Gait-Based Pedestrian Detection for Automated Surveillance

Imed Bouchrika and Mark S Nixon

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

Abstract. In this paper, we explore a new approach for walking pedestrian detection in an unconstrained outdoor environment. The proposed algorithm is based on gait motion as the rhythm of the footprint pattern of walking people is considered the stable and characteristic feature for the classification of moving objects. The novelty of our approach is motivated by the latest research for people identification using gait. The experimental results confirmed the robustness of our method to discriminate between single walking subject, groups of people and vehicles with a detection rate of %100. Furthermore, the results revealed the potential of our method to extend visual surveillance systems to recognize walking people.

Key words: Visual surveillance, motion analysis, people tracking, gait.

1 Introduction

In recent years, automatic visual surveillance has received considerable interest in the computer vision community. This is due to the inability 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. The system should possibly recognize their identities through the use of biometrics technologies.

Detecting and tracking 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 vision-based surveillance system.

Existing surveillance systems are classified into several categories [1] according to their type (single or multiple camera) and their functionality (tracking single , multiple people, etc.). Wren *et al* [2] proposed the PFinder system. It 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 [1] surveillance system employs an appearance model to track people whereby single or group are distinguished using a projection histogram. Each person in the

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

group is located through the tracking of his/her head. Lipton *et al* [3] proposed a real time vision-based system to classify moving objects into either human or vehicle based on the "dispersedness". In his work, people are assumed to have a dispersedness value smaller than vehicles, however shape metrics can vary depending on image size and distance from camera. Furthermore, Gavrila *et al* [14] proposed a method detecting pedestrians from a moving vehicle based on shape matching.

Wang [4] surveyed two type of features used for people detection in surveillance systems: shape-based or motion-based features. The first type relies on the shape of human silhouettes such as dispersedenss [3], 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 [5] described a real time method for measuring periodicity for periodic motion based on self-similarity. Javed *et al* [6] proposed a simple measurement based repeated internal motion.

In this paper, we propose a multi-object tracking method based on features 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 allocation criteria 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 pedestrian detection based on the rhythmic pattern of their gait motion. The novelty of our approach is motivated by the latest research for people identification using gait [7]. 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. In our method, people are detected through the extraction of their heel strikes, whereby stride and cadence can be estimated easily. In contrast to earlier methods [13], our approach does not require the subject to walk in sagittal view, as the gait pattern can be extracted from different viewpoints. This proposed method has a great potential for visual surveillance systems to incorporate a biometric system to recognize people based on the results of [8]. In their work [8], a view-invariant biometric system for people identification is based on the extraction of the stride and cadence of walking people.

This paper is structured as follows: the next section is devoted to the discussion of temporal tracking of detected moving regions. Section 3 describes the proposed method for moving object classification. Finally, the experimental results on a set of processed videos from different databases are drawn in the fourth section.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

2 Foreground Segmentation and Tracking

The first problem for automated surveillance is the detection of moving objects in the scene. This is often performed via some form of background subtraction. Moving objects are detected by taking the difference between the current image and background image in a pixel by pixel fashion. The approach we used for the segmentation of moving objects, is the adaptive background subtraction proposed by Stauffer and Grimson [9]. 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 [10] to evaluate whether a foreground pixel corresponds to shadow based on brightness and color distortion. Morphological operators are used to remove noise produced from foreground segmentation.

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 mainly 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 three 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 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 feature 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.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

– $|s_i - S| < 3\sigma_i$ 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.

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 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 simultaneously. The video scene consists of two subjects walking separately, and one moving vehicle. The three moving objects are tracked successfully during their lifespan via the use of layers. Layers are visualized by overlapping moving objects into one image.

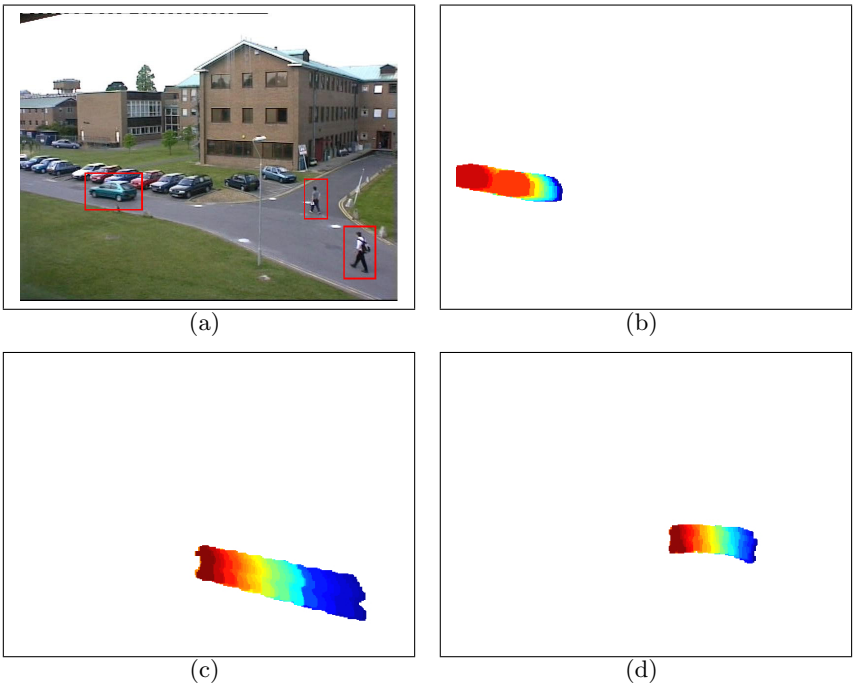


Fig. 1. Tracking Multiple Objects (a) Frame from the Video Sequence. (b) Layer 1: Moving Vehicle (c) Layer 2: Walking Person (d) Layer 3: Walking Person

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

3 Moving Objects Classifications

Our method classifies moving objects into either person, group of people or undefined objects (such as vehicles). The classification procedure is based on the rhythmic pattern of gait motion. Because gait is symmetric and periodic motion, the distances between two close strikes i.e. step lengths should be the same during all gait cycles. This cue is considered the main feature to distinguish walking subjects from other objects.

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. Therefore, if we use a low-level feature extraction method (edges or corners), then a dense region will be accumulated at the heel strike regions. We have chosen to use corners instead of edges, as they maintain enough information to perceive the human motion, in contrast to edges which may cause ambiguity in the extraction process due to the excess data they may contain. Furthermore, a robust vision system based on corner detection can work for low-resolution applications. We have applied the Harris corner detector on every frame t from the video sequence and then accumulated all the corners into one image using equation (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, such that the numeric value zero is considered false, and true otherwise. 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 is an iterative process starting from a radius r. The process then iterates to accumulate proximity values of corners for point p.

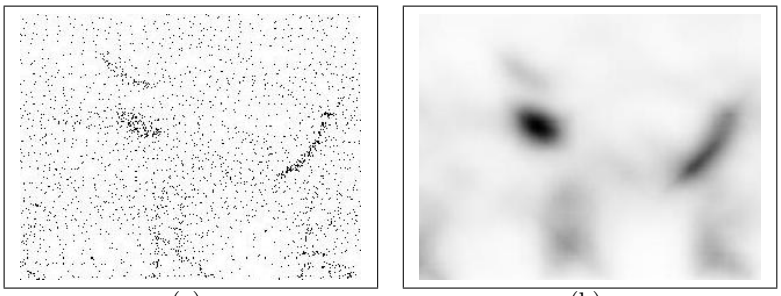
$$\begin{cases} d_p^r = \frac{N_r}{r} \\ d_p^i = d_p^{i+1} + \frac{N_i}{i} \end{cases} \tag{3}$$

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).

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

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

where d_p is the corners 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 for an example image is shown in Figure (2). The input image contains points spread all over the image with a number of dense regions. The resulting image has darker areas which correspond to the crowded regions in the input image.



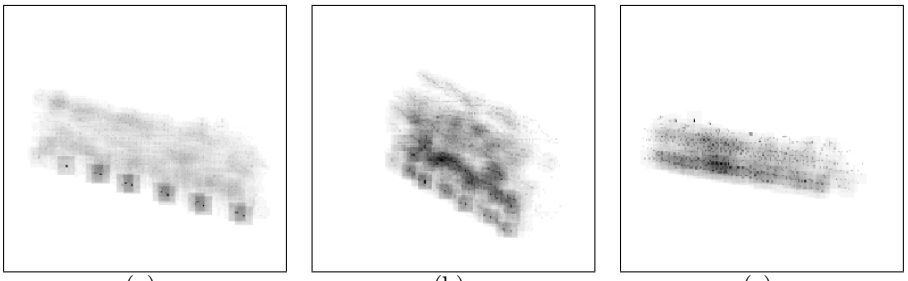
(a) (b)

Fig. 2. Example Results for the Corner Proximity Measure: (a) Input Image, (b) Corner Proximity Image

For the application context of this research, we have applied the corner proximity measure on different moving objects being captured using a surveillance camera. Moving objects include single walking person, a group of people and a vehicle. The results are shown in Figure (3). Clearly, the corner proximity image for single walking subject shown has darker spots being detected at the bottom part of the image as the leg strikes the ground. Moreover, These darker regions are observed to have mostly the same level of darkness with consistent distance between two consecutive regions. On the other side, the proximity image for moving vehicle has an almost flat pattern with arbitrary peaks located in the image. A similar algorithm to [11] is used to derive the positions of the peaks as local maxima.

Clearly, the corners 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 i.e. peaks should tend to zero. For the classification of moving objects, we define the feature vector $< \sigma, b, \alpha >$ where σ is the standard deviation value of distances between two successive peaks extracted from the corner proximity image. The value of σ should tend to zero for walking subjects and gets larger for moving vehicles. b is the proportion of the lower part of the proximity image which should be larger for both single subject and a group of people as most peaks are located at the lower side of the proximity image. α is the aspect ratio of height to width of the bounding box. This value is mainly used as discriminative feature between single subject and

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de



(a) (b) (c)

Fig. 3. The Corners Proximity Images for : (a) Single Walking Person, (b) Group of People, (c) Moving Vehicle

group of people. The k-nearest neighbor rule is used to classify moving objects based on the feature vector.

4 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 is applied on a set of four videos provided by PETS 2001 of which Figure (1) is an example. Videos are filmed in an unconstrained outdoor environment with walking people and moving vehicles. The size of of video frames is reduced to 384x288.

The presented algorithm for tracking multiple objects is tested on the set of video sequences. Moving objects are tracked successfully during their life span in the monitored scene. Furthermore, the system can handle occlusion efficiently, and reallocate the occluded 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). 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 Figures 4(d), 4(e) and 4(f).

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 video data containing a total of 26 moving objects. The leave-one-out validation rule is used to evaluate to performance of the classification 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 classification rate of %100. The results of the classification are detailed in Table

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

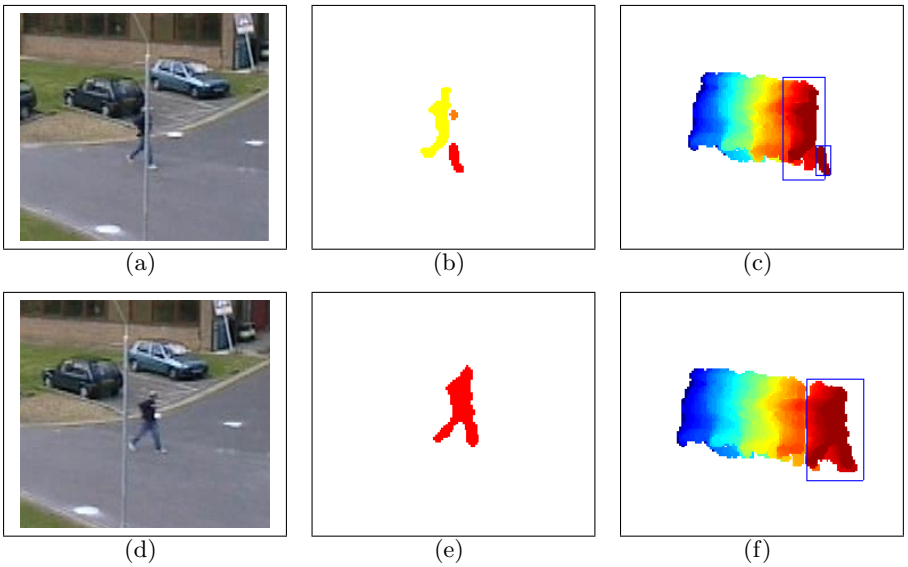


Fig. 4. Experimental Results for Handling Occlusion : (a) Walking subject being occluded. (b) Foreground segmentation of a. (c) Allocation of moving regions into the layer corresponding the walking person .(d) Walking subject after occlusion. (e) Foreground segmentation of d. (f) Tracking recovery results of the walking subject after occlusion.

(1). The feature vectors for the moving objects are projected into the feature space shown in Figure (5) whose dimensions are: bounding box aspect ratio, standard deviation of distances between two successive peaks and the lower part proportion of the corner proximity image. This shows clearly that the standard deviation, i.e. gait periodicity, is a strong cue to distinguish between walking people and vehicle.

Table 1. Moving Objects Classification Results

Type of Object	Number of instances	Instances Correctly Classified
Single Person	15	15
Group of People	4	4
Moving Vehicles	7	7

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 also run the algorithm on a set of 100 different subjects from the SOTON database [12]. The proposed method extracted successfully %99.2 of the strikes from a total of 514 strikes. The mean error for the positions

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

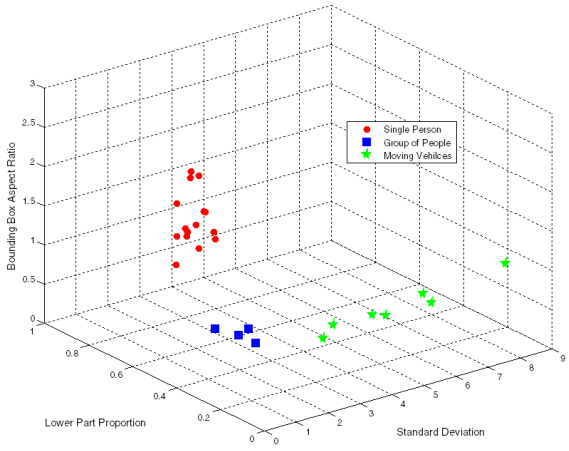


Fig. 5. Feature Space

of 65 strikes extracted by the algorithm compared to strikes manually labelled is %0.52 of the person's height. The error is measured by Euclidean distance between the two strikes, normalized to a percentage of the person's height as the is the most reliably extracted norm. Figure (6) shows the results of heel strike extraction by the described method compared with the data obtained manually for one video sequence.

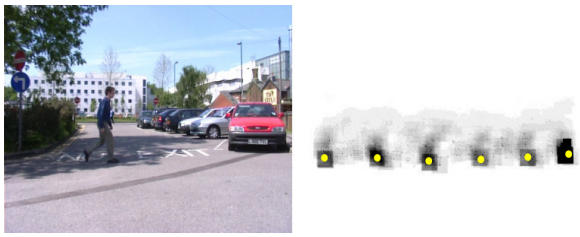


Fig. 6. Experimental Results for Heel Strikes Extraction.

5 Conclusions and Future Work

We have proposed a new method to classify moving objects for automated visual surveillance. Multiple objects are tracked successfully through the use of shape-based parameters to allocate them to different layers. Problems encountered during tracking such us background clutter, appearance of uninteresting objects

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

and entry and exit of objects are handled efficiently. Finally moving regions are classified into either single walking person, group of people or undefined object such as vehicle. In contrast to approaches that employ shape-based parameters for classification, we have explored an alternative technique for walking people detection based the rhythmic pattern of their gait motion. The experimental results confirmed the robustness of our method to discriminate between single walking person, group of people and vehicle with a classification rate of %100.

For future research work, our proposed method for detecting people based on their gait will be extended to recognize walking people based their cadence and stride.

References

1. Haritaoglu, I., Harwood, D., Davis, L. S.: W4: real-time surveillance of people and their activities IEEE TPAMI (2000) 809–830
2. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-Time Tracking of the Human Body, IEEE TPAMI (1997) 780–785
3. Lipton, A. J., Fujiyoshi, H., Patil, R. S.: Moving target classification and tracking from real-time video , Proceedings of IEEE Workshop on Application of Computer Vision (1998) 8–14
4. Hu, W., Tan, T., Wang, L., Maybank S.: A Survey on visual surveillance of object motion and behaviors, IEEE TSMC (2003) 585–601
5. Cutler, R., Davis, L. S.: Robust real-time periodic motion detection, analysis, and applications, IEEE TPAMI (2003) 781–796
6. Javed, O., Shah, M.: Tracking and object classification for automated surveillance, In Proc of the Seventh European Conference on Computer Vision (2002) 343-357
7. Nixon, M. S. and Carter, J. N.: On gait as a biometric: progress and prospects, In Proceedings of Proc. EUSIPCO (2004)
8. BenAbdelkader, C., Cutler, R., Davis, L.: Stride and cadence as a biometric in automatic person identification and verification, In Proc of the 5th International Conference on Automatic Face and Gesture Recognition (2002)
9. Stauffer, C., Grimson, W.: Learning patterns of activity using real-time tracking, IEEE TPAMI (1999) 246–252
10. Horprasert, T., Harwood, D., Davis, L.: A statistical approach for real-time robust background subtraction and shadow detection In Proc IEEE ICCV (1999) 1-19
11. Fujiyoshi, H., Lipton, A. J. Kanade, T.: Real-time human motion analysis by image skeletonization IEICE Trans on Information and Systems (2004) 113-120
12. Shutler, J. D., Grant, M. G., Nixon, M. S., Carter, J. N.: On a large sequence based human gait database In Proc of Recent Advances in Soft Computing (2002) 6671
13. Heisele, B., Woehler, C.: Motion-based recognition of pedestrians In Proc. Fourteenth International Conference on Pattern Recognition (1998)
14. Gavrila, D. M.: Pedestrian detection from a moving vehicle Proc. 6th European Conf. on Computer Vision (2000)

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de