22(8):781–796, 2003.

[5] H. Fujiyoshi, A. J. Lipton, and T. Kanade. Real-time human motion analysis by image skeletonization. *IEICE Trans on Information and System*, pages 113–120, 2004.

[6] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: real-time surveillance of people and their activities. *IEEE TPAMI*, 22(8):809–830, 2000.

[7] T. Horprasert, D. Harwood, and L. Davis. A statistical approach for real-time robust background subtraction and shadow detection. *In Proc IEEE ICCV*, pages 1–19, 1999.

[8] N. Huazhong, T. Tan, L. Wang, and W. Hu. People tracking based on motion model and motion constraints with automatic initialization. *Pattern Recognition*, 37(7):1423–1440, 2004.

[9] I. A. Karaulova, P. M. Hall, and A. D. Marshall. A hierarchical model of dynamics for tracking people with a single video camera. In *Proc of the 11th BMVC*, pages 262–352, Septemeber 2000.

[10] S. Kurakake and R. Nevatia. Description and tracking of moving articulated objects. In *Proceedings. 11th IAPR ICPR*, volume 1, pages 491–495, Octobor 1992.

[11] A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target classification and tracking from real-time video. *4th IEEE Workshop on Application of Computer Vision*, 1998.

[12] A. Shio and J. Sklansky. Segmentation of people in motion. In *IEEE Workshop on Visual Motion*, volume 2, pages 325–332, Octobor 1991.

[13] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE TPAMI*, 22(8):747–757, 2000.

[14] L. A. Wang, W. M. Hu, and T. N. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.

[15] D. A. Winter. *The biomechanics and motor control of human gait: Normal, elderly and pathological.* 2nd Eds, Waterloo Biomechanics, 1991.

[16] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE TPAMI*, 19(7):780–785, 1997.

[17] C. Y. Yam, M. S. Nixon, and J. N. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5), 2004.