Detecting and Counting People in Surveillance Applications

X. Liu P. H. Tu J. Rittscher A. G. A. Perera N. Krahnstoever
GE Global Research
Niskayuna, NY 12309, USA

Abstract

A number of surveillance scenarios require the detection and tracking of people. Although person detection and counting systems are commercially available today, there is need for further research to address the challenges of real world scenarios. The focus of this work is the segmentation of groups of people into individuals and tracking them over time. The relevant applications of this algorithm are people counting and event detection. Experiments document that the presented approach leads to robust people counts.

1. Introduction

A number of surveillance applications require the detection and tracking of people to ensure security, safety, and support site management. Examples include the estimation of queue length in retail outlets, the monitoring of entry points, bus terminals, or train stations as shown in Figure 1. Substantial progress [5, 3, 11, 7, 4, 8] has been made to detect people in constraint settings. Often it is for example assumed that the people in the scene are well separated and that it is possible to identify foreground objects using a statistical background model. Certain actions can only be detected if the location of all individuals in the scene is known. However, in all of the scenarios just mentioned we have to anticipate that people can appear in *groups*. In addition it is often necessary to know how many people are present.

Recently [12] we propose a model based segmentation algorithm which allows to partition a group of people into individuals. Although this method is capable to segment a region of interest into individuals, it needs to be embedded into a comprehensive system which supports the detection, tracking, and detection of specific events. One possible design of such a system is the focus of the presented work. Counting of people entering and leaving a site and the detection of events are presented as potential applications of this system. It will be demonstrated how this system allows the detection of certain events in the underground scenario illustrated in Figure 1.

A summary of the related work is given in the following section. An overview of our system is given in Section 3. Two main components, autocalibration and crowd segmentation, are described in Section 4 and 5 respectively. The idea of counting the number of people entering and leaving through a virtual gate is outlined in Section 6. Experimental results of the counting system are presented in Section 6.2. The recognition of specific events is discussed in Section 7.

Figure 1: **Typical underground scenario.** *Automatic monitoring of a scenario as shown here requires automatic detection and tracking of all individuals in the scene. Occlusions obscure the detection of certain event. It will be shown that by segmenting groups of people into individuals their position can be estimated more accurately.*

2. Related Work

Various techniques [11, 3, 4] have been applied to construct fast and reliable person detectors for surveillance applications. Classification techniques can for example be applied to decide if a given image region contains a person. Amongst others Nakajimia *et al.* [11] use Support Vector Machines to approach this problem. Gravrila [3] uses a tree based classifier to represent possible shapes of pedestrians. Griebel et al. [4] use dynamic point distribution models. An alternative to modeling the appearance of an entire person is to design detectors for specific body parts and combine the result of those. The idea of learning part detectors using Ada-Boost and a set of weak classifiers is presented in [8]. A learning approach is then being used to combine the set of weak classifiers with body part detectors, which are further combined using a probabilistic person model. All these approaches require a fair amount of training data to learn the

parameters of the underlying model. Although these classifiers are robust to limited occlusions, they are not suitable to segment a group of people into individuals.

One possibility of segmenting a group of people is to use the information of various different camera views. The M2-tracker presented in [10] explicitly assigns the pixels in each camera views to a particular person using colour histograms. Zhao and Nevatia [16] make clever use of the fact that they know the camera calibration and can find possible head locations using a head detector. The locations of all individuals in the scene are estimated by maximizing an observation likelihood using Markov Chain Monte Carlo. Their results clearly show that it is extremely helpful to know the location of the ground plane and the camera parameters. The head detector is based on edge information. However, under certain imaging conditions it can be challenging to extract clean edge maps. In order to overcome this limitation we developed a model based segmentation algorithm [12] that simultaneously estimates the position and size of all individuals. The details of our approach will be, as mentioned before, presented in Section 5.

One traditional way to perform people counting is to install turnstiles. However, it has the drawback of high cost and low flexibility. Video-based people counting is a good alternative. Group segmentation alone is not sufficient to count the number of people entering or leaving a site. In order to achieve that it is necessary to track each individual in the group and extract the direction of travel. In this paper, we introduce the idea of a *virtual gate*. Yang *et al.* [15] propose a people counting method that makes use of different views. Here we demonstrate that it is possible to obtain reliable counts from a single view.

3. System Overview

The system (see Figure 2) consists of four main components: a standard low-level foreground estimation algorithm [14], an autocalibration module, a template-based tracker, as well as the crowd segmentation. All four components are combined into a tightly coupled framework.

For each frame, the foreground estimation algorithm detects a set of consistent foreground rectangles, where each of them is assumed to be either a person or a group of people. The tracker maintain the trajectory of each person over time, whose head and foot location are sent to the autocalibration module. The resulting calibration parameters are utilized by the crowd segmentation module in segmenting a group of people into individuals. Furthermore, the segmented individuals are tracked via the data association. Eventually, these trajectories are used in the people counting and event detection applications. In the following, we will describe the autocalibration, the tracker and the integrated crowd segmentation components in more detail.

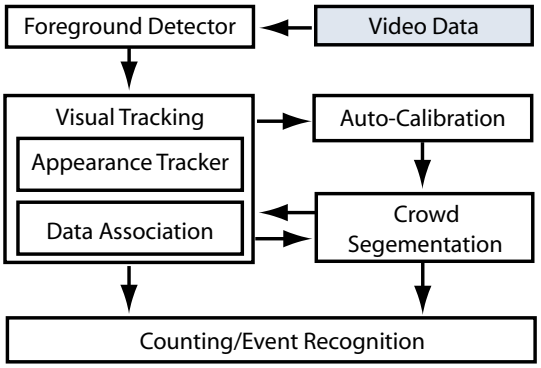

Figure 2: **System overview.** *There are four components: the foreground detector, the tracker, the autocalibration, and the crowd segmentation. See text for details.*

3.1. Autocalibration

Knowledge about the site geometry and the camera parameters makes it possible to establish a connection between image and world measurements. This can, as discussed in Section 1, constrain the problem at hand and make solutions more accessible. Unfortunately, geometric information is rarely available and difficult to obtain after a surveillance system has been installed. Hence, autocalibration approaches that utilize information from an observed scene are attractive for practical applications. For the system presented in this work, a method for camera autocalibration based on information gathered by tracking people is utilized. Basically the head and foot location measurements of the observed individuals are used to compute the camera parameters.

3.2. Tracker

The tracker uses an adaptive appearance based approach similar to [13, 16]. The tracker is adaptive and can track people and other targets such as vehicles alike. Various algorithms are in place for initiating, merging, splitting and deleting tracks. Each track is modeled by a color signature, an appearance template, as well as a probabilistic target mask. The foreground mask is an autoregressive estimate of the foreground information as obtained in the previous stage. The tracker handles short term occlusions between isolated tracks, but groups closely spaced targets together into *group tracks*. Only foreground regions which are large enough to contain a number of people and image regions that contain closely spaced tracks are forwarded to the crowd segmentation algorithm for further analysis. In addition, an improved foreground region image is composed based on the information maintained by the tracker and also supplied to the crowd segmentation algorithm. The moti-

vation for this is the following: The properties of the target masks compare favorably to the direct estimate of the foreground. First, the autoregressive process used to maintain the target masks suppresses high frequency variations and noise in the foreground image. Second, since the target masks are estimated from the foreground image relative to the moving tracks, foreground region information is effectively integrated across multiple images along the motion paths of targets, hence resulting in more accurate overall estimates.

3.3 Crowd Segmentation Component

The crowd segmentation algorithm processes all regions in the image that are likely to contain more than one person. The resulting segmentation observation $\hat{S}^t$ at frame t contains information about the detected number of people and their location in the image $\hat{X}^t = \{\hat{n}^t, (\hat{x}_i^t, \hat{y}_i^t), i = 0, \ldots, \hat{n}^t\}$. As discussed above, noise in the feature extraction process as well as inherent ambiguities will inevitably lead for the estimate $\hat{X}^t$ to deviate from the true state S^t. To reduce the error in the resulting segmentation, the estimated values are processed by a simplified multiple hypothesis tracker. Within each group individual tracks are smoothed using a constant velocity Kalman filter.

4. Site Geometry

One possible approach to autocalibration is based on vanishing points and vanishing lines that can be obtained from tracking human targets in video. Unfortunately it can be shown that this approach is very sensitive to measurement errors which makes existing approaches unsuitable for practical applications. Our approach addresses the problem on two fronts. First, as shown in [9] we perform calibration via the estimation of the so called *foot to head plane homology*, which is to obtain the internal and external calibration parameters of the camera from head and foot location measurements. Second, we perform the estimation of this homology using a Bayesian approach, that can elegantly handle measurement uncertainties, outliers, as well as prior information. The full posterior distribution of calibration parameters given the measurements is estimated, which allows making statements about the accuracy of both the calibration parameters and the measurements involving them.

When observing people, each (foot) location on the ground plane corresponds to exactly one location in the so-called *head plane*, which is located at a height h parallel to the ground plane (i.e., we assume that all observed people have the same average height h). It can be shown, that the homography that maps the images of ground planes to the images of the corresponding points in the head plane is in fact a homology H and is given by

$$\mathrm{H} = \mathbb{I} - \frac{h}{z} \frac{\tilde{\mathrm{v}}^{\infty} (\tilde{\mathrm{l}}^{\infty})^{\mathrm{T}}}{(\tilde{\mathrm{v}}^{\infty})^{\mathrm{T}} \tilde{\mathrm{l}}^{\infty}}, \tag{1}$$

with z the height of the camera above the origin of the ground plane, $\tilde{\mathrm{v}}^{\infty}$ the vanishing point and $\tilde{\mathrm{l}}^{\infty}$ the horizon line. It can furthermore be shown that the horizon line is given by $\tilde{\mathrm{l}}^{\infty} = \begin{bmatrix} \sin(\rho) & -\cos(\rho) & \frac{f}{\tan(\theta)} \end{bmatrix}$ and the vanishing point by $\tilde{\mathrm{v}}^{\infty} = \begin{bmatrix} f\sin(\rho)\sin(\theta) & -f\cos(\rho)\sin(\theta) & \cos(\theta) \end{bmatrix}$, with ρ the roll angle of the camera, θ the tilt towards the ground plane and f the focal length. Making standard assumptions about the remaining parameters of the camera [9], knowledge of the foot to head homology yields complete metric calibration of a camera with respect to the ground plane.

The overall autocalibration approach that is described in detail in [9] now proceeds as follows: Given a sufficient number of isolated people observations, consisting of foot and head image location measurements with associated measurement uncertainties, an initial foot to head homography is estimated using a standard DLT approach [6]. Then, the eigenvalue structure of the targeted homology is exploited to obtain the closest foot to head homology consistent with the data. Finally the initial homology estimate is refined in a Bayesian framework (taking the noise and all nuisance variable into due consideration) and the posterior distribution of the camera parameters given the measurements is estimated.

5. Model Based Segmentation

In [12] a model based approach to crowd segmentation is proposed. Given a foreground segmentation, a set of low level image features $Z = \{z_i\}$ are extracted. In addition, an exhaustive set of feature groupings or cliques $C = \{c_i\}$ is hypothesized. Each grouping corresponds to a potential person (see Figure 3). These groupings are constrained by a geometric shape model which is parameterized by $X = \{x_i\}$ (see prior section). Each feature must be assigned to a single grouping and the shape parameters of each grouping must be estimated. An assignment vector $V = \{v_i\}$ establishes the feature assignments. A likelihood function $P(Z,V;X)$ is defined based on pairwise and single assignments of features to groupings with shape parameters X. The goal is to determine maximum likelihood estimates of both V and X.

A formulation based on EM is used, where V is viewed as a hidden variable. EM provides a method to estimate a distribution $\tilde{P}(V)$ as well as an estimate of X. Once this has been achieved, likely values of V can be selected by sampling $\tilde{P}(V)$. Estimates of $\tilde{P}(V)$ and X are found by maximizing the free energy equation:

$$F(\tilde{P}, X) = E_{\tilde{P}}[\log P(V, Z; X)] + H(\tilde{P}), \tag{2}$$

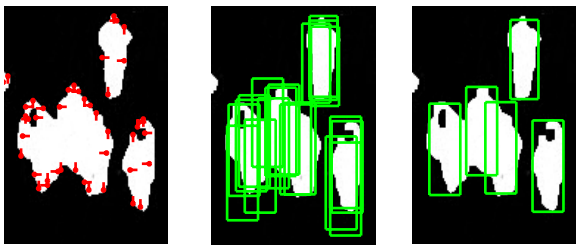

Image Features Z **Clique Set $\mathscr{C}$** **Segmentation X^***

Figure 3: **Image features, cliques, and shape parameters.** *The feature extraction, the computation of the set of cliques $\mathscr{C}$ and segmentation results are shown for one example image. Note that a standard probabilistic background model was used to segment the image foreground. The set of image features Z illustrates that each feature z_i is labeled as being at the top side or bottom of a person. The set of cliques $\mathscr{C}$ illustrates that the algorithm generates a large number of cliques. The segmentation on the right shows a segmentation for the MAP estimate X^* and one sample of V. Relevant details are described in Section 5.*

In order to regularize the optimization process a temperature term T is introduced:

$$F(\tilde{P},X) = E_{\tilde{P}}[\log P(V,Z;X)] + TH(\tilde{P}). \quad (3)$$

Initially T is set to a large value and this favors the entropy term. As a result an initial estimate of $\tilde{P}(V)$ can be set to a uniform distribution. An annealing process is performed by iteratively decreasing T. At each iteration, both an E-step and an M-step is performed. In the E-step, X is fixed to its current value and the free energy is optimized with respect to $\tilde{P}(V)$. In the M-step, $\tilde{P}(V)$ is fixed and optimization is performed with respect to X. The application of the mean field approximatiton to $\tilde{P}(V)$ allows for gradient accent in the E-step. The use of a simplistic shape model allows for the use of exhaustive search in the M-step. As T approaches 0, the estimate of $\tilde{P}(V)$ converges to a delta function centered on a local maxima of the likelihood function $P(Z,V;X)$. This form of optimization is similar to soft assign [2].

The benefits of this approach are:

- The final solution is based on a global optimization scheme which effectively propagates information from regions of high to low certainty.
- No prior information regarding the number of people in the scene is needed.
- Initialization is trivial and optimization can be achieved in an efficient manner.

6. Counting People

A *virtual gate* as proposed here can be used to estimate how many people enter or leave a particular site through any given point. Therefore it can, as discussed in Section 2, be used to replace traditional turnstiles. Given a scene captured by a surveillance camera, the user could simply draw a line or curve at any location in the field of view. Compared to using turnstiles, this approach is more flexible as there is no need to install any specific hardware.

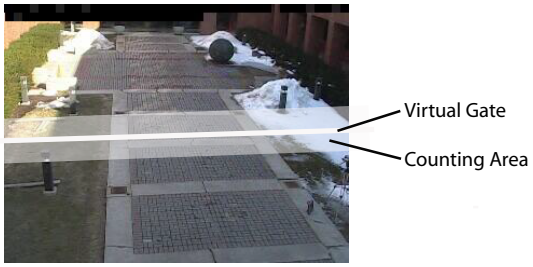

Figure 4: **Virtual gate.** *A line (or a curve) is drawed as the virtual gate in the field of the view. The area within certain distance to the gate is defined as the counting area. The relation between the individual's trajectory and the normal of the gate indicates the travel direction.*

As people can pass through a virtual gate in small groups, it is critical to be able to robustly count the number of individuals at any given time. In addition it is necessary to determine whether individuals enter or leave the site. The combination of the visual tracking module with the crowd segmentation module, as presented in Figure 2, addresses both of these requirements. Whereas the general integration of these two modules was discussed in Section 3, we now give details on how robust data association is achieved to track individuals through the counting area, illustrated in Figure 4.

6.1 Implementation Details

In any given region of interest the crowd segmentation module segments a group of people into individuals. For obvious reasons the accuracy of the person counts depends on the data association needed for tracking. An added benefit of this approach is that the crowd segmentation results are filtered with respect to time and therefore the system effectively reduces segmentation errors.

The tracker is composed of two parts. The first part is a simplified multiple hypothesis tracker, which is described in Section 3.2. The second part is the data association. Given the enhanced segmentation results from neighboring two frames, we use the Hungarian algorithm to find the optimal association. The 2D distance between a pair of segmented individuals from two frames is used to compute the cost matrix for the Hungarian algorithm. Once the data association is performed for every neighboring two frames, we can build the trajectory of each segmented person over time.

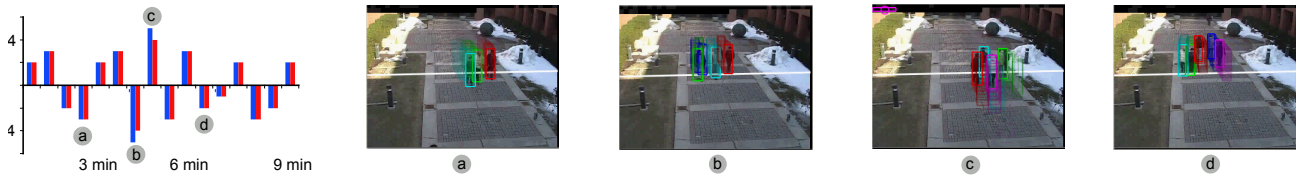

Figure 5: **Experimental results of people counting.** *For each instance long the time axis, the true number of people passing the gate (blue bar), and counting from our algorithm (red bar) is displayed. The snapshots of four instances are plotted on the right and please read the text for detail.*

The system continuously monitors the counting area (see Figure 4). Once people appear, the system estimates the position of all individuals and their direction of travel. The projection of the direction vector onto the normal of the virtual gate is used to determine whether they are entering or leaving. An identification number is then added to a list of people leaving L_o or people entering L_i depending on the above determined direction. The purpose of these two lists is to avoid multiple counting when the person remains inside the counting area in the future consecutive frames. Presently the system is being extended such that appearance information of every individual entering the gate is stored. Thus it will be possible to estimate the amount of time each person stays within a particular area of interest.

6.2 Experiments

The system was set up to monitor a side entrance to our facility. A single surveillance camera is mounted at about 6 meters above the ground. Images taken from this camera are shown in Figure 5. In order to test a number of scenarios subjects are asked to walk along the entrance in different constellations. While some instances are considered to be easy scenarios for people counting, such as a single person walking, the test set also contains difficult cases. For example, three or five people walk together as a group; two groups of people walk in the opposite direction.

One 10 minute long video sequence is used to quantify the results of the proposed system which are summarized in Figure 5. Each time a person or a group of people enters or leaves the gate, we plot the true number of people, and the counts computed by our algorithm. Positive numbers refer to going in the gate, and negative numbers refer to going out the gate. In most instances the estimated number of people entering and leaving as calculated by the algorithm corresponds to the ground truth. In Figure 5 four particular instances are presented. The trajectory of each person is illustrated by colored boxes. Fading positions illustrate the positions in previous frames. The algorithm obtains the right number of people in cases a and d, even though people are potentially occluded while passing the gate. For both cases b and c, the algorithm counts one person fewer than the ground truth, due to the heavy occlusion. For example, the fifth person in case c is almost fully occluded when he is passing the whole counting area. Given the viewpoint, the system is of course not able to take fully occluded people into account. If it is necessary to resolve all of these cases, the viewpoint of the camera needs to be changed or images from multiple cameras need to be considered.

7. Event Detection

Our model based approach, which makes effective use of geometric knowledge, has some specific advantages. In particular our system allows the estimation of distances between objects in the scene and therefore allows the consideration of spatial context. In the following it will be illustrated on how this approach can be used to detect specific events in the set of sequences provided by the Real-Time Event Detection Solutions Challenge [1]. One particular task which needs to be addressed here is the monitoring of passengers for the purpose of ensuring passenger safety on a train platform.

In this context, it is necessary to estimate the position of each passenger on the platform to avoid collisions with incoming trains. As passengers occlude each other, this problem cannot be solved by detecting foot positions alone. To be exact, it is necessary to estimate the positions of all people, even if they are occluded.

In this particular dataset, we can get a relatively good person/background segmentation by simply thresholding the infra-red video sequence. We automatically determine the threshold by analyzing the whole frame to determine the ambient infra-red energy. These foreground regions are then processed by the crowd segmentation module. Given the segmentation of the scene we can then estimate the position of each person on the ground plane.

In order to illustrate the benefits of our approach we compare our results with a naive implementation of a foot detector. For a given image location, we can compute a "foot strength" as the response to a corner-detection template. To detect proximity to the platform edge, we measure

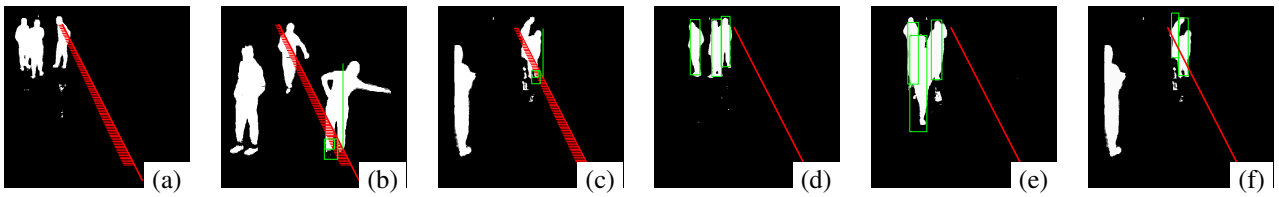

Figure 6: **Segmented frames from the subway sequence.** *The long red line shows the platform edge. In (a)-(c), the horizontal lines show the sampling positions for the foot detector, while the green lines show detected feet and the vertical "event threshold" line. In (d)-(e), the green boxes show the crowd segmentation separating overlapping people.*

foot strength along the platform image line at the points illustrated by the short red lines of Figure 6(a)-(c). Non-maximal suppression and thresholding then yield foot locations, illustrated by the green annotations. Detecting the leaning-over event is then simply a matter of determining if a part of the person segmentation falls to the right of the vertical green line, as happens in Figure 6(b). The comparison of these results with the crowd segmentation results shown in Figure 6(d)-(f) demonstrates the advantages of the proposed approach.

8. Summary and Conclusions

To conclude we present a surveillance system that consists of the following four components visual tracking, autocalibration, crowd segmentation, and a counting/event recognition module. Our experimental results document that there is a significant benefit of making extensive use of the site geometry to constrain the people detection problem and to extract relevant scene information.

The system is capable of segmenting groups of people into individuals and track these over time. Therefore we are, for example, able to count the number of people entering or leaving a particular site. This model based approach also allows to make effective use of spatial context which enables the system to detect certain events automatically.

References

[1] AVSS 2005. Real time event detection solutions for enhanced security and safety in public transportation. http://www-dsp.elet.polimi.it/avss2005/.

[2] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(3):114–141, March 2003.

[3] D. Gavrila. Pedestrian detection from a moving vehicle. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, volume 2, pages 37–49, 2000.

[4] J. Giebel, D.M. Gavrila, and C. Schnörr. A bayesian framework for multi-cue 3d object tracking. In *Proc. 8th European Conf. Computer Vision, Prague, Czech Republic*, pages 241–252, 2004.

[5] I. Haritaoglu, D. Harwood, and L. S. Davis. HYDRA: Multiple people detection and tracking using silhouettes. In *In IEEE International Workshop on Visual Surveillance*, pages 6–13, 1999.

[6] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[7] S. Ioffe and D.A. Forsyth. Probabilistic methods for finding people. *Int. J. Computer Vision*, 43(1):45–68, June 2001.

[8] C. Schmid K. Mikolajczyk and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. 8th European Conf. Computer Vision, Prague, Czech Republic*, volume 1, pages 69–82, 2004.

[9] N. Krahnstoever and P. Mendonca. Bayesian autocalibration for surveillance. In *Proc. of IEEE International Conference on Computer Vision (ICCV'05), Beijing, China*, October 2005.

[10] A. Mittal and L.S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In *Proc. 7th European Conf. Computer Vision, Kopenhagen, Danmark*, volume X, pages 18–33, 2002.

[11] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. People recognition in image sequences by supervised learning. In *MIT AI Memo*, 2000.

[12] J. Rittscher, P. Tu, and N. Krahnstoever. Simultaneous estimation of segmentation and shape. Technical report, 2005.

[13] A. W. Senior. Tracking with probabilistic appearance models. In *ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems*, pages 48–55, 2002.

[14] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc 12th IEEE Computer Vision and Pattern Recognition, Santa Barbara, CA*, volume 2, pages 246–252, 1998.

[15] Danny B. Yang, Héctor H. González-Baños, and Leonidas J. Guibas. Counting people in crowds with a real-time network of simple image sensors. In *Proc. 9th Int. Conf. on Computer Vision, Nice, France*, pages 122–129, 2003.

[16] T. Zhao and R. R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208– 1221, September 2004.