

Observation Functions in an Information Theoretic Approach for Scheduling Pan-Tilt-Zoom Cameras in Multi-Target Tracking Applications

Tiago Marques, Luka Lukic, and José Gaspar

Institute for Systems and Robotics (ISR/IST), LARSyS, Univ. of Lisbon, Portugal
tiago.d.oliveira.marques@tecnico.ulisboa.pt, luka.lukic@epfl.ch,
jag@isr.ist.utl.pt

Abstract. The vast streams of data created by camera networks render unfeasible browsing all data, relying only on human resources. Automation is required for detecting and tracking multiple targets by using multiple cooperating cameras. In order to effectively track multiple targets, autonomous active camera networks require adequate scheduling and control methodologies. Scheduling algorithms assign visual targets to cameras. Control methodologies set precise orientation and zoom references of the cameras. We take an approach based on information theory to solve the scheduling and control problems. Each observable target in the environment corresponds to a source of information for which an observation corresponds to a reduction of the uncertainty and, as such, a gain in the information. In this work we focus on the effect of observation functions within the information gain. Observation functions are shown to help avoiding extreme zoom levels while keeping smooth information gains.

Keywords: Active Cameras Scheduling, Multi-target Tracking, Information Gain

1 Introduction

The increased need for surveillance in public places and recent technological advances in embedded video compression and communications have made camera networks ubiquitous. However, at the moment, there are still missing suitable algorithms that are capable of automatically processing so much data captured by so many cameras having few staff members.

One of the issues associated with this problem is the decision on which control action to send to the pan-tilt-zoom cameras (PTZ) in order to successfully carry on the desired surveillance tasks. In simple words, the cooperative problem of controlling a network of PTZ cameras for the purpose of active surveillance in a dynamic scenario is that one wants to maintain high zoom levels without losing track on the targets in the scene.

The first autonomous surveillance systems were composed of multiple static cameras, working together to solve some practical tasks of interest such as tracking moving objects. The need for considering overlapping and wide fields (resulting in low-resolution images) of views of deployed cameras has led to the deployment of pan-tilt-zoom cameras in modern surveillance systems.

A number of new architectures appeared, such as master-slave camera configurations, and cooperative smart networks [11]. In the master-slave configuration, static cameras are used for event detection in order to direct the PTZ camera to the target of interest. In contrast, with more complex architecture, both static and PTZ camera streams are used for event analysis. This way, the global state of these systems is composed of both the individual state of each target and the camera state. In both architectures, there is a need for efficient and reliable target tracking methodologies and, in more complex cooperative architectures, there is also a need for camera management methodologies, in order to compute the optimal configuration of the network in order to coordinately maximize the coverage of tracked visual targets in the scene.

1.1 Related Work

The surveillance problem in active camera networks can be divided into two main components. The detection and tracking of visual targets of interest in the scene and the computation of a scheduling policy to control each of the cameras' degree of freedom, in order to take into account the dynamics of the scene.

There are many generic methodologies used for target tracking. The commonly taken approach, based on the Bayes filter [1], [11], [10], is characterized at each iteration by the update of the state estimate based on the predicted state given by the motion model of the target and the observations given by the sensor. Another approach is taken in [7], [6], [3] and [2], where, contrary to the recursive approach, the trajectory is estimated in batches, making available both past and future observations for the estimation of the trajectory at a given time instance.

Regarding camera scheduling, Starzyk *et al.* [12] proposed a complete system for tracking multiple targets using cooperative cameras. The conflicts in behaviors are resolved using a central processor, which combines the individual desired behaviors in a single behavior, which reflects the best compromise between all of them.

The Multi-armed bandit algorithm is introduced in [13], [9] and [8], primarily not being a camera management system, but a decision methodology for coordinately allocating *resources* to *projects*, e.g. robots that can travel to certain locations in order to discover events or network packets that can be routed to various channels in order to maximize the throughput. This framework can be used to model the problem of camera management, as well. Each camera is considered as a resource and each target as a project. The objective is to allocate cameras (resources) to targets (projects) in order to maximize some measure of reward over time.

Another approach, based on the information theory framework, is applied to the problem of camera management in [11] and [10]. This approach is based on the previous work on automatic zoom selection in [5], in which the camera parameters (i.e. the target-camera assignment) are chosen based on the mutual information gain between the state estimate and the estimate given an observation. In this approach, the Extended Kalman Filter is used as the selected tracking algorithm.

The aforementioned approach is the one adopted in this work. However, it is noteworthy that the design choice of the observation functions is still an open and ongoing challenge with the family of the information theory approaches. In the following sections, we will present our particular approach.

1.2 Problem Formulation

A set of pan-tilt-zoom cameras is supposed to track and maintain trajectories of as many targets as possible. The targets are circulating in the environment.

We use an Information Theory framework in which the optimal control policy a^* for each camera is defined by the following maximization problem

$$a^* = \arg \max_a I(x; o) = \arg \max_a H_a(x) - H_a(x|o), \quad (1)$$

where $I(x; o)$ denotes the mutual information gain between the state estimate and the observation, and $H_a(x|o)$ is the conditional entropy of the state estimate given the observation, given that some control a was sent to the camera.

Each target has an assigned Extended Kalman Filter, with the motion model dependent on the targets being tracked and the pinhole camera model. When a target is in the field of view of a camera, a new observation is available and its value is used to update the EKF. In this scenario, there is a reduction of the uncertainty in the target's state and, as such, a positive information gain. On the other hand, when a target is not in the camera's field of view, the observation will not contribute to the reduction of the state uncertainty. In other words, the entropy $H(x) = H(x|o)$ and the information gain will be zero. This way, by maximizing the mutual information gain, the cameras will effectively be in configurations in which more targets are present in the field of view.

The aforementioned problem boils down to how to compute $H(x|o)$ before making an actual observation. This entropy can be computed by taking into account that the tracking is being made by using an EKF. In the Kalman Filter (and so in the Extended Kalman Filter), the assumption is that the state distribution follows a Gaussian distribution with the mean x (the state estimate) and covariance P . Under this property, $H(x)$ and $H(x|o)$ are both differential entropies of Gaussian distributed variables, given by

$$H(x) = \frac{k}{2}(1 + \log(2\pi)) + \frac{1}{2}\log(|\Sigma|), \quad (2)$$

where k is the dimension of x and Σ is the covariance matrix of $p(x)$.

The first result consists in that the entropy depends only on the covariance matrix and thus the problem of computing $H(x)$ and $H(x|o)$ can be reduced to computing the covariance of the state estimate P_k and the state estimate after the observation P_{k+1} .

The equations of the Extended Kalman Filter show that the observation z_k is only incorporated in the innovation equation and later used in the state update equation. Therefore, the update covariance matrix $P_{k|k}$ can be computed prior to any observation. The same applies to the conditional entropy $H(x|o)$, as well.

In the most general case, the optimization can be achieved by an exhaustive search procedure over the parameter space of the camera. However, the structure of the sensor can be considered in order to find better optimization strategies.

2 System Overview

The surveillance system consists of a set of active pan-tilt-zoom cameras which acquire images to be processed by a central controller. The controller is responsible for image processing tasks and for keeping track of each of the targets moving in the environment.

Each camera feeds the controller with new image frames and its corresponding state. The controller is responsible for processing the image and extracting target observations by fusing the available information and updating the targets state estimates (see Fig 1).

The result of the update is then fed to the decision controller, which is responsible for computing a new control policy for each camera.

The detection of new targets can be made both by carefully placed static cameras, or by the same active cameras used in the tracking process. The latter approach is the one followed in [10], however, this topic will not be addressed in this work.

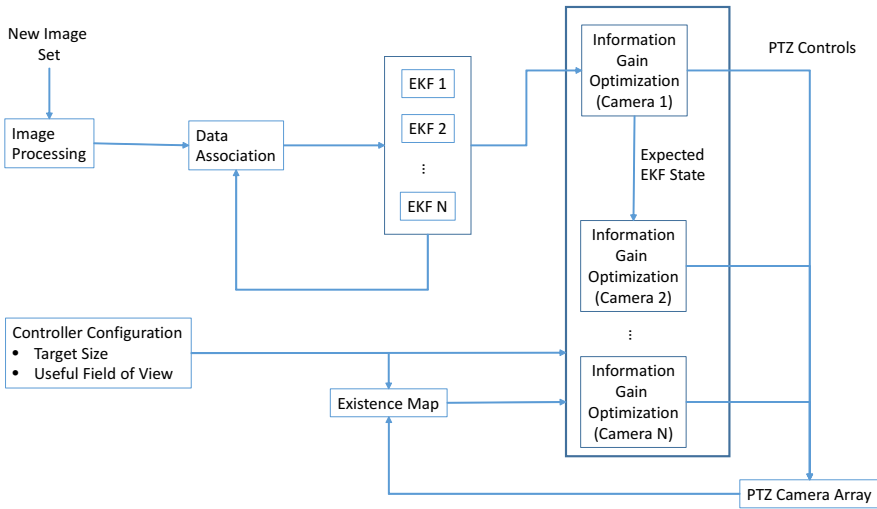


Fig. 1. System architecture

3 Scheduling Cameras

Using the Information Theoretic approach proposed in [10], the camera scheduling problem is modeled as an information gathering problem. For each camera, a choice on the pan-tilt-zoom parameters is made based on the information gain in the state distribution of all targets given possible observations. The optimal pan-tilt-zoom configuration is the one which maximizes the information gain $I(x; o)$ between the state estimate and the observation. In other words, it is the one that leads to a greater increase in the certainty of the state estimate.

3.1 Single Camera Scenario

In a single camera scenario, a single camera is responsible for tracking and maintaining the trajectories of various targets which move freely across the environment.

Each target is tracked using an Extended Kalman Filter (EKF) with the motion model depending on the kind of target being tracked (e.g. pedestrian or vehicle) and the pinhole camera model as the observation model. Let $h(x)$ denote the sensor model and H_k denote its Jacobian around the predicted state. Converting the pinhole camera model in projective coordinates

$$\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} = P_{3 \times 4} \begin{pmatrix} x' \\ y' \\ z' \\ h \end{pmatrix} \Rightarrow \begin{cases} u = \frac{f_1(x,y,z)}{f_3(x,y,z)} = \frac{p_{11}(\frac{x}{h}) + p_{12}(\frac{y}{h}) + p_{13}(\frac{z}{h}) + p_{14}h}{p_{31}(\frac{x}{h}) + p_{32}(\frac{y}{h}) + p_{33}(\frac{z}{h}) + p_{34}h} \\ v = \frac{f_2(x,y,z)}{f_3(x,y,z)} = \frac{p_{21}(\frac{x}{h}) + p_{22}(\frac{y}{h}) + p_{23}(\frac{z}{h}) + p_{24}h}{p_{31}(\frac{x}{h}) + p_{32}(\frac{y}{h}) + p_{33}(\frac{z}{h}) + p_{34}h} \end{cases} \quad (3)$$

where $f_k(x,y,z)$ is the internal product between the k^{th} row of the projective matrix and the real world position in projective coordinates and thus p_{ij} is the value of the projective matrix in the i^{th} row and j^{th} column.

By taking into account the structure of the sensor model, the Jacobian $H(x,y,z) \in \mathbb{R}^{2 \times 3}$ can be easily computed. Let x_i be the i^{th} state variable of the set $\{x,y,z\}$, and g_i the i^{th} observation variable from the set u,v , then the partial derivative in the position (i,j) of the matrix is given by

$$\frac{\partial g_i}{\partial x_j}(x,y,z) = \frac{\frac{\partial f_i}{\partial x_j}(x,y,z)f_3(x,y,z) - \frac{\partial f_3}{\partial x_j}(x,y,z)f_i(x,y,z)}{f_3(x,y,z)^2}. \quad (4)$$

The decision on which pan-tilt-zoom command to send to each camera is done by maximizing the mutual information gain $I(x_t; o_t)$ for all targets, with x_t being the target state estimate at time t and o_t the observation made.

Using the definitions of mutual information gain $I(x; o)$ and conditional entropy $H(x|o)$ [4], the cost function can be expanded

$$a^* = \arg \max_a I(x; o) = \arg \max_a H_a(x) - H_a(x|o) \quad (5)$$

$$= \arg \max_a H_a(x) + \int p(o) \int p(x|o) \log p(x|o) \partial x \partial o \quad (6)$$

$$= \arg \max_a H_a(x) + \int_v p(o) \partial o \int p(x|o) \log p(x|o) \partial x \\ + \int_{\bar{v}} p(o) \partial o \int p(x|o) \log p(x|o) \partial x, \quad (7)$$

where distribution $p(o)$ denotes the probability of making an observation of the target. Its domain of integration can be divided into two subdomains: v , which denotes all the camera configurations in which the target is visible and \bar{v} , which represents all the configurations in which the target is not visible.

Without further assumptions, this problem is hard to solve. However, recalling the EKF structure, predict and update steps, one obtains the following properties: (i) The

probability distribution $p(x|o)$ corresponds to the state distribution after the update step. In other words, $p(x|o) \sim \mathcal{N}(x_{k|k}, P_{k|k})$. (ii) All state variables are gaussian distributed. This means all differential entropies can be computed in closed form according to equation (2). (iii) When there is no observation of the target, the update step is skipped. This means that the state is independent from the observation and $p(x|o) = p(x) \sim \mathcal{N}(x_{k|k-1}, P_{k|k-1})$. Applying these properties into equation (7), the objective function is simplified:

$$a^* = \arg \max_a (H_a(x) + \omega(a)H(x^+) + (1 - \omega(a))H(x^-)), \quad (8)$$

where

$$\omega(a) = \int_{\nu} p(o) do \quad (9)$$

represents the observation function, which depends on the action a .

Note that all the state variables are Gaussian distributed and, by definition, the differential entropy for Gaussian distributed variables depends only on the variable covariance matrix. By observing the EKF equations, it can be seen that the observation o_k is only used in the innovation equation and in the state update equation. This enables the computation of P^+ without having an observation. Using the definition of differential entropy for a multivariate gaussian distribution (2) with the covariance update equation in the Extended Kalman filter, the cost functional can be simplified

$$a^* = \arg \min_a \omega(a) (\log |P^+| - \log |P^-|) \quad (10)$$

$$= \arg \min_a \omega(a) \log(I - K_k H_k). \quad (11)$$

The choice of the optimal configuration is made by optimizing the sum of the mutual information gains $I(x; o)$ for all targets. Despite the elegance of this cost function, in the general case its evaluation is intractable because it requires the exhaustive search over the configuration space of the camera. For each pan-tilt-zoom configuration of the camera, the observation model must be linearized anew and an EKF update must be performed to obtain the new Jacobian matrix H_k and K_k . In order to overcome that problem, in this work each target is modeled as a circle projected into the ground plane. The visible region ν of each camera, in the camera coordinates, is also modeled as an ellipse around the center of the image. The term $w(a)$ in equation (11) is computed by projecting the target ellipse onto the image plane and computing its intersection with the visible region ν of the camera.

3.2 Multiple Camera Scenario

In a multiple camera scenario, multiple cameras can have observations of the same target. The incorporation of these observations is done using a sequential Kalman Filter. In this variation, a single prediction step is made and each observation is used to make an update so that the update from the observation k uses the estimated position and covariance of the update from the observation $k - 1$.

Each observation contributes to the covariance matrix with a factor of $(I - K_c H_c)$, where K_c is the Kalman gain from the observation of camera c and H_c is the Jacobian of the observation model for camera c .

The mutual information gain of a target for multiple observations

$$I(x; o_1, \dots, o_C) \propto \sum_{c \in C} \log |I - \omega(a) K_c H_c|, \quad (12)$$

is then obtained by combining equation (10) with the new covariance matrix update

$$P^+ = \left(\prod_{c \in C} (I - K_c H_c) \right) P^-. \quad (13)$$

3.3 Observation function

The observation function $\omega(a)$, equation (11), is a central component of the target tracking methodology as it effects on the convexity of the information gain.

The observation function can be just a flag indicating whether a point representing the target location, in world coordinates, is visible or not in the image:

$$\omega(a) = \begin{cases} 1, & \mathcal{P}(X_{gnd}; a) \in E_{img} \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

where X_{gnd} is a point in the ground plane representing the target position, $\mathcal{P}(\cdot; a)$ is the projection operator on a set of points, which is defined by action a , i.e. in (3) the projection matrix $P_{3 \times 4}$ is modified by action a and E_{img} is the ellipse in the image plane concentric with the image rectangle. However, this observation function does not perform a smooth regularization of the cost function, making difficult to design iterative optimization algorithms.

By changing the model of the target from a single point in ground plane into a surface in the ground plane, one obtains a smoother optimization function. Modeling the shape of the target as a rectangle, taking into account the visible part normalized by the area of the imaged shape not truncated by the field of view of the camera, the observation function takes the form

$$\omega(a) = \frac{A(\mathcal{P}(R_{gnd}; a) \cap E_{img})}{A(\mathcal{P}(R_{gnd}; a))} \quad (15)$$

where $A(\cdot)$ indicates area of a convex hull of points and R_{gnd} is a rectangle in the ground plane. Alternatively, modeling the target as an ellipse in the ground plane, the function becomes

$$\omega(a) = \frac{A(\mathcal{P}(E_{gnd}; a) \cap E_{img})}{A(\mathcal{P}(E_{gnd}; a))} \quad (16)$$

where E_{gnd} is an ellipsis in the ground plane.

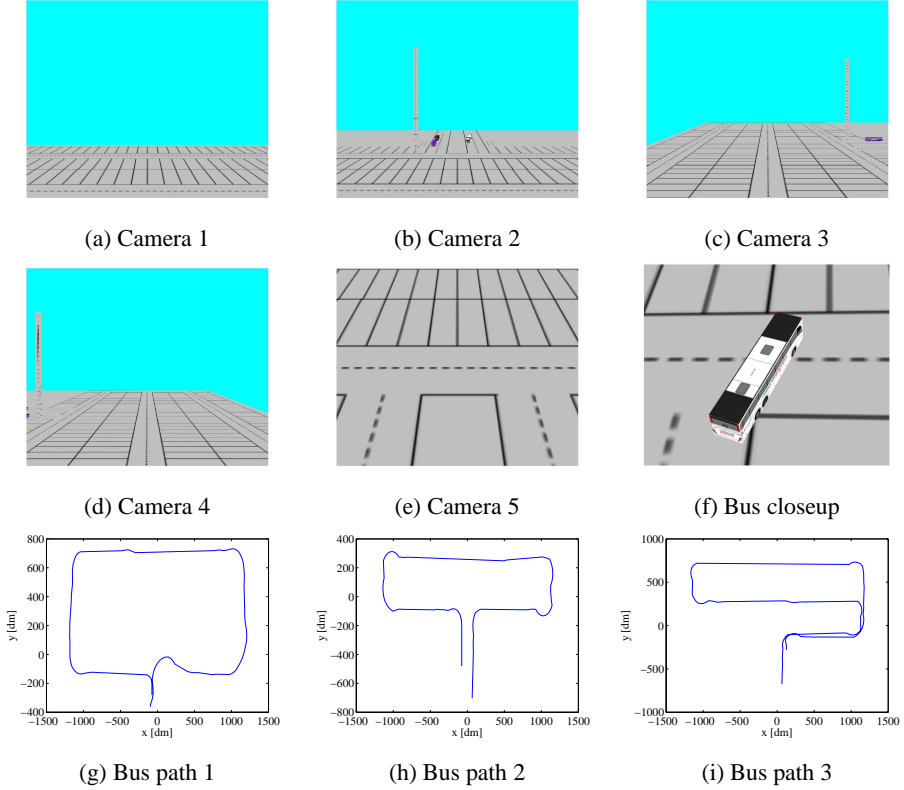


Fig. 2. Field of view of the cameras at rest orientation and typical bus trajectories.

4 Experimental Results

The experiments described in this section are based in a virtual reality environment simulating a parking lot (see Fig. 2(a-f)). A number of buses, with the dimension of $2 \times 10 \times 2 \text{m}$, cross the scene according to predefined trajectories shown in Fig. 2(g-i). These trajectories are generated using the car model $\dot{x} = \cos(\theta)\sin(\theta)V$, $\dot{y} = \sin(\theta)\cos(\theta)V$, $\dot{\theta} = \frac{\sin(\phi)}{L}V$ and $\dot{\phi} = \omega_s$, where V is the linear velocity of the bus and ϕ is the steering angle, both set using a joystick interface. Tracking is performed using five pan-tilt-zoom cameras located at fixed positions on the sides of the parking lot and in the entrance.

In order to evaluate the geometry of the cost function using different observation functions, the cost function was evaluated from a set of values from the pan-tilt-zoom space for a camera. This was achieved by placing a target in a fixed location in front of the camera and making an observation. The camera model and its Jacobian were computed for the new pan-tilt-zoom configuration and, along with the new observation, used to update an existing EKF. The resulting Kalman gain and the linearized observation model were obtained and used in the cost function (11).

Figures 3(a) show the cost function evolution with both the zoom level and the pan-tilt values using the observation function (14). Figures 3(b) show the same evolution as in the previous case, but now using the observation function (16). The term $w(a)$ defined as in (11) changes the cost minimum from extreme values of zoom, pan and/or tilt to within-range, central, values.

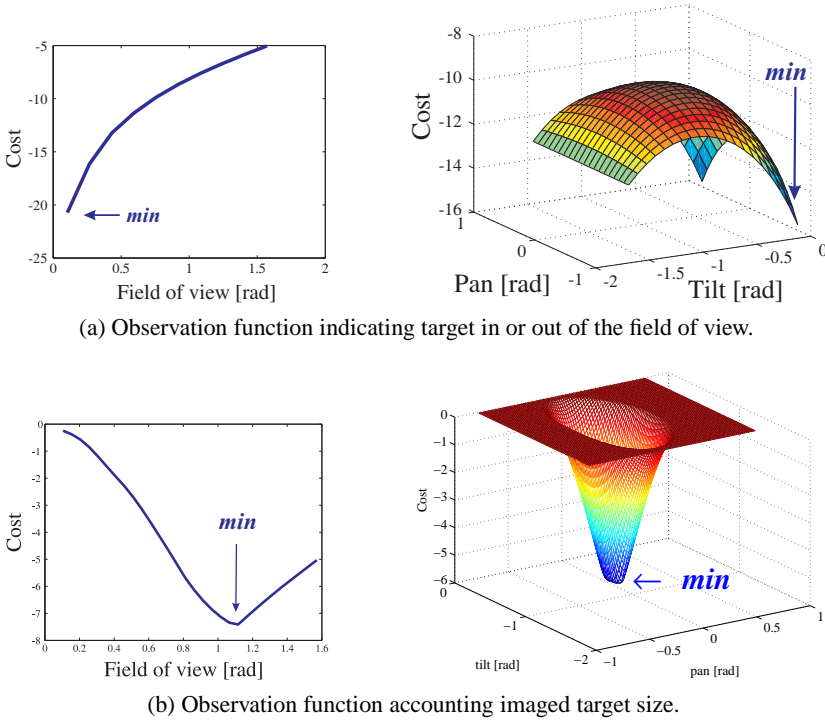


Fig. 3. Effect of the observation function into the cost function. In case (a) the observation function $\omega(a)$ is defined by (14) and therefore just indicates the target in or out the field of view. In case (b) $\omega(a)$ is defined by (16) and thus takes in account the imaged size of the target and the field of view. Two plots for each one of the cases, cost vs zoom (left) and cost vs pan and tilt (right).

The second set of figures was generated in a similar way from the former ones, however sampling the whole camera parameters space. Figure 4 shows slices of the cost function in the camera parameter space (pan-tilt-zoom) using different observation functions as described in Section 3.3. Each row represents a fixed zoom configuration, with lower rows representing configurations with higher field of view (less zoom).

In column (a) of Fig. 4, is used $w(a)$ defined in (11). In column (b) is considered the same $w(a)$ however, in this experiment, there are two targets in the scene in different locations and with different state estimate covariances. In columns (c) and (d) the term $w(a)$ is the one defined by (15) and (16), respectively.

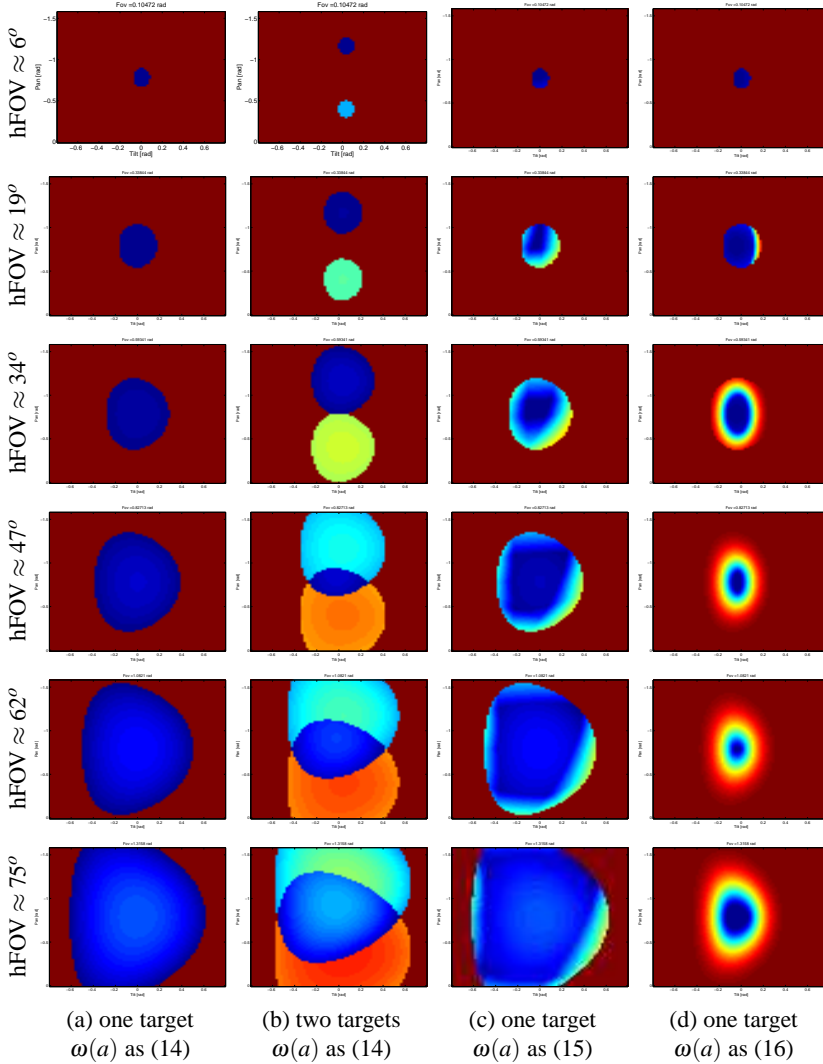


Fig. 4. Cost function slices at fixed field-of-view (zoom). Each row represents a different zoom level (field of view, top row $6[\text{deg}] \approx 0.1 [\text{rad}]$, bottom row $75[\text{deg}] \approx 1.3 [\text{rad}]$). Colder colors represent lower cost functions. Configurations in which the target was not visible were assigned high cost. In cases (a,c,d) one camera observes one target. In case (b) one camera observes two targets. The observation function $\omega(a)$ is defined by (14) in cases (a) and (b), and is defined by (15) or (16) in cases (c) or (d), respectively.

Figure 4(a) shows that the cost function is generally lower in configurations which make the target appear in the image plane with higher resolution, that is, when the camera is at its highest zoom or when the camera has the target of interest in the corner of the image. This result comes from the influence of the observation model (in particu-

lar, its Jacobian) used in the EKF in the cost function. While this is a good result in the sense resolution is maximized, uncontrolled zoom onto the target is not the desired behavior, since a minimum movement can make the target disappear from the camera field of view. Term $\omega(a)$, as defined in (15) or (16), acts a regularizing factor, managing the optimal zoom level to observe the target at constant zoom level and preventing observations just using the “corner of the eye”.

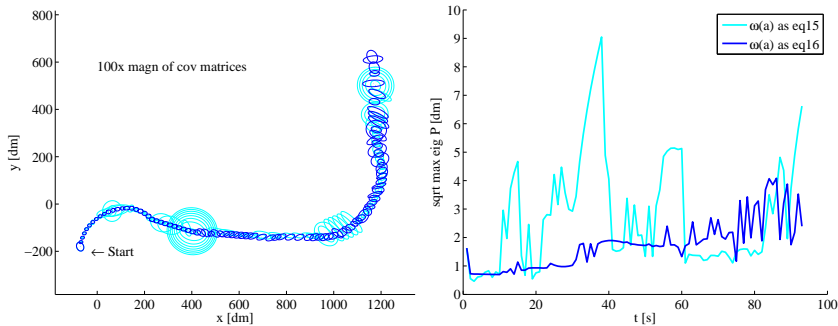


Fig. 5. Scheduling experiment. Uncertainty at the ground plane (left) and uncertainty along time (right), for observation functions defined in (15) or (16).

In the last experiment four cameras are active and four targets (buses) enter the scene sequentially, separated by approximately 10sec, following different trajectories. The proposed scheduling methodology is tested with two alternative observation functions, (15) or (16). The pan, tilt and zoom parameters are estimated for all cameras in a round-robin manner (Fig. 1). Each search of the parameters is based in Nelder-Mead simplex direct search and is limited in the number of iterations.

Figure 5 shows the tracking of the first bus along the first 100sec. The plot in the left shows the uncertainty of the EKF of the first bus, on the ground plane, as ellipses corresponding to a 50% confidence level. Covariances are magnified $100\times$ for readability. The second plot, Fig. 5(right), shows the square root of the maximum eigenvalue of the covariance matrix, along time, for both observation functions. Results show that (16) allows for lower and smoother in time localization uncertainty in presence of distractors, as the other buses entering the fields of view of the cameras, due to its smoother nature allowing for more effective searches of the pan, tilt and zoom parameters.

5 Conclusions and Future Work

The results show the influence of the target modeling in the cost function proposed by [10]. By modeling the buses as ellipses projected onto the ground plane and the visible region of the camera as an ellipse in the image plane, the cost functions gain a defined structure, which simplifies the design of fast algorithms to search for the optimal policy.

When a single target is present inside the range of the active camera, the optimal command is to zoom on the target, according to the cost function. The regulator term controls the optimal zoom on the target, independently of the target location.

In a multi-target scenario, the overlapping cost functions can make the optimal command to keep both targets inside the camera's field of view.

The challenges to be addressed in the future include the integration with policies for exploration of unobserved regions, making the system independent of the need to have carefully placed static cameras.

Acknowledgments

This work has been partially supported by the FCT project [UID / EEA / 50009 / 2013], and by the EU Project POETICON++ EU-FP7-ICT-288382.

References

1. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., Arulampalam, S.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing, IEEE Trans. on* 50(2), 174–188 (2002)
2. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: *IEEE Int. Conf. on Computer Vision (ICCV)*. pp. 137–144. IEEE (2011)
3. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. *Pattern Analysis and Machine Intelligence, IEEE Trans. on* 33(9), 1806–1819 (2011)
4. Cover, T.M., Thomas, J.A.: *Elements of information theory*. John Wiley & Sons (2012)
5. Denzler, J., Zobel, M., Niemann, H.: Information theoretic focal length selection for real-time active 3d object tracking. In: *IEEE Int. Conf. on Computer Vision (ICCV)*. pp. 400–407. IEEE (2003)
6. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. *Pattern Analysis and Machine Intelligence, IEEE Trans. on* 30(2), 267–282 (2008)
7. Jorge, P.M., Abrantes, A.J., Marques, J.S.: Automatic Tracking of Multiple Pedestrians with Group Formation and Occlusions. In: *VIIIP*. pp. 613–618 (2001)
8. Liu, K., Zhao, Q.: Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Trans. on Information Theory* 56(11), 5547–5567 (2010)
9. Ny, J.L., Dahleh, M., Feron, E.: Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In: *American Control Conf.* pp. 4220–4225. IEEE (2008)
10. Sommerlade, E., Reid, I.: Probabilistic surveillance with multiple active cameras. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. pp. 440–445. IEEE (2010)
11. Sommerlade, E., Reid, I., et al.: Cooperative surveillance of multiple targets using mutual information. In: *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008* (2008)
12. Starzyk, W., Qureshi, F.Z.: Multi-tasking smart cameras for intelligent video surveillance systems. *8th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)* pp. 154–159 (Aug 2011)
13. Whittle, P.: Restless bandits: Activity allocation in a changing world. *Journal of applied probability* pp. 287–298 (1988)