# Chapter 2
# Beyond the Static Camera: Issues and Trends in Active Vision

**Murad Al Haj, Carles Fernández, Zhanwu Xiong, Ivan Huerta, Jordi Gonzàlez, and Xavier Roca**

**Abstract** Maximizing both the area coverage and the resolution per target is highly desirable in many applications of computer vision. However, with a limited number of cameras viewing a scene, the two objectives are contradictory. This chapter is dedicated to active vision systems, trying to achieve a trade-off between these two aims and examining the use of high-level reasoning in such scenarios. The chapter starts by introducing different approaches to active cameras configurations. Later, a single active camera system to track a moving object is developed, offering the reader first-hand understanding of the issues involved. Another section discusses practical considerations in building an active vision platform, taking as an example a multi-camera system developed for a European project. The last section of the chapter reflects upon the future trends of using semantic factors to drive smartly coordinated active systems.

M. Al Haj (✉) · C. Fernández · I. Huerta
Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain
e-mail: malhaj@cvc.uab.es

C. Fernández
e-mail: perno@cvc.uab.es

I. Huerta
e-mail: ivan.huerta@cvc.uab.es

Z. Xiong · J. Gonzàlez · X. Roca
Computer Vision Center and Departament de Ciències de la Computació, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain

Z. Xiong
e-mail: zhanwu@cvc.uab.es

J. Gonzàlez
e-mail: jordi.gonzalez@uab.cat

X. Roca
e-mail: xavier.roca@uab.cat

T.B. Moeslund et al. (eds.), *Visual Analysis of Humans*,
DOI 10.1007/978-0-85729-997-0_2, 

## 2.1 Introduction

Many applications in the computer vision field benefit from high-resolution imagery. These include, but are not limited to, license-plate identification [4] and face recognition, where it has been observed that higher resolution improves accuracy [27]. For other applications, such as identifying people in surveillance videos, having highly zoomed images is a must. The problem with zoom control is that two opposing aims are desirable: the first one is obtaining a maximum resolution of the tracked object, whereas the second is minimizing the risk of losing this object. Therefore, zoom control can be thought of as a trade-off between the effective resolution per target and the desired coverage of the area of surveillance.

With a finite number of fixed sensors, there is a fundamental limit on the total area that can be observed. Thus, maximizing both the area of coverage and the resolution of each observed target requires an increase in the number of cameras. However, such an increase is highly costly in terms of installation and processing. Therefore, a system utilizing a smaller number of Pan–Tilt–Zoom (PTZ) cameras can be much more efficient if it is properly designed to overcome the obvious drawback of having less information about the target(s).

Toward this end, different works have investigated the use of PTZ cameras to address this problem of *actively* surveying a large area in an attempt to obtain high-quality imagery while maintaining coverage of the region [25]. Starting two decades ago, the area of active vision has been gaining much attention, in an attempt to: i) improve the quality of the acquired visual data by trying to keep a certain object at a desired scale, and ii) react to any changes in the scene dynamics that might risk the loss of the target.

Accurate reactive tracking of moving objects is a problem of both control and estimation. The speed at which the camera is adjusted must be a joint function of current camera position in pan, tilt and focal length, and the position of the tracked object in the 3D environment.

This chapter deals with active vision systems, offering the reader hands-on experience and insights into the problem. Section 2.2 discusses the different design alternatives for active cameras configurations, such as the autonomous camera approach, the master-slave approach and the active camera network approach, in addition to touching upon the advantages that environment reasoning lends to the problem. In Sect. 2.3, an autonomous camera system is designed, where the problem of jointly estimating the camera state and 3D object position is formulated as a Bayesian estimation problem and the joint state is estimated with an extended Kalman filter. The authors of this chapter had the opportunity to be part of a dedicated consortium working on a European project, called HERMES, where an integrated platform involving active cameras was built. Therefore, in Sect. 2.4, practical considerations involved in building real-time active camera systems are discussed taking the HERMES platform as a case study. This chapter is concluded in Sect. 2.5, where the lessons learned are summarized and the future directions are noted.

## 2.2 Active Camera Configurations

The interest in active camera systems started as early as two decades ago. Beginning in the late 1980s, Aloimonos et al. introduced the first general framework for active vision in order to improve the perceptual quality of tracking results [3]. Since then, numerous active camera systems have been developed. In this section, we take a look at different approaches for configuring these systems.

### *2.2.1 The Autonomous Camera Approach*

Autonomous cameras are those that can self-direct in their surrounding environment. Recent work addressing this topic includes that of Denzler et al., where the motion of the tracked object is modeled using a Kalman filter. The camera focal length that minimizes the uncertainty in the state estimation is selected [12]. The authors used a stereo set-up, with two zoom cameras, to simplify the 3D estimation problem.

A newer approach is described by Tordoff et al., which tunes a constant velocity Kalman filter in order to ensure reactive zoom tracking while the focal length is varying [26]. Their approach correlates all the parameters of the filter with the focal length. However, they do not concentrate on the overall estimation problem, and their filter does not take into account any real-world object properties.

In the work by Nelson et al., a second rotating camera with fixed focal length is introduced in order to solve the problem of lost fixation [19].

The latter two works are primarily focused on zoom control and do not deal with total object-camera position estimation and its use in the control process. An attempt to join estimation and control in the same framework can be found in the work of Bagdanov et al., where a PTZ camera is used to actively track faces [5]. However, both the estimation and control models used are ad hoc, and the estimation approach is based on image features rather than 3D properties of the target being tracked.

### *2.2.2 The Master/Slave Approach*

In a master/slave configuration, a supervising static camera is used to monitor a wide field of view and to track every moving target of interest. The position of each of these targets over time is then provided to a foveal camera, which tries to observe the targets at a higher resolution. Both the static and the active cameras are calibrated to a common reference, so that data coming from one of them can be easily projected onto the other, in order to coordinate the control of the active sensors.

Another possible use of the master/slave approach consists of a static (master) camera extracting visual features of an object of interest, while the active (slave) sensor uses these features to detect the desired object without the need of any training data. In this case, features should be invariant to illumination, viewpoint, color

distribution and image resolution, and usually consist of any kind of coarse-to-fine region descriptors, as in [31].

The master/slave approach is a simple but effective formulation that has been repeatedly used for solving many active vision problems [16, 20, 31]. Nonetheless, the use of supervising cameras has the disadvantage of requiring a mapping of the image content to the active cameras. This mapping needs to be obtained from restricted camera placements, movements or observations extended over time [6, 13].

### 2.2.3 The Active Camera Network Approach

In recent years, interest has grown in building networks of active cameras and optional static cameras, in order to cover a large area while also providing high-resolution imagery of multiple targets [7, 11, 17, 21]. An active camera network is a scaling up of a basic active camera approach, which can be either an autonomous active camera or a master/slave configuration, depending on whether fixed master cameras are deployed or not.

Due to the fact that an active camera network involves multiple cameras and is usually required to accomplish multiple tasks, the challenges of this approach mainly arise from two aspects: i) *task assignment* and ii) *task hand-over*.

Task assignment is the problem of deciding which camera resources are to be allocated to which task, or in other words, the problem of camera scheduling. On the other hand, task hand-over describes model transferring from one camera to another.

Furthermore, like the master/slave configuration, active camera networks also require calibration information, as well as extensive networking infrastructure. Communications within such systems require clever networking algorithms for routing and decision making. Though theoretically appealing, active camera networks are expensive to build and maintain, and do not scale well.

### 2.2.4 Environmental Reasoning

In some cases, low-level approaches such as those described above are not enough to address ambitious applications requiring more complex strategies toward sensors collaboration. Smart coordination among camera sensors requires exploiting resources that are often related to artificial intelligence and symbolic models, including techniques for camera selection according to the given task, protocols for allocating such tasks, tools for reasoning about the environment and mechanisms to resolve conflicts.

Some examples in which such techniques are used to enhance the collaboration among sensors in a camera network include constraint satisfaction formulations [22], Situation Graph Tree (SGT) [14] and Petri net coordination models [29].

## 2.3 The Autonomous Camera: A Hands-on Experience

This section is aimed at providing the reader with a hands-on experience to develop an autonomous active camera system that is able to track a moving object, taking proper decisions on when to zoom in, to maximize the resolution, and when to zoom out, to minimize the risk of losing the object. It is dedicated to an exemplary system showing all the design decisions that has been taken in the process, namely the camera-world model, the estimation process and the control process. Some performance indicators of the system are shown at the end. This section is based on the paper "Reactive object tracking with a single PTZ camera" by Al Haj et al., which appeared in the 20th International Conference on Pattern Recognition [2], ©2010 IEEE.

### *2.3.1 Camera-World Model*

We use a pinhole camera model as shown in Fig. 2.1. The camera center is located at the origin of the world coordinate system. The principal point is at the origin of the plane of projection at zero pan and tilt. The axis of projection is aligned with the $z$-axis.

The object being tracked is assumed to be a rigid rectangular patch perpendicular to the axis of projection. It is located at world position $(X, Y, Z)$ with known width $W$ and height $H$. It is important to note here that upper-case characters, $(X, Y, Z, W, H)$, will be used to denote values in the real-world while lower-case characters, $(x, y, w, h)$, will be used to denote values in the image projection plane.

Changes in camera orientation due to panning and tilting are modeled as pure rotations of the coordinate system:

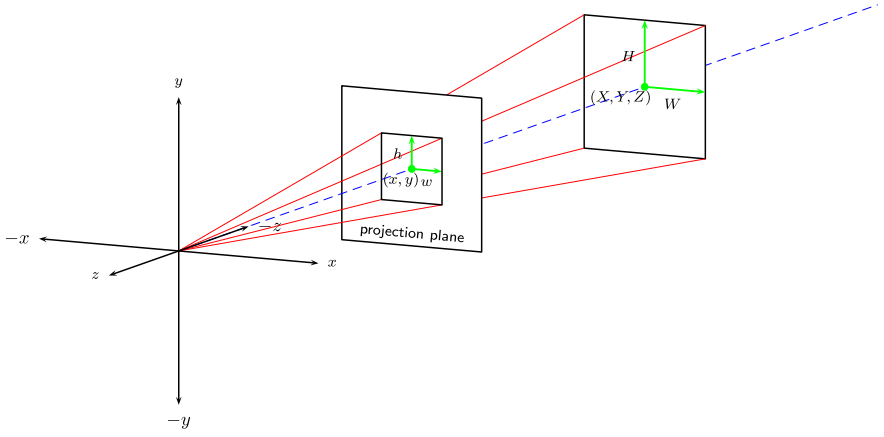


**Fig. 2.1** The pinhole camera model with the camera positioned at the origin of the world coordinates

$$\mathbf{M}(\phi,\theta)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\theta & -\sin\theta\\ 0 & \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\cos\phi & 0 & -\sin\phi\\ 0 & 1 & 0\\ \sin\phi & 0 & \cos\phi\end{bmatrix}, \tag{2.1}$$

where $\phi$ and $\theta$ represent the pan and tilt angles, respectively.

We assume that the camera projection is reasonably approximated using equal scaling in the $x$ and $y$ directions (i.e. square pixels). The center of projection is also assumed to be at the origin of the world coordinate system. Then, the camera matrix, $\mathbf{N}$, is fully parameterized by the focal length parameter $f$:

$$\mathbf{N}(f)=\begin{bmatrix}f & 0 & 0\\ 0 & f & 0\\ 0 & 0 & 1\end{bmatrix}. \tag{2.2}$$

The projection of the object at position $\mathbf{o}=[X,Y,Z]$ onto the plane of projection can now be written as

$$\mathbf{p}(\phi,\theta,f,\mathbf{o})=\begin{bmatrix}\frac{X'}{Z'} & \frac{Y'}{Z'}\end{bmatrix}, \tag{2.3}$$

where $X'$, $Y'$ and $Z'$ are given by the transformation

$$\begin{bmatrix}X'\\ Y'\\ Z'\end{bmatrix}=\mathbf{N}(f)\mathbf{M}(\phi,\theta)\mathbf{o}^{\top}. \tag{2.4}$$

The camera model relates the geometry and position of the tracked object in the 3D world to the internal camera parameters. In the next section, we describe how the estimation problem can be formulated.

### *2.3.2 Estimation*

In this section we formulate the problem of jointly estimating the camera and world parameters in a recursive Bayesian filter framework.

At time $t$, the state configuration of the joint camera/object model is represented by the spatial coordinates of the tracked object in the real-world, the camera intrinsics and the velocities corresponding to the object position and camera intrinsics:

$$\mathbf{s}_t=[\mathbf{o}_t\,|\,\mathbf{c}_t\,|\,\dot{\mathbf{o}}_t\,|\,\dot{\mathbf{c}}_t]^{\top}, \tag{2.5}$$

where each component is defined by

$$\mathbf{o}_t=[X_t,Y_t,Z_t], \tag{2.6}$$

$$\mathbf{c}_t=[\phi_t,\theta_t,f_t], \tag{2.7}$$

$$\dot{\mathbf{o}}_t=[\dot{X}_t,\dot{Y}_t,\dot{Z}_t], \tag{2.8}$$

$$\dot{\mathbf{c}}_t=[\dot{\phi}_t,\dot{\theta}_t,\dot{f}_t]. \tag{2.9}$$

$[X_t,Y_t,Z_t]$ is the position of the planar patch in world coordinates at time $t$, and $[\phi_t,\theta_t,f_t]$ represent the camera pan angle, tilt angle and focal length at time $t$,

respectively. The remaining elements, $[\dot{X}_t, \dot{Y}_t, \dot{Z}_t, \dot{\phi}_t, \dot{\theta}_t, \dot{f}_t]$, represent the velocities of the previously mentioned components.

From time $t-1$ to time $t$, the state is updated by the linear matrix $\mathbf{U}$:

$$\mathbf{s}_t = \mathbf{U}\mathbf{s}_{t-1} + \mathbf{v}_{t-1}, \tag{2.10}$$

where $\mathbf{U}$ is defined by

$$\mathbf{U} = \begin{bmatrix} \mathbf{I}_6 & \mathbf{I}_6 \\ \mathbf{0}_6 & \mathbf{I}_6 \end{bmatrix}, \tag{2.11}$$

and where $\mathbf{I}_n$ and $\mathbf{0}_n$ are the $n \times n$ identity and zero matrices, respectively. The term $\mathbf{v}_{t-1}$ in (2.10) is considered to be a zero-mean, Gaussian random variable adding noise to the system update.

At each time $t$, an observation $\mathbf{z}_t$ of the unknown system $\mathbf{s}_t$ is made:

$$\mathbf{z}_t = [x_t, y_t, w_t, h_t, \hat{\phi}_t, \hat{\theta}_t, \hat{f}_t], \tag{2.12}$$

where $(x_t, y_t)$ is the center of the object in the image plane measured in pixels, $(w_t, h_t)$ are the width and height of the object in the image plane, also measured in pixels, please refer again to Fig. 2.1. $(\hat{\phi}_t, \hat{\theta}_t, \hat{f}_t)$ are the camera parameters arriving from the camera imprecise measurements of the pan angle, tilt angle and focal length.

The measurement equation, against which the observation $\mathbf{z}_t$ is compared, is given by:

$$\mathbf{h}(\mathbf{s}_t) = \left[\mathbf{p}(\phi_t, \theta_t, f_t, \mathbf{o}_t) \,|\, \mathbf{p}\big(0, 0, f_t, [W, H, Z'_t]\big) \,|\, \mathbf{c}_t\right]^\top + \left[\mathbf{n}_t^o \,|\, \mathbf{n}_t^c\right]^\top, \tag{2.13}$$

where $\mathbf{n}_t^o$ and $\mathbf{n}_t^c$ are zero-mean Gaussian noise processes on the object and camera measurements, respectively. $Z'_t$ is the projection of the depth $Z_t$ in the new coordinate system resulting from the pan and tilt of the camera. $\mathbf{p}(\phi_t, \theta_t, f_t, \mathbf{o}_t)$ represents the projection of the object position $\mathbf{o}_t$ into the image plane and, similarly, $\mathbf{p}(0, 0, f_t, [W, H, Z'_t])$ is the projection of the known object size $W \times H$ into the image plane. The camera vector $\mathbf{c}_t$ consists of the pan angle, tilt angle and focal length, as estimated by the state vector.

Given the system update and measurement processes defined in (2.10) and (2.13), the Bayesian estimation problem is to find an estimate of the unknown state $\mathbf{s}_t$ that maximizes the posterior density $p(\mathbf{s}_t|\mathbf{z}_{1:t})$.

Toward this end, an Extended Kalman Filter (EKF) is used to recursively solve this estimation problem [28]. The EKF approximates the likelihood as a Gaussian density with argument $\mathbf{s}_t$, mean $\mathbf{m}_t$ and covariance $\mathbf{P}_t$:

$$p(\mathbf{s}_t|\mathbf{z}_{1:t}) \approx \mathcal{N}(\mathbf{s}_t; \mathbf{m}_t, \mathbf{P}_t). \tag{2.14}$$

Defining $\hat{\mathbf{H}}_t$ as a local linearization, given by the Jacobian, of the non-linear measurement function, $\mathbf{h}(\mathbf{s}_t)$:

$$\hat{\mathbf{H}}_t = \left.\frac{\partial \mathbf{h}(\mathbf{s}_t)}{\partial \mathbf{s}_t}\right|_{\mathbf{s}_t = \mathbf{m}_{t|t-1}}, \tag{2.15}$$

the update from time $t-1$ to time $t$ is given by the set of equations

$$\mathbf{m}_{t|t-1} = \mathbf{U}\mathbf{m}_{t-1}, \tag{2.16}$$

$$\mathbf{P}_{t|t-1} = \mathbf{Q} + \mathbf{U}\mathbf{P}_{t-1}\mathbf{U}^{\top}, \tag{2.17}$$

$$\mathbf{m}_t = \mathbf{m}_{t|t-1} + \mathbf{K}_t\big(\mathbf{z}_t - \mathbf{h}(\mathbf{m}_{t|t-1})\big), \tag{2.18}$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{K}_t\hat{\mathbf{H}}_t\mathbf{P}_{t|t-1}, \tag{2.19}$$

$$\mathbf{S}_t = \hat{\mathbf{H}}_t\mathbf{P}_{t|t-1}\hat{\mathbf{H}}_t^{\top} + \mathbf{R}, \tag{2.20}$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\hat{\mathbf{H}}_t^{\top}\mathbf{S}_t^{-1}. \tag{2.21}$$

$\mathbf{S}_t$ is the covariance of the innovation term $\mathbf{z}_t - \mathbf{h}(\mathbf{m}_{t|t-1})$ and $\mathbf{K}_t$ is the Kalman gain. $\mathbf{Q}$ and $\mathbf{R}$ are the covariance of the Gaussian noise added to the system update and measurement, respectively.

### 2.3.3 Control

The estimated state outputted at each step of the filter is used to control the movement of the camera. Two PID controllers are used: one for controlling the pan and tilt and another one for the zoom. The control signal, outputted by a PID controller, is given by

$$\mathbf{u}(t) = K_p\mathbf{e}(t) + K_i\int_0^t \mathbf{e}(\tau)\,d\tau + K_d\frac{d}{dt}\mathbf{e}(t), \tag{2.22}$$

where $\mathbf{e}(t)$ is the error signal, $K_p$ is the proportional gain, $K_i$ is the integral gain and $K_d$ is the derivative gain.

In our case, and at each time $t$, the *error in pan* is defined as the difference between the estimated pan angle and the estimated horizontal angle that the object forms with the world coordinate system, while the *error in tilt* is defined as the difference between the estimated tilt angle and the estimated vertical angle of the object:

$$e_{\text{pan}} = \arctan(X_t/Z_t) - \phi_t, \tag{2.23}$$

$$e_{\text{tilt}} = \arctan(Y_t/Z_t) - \theta_t. \tag{2.24}$$

The gains are experimentally set to: $K_p = 1$, $K_i = 0$ and $K_d = 0.2$.

To calculate the *error for the zoom controller*, we define the desired area $D_a$, which is the maximum area in pixels we aim to have and which is usually achieved when the object is static. The error is then defined, at each time $t$, as:

$$e_{\text{zoom}} = D_a - w_{\text{proj}} * h_{\text{proj}}, \tag{2.25}$$

where $w_{\text{proj}}$ and $h_{\text{proj}}$ are the projections of the width $W$ and height $H$ of the object in the image plane. The gains are experimentally set to: $K_p = 0.01$, $K_i = 0$ and $K_d = 0$.

The integral phase was bypassed in both controllers, by setting $K_i$ to 0, because the output of the filter was found to be accurate at steady state, i.e. when the object is centered with maximum zoom.

The error $e_{\text{zoom}}$ is considered only when both $|e_{\text{pan}}|$ and $|e_{\text{tilt}}|$ are constant or decreasing; otherwise, a zoom out operation is executed.

### *2.3.4 System Performance*

In this section, we will show the reader the performance of the system on both simulated scenarios and live scenes of a PTZ camera. The simulated scenario consisted of a random motion of an object whose size is $10 \times 10$ cm, and the error was averaged over many runs. The camera used in the live scenes was an Axis 214 PTZ network camera.

#### 2.3.4.1 Simulated Data

The error metric we used in all model parameters estimation is the root mean square deviation (RMSD) defined as:

$$\text{RMSD}(\eta_i) = \sqrt{E\left((\bar{\eta}_i - \eta_i)^2\right)}, \tag{2.26}$$

where $\eta_i$ is one of the model parameters, $[X, Y, Z, \phi, \theta, f, \dot{X}, \dot{Y}, \dot{Z}, \dot{\phi}_t, \dot{\theta}_t, \dot{f}_t]$, composing the state vector in (2.5), and $\bar{\eta}_i$ is the estimated model parameter. The expectation, $E$, is taken over the entire sequence. The RMSD is measured for several runs of the simulation (we used 100 runs in our experiments), and the average RMSD is used as a measure of estimation performance.

Figure 2.2a shows a box-and-whisker summary of the RMSD for a simulation where a moving object is tracked by a moving camera. In these experiments, we

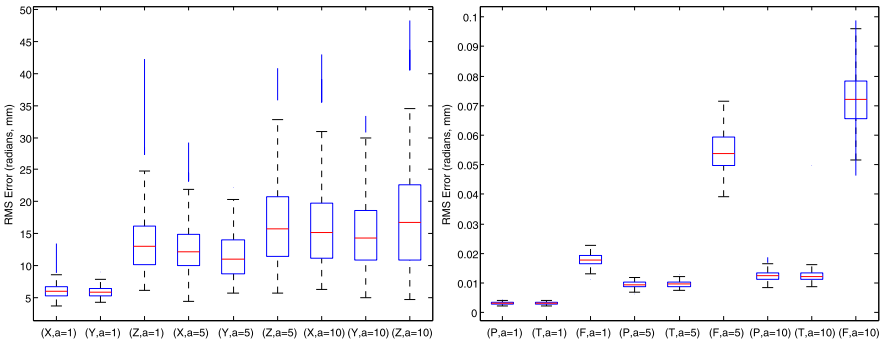


**Fig. 2.2** **a** Error in 3D position parameters $(X, Y, Z)$, measured in millimeters. **b** Error in pan angle, tilt angle and focal length. Angles are measured in radians, focal length in millimeters

simulate the motion the camera would execute due to corrections coming from the PID controllers described in the previous section. Also, some noise is introduced in the different state parameters. To investigate sensitivity to varying measurement noise, this value is scaled by a constant $a \in \{1, 5, 10\}$. Similar results can be seen in Fig. 2.2b for camera parameters estimation. From these figures, one can conclude that scaling the uncertainty, by $a = 5$ and $a = 10$, predictably scales the RMSD error as well as the spread (most notably in $Z$ and $f$) and increases outliers. However, even with such increase, the estimates of both the object position and the camera parameters are very good.

#### 2.3.4.2 Live Cameras

A commodity PTZ camera (Axis 214) was used for tracking different objects. Simple assumptions about object sizes were made: the cup tracked in Fig. 2.3 is assumed to be $8 \times 12$ cm, while the faces in Figs. 2.4 and 2.5 are assumed to be $18 \times 18$ cm. For the detection of the blue cup, a simple heuristics-based classifier for detecting blue regions in the normalized RGB colorspace was used; while for face detection, we used the method developed in [1]. The two red dots represent the center of the object and the upper left corner, outputted by the detection process. The green dots represent the projection of the estimates of the center and the bounding box position. The tracker was able to successfully follow the objects taking correct decisions on when to zoom in and when to zoom out.


**Fig. 2.3** Reactive tracking of a stationary object


**Fig. 2.4** Reactive tracking of a moving face

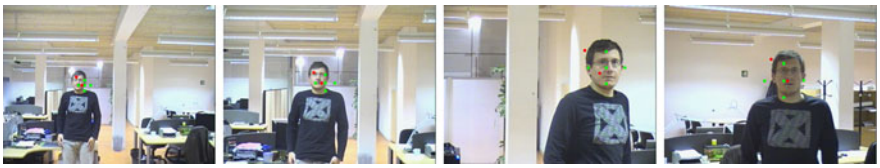


**Fig. 2.5** Another example of reactive face tracking

### *2.3.5 Closing Remarks*

In this section, a method for reactive object tracking has been described. The system uses a single PTZ camera and jointly estimates, in a Bayesian framework, the orientation and focal length of the camera and the position of the tracked object in the 3D environment. The output of the estimation process is used to drive the control process, allowing the camera to reactively track the moving target. The main limitation of this method is that the EKF output is dependent on the detection, i.e. the measurement process; therefore, and although the method is tolerant to measurement noise, continuous erroneous detection leads to inaccurate tracking. Also, this method does not support multiple objects tracking. Other than that, the estimates are robust in the presence of camera motion and increased measurement noise.

## 2.4 Active Vision in Practice: A Case Study

Imagine a user communicating with a set of distributed PTZ cameras, as if they were humans reporting what they see. This would require converting a video stream into a textual description of temporal events. The user should, then, be able to request summaries of recent developments in chosen languages, to obtain responses to his questions for details, and to send commands, e.g., to zoom in on a particular body.

The challenge of building a cognitive system showing the aforementioned behavior involves addressing multiple research areas, such as computer vision, artificial intelligence and computational linguistics, to cite only a few. The term Human Sequence Evaluation (HSE) was coined to refer to this set of requirements, modules and flows of knowledge (numeric or semantic) that is essential for designing such a complex system [15]. As a result, HSE provided a theoretical framework upon which a European project called HERMES[1] was conceived and subsequently implemented thanks to the European Commission. HERMES was a consortium project that concentrated on extracting descriptions of people behavior from videos in re-

[1] http://www.hermes-project.eu

stricted discourse domains, such as pedestrians crossing inner-city roads, approaching or waiting at bus stops and even humans in indoor locations like halls or lobbies.

In this section, we present the resulting HERMES system that uses active cameras to help researchers in exploring a coherent evaluation of human movements and facial expressions across a wide variation of scale. The challenging objectives were the integration, demonstration and validation of different image processing techniques: in essence, the outputs of such techniques were pooled and integrated to build a coherent hardware and software system that can extract a subset of semantically meaningful behaviors from a scene.

To meet such objectives, the system uses active cameras to take high-resolution images of subjects, while still monitoring a large area in a manner similar to [24]. In such settings, the main issues tackled are:

- How to direct an active camera over a moving target while zooming on it. This process is referred to as foveation. Existing solutions to this problem only employ active tracking of motion and appearance to drive the camera motors [18].
- In the presence of several targets, how to select the most semantically relevant one to foveate on, according to a user-determined definition of relevance.

Solutions to this last problem select targets based on *Earliest Deadline First* policies, assigning higher relevance to those subjects that are going to leave the scene sooner [10]. Also, a solution based on the *Dynamic Traveling Salesperson Problem* has been proposed in [5]. While these approaches attempt to maximize the number of targets acquired over time, little effort has been made, yet, to maximize the quality, defined in semantically meaningful terms, of acquired images.

In the literature, active cameras are commonly used to provide a distribution of bodies and faces in the scene that can be exploited to select the best, most meaningful view. In [8], the technology to support tracking in a multiple-camera system is defined and is exploited for extracting and comparing the best view of each detected agent. Also, camera zoom allows active camera systems to supply imagery at the appropriate resolution for motion analysis of the human body and face, thus facilitating expression analysis [9, 30].

However, in our proposed framework the use of active sensors enhances the process of cognition via controlled responses to uncertain or ambiguous interpretations. In particular, the use of zoom provides a unification of interpretations at different resolutions, and bestows the ability to switch the sensing process between different streams in a controlled fashion. This integration of the cycle of *perception–knowledge acquisition–abstraction–reasoning–action generation* was also an interesting avenue of research.

In the rest of this section, we describe the resulting prototypical system that covers the aforementioned requirements. As a result, a slimmed-down demonstrator system, with both fixed and PTZ cameras, was able to generate natural language text based on activities of a particular agent (human or road vehicle) from schematic conceptual representations inferred using trajectory data.

### 2.4.1 Practical Considerations

An integrated hardware platform was designed, built and installed. This hardware platform consists of two high-speed cameras (one fixed and one PTZ) and three dedicated servers to host HERMES systems: for analysis of agent motion, active camera control and inferring high-level descriptions of agent behavior in the scene.

As noted before, the main objective of the HERMES project was set to improve active camera foveation ability by introducing semantics into active sensor guidance systems. This leads to the following specific sub-objectives:

- selecting the most appropriate low-level tracking techniques capable of focusing on specific aspects of agent motion like whole-body, limbs and face;
- constructing systems capable of classifying specific scene trajectories and human actions which form the basic attentive vocabulary for low-level scene description;
- improving active camera control systems to maximize the quality of acquired imagery based on low-level features;
- incorporating semantic feedback and requests from high-level scene description and reasoning into the active sensor control system, enabling it to acquire knowledge used for a robust and accurate description of the scene;
- designing active camera controllers capable of responding to uncertain or ambiguous interpretations;
- controlling active cameras in a manner that allows the analysis of three different degrees of human motion: agent, body and face, depending on the recognized behaviors;
- controlling active cameras to supply visual data while directing camera attention to those agents whose behaviors are deemed interesting.

These objectives allow a cognitive vision system to provide sensor data for each of the modules considered in HERMES, but more importantly, to bring all of the system modules together in a sensor perception/action cycle. Cooperating PTZ sensors enhance the process of cognition via controlled responses to uncertain or ambiguous interpretations. As a result, the use of zoom provides a unification for interpretations at different resolutions while exploiting the ability to switch the sensing process between different streams in a controlled fashion.

### 2.4.2 HERMES Hardware Platform

The HERMES-outdoor demonstrator platform was installed on top of the Computer Vision Center (CVC) building at the Universitat Autònoma de Barcelona, see Fig. 2.6. Based on a design of a demonstrator for indoor active surveillance scenarios [7], the demonstrator at CVC extends such a prototype to an outdoor scenario.

The hardware integration architecture is illustrated in Fig. 2.7. The hardware platform for the HERMES demonstrator consists of a fixed camera, another camera mounted in a pan/tilt platform and fitted with a zoom lens, three dedicated servers

**Fig. 2.6** The view from atop of the Computer Vision Center

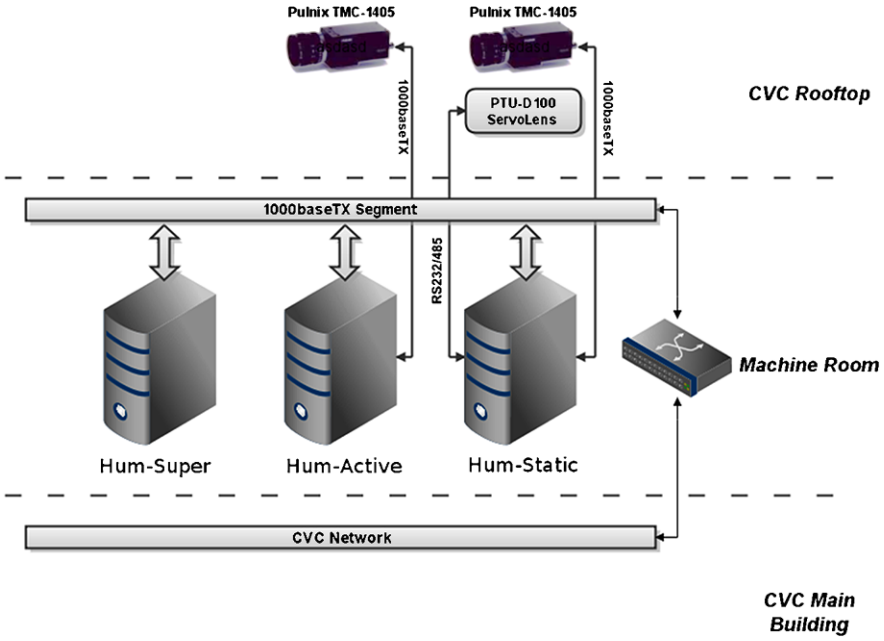


**Fig. 2.7** The HERMES demonstrator hardware architecture

to provide raw computational power and a fast 10 Gb Ethernet switch. The main components of the hardware infrastructure are:

**Cameras** Two Pulnix TMC-1405 cameras are used. These cameras are GigE-compatible and deliver high-resolution images ($1392 \times 1040$ pixels) at high framerate (30 frames per second). Each camera is connected by a dedicated, 100base-TX Ethernet connection to ensure constant, high-framerate streaming.

**PTZ Platform** One of the Pulnix cameras is mounted in a Directed Perception PTU-D100 pan/tilt platform that allows complete 360-degree pan and 180-degree tilt surveillance of the scene. The active camera is also fitted with a ServoLens zoom lens adjustable to focal lengths from 12.5 to 75 mm. Both the zoom lens and the PT platform are connected by direct RS-232/435 serial connections.

**Compute Servers** Three dedicated servers are used. Two of them are directly connected to the Pulnix cameras and are primarily dedicated to video acquisition. The third server is used for components not requiring direct access to the cameras, such as the supervisor tracker and SGT reasoning subsystems (explained later). These three machines are referred to as `hermes-super`, `hermes-fixed`, and `hermes-active` to emphasize their roles in the demonstrator platform.

**Network Infrastructure** The three servers are switched onto a 100baseTZ gigabit Ethernet segment in order to ensure the maximum possible bandwidth for communication among the demonstrator components.

### *2.4.3 HERMES Software Platform*

Here we discuss the software integration of the demonstration platform. A modular software architecture was designed, see Fig. 2.8, which illustrates how the software components are distributed across the HERMES demonstrator machines.

The aim is to support a set of distributed static and PTZ cameras and visual tracking algorithms, together with a central supervisor unit. Each camera (and pan–tilt

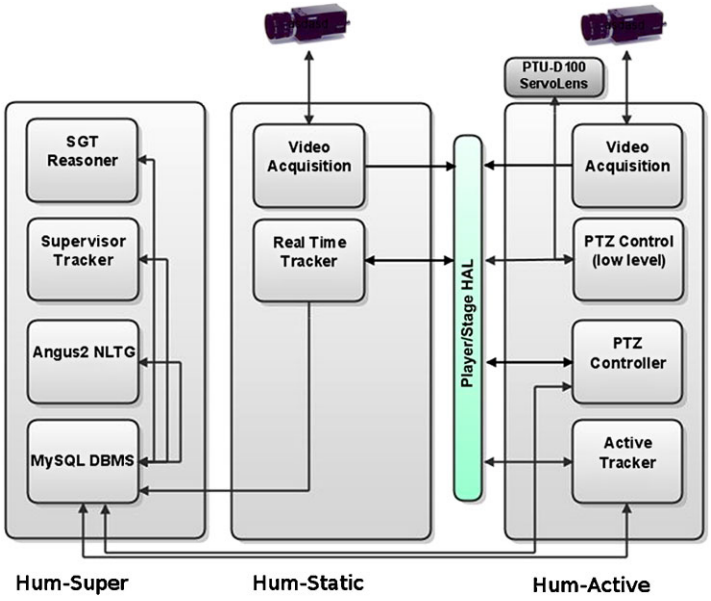


**Fig. 2.8** The HERMES demonstrator software infrastructure

device) has a dedicated process and processor. Asynchronous interprocess communications and archiving of data are achieved in a simple and effective way via a central repository, implemented using a MySQL database. Visual tracking data from static views are stored dynamically into tables in the database via client calls to the SQL server. A supervisor process running on the SQL server determines if active zoom cameras should be dispatched to observe a particular target, and this message is sent via writing demands into another database table.

**Video Acquisition** Video acquisition is performed using the JAI Advanced Imaging SDK. Frames captured by the two cameras at high resolution ($1392 \times 1040$ pixels) are scaled to the desired resolution ($640 \times 480$ pixels) for processing. Scaled frames are made available to the Player/Stage architecture through a shared memory interface. Video is delivered to the HERMES demonstrator components at a constant 25 frames per second.

**PTZ Controller** The Directed Perception P/T platform and ServoLens zoom lens were integrated into the Player/Stage driver system. A custom driver for the ServoLens was created in order to control the zoom lens through the standard Player/Stage PTZ interface.

**Real Time Tracker** The Real Time Tracker (RTT) is one of the fundamental components in the demonstrator platform [23]. The RTT tracks multiple moving targets in the fixed camera view and writes its observations into a table on the MySQL server.

**Supervisor Tracker** A Supervisor Tracker (SVT) is responsible for performing data fusion, smoothing and association based on observations made by the RTT; it is also responsible for issuing commands for the PTZ controller to actively track targets in order to acquire high-resolution imagery of active targets in the area of surveillance.

**SGT Reasoner** In order to demonstrate high-level reasoning and to support generation of natural language text from surveillance scenes, a SGT traversal system was integrated into the demonstrator platform [14]. The SGT reasoner listens for fused, smoothed observations coming from the SVT and records its inferences in a dedicated table on the MySQL database.

**Angus2 NLTG** The Angus2 system for natural language text generation has been adapted to read inferences generated by SGT traversal from the database [14].

**Player/Stage Hardware Abstraction Layer** Integration and communication between low-level components in the demonstrator system is achieved through the use of the Player/Stage system which provides a level of abstraction, allowing the video consuming components to receive streaming video without having to deal with the low-level details of the camera device itself.

**MySQL DBMS** At a very fundamental and low level, communication between the high-level components of the demonstrator is accomplished through a central MySQL database.

A user interface for controlling the real-time demonstrator was also built, allowing the user to administer and monitor the components of the demonstrator platform, see Fig. 2.9.

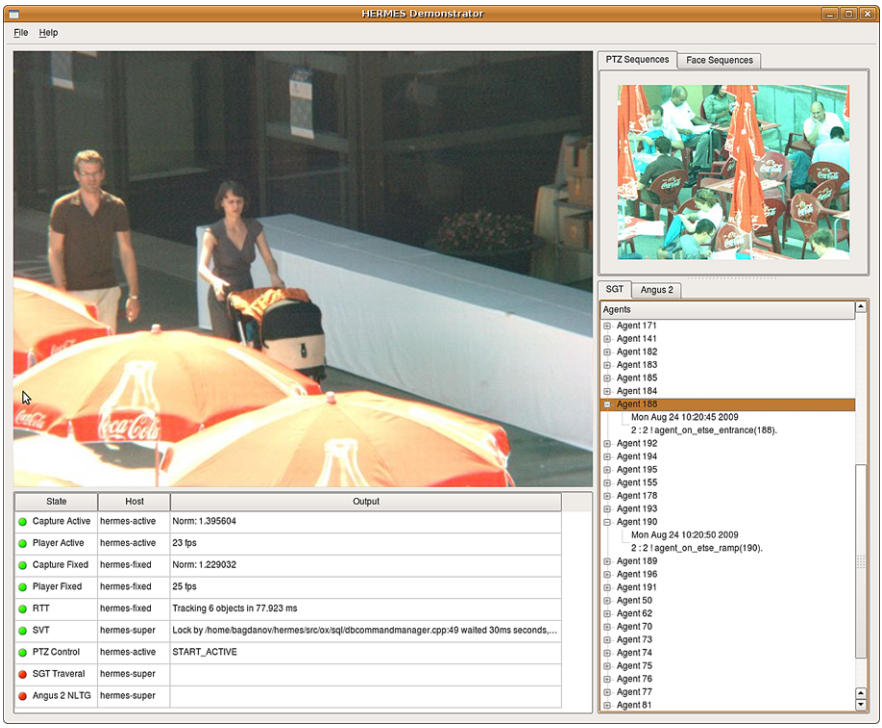


**Fig. 2.9** The GUI for the real-time active surveillance demonstrator

Working with images of size $640 \times 480$ pixels at 25 frames per second, our system can track up to 8 agents at a time. The PTZ camera can easily shift from one target to another; the time needed for the shift is between half a second and one second, depending on the angular distance between targets. The textual descriptions that can be generated include terms that describe status/gesture, such as: Run, Walk, Hand Wave, etc., and terms describing contextualized events, such as: Join, Appear, Disappear, Cross, Meet, Enter, etc.

The use of Player/Stage as an interface for low level modules and MySQL as an interface for high level modules was crucial to the success of the system.

## 2.5 Conclusions: Learning from the Past to Foresee the Future

The problem of covering relatively large scenarios with surveillance cameras, in such a manner that the targets of interest are captured with sufficient resolution, is still nowadays an open and active research field. Whereas camera networks are expensive and hard to manage and scale, active vision appears as a more natural solution to minimize the number of sensors while tackling the aforementioned goal.

Nevertheless, balancing the trade-off between area coverage and resolution per target calls for sensible techniques to control, integrate, and coordinate the possible passive and active components of an active vision system. Moreover, vision systems should be capable of providing human-interpretable descriptions of the occurrences being observed, and react according to certain policies. Thus, integrating the cameras with high-level semantic reasoning seems to be required for the intelligent capture and description of the interactions in a scene.

Toward extracting a semantic inference of what is happening in a scene and/or identifying semantically meaningful attentional factors, the possible future lines of research are:

- how the process of interpretation can be enhanced by PTZ sensors via semantic-based controlled responses to uncertain or ambiguous interpretations of human behaviors;
- how PTZ sensors can work on three different degrees of human motion analysis, i.e. agent, body and face, depending on the recognized behavior;
- how PTZ sensors can optimize transitions between these three degrees of resolution to supply visual data at the coarsest resolution, while subsequently directing the camera's attention to those agents deemed *interesting*;
- how the use of standard body and face detection algorithms can provide an attentional mechanism for controlling PTZ sensors.

The most interesting goal in the future is the control of zoom based on semantics and responding to uncertainty, in particular uncertainties and ambiguities due to high-level interpretations. Toward this goal, one could generate Natural Language descriptions for the active camera itself: a description and justification of what the camera is doing not only numerically but also semantically at a conceptual level.

In this context, semantics-driven control of active cameras will allow a computer vision system to better detect, track and reason about human motion using behavior models. Since these models are semantically rich, inference over them allows us to derive more meaningful targets toward which the active sensors should focus.

All in all, the automatic acquisition and exploitation of scene motion and context is compulsory to enhance the richness and expressiveness of semantic descriptions of human behaviors. By recognizing and labeling regions and objects associated with human activity, active systems will be able to reason about areas where humans are likely to be present and about the expected interactions with scene elements. Contextual reasoning will help to bridge the semantic gap by improving the ability to articulate high-level requests about human behaviors and send them to the active sensors acquiring low-level descriptions of the scene.

**Acknowledgements** This work has been supported by the European Project FP6 HERMES IST-027110. The authors wish to thank the rest of the partners in the HERMES consortium, namely AVL at Oxford University, BiWi at ETH Zurich, CVMT at Aalborg University and IAKS at Universität Karlsruhe. Also, the authors acknowledge the support of the Spanish Research Programs Consolider-Ingenio 2010: MIPRCV (CSD200700018); Avanza I+D ViCoMo (TSI-020400-2009-133); CENIT-IMAGENIO 2010 SEGUR@; along with the Spanish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02. Moreover, Murad Al Haj acknowledges the support from the Generalitat de Catalunya through an AGAUR FI predoctoral grant (IUE/2658/2007).

## References


1. Al Haj, M., Bagdanov, A.D., Gonzàlez, J., Roca, F.X.: Robust and efficient multipose face detection using skin color segmentation. In: Pattern Recognition and Image Analysis. Lecture Notes in Computer Science, vol. 5524, pp. 152–159. Springer, Berlin (2009) [20]
2. Al Haj, M., Bagdanov, A.D., Gonzàlez, J., Roca, F.X.: Reactive object tracking with a single PTZ camera. In: International Conference on Pattern Recognition, pp. 1690–1693 (2010) [15]
3. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. Int. J. Comput. Vis. **1**(4), 333–356 (1988) [13]
4. Anagnostopoulos, C.K., Anagnostopoulos, I.E., Psoroulas, I.D., Kayafas, E.: License plate recognition from still images and video sequences: A survey. IEEE Trans. Intell. Transp. Syst. **9**(3), 377–391 (2008) [12]
5. Bagdanov, A.D., Del Bimbo, A., Nunziati, W.: Improving evidential quality of surveillance imagery through active face tracking. In: International Conference on Pattern Recognition, pp. 1200–1203 (2006) [13,22]
6. Bashir, F., Porikli, F.: Collaborative tracking of objects in Eptz cameras. In: Visual Communications and Image Processing, vol. 6508, p. 2007 (2007) [14]
7. Bellotto, N., Sommerlade, E., Benfold, B., Bibby, C., Reid, I., Roth, D., Gool, L.V., Fernández, C., Gonzàlez, J.: A distributed camera system for multi-resolution surveillance. In: International Conference on Distributed Smart Cameras (ICDSC), Como, Italy (2009) [14, 23]
8. Calderara, S., Cucchiara, R., Prati, A.: Bayesian-competitive consistent labeling for people surveillance. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 354–360 (2008) [22]
9. Cohen, I., Sebe, N., Garg, A., Chen, L., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Comput. Vis. Image Underst. **91**(1–2), 160–187 (2003) [22]
10. Costello, C.J., Diehl, C.P., Banerjee, A., Fisher, H.: Scheduling an active camera to observe people. In: International Workshop on Video Surveillance and Sensor Networks (VSSN) (2004) [22]
11. Del Bimbo, A., Dini, F., Lisanti, G., Pernici, F.: Exploiting distinctive visual landmark maps in pan–tilt–zoom camera networks. Comput. Vis. Image Underst. **114**(6), 611–623 (2010). http://www.micc.unifi.it/publications/2010/DDLP10/DDLP10.pdf [14]
12. Denzler, J., Zobel, M., Niemann, H.: Information theoretic focal length selection for real-time active 3-d object tracking. In: International Conference on Computer Vision, pp. 400–407. IEEE Comput. Soc., Los Alamitos (2003) [13]
13. Erdem, U.M., Sclaroff, S.: Look there! Predicting Where to look for motion in an active camera network. In: International Conference on Advanced Video and Signal-based Surveillance (AVSS), pp. 105–110. IEEE, New York (2006) [14]
14. Gerber, R., Nagel, H.-H.: Representation of occurrences for road vehicle traffic. Artif. Intell. **172**(4–5), 351–391 (2008) [14,26]
15. Gonzàlez, J., Rowe, D., Varona, J., Roca, X.: Understanding dynamic scenes based on human sequence evaluation. Image Vis. Comput. **27**(10), 1433–1444 (2009) [21]
16. Hampapur, A., Pankanti, S., Senior, A., Tian, Y.L., Brown, L., Bolle, R.: Face cataloger: Multi-scale imaging for relating identity to location. In: International Conference on Advanced Video and Signal-based Surveillance (AVSS), pp. 13–20. IEEE, New York (2003) [14]
17. Ilie, A., Welch, G., Macenko, M.: A stochastic quality metric for optimal control of active camera network configurations for 3D computer vision tasks. In: International Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2), Marseille, France (2008) [14]
18. Murray, D.W., Bradshaw, K.J., McLauchlan, P.F., Reid, I.D., Sharkey, P.: Driving saccade to pursuit using image motion. Int. J. Comput. Vis. **16**(3), 205–228 (1995) [22]
19. Nelson, E.D., Cockburn, J.C.: Dual camera zoom control: A study of zoom tracking stability. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Comput. Soc., Los Alamitos (2007) [13]

20. Peixoto, P., Batista, J., Araujo, H.: A surveillance system combining peripheral and foveated motion tracking. In: International Conference on Pattern Recognition, vol. 1, pp. 574–577. IEEE, New York (2002) [14]
21. Qureshi, F.Z., Terzopoulos, D.: Surveillance in virtual reality: System design and multi-camera control. In: Computer Vision and Pattern Recognition, pp. 1–8 (2007) [14]
22. Qureshi, F.Z., Terzopoulos, D.: Multi-camera control through constraint satisfaction for persistent surveillance. In: International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 211–218. IEEE, New York (2008) [14]
23. Roth, D., Koller-Meier, E., Rowe, D., Moeslund, T.B., Gool, L.V.: Event-based tracking evaluation metric. In: International Workshop on Motion and Video Computing (WMVC), Copper Mountain, Colorado, USA (2008) [26]
24. Smith, P., Shah, M., da Vitoria Lobo, N.: Integrating multiple levels of zoom to enable activity analysis. Comput. Vis. Image Underst. **103**(1), 33–51 (2006) [22]
25. Sommerlade, E., Reid, I.: Information-theoretic active scene exploration. In: Computer Vision and Pattern Recognition (2008) [12]
26. Tordoff, B.J., Murray, D.W.: A method of reactive zoom control from uncertainty in tracking. Comput. Vis. Image Underst. **105**(2), 131–144 (2007) [13]
27. Wang, J., Zhang, C., Shum, H.: Face image resolution versus face recognition performance based on two global methods. In: Asian Conference on Computer Vision (2004) [12]
28. Welch, G., Bishop, G.: An introduction to the Kalman filter. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA (1995) [17]
29. Wrede, S., Hanheide, M., Wachsmuth, S., Sagerer, G.: Integration and coordination in a cognitive vision system. In: International Conference on Computer Vision Systems (ICVS), IEEE Comput. Soc., Los Alamitos (2006) [14]
30. Zhang, Y., Ji, Q.: Facial expression understanding in image sequences using dynamic and active visual information fusion. In: International Conference on Computer Vision (2003) [22]
31. Zhou, X., Collins, R.T., Kanade, T., Metes, P.: A master-slave system to acquire biometric imagery of humans at distance. In: International Workshop on Video Surveillance (VS), pp. 113–120, ACM, New York (2003) [14]

http://www.springer.com/978-0-85729-996-3

Visual Analysis of Humans
Looking at People
Moeslund, Th.B.; Hilton, A.; Krüger, V.; Sigal, L. (Eds.)
2011, XXII, 634 p., Hardcover
ISBN: 978-0-85729-996-3