REAL-TIME ORIENTATION OF A PTZ-CAMERA BASED ON PEDESTRIAN DETECTION IN VIDEO DATA OF WIDE AND COMPLEX SCENES

T. Hoedl *, D. Brandt, U. Soergel, M. Wiggenhagen

IPI, Institute of Photogrammetry and GeoInformation, Leibniz Universitaet Hannover, Germany
(hoedl, soergel, wiggenhagen)@ipi.uni-hannover.de

Intercommission Working Group III/V

KEY WORDS: Computer Vision, Detection, Close Range Photogrammetry, Absolute Orientation, Urban Planning, Tracking, Multisensor, Real-time

ABSTRACT:

Object detection and tracking is the basis for many applications in surveillance and activity recognition. Unfortunately the utilized cameras for the observation of wide scenes are mostly not sufficient for detailed information about the observed objects. We present a two-camera-system for pedestrian detection in wide and complex scenes with the opportunity to achieve detailed information about the detected individuals. The first sensor is a static video camera with fixed interior and exterior orientation, which observes the complete scene. Pedestrian detection and tracking is realized in the video stream of this camera. The second component is a single-frame PTZ (pan / tilt / zoom) camera of higher geometric resolution, which enables detailed views of objects of interest within the complete scene. For this reason the orientation of the PTZ-camera has to be adjusted to the position of a detected pedestrian in real-time in order to caption a high-resolution image of the person. This image is stored along with time and position stamps. In post-processing the pedestrian can be interactively classified by a human operator. Because the operator is only confronted with high-resolution images of stand-alone persons, this classification is very reliable, economic and user friendly.

KURZFASSUNG:

Die Erkennung und Verfolgung von Objekten ist die Grundlage für zahlreiche Überwachungsaufgaben und Anwendungen zur Aktivitätserfassung. Leider liefern die für weiträumige Szenen eingesetzten Kameras nur selten detaillierte Information über die beobachteten Objekte. In dieser Arbeit wird ein Zwei-Kamera-System zur Passantenerkennung in weiträumigen, komplexen Szenen vorgestellt, welches die Möglichkeit bietet, detaillierte Information über die beobachteten Personen zu erfassen. Bei der ersten Kamera handelt es sich um eine Videokamera mit fester innerer und äußerer Orientierung, welche die komplette Szene erfasst. Die Erkennung und Verfolgung der Personen erfolgt in den Bildsequenzen dieser Kamera. Die zweite Komponente ist eine PTZ (pan / tilt / zoom – schwenk / neige / zoom) Einzelbildkamera mit hoher Auflösung, die eine detaillierte Aufnahme des interessierenden Objekts innerhalb der Szene ermöglicht. Zu diesem Zweck wird die PTZ-Kamera in Echtzeit auf eine detektierte Person ausgerichtet und ein hoch aufgelöstes Bild der Person aufgenommen. Das Bild wird mit Positions- und Zeitstempel gespeichert um eine eindeutige Zuordnung zu der berechneten Trajektorie zu gewährleisten. Die gewünschten Personenmerkmale können von einem Auswerter im Post-processing erfasst werden. Da der Auswerter hoch aufgelöste Bilder von unverdeckten Personen zur Verfügung hat, ist diese Klassifikation sehr zuverlässig, wirtschaftlich und benutzerfreundlich.

1. INTRODUCTION

1.1 Motivation

The work presented here is embedded in the framework of an interdisciplinary research project aiming at the assessment of the quality of shop-locations in inner cities. In this context the number, the behaviour (e.g., walking speed and staying periods), and the kind (e.g., in terms of gender and age) of pedestrians passing by are crucial issues. In general there are different options for achieving the desired information. On the one hand sensors are feasible, which are carried by persons and which deliver their position based on existing infrastructure (mobile phones, GPS, RFID, Bluetooth). On the other hand such information can be derived entirely from observations from outside (cameras). To be independent from active cooperation of the individuals and to be able to collect information about all pedestrians, cameras were chosen as the appropriate sensors for this project. The task requires both the surveillance of a large and complex scene and at the same time the need to gather high-resolution data of individuals, which can hardly be fulfilled by a single camera system. Hence, a two-camera-system set-up is used in this approach.

The first one, the observation camera, is a static video camera with fixed interior and exterior orientation. Pedestrian detection and tracking must occur in real time in the video stream of this camera. The positions of the detected individuals in object space are passed to the second camera. This camera is a PTZ (pan / tilt / zoom) camera of higher geometric resolution, which enables to focus on objects of interest within the complete scene. Hence a detailed analysis of the individuals is possible.

* Corresponding author

1.2 Related Work

Due to the broad range of applications (surveillance, activity recognition or human-computer-interaction) human motion analysis in video sequences has become one of the most active fields in computer vision in recent years. Latest surveys of the numerous publications were issued by Moeslund et al. (2006) and Yilmaz et al. (2006).
One focus of research is automatic detection and tracking of humans in uncontrolled outdoor environments. The tracking of articulated objects such as human bodies is much more complex, than the tracking of solid objects, as e.g., cars, due to the fact that the relation of the limbs changes by time. Nevertheless these approaches show already promising results, especially for simple scenes populated by only a few individuals.

The initial step in many approaches is background subtraction. For many years background subtraction was only used for controlled indoor environments, but with the adaptive Mixture of Gaussian (MoG) method by Stauffer & Grimson (1999) it also became a standard for outdoor environments. Recent advances in background subtraction, which are mostly based on the MoG-Algorithm, deal with minimizing false positives or negatives, for example due to shadows, or background updating.

Moeslund et al. (2006) categorize approaches for object detection based on the segmentation methods: motion, appearance, shape and depth-based. Any use of just one of these methods is only successful to a certain point in complex scenes. For this reason newer approaches combine several segmentation methods, e.g., Viola et al. (2005) combine motion and appearance based segmentation, Leibe et al. (2005) integrate colour information in their shape based approach.

Establishing temporal correspondences of the segmented objects is an essential task for object tracking, especially for scenes with heavy occlusions. One efficient approach is the application of a Kalman-filter (Zhao & Nevatia, 2004; Rosales & Sclaroff, 1998). Weaknesses of this approach appear in the case of abrupt variation in speed or moving direction. Improvements can be achieved by integration of microscopic motion models for the objects, that are adapted to typical events or positions in object space (Antonini et al., 2006).

Automatic extraction of additional attributes of humans, such as gender and age, is a difficult pattern recognition problem, since dressed persons have a big variety in shape and dress style, the facial expression of individuals may change significantly depending on mood, and variations in lighting conditions may occur. So far this task was addressed only in few approaches.
In general these methods rely on high-resolution images of faces or are analysing the gait of persons. The opportunity for classifying the pedestrians by analysis of their gait is not given in this case, because it requires continuous observation of the whole body. In complex scenes this is not realizable.
Good results for gender classification of faces are reported by Baluja & Rowley (2007). They are using AdaBoost and achieve correctness over 90%. Lanitis et al. (2004) determine the age of person with a variance of about 5 years. Both approaches underlie the constraints of a close defined viewing angle, uniform lighting conditions, and the absence of occlusions.
These constraints are not met in this project, because we have to deal with wide and complex scenes. Individual pedestrians may wear different kinds of clothing, can appear with arbitrary orientation with respect to the camera, and differ in size and shape. An automatic extraction of the features age and gender is therefore hardly possible. Instead, a semi-automatic approach is developed in which such decisions are made by an operator.

2. SYSTEM SET-UP

As described above, the system consists of two cameras. To ensure, that both cameras cover the same area and have a similar viewing angle, they were placed close to each other. The first one, the observation camera, is a static video camera with fixed interior and exterior orientation. For our initial experiments we used a standard webcam with a resolution of 640*480 pixels. The second camera is a single-frame PTZ-camera of higher geometric resolution. Due to the fact, that common PTZ-cameras are video cameras with pal-resolution, we developed own prototypes. These prototypes allow the use of standard high-resolution single-frame cameras. The frameworks of these prototypes were realized with Lego NXT. The employment of Lego NXT for this task has two major advantages: The framework can be adjusted on demand, so the projection centre of the cameras was positioned into the origin of the rotation axis. Furthermore, the control of motors for the rotations can be realized in C++ Code.
Two framework prototypes were developed, one for lightweight compact cameras and one for heavier single-lens reflex cameras.

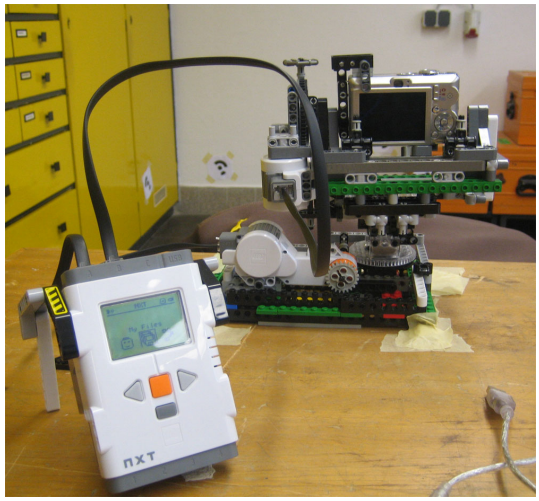

Figure 1. Prototype for lightweight compact cameras.

3. WORKFLOW

The complete scene of interest is observed by the video camera. Pedestrians appearing in the scene are detected and tracked in the video stream. The positions of the individuals are transformed from image to object space using projective transformation. Based on the positions of the pedestrians in object space the orientation parameters of the PTZ-camera are computed and high resolution images of the tracked persons are acquired. These high-resolution images are classified interactively. In the following the major steps are described in detail.

3.1 Video stream analysis

This project requires real-time analysis of the video stream for pedestrian detection and tracking. We used the free available OpenCV-library, which is implemented in C and C++ Code. The library offers a broad range of computer vision functions and allows an easy link to our PTZ-camera prototypes. For the graphical user interface and to enable a continuous observation of the scene in the video camera while acquiring images with the PTZ-camera we used the Qt-library.

Background subtraction

In order to reduce the false positive rate, which usually is a problem in different approaches for people detection in complex scenes, background subtraction was chosen as first step. In background subtraction the static background is separated from all moving objects, which are segmented as foreground. In the following only foreground pixels, which contain the regions of interest for pedestrians, are examined.

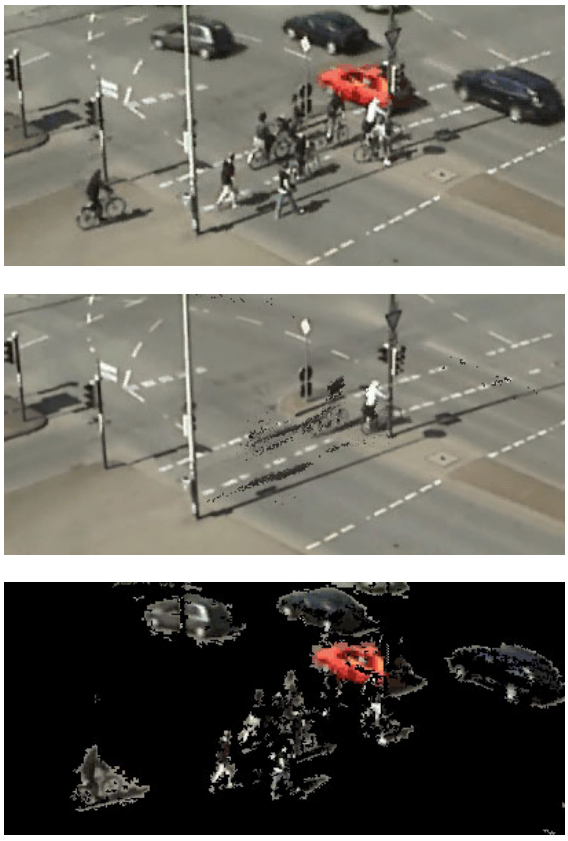

Figure 2. Section of a test scene (upper image), background (middle image) and foreground (bottom image) segmented with MoG-algorithm.

For complex outdoor scenes an adaptive approach for background subtraction is important, due to changing lighting conditions. Moreover, objects that move into the scene, but become static after a while, for example parking cars, should be classified as background.

Three approaches for adaptive background subtraction were investigated:

1) The OpenCV-implementation of a modification of the MoG-Algorithm: The colour of each pixel in the successive frames is described by multiple Gaussian distributions. The dominating distributions are interpreted as background, while weak distributions indicate the existence of a foreground object at the pixel position.
2) The OpenCV-implementation of a Bayes decision framework by Li et al. (2003): The classification decision is based on the statistics of feature vectors. Stationary background objects are described by colour features. Furthermore moving background objects can be classified using colour co-occurrence features.
3) Median filtering: Comparison of the actual pixel grey value and the Median grey value of m pixel values of every n-th preceding frame. If both values match, the existence of background is assumed. Best results with this approach were gained with m = 9 and n = 9.

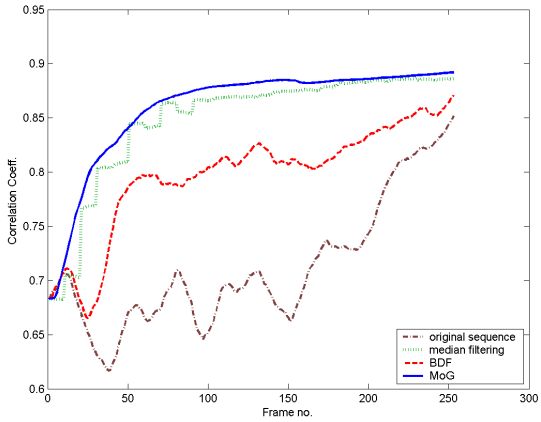

Figure 3. Correlation of the background estimated by three different approaches and the reference background.

After an initialisation phase all tested approaches delivered reasonable results. Problems generally occurred if the background grey value and the object value are very similar.
Figure 3 shows the correlation of the background estimated by the three different approaches and the "true" reference background. Moreover the correlation of the original image and the reference background is displayed. This correlation is an indicator for the number and size of moving objects within the scene.
For the tested scenes best results were achieved with the MoG-Algorithm. One challenge in this approach is, after when a non-moving object should be classified as background (e.g., person waiting at the traffic lights, fig. 2, middle image). Problems also occur in case of shadows of moving objects, since they are also classified as foreground objects.
The Bayes-approach can deal with this challenge and is also able to deal with waving branches. In general the results of this approach are comparable to the MoG-results for close objects, but it showed weaknesses for small and distant objects.
The Median-filtering approach is up to three times faster than the other approaches. The disadvantage of this approach in comparison to the MoG-approach, is that only one span of grey values is treated as background. Hence it does not reach the classification quality of the other approaches in case of quickly changing illumination conditions.
For our further proceeding we used the MoG-Algorithm.

Pedestrian detection

In literature several approaches for pedestrian detection are described. We investigated two approaches, one based on background subtraction while the second approach consists of the application of already trained classifiers for persons. The later can also be applied on the original images.

The first approach is, to group the segmented foreground pixels into connected components. Small components, which result from noise or small non-interesting objects, and which do not pass an area criterion, are eliminated. The remaining components can be divided into three categories:

- Single stand-alone persons
- Groups of persons
- Other moving objects (cars, cyclists, …)

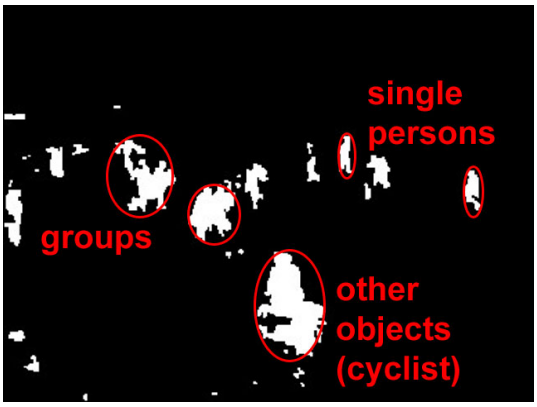

Figure 4. Foreground segmentation result with connected components superimposed with best fitted ellipses.

To enable the acquisition of just one person in one image of the PTZ-camera, only the category of the stand-alone persons is of further interest. This category can be distinguished from the other categories by superimposing the connected component with a best-fitted ellipse. The ratio of the both semi axis and the height of the component in object space are computed. If both criteria fit predefined tolerance ranges, the existence of a stand-alone person is assumed.

The second approach is using classifiers for persons based on AdaBoost (Viola & Jones, 2001). The principle of AdaBoost is, to combine many weak classifiers, in this case based on rectangular features, to form a strong classifier. On the basis of a cascade of the weak classifiers, the algorithm is very fast. It can be applied on the original image data as well as on the segmented foreground data. In the later case a dilatation of the segmented foreground is required to close gaps in the segmented objects.
OpenCV offers already trained classifiers for the detection of faces, upper and lower bodies and whole human bodies. The whole body detector works quite well in non-complex scenes and can also be applied for finding single persons. In case of occlusions of some body parts, it failed. Thus the solely employment of this detector is not sufficient for our tracking purposes. In complex scenes the upper part of the body is far less occluded, than the lower body part. Hence the use of the upper body and the face detector is reasonable. The lack of the face detector is, that it requires a larger scale of the pedestrian (at least 30*30 pixels for the head), and that the person must be frontal to the camera. Efficient pedestrian detection can be achieved by using different detectors depending on the position and orientation of the individuals within the image.

Pedestrian tracking

The challenge in tracking is finding the correct temporal correspondences of detected objects in successive frames. This is a minor problem, as there is just one object or multiple distinct objects in the scene, but it becomes harder in case of splitting and merging objects or in case of occlusions.
For tracking of the connected components we are applying the kernel-based approach by Comaniciu et al. (2003), tracking of the pedestrians detected with AdaBoost is performed using Kalman-filtering. Since the test scenes have only few occlusions, so far these approaches satisfy our demands. For more complex scenes a deeper evaluation is in progress.

3.2 Computation of 3D-Positions

All computations so far were executed in image space of the video sequence. For recording the trajectories and to calculate the orientation of the PTZ-camera a transformation of the positions of the detected individuals to object space is required.

In our case the area under observation is approximately a plane surface. The transformation of coordinates from one plane to another can be achieved by using projective transformation.

$$X = \frac{a_0 + a_1 x' + a_2 y'}{c_1 x' + c_2 y' + 1} \qquad (1)$$

$$Y = \frac{b_0 + b_1 x' + b_2 y'}{c_1 x' + c_2 y' + 1} \qquad (2)$$

where X,Y = object coordinates
x',y' = image coordinates
$a_0,a_1,a_2,b_0,b_1,b_2,c_1,c_2$ = transformation parameters

The eight transformation parameters are determined with at least four planar control points, which should be placed in the corners of the observation area. When using more than 4 control points the parameters are determined by an adjustment.

Since we assume, that all persons are upright, the pixels of the connected components belong to different height coordinates in object space. To achieve the correct position of completely visible persons the bottom pixel of the component must be transformed.
Unfortunately in complex scenes the feet of a person often can not be seen because of occlusions. In this case a mean height of the person of 1.75 meters is assumed and the upper pixel is transformed into a parallel plane.

3.3 Orientation of the PTZ-camera

The parameters of the interior orientation for the PTZ-camera are determined by a testfield camera calibration. With this method the calibrated focal length, principal point, radial and tangential lens distortions were calculated for the later correction of image distortions. The initial values for the exterior orientation of the PTZ-camera are calculated by resection in space. If more than three control points are visible

in the field of view the exterior orientation parameters are determined by an adjustment.

To rotate the PTZ-camera to a point of interest detected by the observation camera two rotation angles have to be calculated. The first (α) lies in the horizontal plane and the second (β) in a plane perpendicular to the ground plane. After consideration of the current rotation angles, the optical axis of the PTZ-camera points to the centre of the detected object. Then the object will be imaged in the middle of the high resolution image of the PTZ-camera.
In the event loop of the motion control following steps have to be calculated with known initial values for the exterior orientation and the calibrated focal length c of the PTZ-camera:

- Calculation of X,Y,Z coordinates of the object location from x',y' image coordinates of the observation camera with equation (1) and (2).
- Calculation of x", y" image coordinates for the PTZ-camera with the collinearity equation and initial exterior orientation parameters.
- Calculation of rotation angles α and β with $\alpha = \text{atan}(x"/ c)$ and $\beta = \text{atan}(y" /c)$.

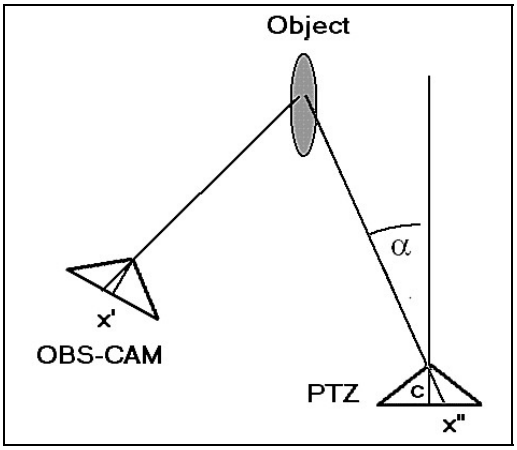

Figure 5. Calculation of the rotation angle α in the horizontal plane.

3.4 Analysis of high-resolution images

The high-resolution images are stored with time and position stamps. A unique assignment of the attributes to the trajectories is feasible, because of the observation of stand-alone pedestrians. Since there is no automatic classification implemented up to now, the pedestrians are interactively classified by a human operator.

4. OUTLOOK

The two-camera system presented in this work is an essential component for the assessment of the quality of shop-locations in inner cities, since it delivers both the trajectories of the pedestrians and information about additional attributes. So far, the project is just in its starting phase. Actual investigations concentrate on the improvement of speed and accuracy of the control for the exterior orientation and on the zoom-control of the PTZ-camera. This will improve the reliability of the assignment of the classified pedestrians to their trajectories. Furthermore the choice of a person from a number of detected individuals up to now is arbitrary. A knowledge-based camera control system will ensure the single acquisition of each individual or the multiple acquisition of the same person if wanted.

Future work will investigate the opportunity for automatic classification of the pedestrians in the high-resolution images. For this purpose different classifiers (e.g., AdaBoost, SVM) will be evaluated. A training data set is gained by acquisition of the high-resolution images.
The integration of a complete 3D-model of the scene allows an improved modelling of the behaviour of pedestrians at obstacles or in case of occlusions. Moreover motion models that specify the probability for a change in direction or speed at certain position will be integrated.

ACKNOWLEDGEMENT

Part of the work presented here was carried out in a student research project by (in alphabetical order): Jonas Bostelmann, Steven Piorun, Falko Schindler, Axel Schnitger, Aiko Sukdolak, Martin Wiedeking,

REFERENCES

Antonini, G., Venegas Martinez, S., Bierlaire, M., Thiran, J. P., 2006. Behavioral Priors for Detection and Tracking of Pedestrians in Video Sequences. *International Journal of Computer Vision,* 69, p. 159-180.

Baluja, S., Rowley, H., 2007. Boosting Sex Identification Performance. *International Journal of Computer Vision,* 71, p. 111-119.

Comaniciu D., Ramesh, V., Meer, P., 2003. Kernel-Based Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 25, p. 564-577.

Lanitis, A., Draganova, C., Christodoulou, C., 2004. Comparing Different Classifiers for Automatic Age Estimation. *IEEE Transactions on Systems, Man and Cybernetics* - Part B: Cybernetics 34, p. 621-628.

Leibe, B., Seemann, E., Schiele, B., 2005. Pedestrian Detection in Crowded Scenes. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),* p. 878-885.

Li, L., Huang, W., Gu, I.Y.H., Tian, Q., 2003. Foreground Object Detection from Videos Containing Complex Background. *Proceedings of the eleventh ACM international conference on Multimedia,*p. 2-10.

Moeslund, T. B., Hilton, A., Krüger, V., 2006. A Survey of Advances in Vision-based Human Motion Capture and Analysis. *Computer Vision and Image Understanding,* 104, p. 90-126.

Rosales, R., Sclaroff, S. 1998. Improved Tracking of Multiple Humans with Trajectory Prediction and Occlusion Modeling. *IEEE Conference on Computer Vision and Pattern Recognition*. Workshop on the Interpretation of Visual Motion.

Stauffer, C., Grimson, W. E. L., 1999. Adaptive Background Mixture Models for Real-time Tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2, p. 246-252.

Viola, P., Jones, M., 2001. Robust Real-time Object Detection. *Second International Workshop on Statistical and Computational Theories of Vision - Modelling, Learning, Computing and Sampling*.

Viola, P., Jones, M. J., Snow, D., 2005. Detecting Pedestrians Using Patterns of Motion and Appearance. *International Journal of Computer Vision,* 63, p. 153-161.

Yilmaz, A., Javed, O., Shah, M., 2006. Object Tracking: A Survey. *ACM Computing Surveys,* 38, Article 13.

Zhao, T., Nevatia, R., 2004. Tracking Multiple Humans in Complex Situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 26, p. 1208-1221.