Chapter 10

# SECURITY AND BUILDING INTELLIGENCE

## *From people detection to action analysis*

G.L.Foresti[1], C.Micheloni[1], L.Snidaro[1], P.Remagnino[2]

[1] *Department of Computer Science (DIMI), University of Udine, Italy*
{Foresti,Michelon,Snidaro}@dimi.uniud.it

[2] *Digital Imaging Research Centre (DIRC), Kingston University, UK*
p.remagnino@kingston.ac.uk

**Keywords:** Ambient intelligence, people detection and counting, tracking, event classification and association

## 1. Introduction

The events occurred over the last few years have increased the importance of security. Computer vision and image processing play a paramount role in the development of surveillance systems and they are commonly used to interpret video data. Video or visual surveillance is employed in applications such as traffic monitoring [12, 19] and the automatic understanding of human activity [6]. Generally, monitoring means keeping under control a wide area by deploying a heterogeneous network of sensors including static cameras [8], ominidirectional cameras [9], motorised pan-tilt-zoom (PTZ) cameras [16] and on-board cameras (mounted on vehicles, for instance). When more cameras are used, data and information fusion algorithms must be devised; this effectively can be seen as a cooperation among sensors to solve the common tasks of monitoring a public or private area [3, 13].

*Ambient Intelligence* (AmI) introduces a new transparent communication layer to support machine intelligent algorithms, ultimately capable of customizing the living environment to best suit the person habits and to improve their quality of life. AmI fosters "environments able to recognize and respond to the presence of individuals in an invisible way", as stated by the IST Advisory Group (ISTAG) in the Final Report "Scenarios for Ambient Intelligence in 2010" [10]. Devising such systems is a major endeavor, that only a multidisciplinary team of researchers and technologists can make possible.

This chapter is concerned with the development of algorithms for the intelligent building. In particular, the focus is on all those technological innovations embedded in a building that enable a continuous response and adaptation to changing conditions, increasing the comfort and security of its occupants, and allowing for a more efficient use of resources. Resource management is a crucial aspect usually addressed to maximize the return of investment and to fulfill the objectives of the organization who owns it [5], [11].

In this work, we describe a security system that follows some of the criteria of AmI, to supply custom information from the monitoring of a wide area such a university building. In particular, a distributed surveillance system able to track people either on or through the floors of the building has been studied. The system is composed of a network of sensors strategically placed to detect people inside the monitored environment. Events are automatically generated once a person takesa predetermined action as entering or exiting the building, taking the stairs etc. A correlation process among the set of events allow to determine the path designed by single persons from the time instant of their ingress till the exiting from the building.

By knowing, for each time instant, the position of a person we can perform a customization of the information supplied by selecting both the event and the type of communication. In particular, we propose a framework in which an identified person can be updated with useful information anywhere inside the building. As an Example, once the event classification module has determined an important action, such as *"person A is looking for person B",* the system, by knowing where *"person B"* is, can send the information *"Person A is looking for you"* to person B by selecting the proper receiving device, i.e. laptop, desktop, PDA, etc. Finally an AmI system, currently under development, is described. It implements an intelligent infrastructure to monitor the training of student nurses in a academic environment, where hospital ward simulations take place and their monitoring is aimed at simplifying the task of the instructor and implement a novel training method.

In the following sections we will focus on the identification and localization of people inside the building by tracking and counting them.

## 2. System Description

The proposed system is based on a network of smart sensors that cooperating aim to understand *who is going where* inside a monitored building (i.e. the case study is represented by a university building). Here, smart sensors are represented by subsystems able to detect and to track people, to extract important features for event detection purposes and to send information for distributed data association. High level nodes fetch information generated by the sensors and associate different simple events, on the bases of rules from a knowledge base, to infer more complex and complete events. In our case, simple events are represented by states of an finite automata, see Figure 10.1, corresponding to simple actions as walking, keeping the stairs etc. Complex events, on the other hand, do not represent simply a sequence of events performed by a single person, but also semantic information. In particular from a sub-sequence, opportunely

selected, of simple events the system is able to recognize particular actions of interest for the customization of the environment.

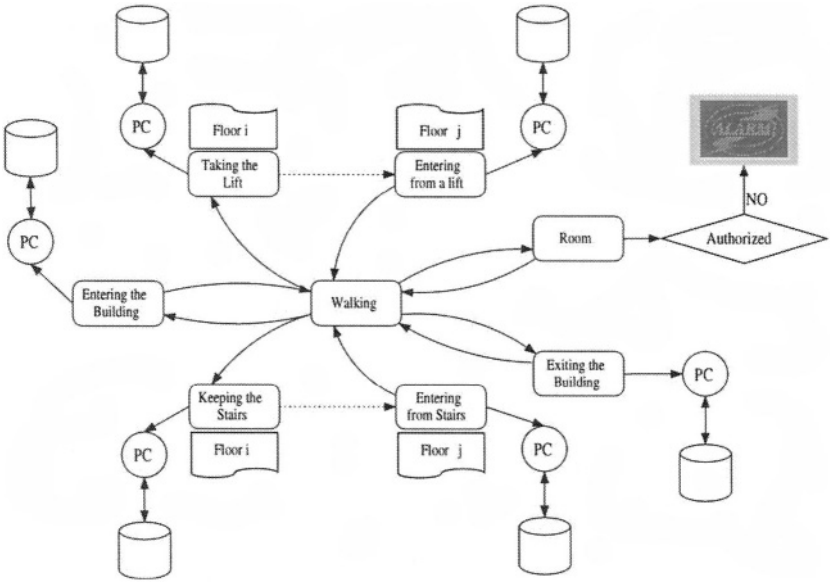


*Figure 10.1.* Event representation for the actions performed by a person in the building

Into the system architecture, smart sensors can be of two types, from sensors for a cooperative multisensor multitarget tracking purposes and sensors used to count people inside a building and on the floors. The former type is composed by cameras that by performing change detection techniques are able to segment moving objects inside the scene and to recognize people from other types of objects. Targets' position on a top-view map of the scene is also computed through a perspective transformation matrix obtained during the initial calibration of each sensor. Therefore, a tracker based on a Kalman Filter is applied on each sensor to compute the trajectories of each person.

The latter is represented by cameras mounted at each gate of the building and of the floors in order to acquire and to count people entering and exiting the floors. People blobs are first identified then tracked by a Kalman filter in order to have information about the direction of the persons moving inside the field of view. Hence, from information about the tracking phase the sensor is able to increment or decrement a shared variable representing the number of persons inside the floor to which the sensor belongs.

Information about movements are associate, in a centralized manner, to an ID related to the person performing the movements in consideration. Such information are used by a event detection and association module in order to

associate a state of a finite automata to the ID and finally to derive semantic information from a sequence of states associated to a single ID.

## 3. People tracking and counting

This module of the system aims to identify the trajectories of the people walking inside the building by maintaining also the number of persons for each floor. The problem is addressed by employing a distributed people tracker which allows to track people through an entire floor. In particular, each camera tracks people inside its field of view and exchange feature information with neighboring cameras once the target is going out of the field.

To maintain track of the number of persons on each floor we have disposed a set of people detectors installed at every entrance or exit. Such detectors are represented by systems composed by a camera placed in order to acquire a top view image which helps to segment people from the background.

### 3.1 People tracking

The system needs to maintain tracks of all objects simultaneously. Hence, this is a typical multi-sensor multi-target tracking problem: measurements should be correctly assigned to their associated target tracks and a target's associated measurements from different sensors should be fused to obtain better estimation of the target state.

A first tracking procedure occurs locally to each image plane. The system then executes an association algorithm to match the current detected blobs with those extracted in the previous frame. A number of techniques are available, spanning from template matching, to features matching [3], to more sophisticated approaches [4]. The approach used in our system was twofold, exploiting the Meanshift predictions [4] and matching blob features (Hu moments, base/height ratio, etc.).

To determine the target trajectories a 2D top view map of the monitored environment is taken as a common coordinates system where the correspondence between an image pixel and a planar surface is given by a planar homography [17, 7]. Measurement gating and assignment is then performed and the Mahalanobis distance can be used to determine the validation region as in [14]. This step reduces the probability of erroneous associations due to noise. The measurements coming from each sensor, for a given object, falling within the gating region are then fused together.

To deal with the multi-target data assignment problem, especially in the presence of persistent interference, there are many matching algorithms available in the literature: Nearest Neighbor (NN), Joint Probabilistic Data Association (JPDA), Multiple Hypothesis Tracking (MHT), and S-D assignment. More details on the employment of such techniques as consequence of the application can be found in [2–1, 15].

The trajectory on the top-view map of every object is modeled through a linear Kalman filter, where the state vector $\hat{x} = (x, v_x, y, v_y)$ is constituted by the position and velocity of the object on the map. At every frame (the

system processes 25 frames per second) a new measurement of the position is received. Finally, position estimates from different sensors are fused in a centralized fashion.

## 3.2 People counting

In this module, cameras are placed in order to have top view of the area around the gate. Hence, from the binary image $B^t(x, y)$, computed on the HSV color space, the system looks for objects in the scene. A search for connected components is then performed by discarding the ones with area below a given threshold. The great advantage of the top view is the nearly constant size of the objects, and this can be exploited for tracking and counting purposes. In fact, knowing the average area in pixels of a person from a given top view, it can be easily determined how many persons form a given connected component. The persons walking in the scene are tracked through the image to maintain a unique ID for each one [2, 15]. Despite the advantages given by the overhead placement of the camera, the tracking of the persons is non trivial due to the small field of view and possible crowded conditions. The following features are extracted for each blob and used by the tracking procedure:

- area
- density
- bounding box coordinates
- centroid coordinates
- centroid deviation from the center of the bounding box
- mean color values
- histograms of the H and S planes,

These features are used along with the Kalman filter [18] and Meanshift [4] predictions in the assignment phase. The system model used for the centroid's coordinates is the following:

$$\mathbf{x}(k) = \mathbf{A}\mathbf{x}(k-1) + \mathbf{w}_{\mathbf{x}}(k) \tag{10.1}$$

$$\mathbf{y}(k) = \mathbf{A}\mathbf{y}(k-1) + \mathbf{w}_{\mathbf{y}}(k) \tag{10.2}$$

where each of the state vectors x and y are composed by the position and velocity components and the state transition matrix is given by:

$$\mathbf{A} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}$$

The tracking algorithm explicitly takes into account merge and split cases generated for example by two or more persons walking side by side: at some point the single blobs can be detected *(merge)* as one big blob and then split again.

## 4. Event detection and association

As shown in Figure 10.1 simple events are identified by states of a finite automata. First of all, when a person enters in the building, it is counted by a people counter installed to monitor the entry. Therefore, a vector $v_{p_i} = (f_1, \ldots, f_m)$ of features characterizing the $i-th$ person is extracted and associated to a unique ID which will identify the person. These features, will be used and updated in the future to allow people recognition during their motion inside the building. It is worth nothing how the vector of features can change its dimension as consequence of new features extracted or as consequence of new trusted feature in antithesis with the old ones.

The event detection and association module has the objective to assign to each active identifier (i.e. identifiers associated to a person still inside the building) a state representing an event. Having to monitor a building composed by multiple floors, the system has to face to the situation in which the target cannot be tracked continuously by sharing information among camera with overlapped views. While the system can determine the most appropriate camera to switch to when a person walks on a floor, this is not possible when a person keep the stairs or an elevator. In this case, the event association has to address an higher level of uncertainty represented by the fact that people are not tracked on the stairs or inside the elevators. In Figure 10.1, this uncertainty is represented by dashed lines. To address this problem a camera pointing to each entrance of each floor is responsible to extract important information for people recognition purposes. In particular, a vector $v_c = (f_1, \ldots, f_n)$ of the current feature is extracted for each detected person and compared with the entire set of feature vectors $v_{p_i}$ associated to active identifier.

Let $ID_i$ be the identifier of a person still inside the building, a distance function can be studied for the computation of the probability $P(ID_I|v_c)$ that the person $ID_i$ is the person represented by the feature vector $v_c$. Therefore, the association of the detected people given the feature vector is performed by maximizing the probability. The event *person $ID_i$ enters from the elevator/stairs on floor $j$* is therefore associated to the event *person $ID_I$ takes the elevator/stairs on floor i.* The scheme of this data association is shown in Figure 10.1.

In the proposed system, event association is performed to extract semantic information from a sequence of single events. A set of inference rules belonging to a knowledge base has been adopted to derive complex actions. As an example from the sequence S={"A exits from room $k$","A walks on floor $i$",....,'A enters on floor $j$ from stairs",'A stops in front of room $l$",'A does not enter room $l$"} the inference engine triggers the consequence "A looks for B" by knowing that the room $l$ is associated to the person B.

## 5. Experimental results

The system has been tested on a real environment represented by a university building having a special room for medical purposes. In particular, the tests conducted have been performed first to test separately the performance of the people tracker and people counting modules.

In Figure 10.2 some frames of a testing sequence for people tracking are reported. In this context, the behavior of the system is resulted very good by extracting with a good reliability the trajectories of the persons moving inside the floor. The mean and standard deviation of the distance (in pixels, 1 pixel $\approx$ 10

*Figure 10.2.* Some frames of a sequence acquired during the test phase. Three people are entering the $3^{rd}$ floor from the stairs and are going to a Lab.

cm) from measured ground truth positions on the map of the walking persons in Figure 10.2 are respectively 8.66 and 4.42. On the map of the $3^{rd}$ floor represented in Figure 10.3 the trajectory computed for the persons of the same sequence are plotted. It is worth nothing how the trajectory represented by the continuous line at a certain point splits in two different trajectories. This is due by an error in the people detection module which is not able to distinguish two persons while they are occluded by a third one. Once the acquisition becomes optimal the system identifies both persons and their relative trajectories. The errors computed take into account also the problem of the miss detection.
Some frames of a test sequence for the people counter module are shown in Figure 10.4. The images are acquired by cameras placed at 3 meters from the floor and wide-angle (2mm focal length) lenses are used. These sequence are constituted by 7000 frames of 384x288 pixels each and lasts 280 seconds. The number of persons, equally subdivided in both directions, was 130 in total. The experiment accounted for a maximum of 4 persons walking simultaneously in the scene in different directions. The system reported a performance falling between 90% and 98%.

## 6. AmI for training environments

In general, all the concepts put forward by the intelligent building can be easily *ported* to professional training environments. Professional training practice

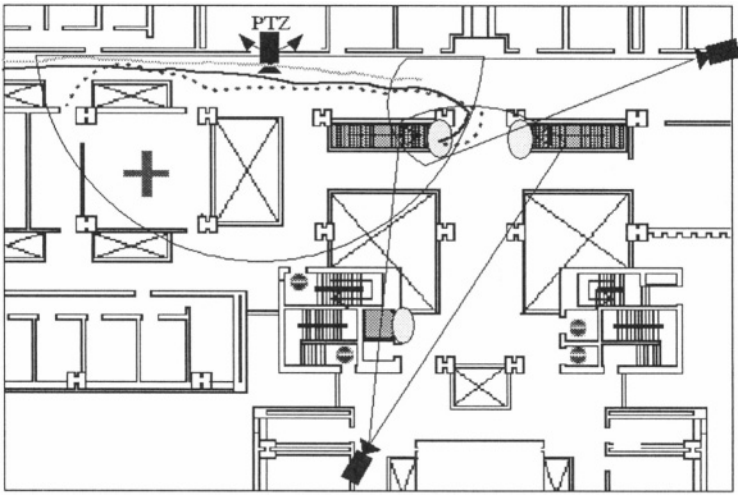


*Figure 10.3.* Floor map of the third floor of the University Building where the medical room is placed. Ellipses on the gates to and from stairs or elevator represent the people counter sensors that count people on each floor of the building. Rooms with "Do not enter " sign are rooms for which a granted access is needed.

| max # persons | Performance 3m |
|---|---|
| 1 | 100% |
| 2 | 100% |
| 3 | 95% |
| 4 | 90% |

*Table 10.1.* Performance of the system against the number of persons inside the view.

usually takes place in more than one environment where one or more instructors can directly or remotely guide or test the trainee. The training of astronauts and nurses are for instance two very different but somewhat also very similar domains for which AmI criteria can be employed. The former domain entails the training of a small number of professionals in tight spaces, the latter usually a large number of students in relatively large spaces. Both require very direct approach and a strict set of rules or protocol to follow. Professionals in both domains might have to work in many environments and under strict control of an instructor.

The Faculty of Health and Social Care Science at Kingston University runs cutting edge training methods for student nurses. Nurse training takes place in laboratories equipped as hospitals wards. The simulation commences with the delivery of a scripted handout to the students. The patient's history, diagnosis and current health status are discussed. The students are also given information on future admissions through out the day and any potential discharges. The team of both nursing and medical students is then given time to consider their

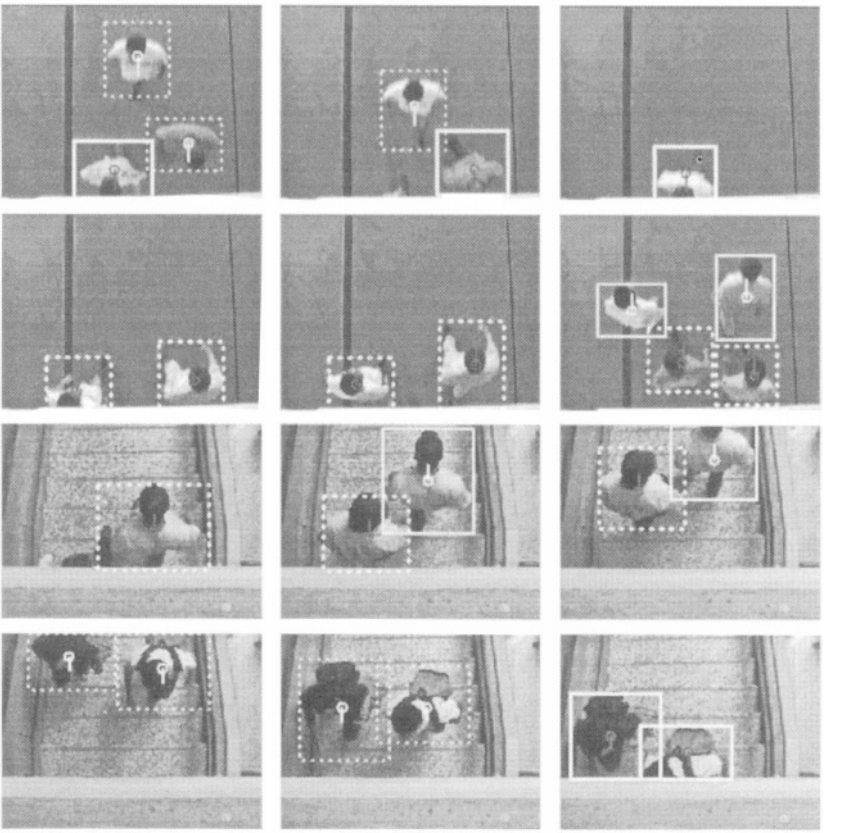

*Figure 10.4.* Some frames of sequences used to test the performance of the people counting module are here shown. The first two rows show images in which people respectively is entering and exiting the elevator of the third floor. The bottom two rows, instead, show images of people going respectively downstairs and upstairs. Continuous bounding box represent a person already counted by the system, while dashed bounding represent a person to be counted as entering or exiting the floor.

priorities and plan care delivery. Verbal feedback and video recording are currently used to identify student progression. However video recording is dependent on the careful placement of a limited number of stationary cameras and as such have proven problematic and limiting. Verbal feedback is conducted throughout the simulation exercises with individuals or in small sub-groups, whereas full feedback is given to the whole group at the end of an exercise. The current approach is limited because both methods of feedback have proven to be time and labour intensive and are heavily dependent on limited staff/lecturer

resources. Figure 10.5 shows two individual skills demonstrated by instructors.

*Figure 10.5.* Two views of one the Skills' laboratories at Kingston University.

The Ideal Objectives of the nurse training application include [1]

- To provide effective and objective feedback for individual students, during and following a simulated exercise, which identifies both best practice and highlights areas for improvement.
- Having a means to allow students to practice a task/skill repeatedly in a non-threatening environment, receiving effective feedback for the duration of practice.
- Providing students with a means of independent practice within the safe parameters of best practice.
- To equip students with effective yet time and resource efficient feedback.

AmI criteria here can be implemented in a system capable of monitoring the cluttered environment of simulation and practice sessions, to enhance the communication between instructor and trainee, including new methods to rehearse skills and replay performance. The system would make use of a sophisticated infrastructure using a network of cameras, a network of high specification computers, and sophisticated user interfaces for the real-time delivery of video contents, trainee performance evaluation and analysis.

The environment at the Faculty of Health and Social Care Sciences is illustrated in Figure 10.6. The environment is currently being equipped with a network of fixed cameras and a pan-tilt-zoom camera. The network will be used to capture visual data for individual practice skills and simulations and the data will be automatically stored in a annotated database to improve the current state of the art in professional practice training. All modules described

[1]This is outcome of a number of meetings held with Susan Rush, Kingston University Principal Lecturer leading the professional skills' laboratory.

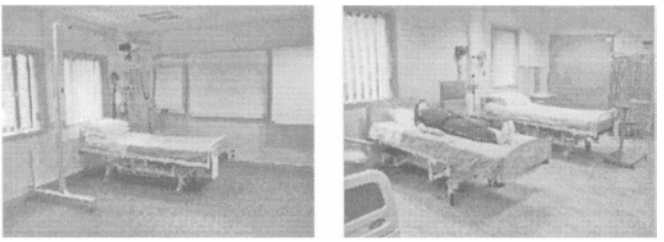

*Figure 10.6.* Two views of one the Skills' laboratories at Kingston University.

in the previous sections of the chapter can indeed be applied to identify trainees, track them throughout the environment and make not - for instance by storing information in a database - of their actions. In practice, in order to develop an intelligent system useful for the users - trainee and instructor - two types of handling information can be devised:

- On one hand a passive system, logging information about individual trainee over short and long period of times, during practice skills' session,
- On the other hand an active system, where information is build and on the fly communicated to the interest user, either real or quasi real-time.

The former is a standard monitoring system, very much in place in many visual surveillance installations. The latter is much more interesting and would have to implement a few of the criteria AmI advocates as necessary to create a real self-sustaining and truly interacting environment. A few problems still hold, including ethics and privacy of the trainee and, perhaps, the existence of a network of sensors, impinging on the performance of the student or professional. All these issues are currently being dealt with.

## 7. Conclusions

In this paper we have presented a cooperative network of smart sensors able to monitor the flow of people inside the building. The surveillance capacity of the proposed system is therefore used to customize the information supply for people inside the building. Two main aspects of current research in computer technology have been employed by using methods for environment security to customize the environment thus giving a sort of intelligence to the building. The domain of training of professionals can indeed benefit from the technologies presented in the chapter. The application domain presented here shows that there is the need for an interdisciplinary approach: technology alone cannot solve all problems, *above all* the human dimension must be taken into consideration.

## Acknowledgments

This work was partially supported by the Italian Ministry of University and Scientific Research within the framework of the project "Distributed systems

for multisensor recognition with augmented perception for ambient security and customization " (2002-2004).The authors wish to thank Susan Rush, leader of the Professional Skills laboratory at Kingston University for her input on the training of student nurses.

## References

[1] S.S. Balckman. *Multiple-target tracking with radar applications.* Artech House, 1986.

[2] Y. Bar-Shalom and X.R. Li. *Multitarget-multisensor tracking: principles and techniques.* YBS Publishing, 1995.

[3] R.T. Collins, A.J. Lipton, H. Fujiyoshi, and T. Kanade. A system for video surveillance and monitoring. In *Proceedings of the IEEE,* volume 89, pages 1456–1477, October 2001.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. on Pattern Analysis Machine Intelligence,* 25(5):564–575, 2003.

[5] T. Derek and J. Clements-Croome. What do we mean by intelligent buildings? *Automation in Construction,* pages 395–400, 1997.

[6] S.L. Dockstader and T. Murat. Multiple camera tracking of interacting and occluded human motion. *Proceedings of the IEEE,* 89(10):1441–1455, October 2001.

[7] O.D. Faugeras, Q.T. Luong, and S.J. Maybank. Camera self-calibration: Theory and experiments. In *Proceedings of the European Conference on Computer Vision (ECCV),* pages 321–334, 1992.

[8] G.L. Foresti. Object recognition and tracking for remote video surveillance. *IEEE Trans. Circuits Syst. Video Technol.,* 9(7):1045–1062, October 1999.

[9] K. Huang and M. M. Trivedi. Video arrays for real-time tracking of person, head, and face in an intelligent room. *Machine Vision and Applications,* 14(2):103–111, June 2003.

[10] ISTAG. Scenarios for ambient intelligence in 2010. Technical Report 10, EC, February 2001.

[11] A. Kell. Intelligent buildings now. *Electrotechnology,* pages 26–27, October/November 1996.

[12] D. Koller, K. Daniilidis, and H. H. Nagel. Model-based object tracking in monocular sequences of road traffic scenes. *International Journal of Computer Vision,* 10:257–281, 1993.

[13] T. Matsuyama and N. Ukita. Real-time multitarget tracking by a cooperative distributed vision system. *Proceedings of the IEEE,* 90(7): 1136–1149, 2002.

[14] I. Mikić, S. Santini, and R. Jain. Tracking objects in 3d using multiple camera views. In *Proceedings of ACCV,* January 2000.

[15] A.B. Poore. Multi-dimensional assignement formulation of data association problems arising from multi-target and multi-sensor tracking. *Computational Optimization and Applications,* 3:27–57, 1994.

[16] B.J. Tordoff and D.W. Murray. Reactive control of zoom while tracking using perspective and affine cameras. *IEEE Trans. Pattern Anal. Machine Intell.*, 26(1):98–112, 2004.

[17] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation,* (4):323–344, 1987.

[18] G. Welch and G. Bishop. An introduction to the kalman filter. Technical report 95-041, University of North Carolina at Chapel Hill, Department of Computer Science, 1995.

[19] Zhigang Zhu, Guangyou Xu, Bo Yang, Dingji Shi, and Xueyin Lin. Visatram: a real-time vision system for automatic traffic monitoring. *Image and Vision Computing,* 18(10):781–794, October 2000.