Vision-Based Human Tracking and Activity Recognition

Robert Bodor Bennett Jackson Nikolaos Papanikolopoulos
AIRVL, Dept. of Computer Science and Engineering, University of Minnesota.

Abstract-- *The protection of critical transportation assets and infrastructure is an important topic these days. Transportation assets such as bridges, overpasses, dams and tunnels are vulnerable to attacks. In addition, facilities such as chemical storage, office complexes and laboratories can become targets. Many of these facilities exist in areas of high pedestrian traffic, making them accessible to attack, while making the monitoring of the facilities difficult. In this research, we developed components of an automated, "smart video" system to track pedestrians and detect situations where people may be in peril, as well as suspicious motion or activities at or near critical transportation assets. The software tracks individual pedestrians as they pass through the field of vision of the camera, and uses vision algorithms to classify the motion and activities of each pedestrian. The tracking is accomplished through the development of a position and velocity path characteristic for each pedestrian using a Kalman filter. With this information, the system can bring the incident to the attention of human security personnel. In future applications, this system could alert authorities if a pedestrian displays suspicious behavior such as: entering a "secure area," running or moving erratically, loitering or moving against traffic, or dropping a bag or other item.*

Index Terms—Human Activity Recognition, Computer Vision, Reconnaissance and Surveillance, Human Tracking.

I. Introduction

The problem of using vision to track and understand the behavior of human beings is a very important one. It has applications in the areas of human-computer interaction, user interface design, robot learning, and surveillance, among others.

At its highest level, this problem addresses recognizing human behavior and understanding intent and motive from observations alone. This is a difficult task, even for humans to perform, and misinterpretations are common.

In the area of surveillance, automated systems to observe pedestrian traffic areas and detect dangerous action are becoming important. Many such areas currently have surveillance cameras in place, however, all of the image understanding and risk detection is left to human security personnel. This type of observation task is not well suited to humans, as it requires careful concentration over long periods of time. Therefore, there is clear motivation to develop automated intelligent vision-based monitoring systems that can aid a human user in the process of risk detection and analysis.

A great deal of work has been done in this area. Solutions have been attempted using a wide variety of methods (e.g., optical flow, Kalman filtering, hidden Markov models, etc.) and modalities (e.g., single camera, stereo, infra-red, etc.). In addition, there has been work in multiple aspects of the issue, including single pedestrian tracking, group tracking, and detecting dropped objects.

For surveillance applications, tracking is the fundamental component. The pedestrian must first be tracked before recognition can begin. Kalman filters have been used extensively for tracking in many domains. In visual surveillance, this method appeared very often in the literature ([3][9][10][13][28][32]). It is worthy of note that most applications used only a linear Kalman filter approach. It seems that this was sufficient for many problems. We believe this is due to the controlled indoor and outdoor environments that were used. Many applications could model 2D or near 2D motion exclusively (camera above the scene outdoors, camera tracking lateral pedestrian motion indoors).

The majority of papers detailed methods that tracked a single person only ([4][6][17][25][29]). Most of these involved indoor domains for purposes of gesture recognition [17], and user interfacing [4]. Both [25] and [29] use pan-tilt rigs to track a single individual indoors.

Tracking groups and their interactions over a wide area has been addressed to a limited extent. Maurin *et. al.* used optical flow to track crowd movements both day and night around a sports arena [1]. Haritaoglu *et. al.* track groups as well as individuals by developing different models of pedestrian actions [34]. They attempt to identify individuals among groups by segmenting the heads of people in the group blob.

II. Description of Work

This research was developed in two parts: tracking pedestrians and capturing pedestrian images and pedestrian activity recognition based on position and velocity. Figure 1 below shows a schematic overview of the processes. The first two components, human detection and human tracking are described in Part A below, while human activity recognition and high-level activity evaluation are described in Part B. For the purposes of this work, we define 'activity' as a set of actions.

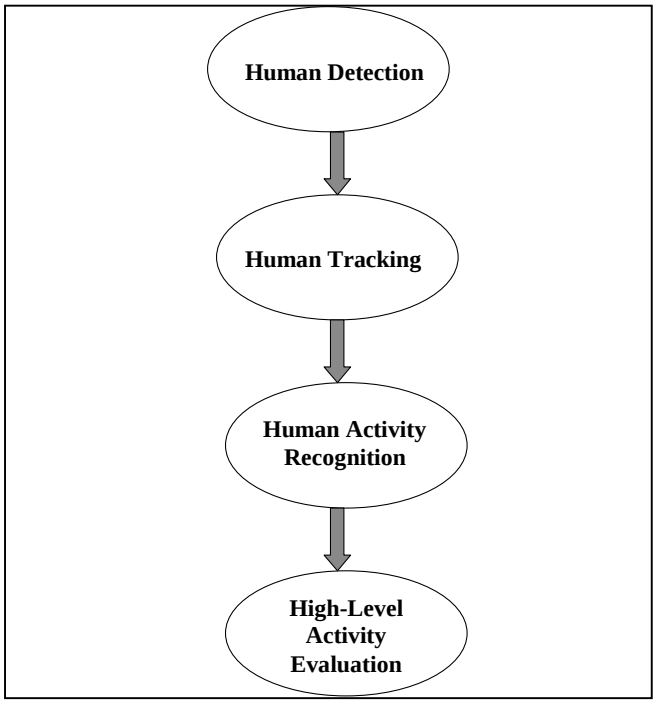

Figure 1. Process overview.

A. Tracking pedestrians and capturing pedestrian images

This work built upon the pedestrian and vehicle tracking work developed in the Robotics and Vision Laboratory. Specifically, we built upon a framework of code developed by Harini Veeraraghavan. This code tracked objects appearing in a digitized video sequence with the use of a mixture of Gaussians for background/foreground segmentation (see [7]) and a Kalman filter for tracking.

All experiments were run on 320x240 pixel resolution images on a computer with a Pentium II 450 MHz single processor and 128MB of RAM. In addition, the computer incorporated a Matrox™ Genesis board for video capture. All images were taken using a Sony™ Digital8 video camera.

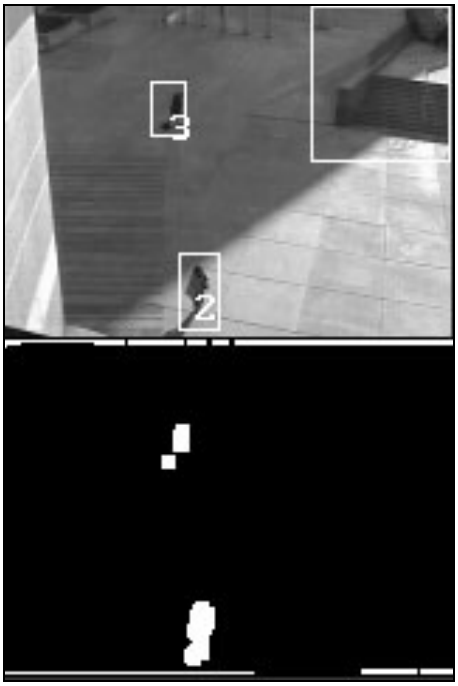

Figure 2. Surveillance image and associated pedestrian 'blobs.'

The goal of this stage was to segment and extract the image of each pedestrian from all appearances in the image sequence. This "pedestrian image sequence" data could then be used in the later stages of the system to provide information to the motion recognition components to classify the pedestrian motion.

We used a Matrox™ video capture board and its Genesis Native Library to be able to augment the tracker and access the pedestrian positions generated. We developed routines to accomplish three things:

1. Establish a stable oversized bounding box around pedestrians tracked smoothly throughout video sequence (see Figure 2 above).
2. Grab the image of the pedestrian within the bounding box and save it (see Figure 3).
3. Combine the individual images into movie files. (see Figure 4).

Figure 3. Pedestrian tracked across 2 frames of image sequence.

This module could then track a pedestrian and generate single image snapshots or movies of the pedestrian's motion. Figure 4 below shows some example image sequences generated.

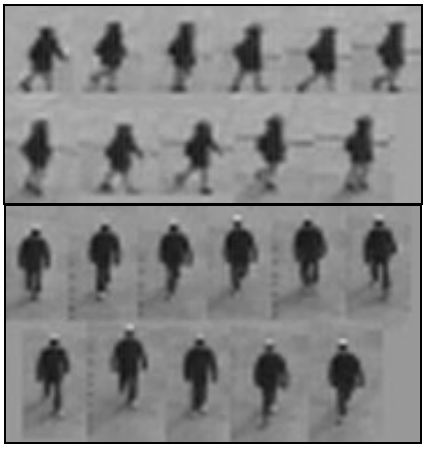

Figure 4. Samples of two sequences of pedestrian images.

B. Human activity recognition based on pedestrian position and velocity

This component estimates the pedestrian motion based on the speed and position of the pedestrian. The basic assumption is that much of the pedestrian's activities can be ascertained by measuring these simple aspects. Measuring these values provides several advantages over articulated motion analysis: these measurements can be made in real time and are far more robust to noise and poor image quality. In addition, for our purposes, if a pedestrian is moving in an area that is off limits, that should be flagged as a warning. In this circumstance, the type of motion is generally irrelevant.

This process had several components:

1. Track each pedestrian throughout scene using the Kalman filter estimates.
2. Record the position and velocity state.
3. Develop a position and velocity path characteristic for each pedestrian. This was done using the Kalman filter prediction of future state.
4. Set a "warning signal" under the following conditions:
    a. Pedestrian enters near a "secure area" (a gate, expensive art display, podium, etc…).
    b. Pedestrian moves above a walking speed .
    c. Pedestrian loiters in the area for a long time.
    d. Pedestrian falls down.

For the purposes of testing, we assigned the upper right corner of the screen (the steps in front of the building entrance) to be a "secure area" and programmed the software to signal a warning if any pedestrians entered that region. In addition, the software calculated the speed of each pedestrian and signaled a warning if any pedestrian exceeded the speed threshold for walking (our experiments indicated a top walking speed of 2.25 meters/second). Falling was detected using a combination of pedestrian velocity and shape. In each of these cases, a pedestrian motion image was captured to record the incident. This image was taken from the wide-angle surveillance footage. In the future, we plan to integrate a pan-tilt mounted zoom camera to capture high-resolution images of incidents.

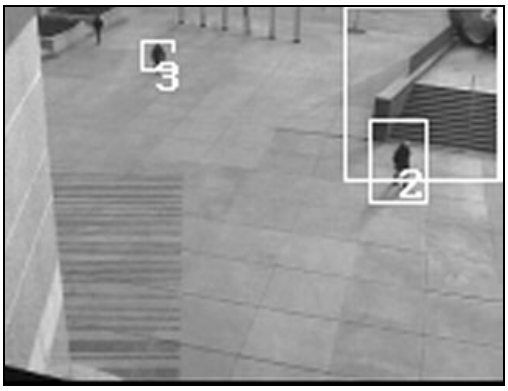

Figure 5. Screenshot of the system in operation. Pedestrian "2" has entered a secure area and created a warning signal.

III. RESULTS

We tested the system in an outdoor courtyard where there was a continuous flow of pedestrian traffic. Figures 6, 8, and 9 below show the surveillance images taken, with the pedestrian motion paths in image space superimposed on them. In each case, the bottom figure shows a map of the pedestrian paths and motion type in world coordinates.

A. Activity recognition

Sequence 1. This sequence tracked two pedestrians crossing the courtyard in different directions. Pedestrian #2 tripped and fell down during the sequence.

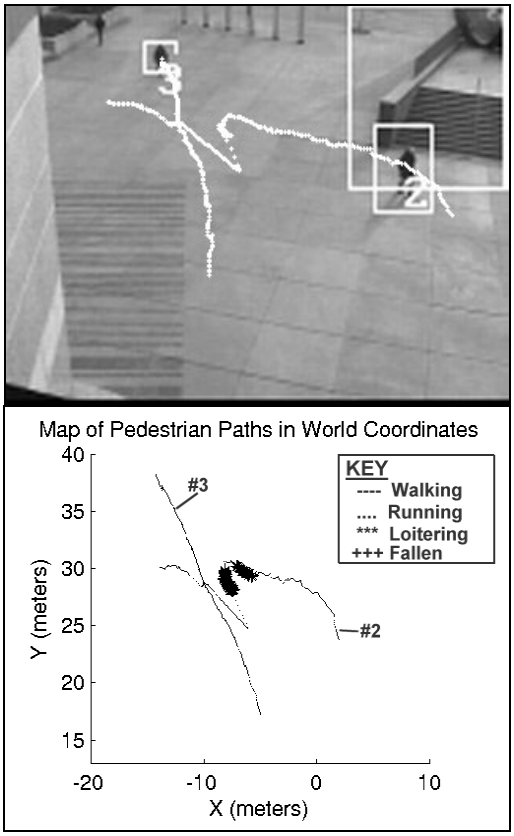

Figure 6. Tracked pedestrian image and map of motion.

Figure 7. Surveillance image of pedestrian "2" fallen down.

In the above example, pedestrian #2 spent 9.4 seconds fallen (as indicated by the dark areas in Figure 6). In addition, pedestrian #2 later trespassed in the "secure area" for a total time of 4.6 seconds. The motion path of pedestrian #2 indicates a large divergence midway through the image (see Figure 6). The pedestrian did not actually take this path. This error results from an artifact introduced by the background separation method we used (see section C below).

Sequence 2. This sequence tracked three pedestrians crossing the courtyard. Pedestrian #3 walked through the "secure area" at the end of the sequence.

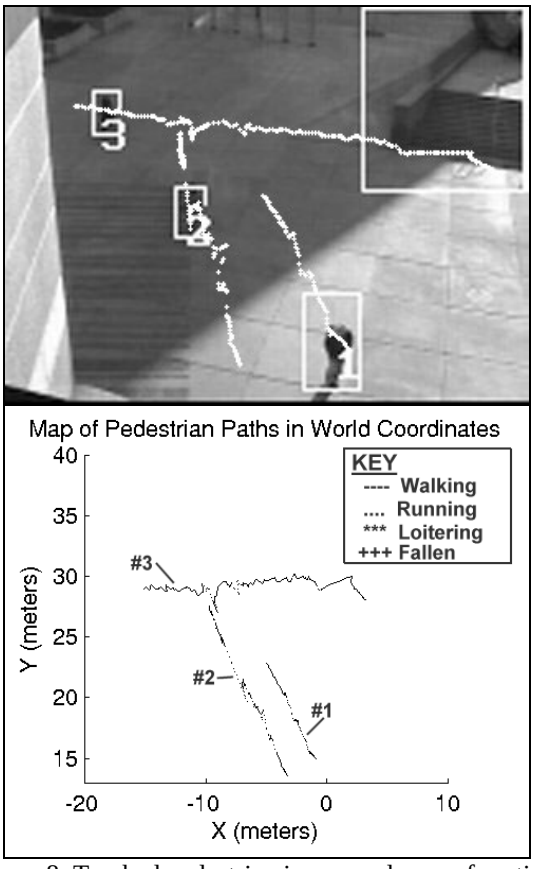

Figure 8. Tracked pedestrian image and map of motion.

In this test, pedestrian #3 spent 5.6 seconds in the "secure area."

Sequence 3. This sequence tracked two pedestrians crossing the courtyard. Pedestrian #3 was on a bicycle.

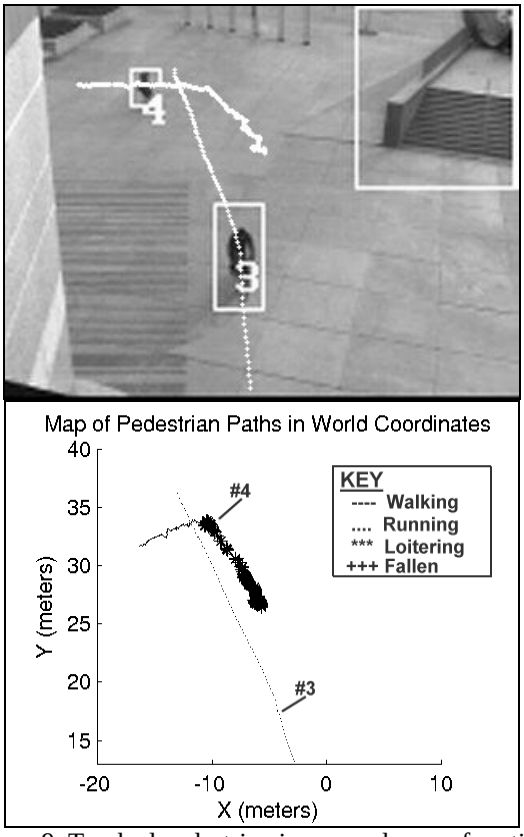

Figure 9. Tracked pedestrian image and map of motion.

The dark area in the above figure shows pedestrian #4 loitering in the courtyard for 22.3 seconds. The motion path for pedestrian #3 shows the pedestrian moving rapidly throughout the sequence.

B. Pedestrian velocity analysis

Each pedestrian's velocity was calculated and used as part of the activity recognition. Figure 10 below shows a sample of the velocity characteristic over time for a walking pedestrian and a running pedestrian.

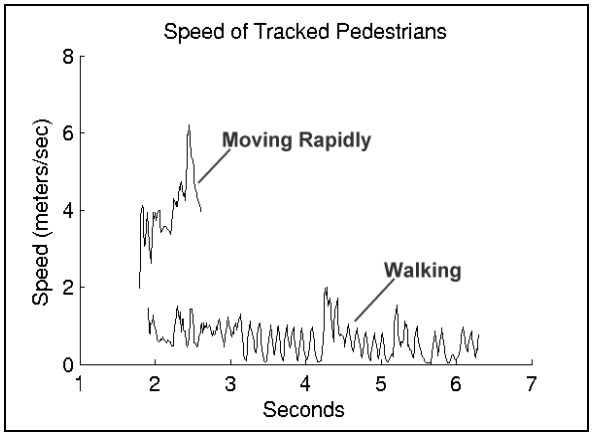

Figure 10. Characteristic sample of measured pedestrian velocity. This velocity is calculated in world coordinates.

C. Problems encountered/limitations of current solution

We encountered significant changes in lighting when we captured the test sequences (compare Figures 8 and 9). Rapid changes in the lighting of the outdoor scene such as those caused by the sun suddenly going behind/emerging from a cloud (dark shadows, harsh edges) introduced some error in our tracking system. The background segmentation method that we used took up to 15 seconds to adapt to large-scale changes of this kind, causing blobs to appear where there was no foreground person or object. This problem became particularly evident in test Sequence 1 (see Figure 6).

One limitation of our system is that since we do not directly observe the motion of the pedestrian through articulated motion analysis, our system does not distinguish between objects moving at the same speed through different means, such as a bicyclist and a runner (see Figure 8 above).

IV. Conclusions/Future Work

To advance the system in the future, we would like to add several components. We would like to expand the pedestrian tracker to consider many more states of pedestrian motion and generate warnings, including erratic pedestrian motions (changes motion suddenly) and pedestrian motion against the flow of traffic (incorporating some optical flow principles). We would also like to add a detection scheme for dropped objects and objects newly appearing in the scene.

In addition, we believe this system would benefit from the addition of multiple cameras of different types, including a pan-tilt mounted zoom camera and an infrared camera.

In the far future, we would like to examine the use of a motion recognition and tracking system on a mobile robotic platform to detect and follow individuals.

V. Acknowledgements

This work has been supported by the National Science Foundation through the grant #IIS-0219863 and the ITS Institute at the University of Minnesota. We would like to thank Harini Veeraraghavan for her help in integrating the existing tracker and Kalman filter code. We would also like to thank Osama Masoud for his review of this paper, and for developing the basis from which we could work.

VI. References

[1] B. Maurin, O. Masoud, and N. Papanikolopoulos, "Monitoring Crowded Traffic Scenes," *Proc. of the IEEE 5th Int. Conf. On Intelligent Transportation Systems (ITSC 2002),* pp 19-24, Singapore, September 3–6, 2002.

[2] O. Masoud, "Tracking and Analysis of Articulated Motion with an Application to Human Motion," Doctoral Thesis, Univ. of Minnesota, March 2000.

[3] C.R. Wren and A.P. Pentland, "Dynamic Models of Human Motion," *Proc. Third IEEE Intl. Conf. Automatic Face and Gesture Recognition*, April 1998.

[4] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, No.7, July 1997.

[5] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, "Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance," *Proc. of the IEEE*, vol. 90, No. 7, July 2002.

[6] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human Activity Recognition Using Multidimensional Indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, No. 8, August 2002.

[7] C. Stauffer and W.E. Grimson, "Learning Patterns of Activity Using Real-Time Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, No. 8, August 2000.

[8] L. Zhao and C.E. Thorpe, "Stereo and Neural Network Based Pedestrian Detection," *IEEE Trans. On Intelligent Transportation Systems*, vol. 1, No. 3, September 2000.

[9] D. Beymer and K. Konolige, "Real-Time Tracking of Multiple People Using Continuous Detection," Artificial Intelligence Center, SRI International Report, 1998.

[10] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions," MIT Media Laboratory Report, 1998.

[11] H. Mori, M. Charkari, and T. Matsushita, "On-Line Vehicle and Pedestrian Detections Based on Sign Pattern," *IEEE Trans. on Industrial Electronics*, vol. 41, No. 4, August 1994.

[12] R. Fablet and M.J. Black, "Automatic Detection and Tracking of Human Motion with a View-Based Representation," *European Conf. On Computer Vision*, ECCV'02, May 2002.

[13] R. Rosales and S. Sclaroff, "3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions," *Proc. of IEEE Conf. On Computer Vision and Pattern Recognition*, June 1999.

[14] R. Rosales and S. Sclaroff, "Trajectory Guided Tracking and Recognition of Actions," *PAMI, Special Issue on Video Surveillance and Monitoring*, 1999.

[15] R. Polana and R. Nelson, "Nonparametric Recognition of Nonrigid Motion," University of Rochester, New York Report, 1994.

[16] Z. Ghahramani, "An Introduction to Hidden Markov Models and Bayesian Networks," *Intl. Journal of Pattern Recognition and Artificial intelligence*," vol. 15, No. 1, 2001.

[17] R. Cutler and M. Turk, "View-Based Interpretation of Real-Time Optical Flow for Gesture Recognition," University of Maryland, College Park Report, 1997.

[18] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73, No. 1, 1999.

[19] B. Heisele, A. Verri, and T. Poggio, "Learning and Vision Machines," *Proc. of the IEEE*, vol. 90, No.7, July 2002.

[20] S.M. Seitz and C.R. Dyer, "Photorealistic Scene Reconstruction by Voxel Coloring," Univ. of Wisconsin, Madison Report, 1998.

[21] I. Pavlidis, J. Levine, and P. Baukol, "Thermal Image Analysis for Anxiety Detection," *Nature*, vol. 415, January 2002.
[22] R. Morris and D. Hogg, "Statistical Models of Object Interaction," School of Computer Science, University of Leeds, England, 1998.
[23] W. Grimson, C. Stauffer, R. Romano, and L. Lee, "Using Adaptive Tracking to Classify and Monitor Activities in a Site," *IEEE Press*, 1998.
[24] K. Konolige, "Small Vision Systems: Hardware and Implementation," Artificial Intelligence Center, SRI International, 1997.
[25] A. Cretual, F. Chaumette and P. Bouthemy, "Complex Object Tracking by Visual Servoing Based on 2D Image Motion," *Proc. of the IAPR Intl. Conf. on Pattern Recognition*, vol 2, pp 1251-1254, Australia, August 1998.
[26] C. Wohler, J. Anlauf, T. Portner, and U. Franke, "A Time Delay Neural Network Algorithm for Real-Time Pedestrian Recognition," *IEEE Conf. on Intelligent Vehicles*, pp 247-252, 1998.
[27] A. Mittal and D. Huttenlocher, "Scene Modeling for Wide Area Surveillance and Image Synthesis," *IEEE Press*, 2000.
[28] A. Azarbayejani and A. Pentland, "Real-Time Self-Calibrating Stereo Person Tracking Using 3D Shape Estimation from Blob Features," *Proc. of the ICPR, IEEE*, 1996.
[29] C. Eveland, K. Konolige, and R. Bolles, "Background Modeling for Segmentation of Video-Rate Stereo Sequences," Report, Computer Science Department, University of Rochester, 1997.
[30] T. Darrell, G.Gordon, M. Harvillem and J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection," Interval Research Corporation Report, 1997.
[31] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, "A Stereo Machine for Video-Rate Dense Depth Mapping and Its New Applications," *IEEE Press*, pp 196-202, 1996.
[32] C. Bregler, "Learning and Recognizing Human Dynamics in Video Sequences," *IEEE Press*, pp 568-574, 1997.
[33] Y. Leclerc, Q. Luong, and P.Fua, "Self-Consistency, Stereo, MDL, and Change Detection," Artificial Intelligence Center, SRI International, October 2000.
[34] I. Haritaoglu, D. Harwood, L. Davis, "W^4: Real-Time Surveillance of People and Their Activities," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol.22, no. 8, August 2000.
[35] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. Buhmann, "Topology Free Hidden Markov Models: Application to Background Modeling," *IEEE Press*, pp 294-301, 2001.
[36] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, "Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information," *IEEE Press*, pp 627-630, 2000.
[37] S. McKenna, S. Jabri, Z. Duric, and H. Wechsler, "Tracking Interacting People," Dept. of Applied Computing, University of Dundee, Scotland, 1999.
[38] J. van Hateren, and A. van der Shaaf, "Independent Component Filters of Natural Images Compared with Simple Cells in Primary Visual Cortex," *Proc. of R. Soc. Lond. B 265*, p. 359-366, 1998.
[39] S. Roweis, and Z. Ghahramani, " A Unifying Review of Linear Gaussian Models," Dept. of Computer Science, University of Toronto, August 1997.
[40] T. Lee, M. Girolami, A. Bell, and T. Sejnowski, "A Unifying Information-Theoretic Framework for Independent Component Analysis," *Intl. Journal of Computers and Mathematics with Applications*, 1999.