Unsupervised Motion Classification by means of Efficient Feature Selection and Tracking

Angel D. Sappa† Niki Aifanti‡ Sotiris Malassiotis‡ Michael G. Strintzis‡

Computer Vision Center†
Edifici O, Campus UAB
08193 Bellaterra - Barcelona, Spain
angel.sappa@cvc.uab.es

Informatics & Telematics Institute‡
1st Km Thermi-Panorama Road
Thermi-Thessaloniki, Greece
{naif, malasiot}@iti.gr strintzi@eng.auth.gr

ABSTRACT

This paper presents an efficient technique for human motion recognition; in particular, it is focused on labeling a movement as a walking or running displacement, which are the most frequent type of locomotion. The proposed technique consists of two stages and is based on the study of feature points' trajectories. The first stage detects peaks and valleys of points' trajectories, which are used on the second stage to discern whether the movement corresponds to a walking or a running displacement. Prior knowledge of human body kinematics structure together with the corresponding motion model are the basis for the motion recognition. Experimental results with different video sequences are presented.

1. INTRODUCTION

High level description of human body displacement is a challenging topic required for many surveillance applications. Generally, several surveillance video cameras exist and human operators are responsible to carefully watch all the recordings. Every day, though, more and more cameras are installed in different environments, so more and more screens need to be permanently surveyed. This boring and tedious work could affect the efficiency of the whole surveillance system. So, in order to make this job efficiently a system that automatically identifies human bodies activities (normal activity, suspicious activity, etc.) is required.

Vision-based human motion modeling approaches usually combine several computer vision processing techniques (e.g. *low level*: video sequence segmentation, object tracking; *high level*: motion prediction, 3D object representation, model fitting, etc.). Different techniques have been proposed to tackle the motion modeling problem. These approaches can be broadly classified into monocular or multi camera approaches.

Delamarre and Faugeras [1] propose a stereoscopic technique, which is able to cope not only with self-occlusions but also with fast movements and poor quality images, using two or more fixed cameras. This approach incorporates physical forces to each rigid part of a kinematics 3D human body model consisting of truncated cones. These forces guide each 3D model's part towards a convergence with the body posture in the image. The model's projections are compared with the silhouettes extracted from the image by means of a novel approach, which combines the Maxwell's demons algorithm with the classical ICP algorithm. Although stereoscopic systems provide us with more information for the scanned scenes, 3D human motion systems with only one camera-view available is the most frequent case.

Motion modeling using monocular image sequences constitutes a complex and challenging problem. Similarly to approach [1], but in a 2D space and assuming a segmented video sequence is given as an input, [2] proposes a system that fits a projected body model with the contour of a segmented image. This boundary matching technique consists of an error minimization between the pose of the

This work has been carried out as part of the ATTEST project (Advanced Three-dimensional TElevision System Technologies, IST-2001-34396). The first author has been supported by *The Ramón y Cajal Program*.

Proceedings of the 2nd Intl Symp 3D Data Processing, Visualization, and Transmission (3DPVT 2004)
0-7695-2223-8/04 $20.00 © 2004 **IEEE**

projected model and the pose of the real body—all in a 2D space. The main disadvantage of this technique is that it finds the correspondence between the projected body parts and the silhouette contour, before starting the matching approach. This means that, it looks for which point of the silhouette contour correspond to a given body part (assuming that the model posture is not initialized). Moreover, this problem is still more difficult to handle in those frames where self-occlusions appear or edges cannot be properly computed.

Differently than the previous approaches, the aspect ratio of the bounding box of the moving silhouette has been used in [3]. This approach is able to cope with both lateral and frontal views. In this case the contour is studied as a whole and body parts do not need to be detected. The aspect ratio is used to encode the pedestrian's walking way. However, although shapes are one of the most important semantic attributes of an image, problems appear in those cases where the pedestrian wears clothes not so tight or carries objects such as a suitcase, handbag or backpack. It is obvious that the carried objects distort the human body silhouette and therefore the aspect ratio of the corresponding bounding box.

Instead of obtaining the exact configuration of the human body, human motion recognition consists in identifying the action performed by a moving person. Most of the proposed techniques focus on identifying actions belonging to the same category. For example, the objective could be to recognize several aerobic exercises or tennis strokes or some everyday actions such as sitting down, standing up, walking.

Action and interaction recognition such as standing, walking, meeting people and carrying objects, is addressed by [4] and [5]. A real-time tracking system, which is based on outdoor monocular grayscale images taken from a stationary visible or infrared camera, is introduced. Grayscale textural appearance and shape information of a person are combined to a textural temporal template, which is an extension to the temporal templates defined by [6]. A novel approach for the identification of human actions in an office (entering the room, using a computer, picking up the phone, etc.) is presented in [7]. The novelty of this approach consists in using prior knowledge about the layout of the room. Action identification is modeled by a state machine consisting of various states and transitions between them. The performance of this system is affected if the skin area of the face is occluded, if two people get too close and if prior knowledge is not sufficient.

In this paper a new approach to the problem of human motion recognition is presented. The main idea is to find a particular kinematics configuration throughout the frames of the given video sequence, and then to use the extracted information in order to decide whether the motion is a walking or a running displacement. The proposed approach consists of two stages. The first stage focuses on motion detection (feature point selection and tracking), while the second stage identifies if the movement corresponds to a walking or running displacement. Walking and running are cyclic motions that consist of the *synchronized* movements of legs and arms. Although each person walks/runs with a particular style, there is a set of curves, obtained from anthropometric studies ([8], [9]), which represents human walking/running. This work is based on the observation that, there are two instants in a walking or a running cycle where every human body structure achieves the same configuration. These instants correspond to peaks and valleys of the feature points' trajectories.

The outline of this work is as follows. Next, a brief description of the 3D body model used to represent the person's movement is presented. The proposed technique is described in section 2. Finally, experimental results using different video sequences are presented in section 3. Conclusions and further improvements are given in section 4.

2. Body Modeling

In this work, similarly to [10], an articulated structure defined by 16 links (superquadrics) is used. However, in order to reduce the complexity, the motion model is simplified to 12 DOF. This simplification assumes that in walking and running, legs' and arms' movements are contained in parallel planes (see illustration in Fig. 1(*left*)). Hence, the final model is defined by two DOF for each arm and leg and four for the torso (three for the position plus one for the orientation). The movements of the limbs are based on a hierarchical approach using Euler angles. The body posture is synthesized by concatenating the transformation matrices associated with the joints, starting from the torso. A complete mathematical description of this volumetric model can be found in [10].

3. Motion Detection

At this stage body motion is detected by studying a selected set of feature points, which are representative of the body movement, instead of considering all body's points. It consists of two steps; initially feature points are selected over the first frame. Then, these points are tracked over the whole video sequence. Although this stage is focussed on motion detection, feature point trajectories, together with prior knowledge of human body movement, are rich sources of information to analyze the kind of movement performed by the body.

3.1. Feature Point Selection

In this work, feature points are used to capture human body movements and they are selected by using a corner detector

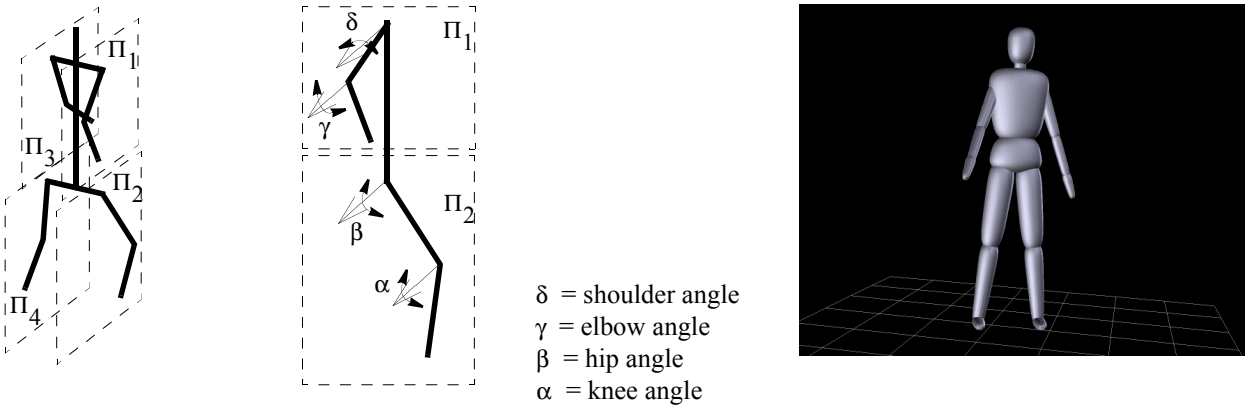

Fig. 1. (*left*) Articulated structure used to model walking and running sequences—legs' and arms' movements are performed on a plane. (right) Illustration of a 22 DOF model built with superquadrics.

algorithm. Let $I(x, y)$ be the first frame of a given video sequence. Then, a pixel (x, y) is a corner feature if at all pixels in a window W_S around (x, y) the smallest singular value of $\boldsymbol{G}$ is bigger than a predefined σ; in the current implementation W_S was set to *5x5* and $\sigma = 0.05$. G is defined as:

$$G = \begin{bmatrix} \Sigma I_x{}^2 & \Sigma I_x I_y \\ \Sigma I_x I_y & \Sigma I_y{}^2 \end{bmatrix}$$

and (I_x, I_y) are the gradients obtained by convolving the image I with the derivatives of a pair of Gaussian filters. More details about corner detection can be found in [11]. Assuming that at the beginning there is no information about the pedestrian's position in the given frame, and in order to enforce a homogeneous feature sampling, input frames are partitioned into 4 regular tiles (*2x2* regions of *120x160* pixels each in the illustration presented in Fig. 2).

3.2. Feature Point Tracking

After selecting a set of feature points and setting a tracking window W_T (*5x5* in the current implementation) an iterative feature tracking algorithm has been used [11]. Assuming a small interframe motion, feature points are tracked by minimizing the sum of squared differences between two consecutive frames.

Points, lying on the head or shoulders, are the best candidates to satisfy the aforementioned assumption. Most of the other points (e.g. points over the legs, arms or hands, are missed after a couple of frames). Fig. 2(*left*) illustrates feature points detected in the first frame of a video sequence defined by 40 frames. Fig. 2*(right)* depicts the trajectories of the feature points when all frames are considered and static feature points are removed. In the current implementation only a single feature point's trajectory was used; further improvements could be to merge feature points' trajectories in order to generate a more robust approach.

4. Motion Analysis

The outcome of the previous stage is the trajectory of a feature point consisting of peaks and valleys. This trajectory, in some way, encodes part of the information required to identify the type of locomotion. The proposed technique consists of two steps. Firstly, the frames corresponding to peaks and valleys are detected—they are called key frames, and then the body posture that best matches the image silhouette is computed. These steps are explained below.

Initially, the first-order derivative of the curve is computed to find peaks' and valleys' positions by seeking the positive-to-negative zero-crossing points. The peaks correspond to those frames where the body structure reaches the maximum height, while the valleys correspond to those frames where the body structure reaches the minimum height. In case the person is walking, a peak occurs in that moment of the half walking cycle when the hip angles are minimum and a valley when the hip angles are maximum. On the contrary, if a running activity is being performed, peaks correspond to maximum hip angles, while valleys to minimum hip angles. These four cases are illustrated in Fig. 3.

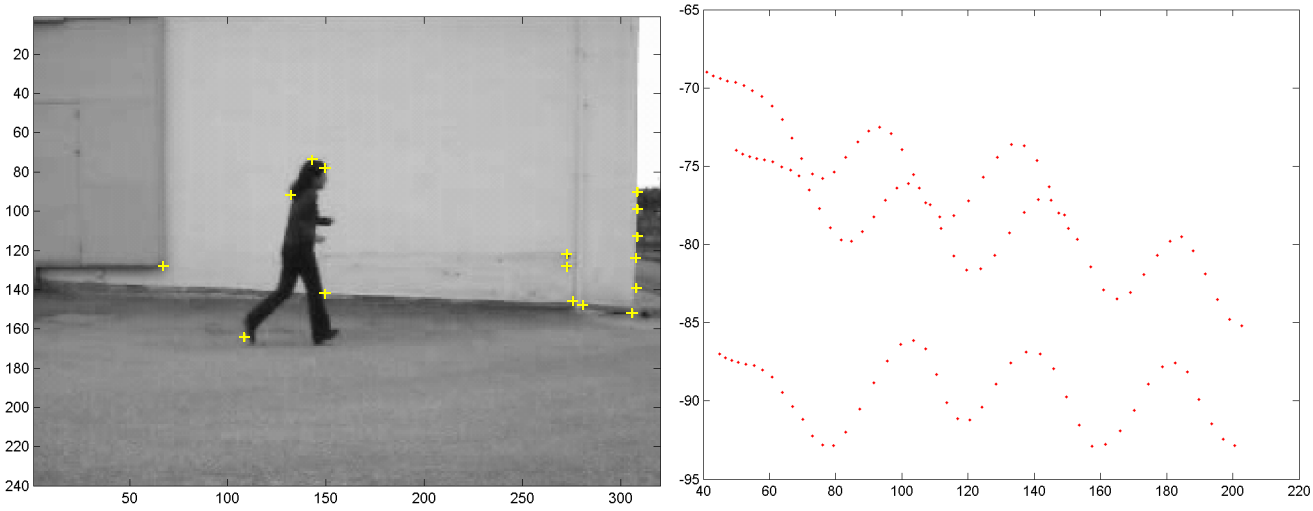

Fig. 2. (*left*) Input frames of a running video sequence (crosses correspond to detected feature points). (*right*) Trajectories of moving feature points.

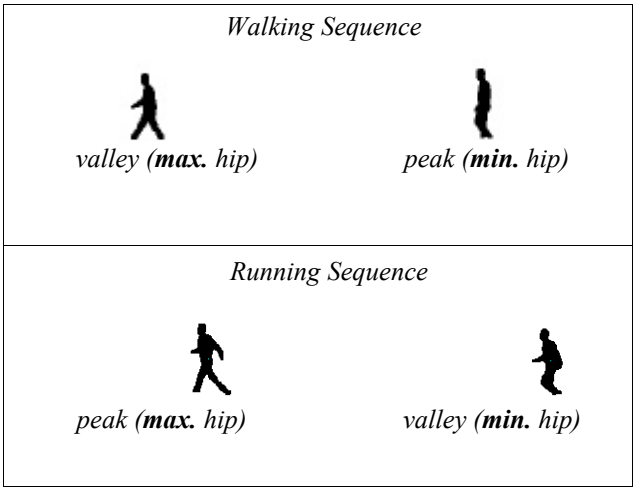

Fig. 3. Illustration of the four postures corresponding to peaks and valleys in walking and running sequences

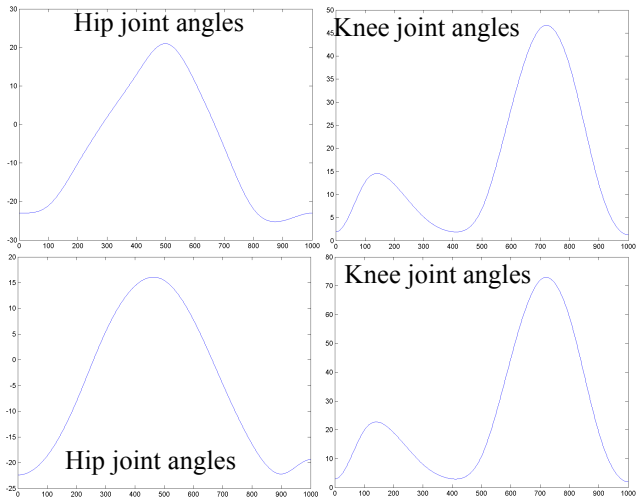

Fig. 4. (*top*) Walking motion curves. (*bottom*) The corresponding running motion curves. Both computed according to [9].

Previous works on walking and running recognition emphasize the existence of periods of *double support* (walking) or *double float* (running). For walking there exists a period where both feet are in contact with the ground (double support), whereas for running, there exists a period where both feet are not in contact with the ground (double float). In addition, walking and running motion model can be used—similar shapes but different scales. In the current work, motion recognition is based on the study of body posture at the key frames of the sequence. Image silhouettes, at each key frame, are computed by means of a background subtraction algorithm ([12]). Fig. 6(*right*) presents two illustrations of body silhouettes, one corresponding to a walking sequence and the other to a running. For each image silhouette two registration errors are computed; one that derives from a projected model with maximum hip angles and another one that derives from a projected model with minimum hip angles (hip and knee angles were computed according to [9], see Fig. 4). The model that corresponds to the lowest registration error and the type of the key frame—peak or valley—are used to identify the type of movement.

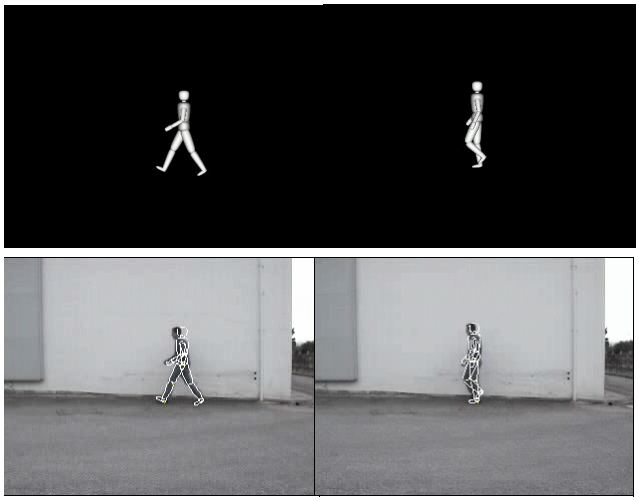

Fig. 5. (*top*) 3D model postures, computed from Fig. 4(*top*), for a valley and peak feature point's trajectory. (*bottom*) Projected model used to compute the *RQI*.

In order to estimate the registration errors, an error measurement (registration quality index: *RQI*) is introduced. The proposed *RQI* measures the quality of the matching between the projected 3D model and the corresponding body silhouette computed from the background subtraction algorithm: $RQI = overlappedArea / totalArea$; where total area consists of the surface of the projected 3D model plus the surface of the body silhouette less the overlapped area, while the overlapped area is defined by the overlap of these two surfaces. Fig. 5(*top*) shows the two 3D model associated with a valley and a peak of the feature points' trajectories presented in Fig. 6(*bottom-middle*). Boundaries of the projected 3D models can be appreciated in Fig. 5(*bottom*); these projected models, together with the body's silhouettes (Fig. 3) are used to compute the corresponding registration errors—*RQI*.

5. EXPERIMENTAL RESULTS

The proposed technique has been tested with different outdoor video sequences containing walking and running movements. The video sequence used as an illustration in Fig. 2 consists of 40 frames of *240x320* pixels each, which has been segmented using the technique presented in [12]. Feature points' trajectories are presented in Fig. 2(*right*). After registering peaks and valleys with a 3D projected model with maximum and minimum hip angles (Fig. 5(*top*)), the computed *RQI* values are presented in Table 1 and indicate that the movement corresponds to a running sequence.

Despite that the trajectory is not orthogonal to the camera direction, the proposed technique is able to recognize

Table 1: Running Sequence

RQI (Fig. 3)	Max. Hip Angle	Min. Hip Angle
Peaks	**0.4283**	0.4242
Valleys	0.4260	**0.4982**

the activity. The RQI values are computed as an average of all peaks and valleys when a maximum and minimum hip angle postures were considered.

Fig. 6 shows a running and a walking displacement. The first video sequence is defined by 40 frames of *240x320* pixels each, while the second is defined by 80 frames of *240x320* pixels each. The *RQI* values associated with each sequence are summarized in Table 1 and Table 2. The first column's value indicates the *RQI* when a model with a maximum hip angle (Fig. 5(*top-left*)) is used; while the second corresponds to the RQI values when the model presented in Fig. 5(*top-right*) (minimum hip angle) is used. As it was expected, a walking movement is identified by maximum hip angles in valleys and minimum hip angles in peaks. On the contrary, a running movement is defined by maximum hip angles in peaks and minimum hip angles in valleys.

Table 2: Walking Sequence

RQI (Fig. 6 (*bottom*))	Max. Hip Angle	Min. Hip Angle
Peaks	0.4193	**0.5293**
Valleys	**0.5012**	0.3227

Table 3: Running Sequence

RQI (Fig. 6 (*top*))	Max. Hip Angle	Min. Hip Angle
Peaks	**0.4961**	0.3205
Valleys	0.4348	**0.5071**

6. Conclusions and Future Work

This paper presents an efficient and robust technique for motion recognition. It is based on key frames detection and body posture projection matching. Key frames are detected by studying feature points' trajectories. These key frames alternate two specific body configuration, assuming the person is performing the same motion through the given video sequence. Each one of this body's configuration (maximum and minimum hip angle) are used to define the motion as a walking or a running.

Future works will include modeling the whole trajectory by means of a 3D body. Matching errors throughout

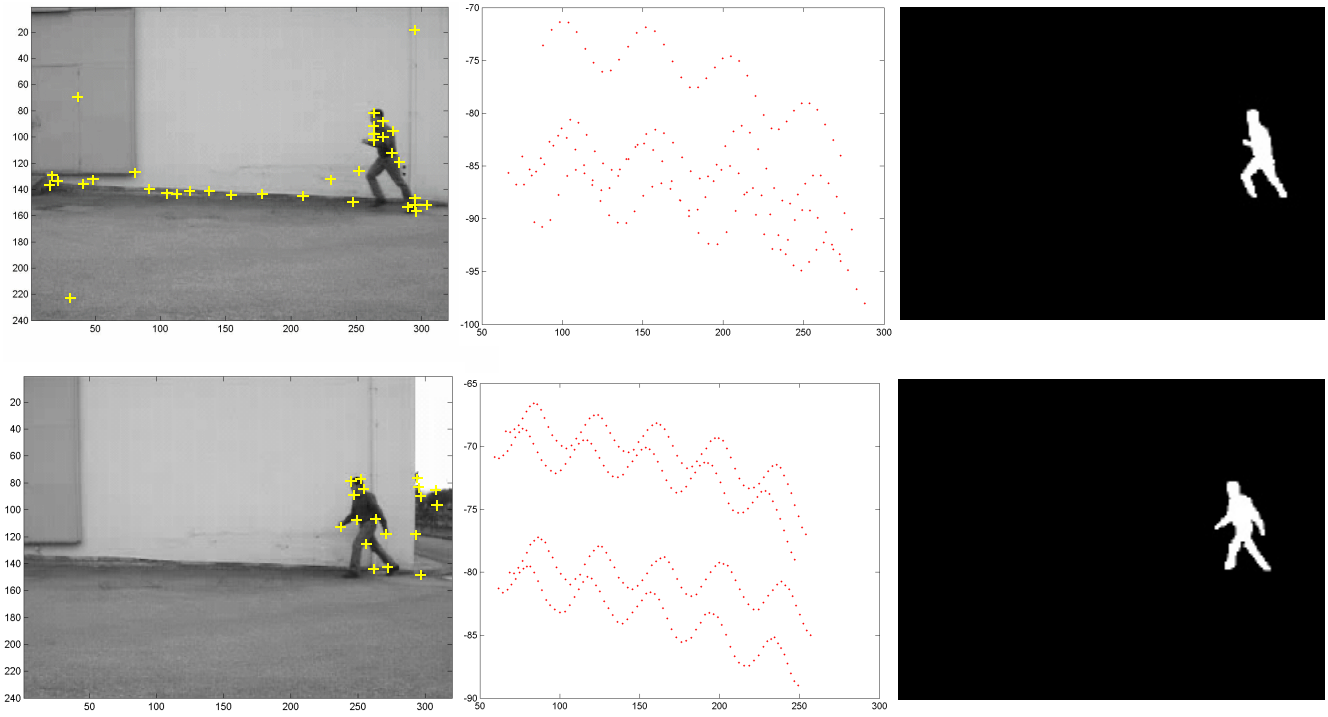

Fig. 6. (*left*) Input frames of two video sequences (running and walking) with feature points. (*middle*) Trajectories of moving feature points. (right) The corresponding segmented frames containing the body silhouette.

the frames will help to tune the motion model and tackle the gait recognition problem.

7. REFERENCES

[1] Q. Delamarre and O. Faugeras, "3D Articulated Models and Multi-View Tracking with Physical Forces", *Special Issue on Modelling People, Computer Vision and Image Understanding*, Vol. 81, 328-357, 2001.

[2] H. Ning, T. Tan, L. Wang and W. Hu, "Kinematics-Based Tracking of Human Walking in Monocular Video Sequences", Image and Vision Computing, Vol. 22, 2004, 429-441.

[3] L. Wang, T. Tan, W. Hu and H. Ning, "Automatic Gait Recognition Based on Statistical Shape Analysis", *IEEE Trans. on Image Processing,* Vol. 12 (9), September 2003, 1-13.

[4] I. Haritaoglu, D. Harwood and L. Davis, "W4: real-time system for detecting and tracking people", *IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, 1998.

[5] I. Haritaoglu, D. Harwood and L. Davis, "W4: Real-time surveillance of people and their activities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (8), 809-830, 2000.

[6] A. Bobick and J. Davis, "Real-time recognition of activity using temporal templates", *3rd IEEE Workshop on Application of Computer Vision*, Sarasota, USA, 1996.

[7] D. Ayers and M. Shah, "Monitoring human behavior from video taken in an office environment", *Image and Vision Computing*, 19(12), 833-846, 2001.

[8] K. Rohr, "Human Movement Analysis Based on Explicit Motion Models", *Chapter 8 in Motion-Based Recognition*, M. Shah and R. Jain (Eds.), Kluwer Academic Publisher, Dordrecht Boston 1997, pp. 171-198.

[9] C. Yam, M. Nixon and J. Carter, "Automated Person Recognition by Walking and Running Via Model-Based Approaches", Patern Recognition, (15 pages) (*in press*).

[10] A. Sappa, N. Aifanti, S. Malassiotis and M. Strintzis, "Monocular 3D Human Body Reconstruction Towards Depth Augmentation of Television Sequences", *IEEE Int. Conf. on Image Processing,* Barcelona, Spain, Sep. 2003.

[11] Y. Ma, S. Soatto, J. Kosecká and S. Sastry, "An Invitation to 3-D Vision: From Images to Geometric Models", Springer-Verlag New York, 2004.

[12] C. Kim and J. Hwang, "Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications", *IEEE Trans. on Circuits and Systems for Video Technology,* Vol. 12 (2), February 2002, 122-129.