Detecting and Counting People in Surveillance Applications

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

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. One relevant application of this algorithm is people counting. 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 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 [4, 2, 10, 6, 3, 7] has been made to detect 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. In all of the scenarios just mentioned we have to anticipate that people appear in groups. Certain actions can only be detected if the location of all individuals in the scene is known. In addition it is often necessary to know how many people are present. Recently [11] we proposed a model based segmentation algorithm which allows to partition a group of people into individuals. It will be demonstrated how this algorithm allows the detection of certain events in the underground scenario illustrated in figure 1. Naturally this algorithm can also be applied to count the number of individuals in the field of view. Here is algorithm is extended to count the number of people entering or leaving the site using the idea of a virtual gate. A summary of the related work is given in the following section. An overview of our approach is given in section 5. The idea of counting the number of people entering and leaving through a virtual gate is outline in section 6. Experimental results of the counting system system are presented in section 7. The recognition of specific events is discussed in section 8.

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 certain actions of people. It will be shown that by segmenting groups of people into individuals their position can be estimated more accurately.*

2. Related Work

Various techniques [10, 2, 3] 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.* [10] use Support Vector Machines to this problem. Gravrila [2] uses a tree based classifier to represent possible shapes of pedestrians. Griebel et al. [3] use dynamic point distribution models. An alternative to modelling the appearance of an entire person is designing detectors for specific objects 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 [7]. A learning approach is then being used to combine the set of weak classifiers to body part detectors which are then combines 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 [9] explicitly assigns the pixels in

0-7803-9385-6/05/$20.00 ©2005 IEEE

each camera views to a particular person using colour histograms. Zhao and Nevatia [14] 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. 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 [11] that simultaneously estimates the position and size. The details of our approach will be, as mentioned before, presented in section 5.

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 extract the direction of travel. Here we introduce the idea of a virtual gate. Yang *et al.* [13] propose a people counting method that makes use of different views. Here we demonstrate that is possible to obtain reliable counts from a single view.

3. Site Geometry

Knowledge about the site geometry and the camera parameters makes 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 much 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. One 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. We address the problem on two fronts. First, the foot to head plane homology is efficiently estimated from head and foot location measurements to obtain the internal and external calibration parameters of the camera. Second a Bayesian solution to the calibration problem that can elegantly handle measurement uncertainties, outliers, as well as prior information is employed. The full posterior distribution of calibration parameters given the measurements can be estimated, which allows making statements about the accuracy of both the calibration parameters and the measurements involving them. See [8] for a detailed description of our approach.

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

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

with z the height of the camera above the origin of the ground plane, $\tilde{\mathrm{v}}$ the vanishing point and $\tilde{\mathrm{l}}$ the horizon line. It can furthermore be shown that the horizon line is given by $\tilde{\mathrm{l}} = \left[\sin(\rho) \quad -\cos(\rho) \quad \frac{f}{\tan(\theta)}\right]$ and the vanishing point by $\tilde{\mathrm{v}} = \left[f\sin(\rho)\sin(\theta) \quad -f\cos(\rho)\sin(\theta) \quad \cos(\theta)\right]$, 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[8], knowlegde of the foot to head homology yields complete metric calibration of a camera with respect to the ground plane.

The overall auto calibration approach that is described in detail in [8] 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 [5]. 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.

4. System Overview

The system (see Figure 2) consists of three main components: A standard low-level foreground estimation algorithm, a template-based tracker as well as the crowd segmentation. All three components are combined into a tightly coupled framework. In the following, the tracker and the integrated crowd segmentation components are described in more detail.

4.1. Tracker

The tracker uses an adaptive appearance based approach similar to [12, 15]. 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

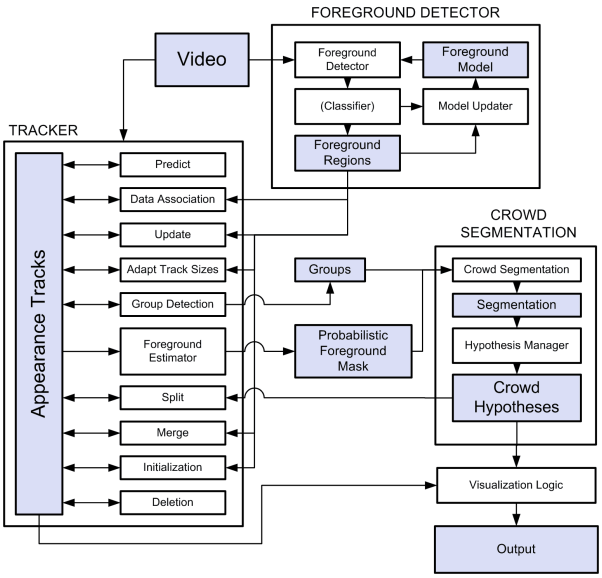

Figure 2: **System overview.** *There are three components: foreground detector, probabilistic tracker, and the crowd segmentation algorithm. See text for details.*

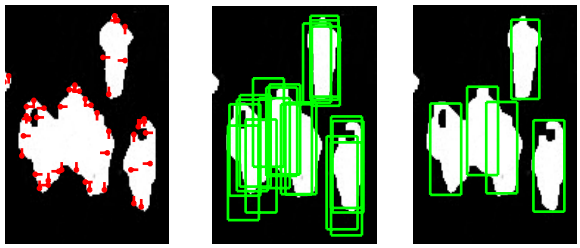

Image Features Z **Clique Set** $\mathscr{C}$ **Segmentation** X^*

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* **??**.

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 motivation 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.

4.2 Crowd Segmentation Component

The crowd segmentation algorithm processes all regions in the image that the tracking component has judged to contain groups of people. 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.

5. Model Based Segmentation

In [11] a model based approach to crowd segmentation was proposed. This method is now sumerized. Given a foreground segmentation, a set of low level image features $Z = \{z_i\}$ are extracted. In addition, an exaustive set of feature groupings or cliques $C = \{c_i\}$ is hypothesised. 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 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:

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

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 favours 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 perfomed. 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 exaustive 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 [1].

The benefits of this approach are:

- The final solution is based ob a global optimization scheme which effectively propogates 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

One traditional way to perform people counting is to install turnstiles. However, it has the drawback of high cost and low flexibility. In this paper, we propose the idea of *virtual gate*. Given a scene captured by a surveillance camera, the user could simply draw a line at any location in the field of the view, and the algorithm continuous counts how many people are going through the line. Comparing to using turnstiles, this approach is highly flexible. Furthermore, it enables some applications that are not feasible using the traditional method. For example, the retail store might be interested in knowing how many people are strolling in a particular area of the store. We could draw a virtual gate, which could be a line or a curve, that covers the area of interests.

Given a video sequence of interest, the crowd segmentation algorithm in the previous section classifies the crowd in *each* frame into multiple people. However, there is no association between the people in the neighboring frames. In order to determine the number of people and their moving direction, the trajectory of each person over time is needed, which means we need to integrate tracking and crowd segmentation. Also, another benefit of integration is that the crowd segmentation results can be stabilized and enhanced. Over a long sequence, the crowd segmentation might not obtain the right number or location of the person for certain frames due to heavy occlusion. In this case, the tracker can smooth out the segmentation result such that it is consistent with the trajectory.

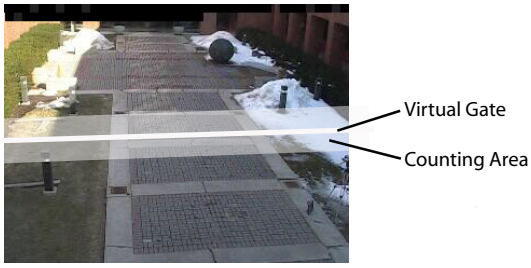

Figure 4: **The virtual gate and the counting area in a scene.**

It is relatively easy for a surveillance system to track individual person if there is no occlusion among people. However, the people tracking becomes difficult when a group of people walking together as a crowd. When this happens at the virtual gate, the crowd segmentation for each individual frame might not obtain the same number of people counting. Thus the main challenge of people counting is to be robust to the errors from the crowd segmentation, by using a integrated tracker.

The tracker is composed of two parts. The first part is a simplified multiple hypothesis tracker, which is described in Section 4.2. 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 person from two frames is used in computing 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.

Having introduced the integration of segmentation and tracking, we will present our approach in people counting. As shown in Figure 4, given a virtual gate in the scene, we consider the area within certain distance to the gate as the *counting area*. We keep monitor the counting area. Once one segmented person enters this area, we start the counting module, which performs two tasks. One is to retrieve the trajectory of this person and determine his/her moving direction with respect to the virtual gate. There are many ways to perform this 2-class classification, *i.e.*, going in the

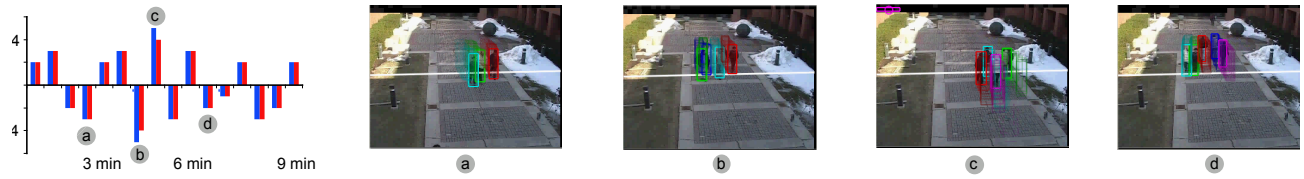

Figure 5: **Experimental results of people counting.** *For each instance long the time axis, the true number of people passing the gate (blue bar), and counting from our algorithm (red bar) is displayed. The snapshots of four instances are plotted on the right.*

gate or going out the gate. We found that simply comparing two instances of the person's location along the direction perpendicular to the gate performs reasonable well. The other is to add the identification (ID) number into the list of people going in the gate (*ListIn*) or going out the gate (*ListOut*) 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. Furthermore, if we associate the person in the ListOut and ListIn based on the appearance information, we can find out how long each person stays in the place of interests.

7. Experiments

To test the performance of our algorithm, we collect a number of sequences as illustrated in Figure **??**. One single camera is mounted on the second floor and pointing to the entrance of a building. A number of subjects are asked to walk along the entrance in different combination of groups. While some instances are considered as easy scenario for people counting, such as one single person walks through, there are also many difficult cases. For example, three or five people walk together as a group; two groups of people walk in the opposite direction.

We use a 10-minutes video clip, where people pass the virtual gate for 16 times, as the testing sequence. Figure 5 illustrates the experimental results on this sequence. At each instance of passing gate, we plot the true number of people, and the counting from our algorithm. Positive numbers refer to going in the gate, and negative numbers refer to going out the gate. We can see for most instances the algorithm can correctly count the people. We also display the snapshot of four instances, where the trajectory of each person is illustrated by colored boxes with fading positions at previous frames. The algorithm obtains the right number of people in case a and d, even though there is occultation while passing the gate. For both case b and c, the algorithm counts one person less than the ground truth, due to the heavy occultation. For example, the fifth person in case c is almost fully occulted when he is passing the counting area. For these extremely difficult cases, using multiple cameras should be very helpful for people counting.

8. Event detection

Automatically detecting behavioural events is a problem of great importance, with applications in many disparate areas such as surveillance, safety, and marketing. A specific challenge problem posed at this conference is detecting dangerous behaviour on a subway platform. One such behaviour is when a person leans over the edge of the platform. The goal here is to automatically detect this event.

The key elements of the solution is to detect the people, detect the proximity to the platform edge, and then to detect when portions of the person goes over the edge. 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.

The platform edge is defined by a line in the image. Because of the oblique view, we cannot simply conclude that a person is near the platform edge when a person segmentation crosses the line. For example, the top person detection in Figure 6(a) crosses the line, but is not near the platform edge. The only reliable way is to detect proximity is to detect when a person's foot is near the line. For this particular scenario, we can model the foot location as the lower-left corner of a person segmentation. 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 foot strength along the platform image line at the points illustrated by the red lines in the top row of Figure 6. 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).

While this approach solves the problem for many cases, it fails when people overlap in the view. In such cases, we can use the crowd segmentation algorithm to tease apart the people. Figure 6(c) shows a situation where the first per-

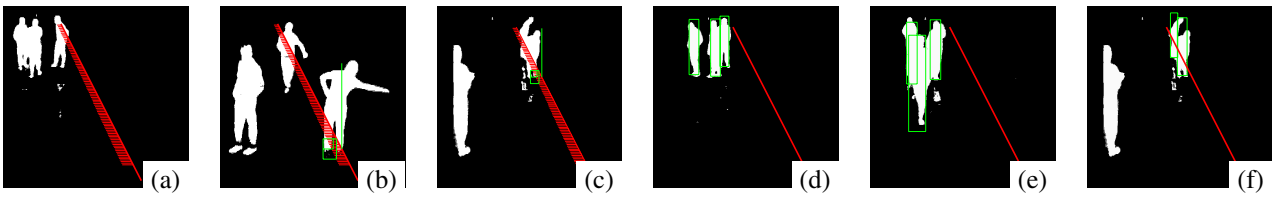

Figure 6: **Segmented frames from the subway sequence.** *The long red line shows the platform edge. In (a)-(c), the horizontal lines show the sampling positions for the foot detector, while the green lines show detected feet and the vertical "event threshold" line. In (d)-(e), the green boxes show the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

son partially obscures the second person, so that the foot location of the second person cannot be easily determined. Thus, with the simple approach above, we cannot detect any behaviour regarding the second person, because we only have a foot location for the first person. After applying crowd segmentation, however, we can determine that the foot of the second person is also at the platform edge, and that his head is beginning to cross into the dangerous area. This is shown in Figure 6(f).

9. Summary and Conclusions

The paper presents two application of a recently developed appraoch to crowd segmentation to two different surveillance problems: people counting and event recognition. The experiments presented demonstrate that our approach of counting people passing through a virtual gate produces reliable counts.

The results presented in section 8 demonstrate that segmenting groups of people into individuals aids the event recognition process by identifying the location of each individual.

References

[1] 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.

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

[3] 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.

[4] 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.

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

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

[7] 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.

[8] N. Krahnstoever and P. Mendonca. Bayesian auto calibration for surveillance. Technical Report GRC Some Number, GE Global Research, 2005.

[9] 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.

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

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

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

[13] 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.

[14] T. Zhao and R. R. Nevatia. Bayesian human segmentation in crowded situations. In *IEEE Computer Vision and Pattern Recognition, Madison, Wisconsin*, volume 2, pages 459–466, 2003.

[15] 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.