# Human Motion Detection, Tracking and Analysis For Automated Surveillance

Eyup Gedikli, Murat Ekinci

Computer Vision Lab.
Department of Computer Engineering
Karadeniz Technical University
Trabzon 61080, Turkiye,
gediklie, ekinci@ktu.edu.tr

**Abstract.** This paper presents a silhouette based human motion detection and analysis for real-time visual surveillance system. In order to detect foreground objects, first, background scene model is statistically learned even the background is not completely stationary. A background maintenance model is also proposed for preventing some kind of falsies. Then, the candidate foreground regions are detected using thresholding, noise cleaning and their boundaries extracted using morphological filters. For human motion detection, object detection and classification approach for distinguishing a person, a group of person from detected foreground objects (*e.g.*, cars) using silhouette shape and periodic motion cues is performed. Finally, the trajectory of the people in motion and several motion parameters produced from the cyclic motion of silhouette of the object under tracking are implemented for analyzing people activities such as walking and running, in the video sequences. Experimental results on the different test image sequences demonstrate that the proposed algorithm has an encouraging real-time background modeling based human motion detection and analysis performance with relatively robust and low computational cost.

## 1 Introduction

Visual analysis of human motion has recently persuaded more studies in the computer vision area. It attempts to detect, track, and identify people, and more generally, to understand human behaviors from image sequences involving humans [1] [2], [12]. Some companies such as IBM, Microsoft, and Mitsubishi are also investing on research on human motion analysis [5], [4], [6]. One of the important research areas is automatically human behavior understanding [7]. Biometrics is also a technology that makes use of the physiological or behavioral characteristics to authenticate the identifies of people [8].

Moving human detection is the first step processes for nearly every system of vision-based human analysis [17]. The aim on moving human detection is to segment the regions corresponding to people from the rest of an image sequence. It is known to be a significant and difficult research problem [13]. Background

subtraction is a particularly popular method for motion segmentation [12][10]. W4 [2] uses dynamic appearance models to track people. A recursive convex hull algorithm is used to find body part locations for single person. Symmetry and periodicity analysis of each silhouette is used to determine if a person is carrying an object. Ricquerberg and Bouthemy [14] proposed tracking people by exploiting spatio-temporal slices. Their detection scheme involves the combined use of intensity, temporal differences between three successive images and of comparison of the current image to a background reference image which is reconstructed and updated on line.

The following process after successfully tracked the moving humans from one frame to another in an image sequence in the video surveillance applications is human behaviors understanding from image sequences. Human behavior understanding is to analyze and recognize the motion region segments by reason of human actions in frames and to produce high-level description of human actions. There has been considerable interest in the area of human motion analysis in recent years [1],[2], [7]. Further works are also focused to human identification based on gait analysis [15],[16].

This paper presents a set of techniques integrated into a low-cost PC based real time visual surveillance system for simultaneously human motion detection, tracking people, and analysis their activities in monochromatic video. Human motion detection and classification approach for distinguishing a person, a group of person from detected foreground objects (*e.g.*, cars) using silhouette shape and periodic motion cues is presented. People tracking in a single camera is performed using background subtraction, followed by region correspondence. This takes into account multiple cues including velocity, sizes, distances of bounding box, and skeleton structures of the silhouette. Objects can be classified based on the type of the cues and their motion characteristics. Finally an algorithm depending on four parameters produced from the skeleton of the silhouette is developed in order to human motion analysis for distinguishing walking and running actions. Then experimental results and discussion are presented in the final section.

## 2 Motion Detection and Classification

The aim on moving human detection is to segment the regions corresponding to people from the rest of an image sequence. Background estimation is a particularly popular method for motion segmentation. The background scene model is statistically learned using the redundancy of the pixel intensities in a training stage, even the background is not completely stationary. This redundancy information of the each pixel is separately stored in an history map shows the intensity variations on the pixel locations. Then the highest ratio of the redundancy (HRR) on the pixel intensity values in the history map in the training sequence is determined to have initial background model of the scene. Then an adaptive background model is updated to accommodate changes to the background while maintaining the ability to detect independently moving objects (person(s)). More details are given in [11][10]. The basis idea on the foreground

region detection based on adaptive background subtraction is to maintain a running statistical background value of the intensity at each pixel. When the value of a pixel in a new image differs significantly from the background value, the pixel is flagged as potentially containing a foreground region. Then foreground objects are segmented from the background in each frame of the video sequence by a four stage process: thresholding, noise cleaning, morphological filters, and object detection. Each pixel is the first classified as either a background or a foreground pixel using the background model, as shown in figure 1.a-b.

People have very distinctive shape, appearance, and motion patterns compared to other objects (car, animal, *etc.*). One can use a static shape analysis, such as aspect ratio, area, size perimeter, or dynamic motion analysis, such as speed, or periodicity of the motion to distinguish people from other objects. All cues in the static shape analysis can easily be obtained from the binary data produced by thresholding and the bounding box parameters. But the cues can be produced from the dynamic motion analysis presents more reliable information for classification, especially for the motion analysis. A skeleton based approach similar with [3] but more developed is presented in this study for the region classification and for the motion analysis.

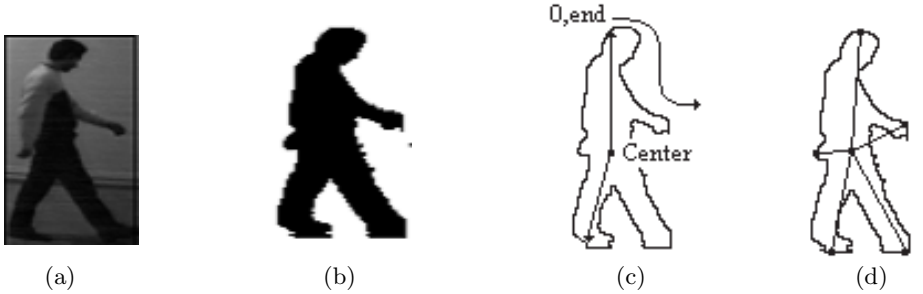



**Fig. 1.** Skeleton production from motion silhouette detected. (a) Detected motion region, (b) Binarized region, (c) Border of the region, and its center of the gravity, (d) local maximal points, then the points projected onto silhouette.

The process is first to binarized to the foreground region then to extract the outline of the silhouette region using morphological processes as shown in figure 1a-c. Then a star skeleton is produced by detecting extremal points on the boundary of the silhouette detected, as shown in figure 1d, see to [11] for more details. The next goal in this section is to classify each moving object visible in the video sequence as a single person, a group of persons, or a vehicle. One of the advantages of video for classification is its temporal component. To exploit this, the static and dynamic features over time are computed in each bounding box as it is detected through the sequence of frames (3-10 frames).

The static shape features are directly measured from the silhouette and its bounding box. They are the bounding box aspect ratio, the axis of second moment of the silhouette, the bounding box dispersedness ($perimeter^2/area$). The dynamic shape features are also produced from the skeleton of the silhouette of

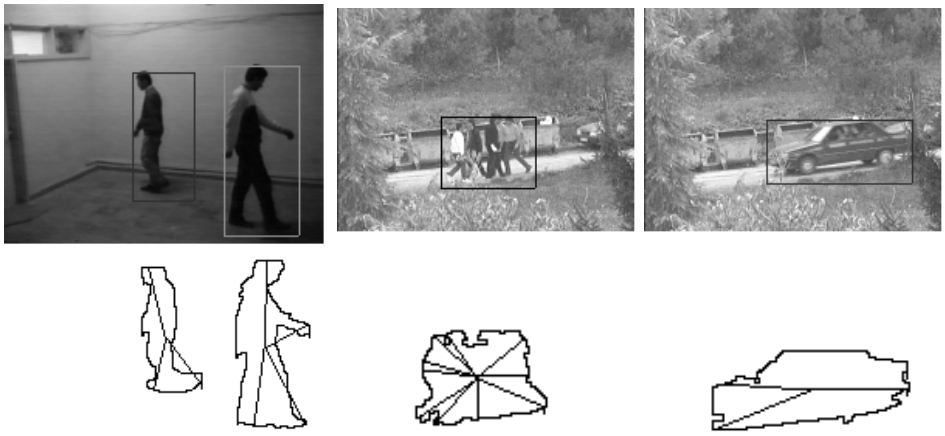


**Fig. 2.** The classification process on detected foreground regions, the dynamic structures of a single human, a group of a person, and a car.

the region detected over time. Both the structure of the skeleton and repetitive changing on the skeleton gives important cues in analyzing different types of objects. Figure 2 shows the structure of the skeleton for different types of objects (a person, a group of person, and a car). It is considered, the repetitive changing on the structures of the skeletons has another good enough feature for classification of the detected foreground regions in the image sequence. The local maximal of the distances (see figure 1 d-f) are determined as the structure of the skeleton. Then this feature can be used for human model criteria for deciding whether the object detected is human or not as shown in figure 2. To more reliable classification another feature in the natural motion of people is also considered. It is that people exhibit periodic motion while they are moving. A periodic motion can be determined by self-similarity of the characteristics of silhouettes over time using the skeleton features of the silhouettes. As a result, static shape cues with a dynamic periodicity analysis and a periodicity of the motion analysis can be combined to distinguish human from the objects, such as a group of person, a car.

## 3 Human Motion Tracking

It is assumed in this study that the regions can enter and exit the scene and they can also get occluded by other regions. Regions carry informations like shape and size of the silhouette, and colors data on a bounding box location estimated for each person. Each region is defined by the 2D coordinates of the centroid, $P$, a ratio between the total number of foreground pixels $(T)$ and the size of the bounding box $(B)$, $R = T/B$, and the color/gray level characteristic, $D$. The regions, for which correspondence has been established, have also an associated velocity, $V$. In frame $t$ of a sequence, there are $M$ regions with centroids $P_i^t$ (where $i$ number of regions) whose correspondences to previous frame are unknown.

There are $K$ regions with centroids $P_L^{t-1}$ (where $L$ is the label) in frame *t-1* whose correspondences have been established with the previous frames. The number of regions in frame $t$ can be less than the number of regions in frame *t-1* due to entries and it might be less due to exits or occlusion. The task is to establish correspondence between regions in frame $t$ and frame *t-1*, and to determine entries and exits in these frames. The minimum cost criteria is used to establish correspondence. The cost function between two regions is defined as

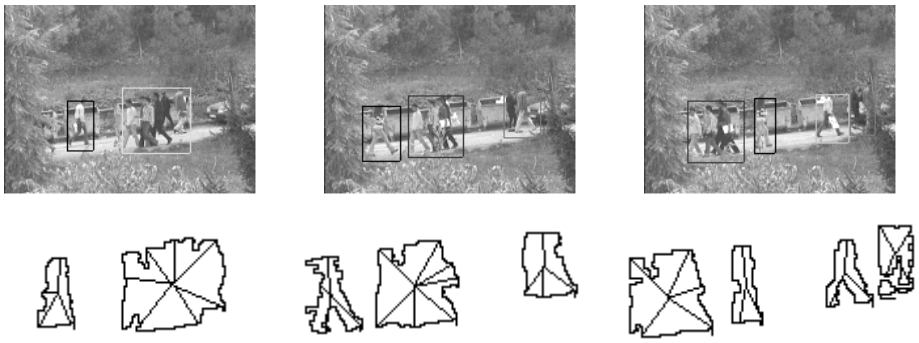
$$C_{Li} = \frac{P_L^{t-1} + V_L^{t-1}}{P_i^t} + \frac{R_L^{t-1}}{R_i^t} + \frac{D_L^{t-1}}{D_i^t} \tag{1}$$

where $L$ is the labels of region in frame *t-1*, $i$ is index of non-corresponded region in frame *t*. The cost is calculated for all *(L, i)* pairs. Correspondence is established between the pair that gives the lowest cost, with the cost being less than a threshold. The all parameters of each region are updated using linear low pass filter prediction models.

The process on correspondence continues till no pairs are left or the minimum cost rises above the threshold. In other words, the correspondences have been found between all regions in frames *t-1* and *t*, or there might be regions in frame *t-1*, which have not been corresponded to in frame $t$ due to exist from the scene or due to occlusion, or there may be regions in frame *t*, which have not been corresponded to regions in frame *t-1*, because they just entered the frame. The position plus predicted velocity of the region exit/enter from/to scene are easily used for determining to have exited/entered the scene. If this is not the case, then a check for occlusion is made. While an occluded is determined, all the regions in occluded have merged in a single region in frame *t*. Now we need to update the parameters of the occluded region. For occluded region same cost function is applied for tracking.

**Fig. 3.** Tracking results on our data set (frames 2555-2606). Two person moving toward to each other, turn around each other, then go back away. Thick and thin white lines separately represent their trajectories produced by tracking algorithm.

In addition, at the tracking process, the center of the detected foreground region in the sequences is stored in a trajectory map. This is shown in figure 3 as white lines. In figure 3, two people moves toward each other. When they are occluded, they turn around each other, then go back away. The thick and thin white lines separately represent the trajectories of each object. The stored data in the trajectory map is used to implement each foreground region motions in the scene. When any foreground object in tracking is hidden to any non-motion area (like passing at the behind of parked car), or temporarily occluded



(a) Frame 809 (b) Frame 821 (c) Frame 836

**Fig. 4.** Occlusion example frames.

by other foreground object as they pass, the detected data of that object may not be obtained at the low level processing. At that or similar situations, high level implementation procedure is activated to estimate the possible position of that object using the previous tracked data obtained from trajectory. At the following frames, if the low level data about it is not obtained, its confidence is reduced. If the confidence of that object drops below a given threshold, it is considered lost, and is dropped from the tracking list stored in the trajectory map. High confidence objects (ones that have been tracked for a reasonable period of time) will persist for several frames, so if an object is momentarily occluded but then reappears, the foreground object tracker will reacquire it. An example experimental results on a test image sequence includes occlusion example between the person and a group of person is shown in figure 4. In figure 4-a, a person and a group of person are tracked in the image sequence. The person under tracking enters to the group of person, another a small group contains two persons exits from the big group, and second person enters to the scene in the following frame sequence, as shown in figure 4-b. Final frame in figure 4-c, first and second persons occluded with the big group separately exit from the big group, and the small group is still in scene and tracked. The skeleton structures of the foreground regions are also illustrated in figure 4, respectively.

## 4 Human Motion Analysis

For the human motion analysis, using the geometrical shape models has the advantage that they have much information than directly obtained features. But the difficulty and cost of calculation in extracting the models from the input frames are disadvantage of using shape models for real time video surveillance applications. Those difficulties prevent researches from concentrating on cognition part of motion analysis process. Consequently, an approach depends on the variations of the features produced from the silhouette motions in frames is presented in this study. The features implemented for human motion analysis are

directly obtained from dynamic variations on the star skeleton structures of the silhouette shape. This basic idea behind of the human motion analysis presented is similar to the study in [3], and is an attempt to make motion analysis more robust for real time applications.

The star skeleton consist of the centroid of a motion blob and three local extremal points that are recovered when traversing the boundary (see section 2). The three local extremal points correspond to head, and two legs. A human is moving in an upright position, it can be assumed that the uppermost skeleton segment represents the torso, and the lower segments determined by two extremal points represent two legs. Then the angle $\theta$ measures between the upper-most extremal point and vertical, the angle $\alpha$ measures between the lower two extremal points, and the angle $\beta$ also measures moving variations in time between end locations of two extremal points in 2D space corresponding to the ankles, as shown in figure 5.a. $(x_c, y_c)$ is the centroid of the motion blob (silhouette of the object under tracking).

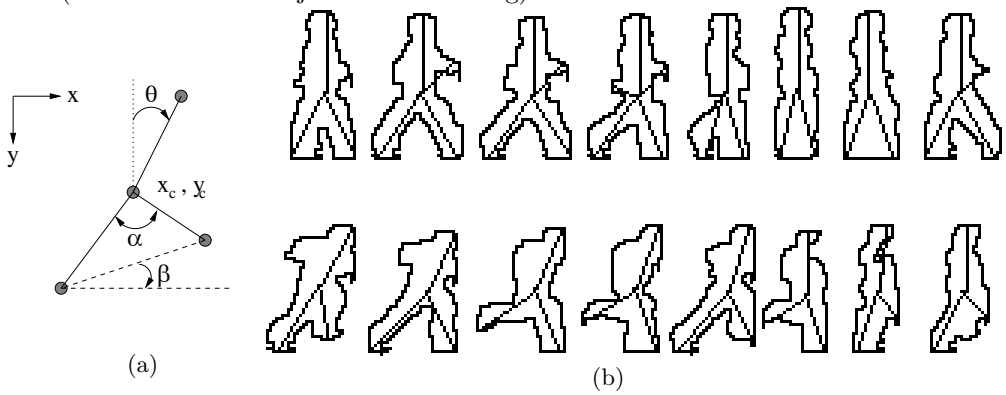



**Fig. 5.** (a) Determining of posture features from the skeleton, (b) Silhouette and skeleton motion sequences of a walking and running person, respectively.

An approach to distinguish the walking and running actions was developed and tested on the different test sequences in our database. The most important features for distinguishing walking and running person can be produced by moving types of the foots in time, the characteristics on the foot cyclic and their speed variations [18]. That features can be easily and simply obtained by manipulating the star skeleton properties. Figure 5.b shows silhouette and skeleton motion sequences for walking and running person. Figure 6 a-d plots the values $\theta_n$, $\alpha_n$, an acceleration of the centroid of the silhouette, and $\beta_n$ over time. The actions characterized in figure 6 in the sequence are walking before frames numbered with 200 and after that running, respectively. Examining the cyclic values of each angles shows that each angle has significant meaning for distinguishing both actions (walking and running), but the more robust human behavior understanding issue can be obtained by fusion the results of all that angles rather than implementing of the features alone. That is, the feature produced by the

$\theta$ angle (represents posture of the person in action) can be manipulated to distinguish the running person from that of the walking person. But not all people lean forward when they run. In other words, there may be no big differencing between on some people lean forward when they run and walk. Figure 6-b plots $\theta$ angle variations in time on the skeleton motion sequence, some samples are shown in figure 5.b, for a walking and then a running people, respectively. There is no big enough significant differencing between two sequences for the posture of the person in action, however, the $\beta$ angle, as shown in figure 6-d, has good enough significant and also presented an encouraged feature to distinguish walking-running actions in the test sequences. The producing of the reliable skeleton from the silhouette is important task because the $\beta$ angle is directly obtained from both end points correspondence to the location of the ankles in the silhouette. Otherwise, the reliability of the implementation depends on the $\beta$ angle might be possible reduced.

When the variation in time on the $\alpha$ angle is considered, it is also giving one of good significant feature for distinguishing both actions. The acceleration on the silhouette in time is also producing the other good enough significant characteristics for analyzing both actions, as shown in figure 6-c. Consequently, to be able to produce more robust and reliable human motion analysis, a fusion task which takes from each feature characteristics then fuses all the features together into a fused motion analysis using a weighted averaging process might be better [9]. This is one of the next studies for human motion analysis in the project presented. Implementation of the signals produced by each features of the skeleton is also another future work.

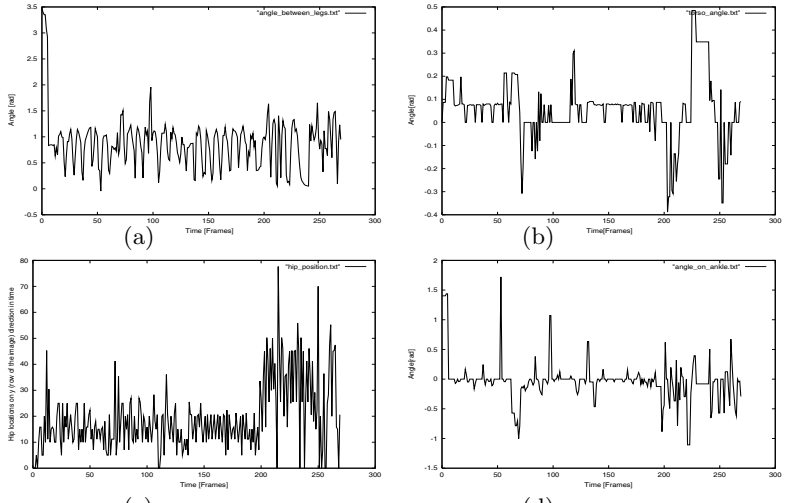



**Fig. 6.** The varations data of the angles produced from star skeleton shape. (a) leg angle $\alpha$, (b) torso angle $\theta$, (c) acceleration on the centroid point, (d) ankle angle $\beta$.

## 5 Experimental Results and Discussion

A video surveillance database is established for our experimental results. The database mainly contains video sequences on different days in outdoor and indoor environments. A digital camera (Sony DCR-TRV355E) fixed on a tripod and a CCD camera fixed on a pan-tilt motor platform are used to capture the video sequences. The algorithm for human motion detection, tracking and analysis presented in this paper has been implemented in C++ and runs under Windows 2K operating system at 96/133 MByte/MHz RAM, 850 MHz Celeron PC without using any special hardware. Currently, for 240 x 180 resolution gray-scale image sequences, the algorithm code without optimizing runs at 13-22 fps depending on the number of people in its field of view. Tests were performed on several sequences (each at least 30 minutes or more) representative of situations which might be commonly encountered in surveillance video.

For object tracking (a person and a group of person), after the object is detected, the tracking algorithm calculates the bounding box, the centroid and correspondence of each object over the frames. The tracking algorithm successfully handled occlusions between people. Entry of a group of people was detected as a single entry, however, as soon as one person separated from the group he was tracked separately as shown in figure 4.

For human motion analysis, an approach similar in [3] but more reliable by adding more important features was presented. Walking and Running actions in the surveillance scene were only considered to differentiate from each other. Four main parameters ($\theta, \alpha, \beta$, acceleration variations on the centroid of the motion blob) are basically implemented to analysis two actions. In addition, the speed of the bounding box surrounding of the silhouette detected could be more considered for analyzing. But the basic approach presented will be developed to extrapolate for the future studies to analysis the other possible people actions such as jumping, sitting, standing, lying, *etc.* For that reasons, the basic idea on the human motion analysis has been focused to the features of the silhouettes. The four parameters basically involve the features of the silhouette for two action (walking, running) analysis [18]. Test results shown in figures 6a-d, encourage to implement this kind of parameters for human motion analysis for real time video surveillance applications. But the shadow is important problem for the silhouette based motion identification and analysis because the structure shape of the silhouette may be fluctuated by the shadow types. Test results produced for human motion detection and analysis were obtained in the surveillance area without having a shadow.

## 6 Acknowledgment

This research work is supported by Karadeniz Technical University Research Foundation grant to (KTU-2002.112.009.1).

## References


1. R.T. Collins, A. J. Lipton, H. Fujiyoshi, T. Kanade.: Algorithms for Cooperative Multi sensor Surveillance. Proc. of IEEE Vol. 89. No.10, (2001).
2. I. Haritaoglu, D. Harwood, L.S. Davis.: W4: Real-Time Surveillance of People and Their Activities. IEEE Trans. on PAMI, Vol. 22, No.8, (2000).
3. H. Fujiyoshi, A. J. Lipton, T. Kanade.: Real-Time Human Motion Analysis by Image Skeletonization. IEICE Trans. Inf.& SYST., Vol.E87-D, No.1, pp.113-120, January 2004.
4. P. Perez, C. Hue, J. Vermaak, M. Gangnet, *Color-Based Probabilistic Tracking,* Proc. of European Conference on Computer Vision, Copenhagen, 27 May- 2 June 2002, Denmark.
5. I. Haritaoglu, M. Flickner, *Detection and Tracking of Shopping Groups in Stores,* Proceeding of the 2001 IEEE Computer Vision and Pattern Recognition, Vol. 1, 8-14 December, 2001.
6. Porikli F., Tuzel O., *Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis*, IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, March, 2003.
7. Mubarak Shah, *Understanding human behavior from motion imagery,* Machine Vision and Applications, Special Issue: Human modeling, analysis, and synthesis, Vol. 14, Issue 4, pp. 210-214, September 2003.
8. A. K. Jain, A. Ross, S. Prabhakar, *An Introduction to Biometric Recognition*, IEEE Transactions on Circuit and Sytems for Video Technology, Special Issue on Image- and Video-Based Biometrics, Vol. 14, No.1, pp. 4-20, January 2004.
9. M. Ekinci, F. W. Gibbs, B. T. Thomas.: Knowledge-Based Navigation for Autonomous Road Vehicles. Turkish Journal of Electrical Engineering and Computer, Vol. 8, No. 1, 2000.
10. M. Ekinci, E. Gedikli, *A New Algorithm of Background Model Initialization and Maintenance for Real-Time Video Surveillance Applications.* To be Appear in Turkish Journal Electrical and Computer Science, June, 2005.
11. M. Ekinci, E. Gedikli, Background Estimation Based People Detection and Tracking for Video Surveillance. Springer LNCS 2869, ISCIS 2003, Computer and Information Sciences, 18*th* Int. Symp., pp. 421-429, Turkey, November, 2003.
12. L. Wang, T. Tan, H. Ning, W. Hu, *Silhouette Analysis-Based Gait Recognition for Human Identification,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, December, 2003.
13. K. Toyama, J. Krumn, B. Brumit, B. Meyers.: Wallflower: Principles and Practice of Background Maintenance. 7th IEEE Inter. Conf. on Computer Vision, Nov., (1999).
14. Y. Ricqueburg and P. Bouthemy, *The Recognition of Human Movements Using Temporal Templates*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.8, August 2000.
15. L. Wang, W. Hu, T. Tan, *Recent developments in human motion analysis*, Pattern Recognition, Vol. 36, pp. 585-601, 2003.
16. B. Bhanu, J, Han, *Individual Recognition by Kinematics-Based Gait Analysis,* in Proceeding of International Conference on Pattern Recognition, pp. 343-346, 2002.
17. J.K. Aggarwal, Q. Cai, *Human Motion Analysis: A Review,* Computer Vision and Image Understanding, vol. 73, n. 3 pp. 428-440, March 1999.
18. T. Mori, K. Tsujioka, T. Sato, *Human-like Action Recognition System on Whole Body Motion-captured File* Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Mani, Hawai, USA, Oct. 29 - Nov. 03, 2001.