# Non-Myopic Information Theoretic Sensor Management of a Single Pan-Tilt-Zoom Camera for Multiple Object Detection and Tracking

Pietro Salvagnini[1][a], Federico Pernici[b], Marco Cristani[a,c], Giuseppe Lisanti[b],
Alberto Del Bimbo[b], Vittorio Murino[a,c]

[a]*Istituto Italiano di Tecnologia, Pattern Analysis & Computer Vision Department, Via Morego 30, 16163 Genova*
[b]*University of Florence, Media Integration and Communication Center, Viale Morgagni 65, 50134 Firenze*
[c]*University of Verona, Department of Computer Science, Strada Le Grazie 15, 37134 Verona*

## Abstract

Automatic multiple object tracking with a single pan-tilt-zoom (PTZ) cameras is a hard task, with few approaches in the literature, most of them proposing simplistic scenarios. In this paper, we present a novel PTZ camera management framework in which at each time step, the next camera pose (pan, tilt, focal length) is chosen to support multiple object tracking. The policy can be myopic or non-myopic, where the former analyzes exclusively the current frame for deciding the next camera pose, while the latter takes into account plausible future target displacements and camera poses, through a multiple look-ahead optimization. In both cases, occlusions, a variable number of subjects and genuine pedestrian detectors are taken into account, for the first time in the literature. Convincing comparative results on synthetic data, realistic simulations and real trials validate our proposal, showing that non-myopic strategies are particularly suited for a PTZ camera management.

*Keywords:* Pan-Tilt-Zoom Camera, Multiple Object Tracking, Sensor Management, Markov Decision Process

## 1. Introduction

Visual Tracking of multiple objects in realistic outdoor scenarios is often performed in wide areas. In these viewing conditions a stationary fixed focal length camera has typically too limited field of view and image resolution with respect to the scene extent. Therefore, a network of cameras is used to sufficiently cover the area at the required resolution [1, 2, 3, 4, 5]. However, this may be unfeasible for the cost associated to the setup and maintenance of the camera network, as well as for the practical impossibility to provide all the necessary resolutions for target biometric recognition at a distance. Similarly, in the case of a vehicle mounted camera [6, 7, 8, 9] it would be difficult to cover a wide area at adequate resolution due to the limited acceleration at which the camera may be moved. Active Vision [10] and specifically Active Pan Tilt Zoom (PTZ) cameras, have promised to solve these limitations, permitting at least in principle the monitoring of a large space at variable image resolutions [11, 4]. However, letting a large number of stationary

[1] Corresponding author. E-mail address: *pietro.salvagnini@gmail.com*

or PTZ cameras operate in a cooperative way is still an expensive and complex solution [12, 13]; for this reason, exploiting a single zooming sensor could be a more reasonable and worthy goal. According to this, in this paper, we propose and show the benefits of an active sensing approach to multiple object detection and tracking using a single pan tilt zoom camera.

Despite the high exploitable potential, when applied for the task of multiple object tracking in world coordinates, a single PTZ camera induces a number of complex problems that must be solved to obtain effective results [14, 15]. Specifically, camera calibration solutions adopting natural landmarks [16, 17] should be preferred with respect to others adopting domain specific scene landmark geometry as in [11, 15, 18]. Since the PTZ camera must also undergo rapid and unpredictable motions to rapidly gaze at any part of the field of view, real time tracking of camera motion should not be based on recursive filtering, but on keyframe based methods [16, 19]. At the same time, the scene background appearance must be continuously updated [16, 20, 21]. Moreover, due to the fact that monitoring is performed in a large area, accurate objects localization in a common 3D world reference frame is needed to track targets at a distance. This requires some form of online camera calibration since the camera parameters change dynamically. The framework we developed in [22] is conceived to support all these requirements and is therefore suitable to be used in task-driven active surveillance of scenes with multiple moving objects.

Starting from this framework [22], we propose a solution for sensor management (i.e. determine the best way to control the visual sensor) in order to enhance multiple target detection and tracking in a wide area. Here the focus is on non-myopic sensor management where the long-term ramifications of taking a particular *sensing action* are accounted for decision making. A sensing action may consist of choosing a particular image processing modality (e.g. pedestrian detection or motion detection), a particular camera pose and focal length, or a combination of the two. Information gain [23] is chosen as performance indicator of decision making, since it has the desirable property that different inhomogeneous sensing actions can be simultaneously optimized in a single metric. This requires to maintain the probability density which capture uncertainty in the current state estimate. In our setting there are multiple actions that can be tasked by evaluating a single global metric, some of which contribute better than others to tracking. The PTZ camera sensor is used to gain information about the kinematic state (e.g. position and velocity) and objects detectability[2]. There are many objectives that the sensor manager may be tuned to meet, e.g. minimization of track loss, probability of target detection, minimization of track error/covariance, and identification accuracy. Each of these different objectives taken alone may lead to a different sensor allocation strategy. As detailed in Sec. 4, we jointly optimize over all these objectives by maximizing the expected amount of information extracted from the scene, namely the expected information gain between the current objects state estimate and the state estimate after a measurement has been made. Since the best sensing action must be selected before actually executing it, what is practically maximized is the expected reduction in entropy (i.e. the expected information gain) that a sensing action will produce. Fig. 1 shows the three main components of the complete multi target tracking system for a single PTZ camera.

The sensor management problem can be approached in a principled way with the Markovian Decision Process (MDP) formalism [24]. However, the long-term (non-myopic) planning solution suffers from combinatorial explosion and may be defined, as in our case, in a continuous state space. Approximate solutions are therefore required and will be discussed in the next

[2] Image object measurements are obtained according to a detector that have a time-varying object response characteristics. For example in the case of pedestrian detection as processing modality, the response characteristic varies depending on the imaged size of the object and on how much the object is occluded.

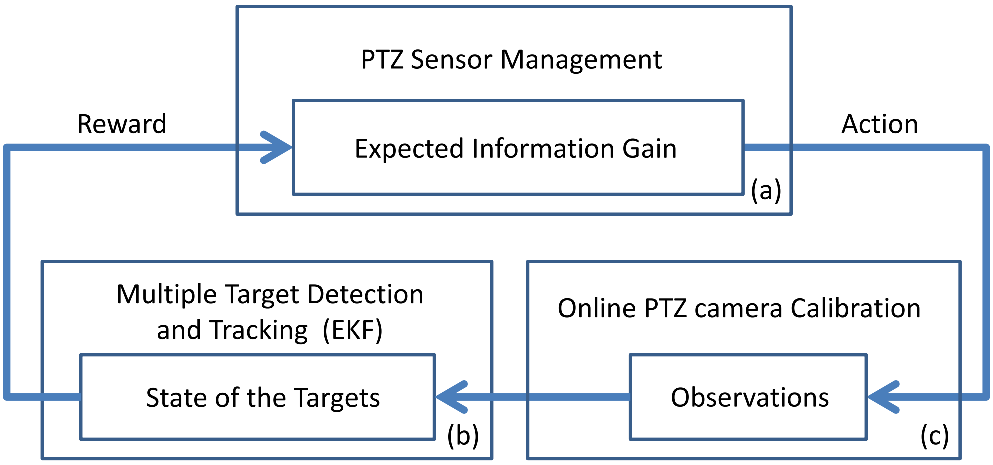
PTZ Sensor Management
Expected Information Gain
(a)
Reward
Action
Multiple Target Detection and Tracking (EKF)
State of the Targets
(b)
Online PTZ camera Calibration
Observations
(c)

Figure 1: The three main components of the system.

section. This work extends the preliminary results we obtained in [25] where we have analyzed the myopic (i.e. greedy) aspects of sensor management. Experimental results show that the non-myopic strategy provides a substantial performance improvement by better capturing the complex space-time trade-off between objects and camera motion. Two motivating examples for which the non-myopic will outperform the myopic strategy are: 1) the case in which an object is repeatedly measured before it gets occluded so as to sharpen its uncertainty when it reappears; 2) the case in which objects are measured exploiting the calibrated zoom[3] so as to sharpen their uncertainties. The underlying assumption is that if the operative scenario evolves with reasonable temporal coherence, it is possible to predict the ability of gathering information of a future action.

Synthetic and real experiments are shown confirming the suitability of our approach for realistic scenarios. Fig. 2 shows few frames from the three sets of experiments.

The rest of the paper is organized as follows. We give and overview of related work in Sec. 2 while we summarize our contributions in Sec. 3. The information theoretic formulation based on MDP for the myopic version is presented in Sec. 4, and the modeling of the real world challenges such as missed detections and occlusions is presented in Sec. 5. The non-myopic version is described in Sec. 6. In Sec. 7 we give a detailed discussion about how the proposed solution can be extended to a network of multiple cameras. Some implementation and evaluation details are given in Sec. 8. Experiments for the myopic framework are reported in Sec. 9 while experiments for the non-myopic version are reported in Sec. 10. Finally the conclusions are drawn in Sec. 11.

## 2. Related work

Automatic multiple object tracking with a single pan-tilt-zoom camera is a hard task with few approaches present in the literature, most of which propose simplified scenarios. One of

[3] Calibrated zoom allows increasing measurement accuracy in world plane coordinate object localization. In the Appendix a formal proof of this result is presented.

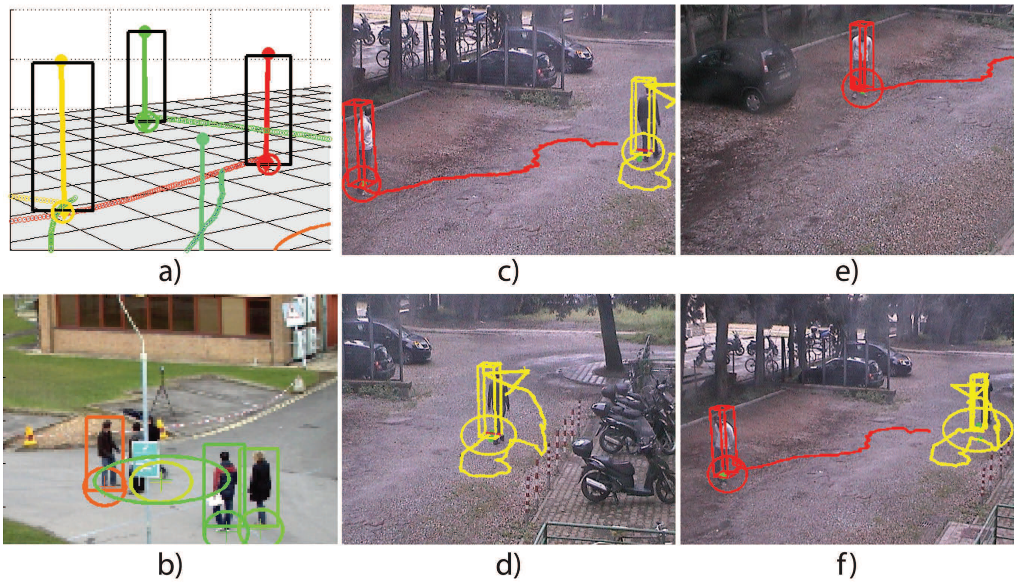

Figure 2: (a) Synthetic scenario; (b) realistic simulation; (c-f) real trial (best viewed in colors).

the most challenging part is that the evaluation with real data require the development of a real time system since it is not possible to work offline with pre-recorded[4] videos. To deal with this issue, in [26] a completely simulated environment is created through computer graphics and different strategies for camera to target assignments are proposed and compared. In [27] the authors propose a system for cooperative tracking between multiple Active Vision Agents (AVA). In this solution each AVA agent manages visual perception, camera action and network communication to perform cooperative tracking. Different scheduling policies for a network of PTZ cameras in a master-slave configuration were tested in [28, 29]. However, the strategies described above [26, 27, 28, 29] are mainly hand-crafted, and require precise information on the targets' position from other sensors. An overview of recent methods for managing PTZ camera networks can be found in [30, 31].

Principled information theoretic frameworks exploiting the concept of information gain for single object tracking are introduced in [32, 33, 34]. In [32] optimal selection of the focal lengths of two cameras during active 3D object tracking is proposed. This is the first work on active focal length selection for improving accuracy in 3D object tracking. Despite the promising results observed in a controlled laboratory test, the system is not yet mature to work in unconstrained video sequences. In [33], the authors propose a non-myopic solution for optimal focal length selection based on the minimization of the expected entropy of a tracked object. Tracking is performed in 2D, only on simulated data, using an extended Kalman filter. In [34] the authors suggests a method to control the zoom in order to obtain maximum resolution by placing a limit in the innovation of a constant velocity Kalman filter. In particular, the zoom is used to modify

[4] If the video is recorded at high resolution it is possible to crop and downsample the image to get the desired field of view. However the level of detail and the quality of the image that can be captured with optical zoom is still orders of magnitude larger than the one achievable with the digital zoom.

Table 1: Overview of the main characteristics for state of the art methods and our solution.

| **Method** | **Sync.** | **Calibration** | **# Tested Cameras** | **Multi target** | **Occlusions** | **Optimized cost** |
|---|---|---|---|---|---|---|
| [12] | Yes | Yes (offline) | 4 fixed and 4 PTZ | Yes, in the fixed cameras | Yes, in the fixed cameras | Highest target resolution |
| [13] | Yes | Yes (offline) | 9 PTZ | Yes | No | Tracking accuracy and highest target resolution |
| [35, 36, 37] | No | Yes (online) | 1 PTZ | No | No | Costant object imaged size |
| [38] | No | No | 1 PTZ | Yes | No | Tracking accuracy |
| [39] | Yes | Yes (offline) | 2 PTZ | Yes | No | Tracking accuracy and highest target resolution |
| [40] | Yes | Yes (offline) | 2 PTZ | Yes | Yes | Tracking accuracy |
| Ours | No | Yes (online) | 1 PTZ | Yes | Yes | Tracking accuracy |

the measurement process. Other works that control the focal length to keep the imaged size of a single object constant were proposed in [35, 36, 37], but no 3D localization uncertainty is taken into account and the extension to multiple objects is not trivial.

Sensor management for the task of multiple object tracking is addressed in [38]. Here tracking is performed in the image plane and therefore focal length selection cannot be used to improve the accuracy in 3D object localization. Multiple zooming cameras which give a 3D representation of target positions are considered in [39] for the multi-target scenario. In both [38] and [39] the evaluation is carried out with ground truth data (i.e. sources of error from the detection, tracking and and data association stage are ignored). All the works described above optimize over a single step look-ahead (i.e. myopic) except for the method in [33] which optimize over multiple step ahead for the task of tracking a single moving object. However testing is conducted with single object in a constrained simulated environment and only the zoom is managed.

Recently novel sensor management approaches with real-time implementations has been reported in [12] and [13] with convincing results. The network camera system described in [12] comprises a total of eight cameras, four fixed and four PTZ. The fixed cameras are processed at a resolution of $320 \times 240$ while no image processing is performed on the PTZ views. The sensor network in [13] includes nine PTZ network IP cameras with a resolution of $320 \times 240$ pixels and $12\times$ optical zoom. In this system the control of the PTZ parameters is modeled as a multiplayer game where the cameras gain by reducing the error covariance of the tracked targets or through higher resolution feature acquisition, which, however, comes at the risk of losing the target. The work in [40] proposes a distributed approach to optimize various scene analysis performance criteria through distributed control of a dynamic camera network including the uncertainty of the targets. All these works adopt a large number of stationary or PTZ cameras operating in a coordinated way (typically in master-slave configuration [41, 42]). Although these approaches could be in principle applied to the case of a single camera we are not aware of any work investigating in this regard.

In Tab. 1 we give an overview of the main characteristics for some of the methods described above and our solution. In particular, we highlight the main differences in terms of: necessity of synchronization between the sensors involved in the network, calibration of each sensor, number of cameras, number of tracked targets, occlusions management and the cost to be optimized. In particular, the cost to be optimized can be: the accuracy in tracking the targets (Tracking accuracy), the necessity to maintain constant the size at which the object is observed (Costant object imaged size) or the necessity to obtain the highest resolution for the object of interest (Highest target resolution).

*Computational Model.* The sensor management problem is generally approached as planning under uncertainty according to Markov Decision Processes (MDPs) [24]. Such framework explicitly models the temporal state evolution and designs a policy for selecting the action based on a reward function. However, optimal long-term solutions suffer from combinatorial explosion, for this reason suboptimal approximate methods must be applied.

The non-myopic strategy can be optimized with a Monte Carlo rollout strategy as described in [43, 44]. These approaches address the solution for large MDPs while small problems can be directly solved with Dynamic Programming [45]. There are two basic variants for estimating (online) an approximate strategy of a MDP and both these variants can be classified based on the length of the planning horizon, namely: Monte Carlo Tree Search methods (MCTS) [46] and Reinforcement Learning (RL) based methods [47]. The former guide the search using results from rollouts in the decision tree of the actions and are appropriate for the finite horizon case. The latter are most indicated for finding approximate solutions in the infinity horizon case. The method that we investigate here is focused on finite horizon and includes sparse sampling techniques for direct approximation of the Bellman equation, as described in [48]. A relevant application of this technique has been recently presented in [49] for the task of tracking vehicles from radar imagery. The rollout approach driven by information metric is exploited to capture the long-term reward due to expected visibility and occlusion of objects.

Another application of MDP, hidden MDP (hMDP), is proposed in [50], where a target moving in the scene is modeled as an agent for which the state is the position on the plane and the action is its future direction. In this work the goal is to estimate the policy it is following in order to forecast its future behavior.

## 3. Contributions

Our contributions with respect to the related work are:

- A well-founded theoretical solution for the information theoretic management of an active camera, which keeps into account all the typical sources of error of a tracking system: detector performance, limited field of view, occlusions among targets, variable number of targets. The solution has been divided in two techniques: one myopic, introduced in[25] and here fully detailed in all the mathematical derivations; the other technique is a non-myopic minimization strategy that can more effectively deal with a high number of targets, occlusions among them and the mechanical constraints of the camera.

- We adopt the sampling method in [48] to handle large Partially Observable MDP (POMDP) and modify it to further limit the computational cost.

- We improve the evaluation of the whole method with respect to previous works [38] and [39] in which ground truth data are used as objects measurements, and use standard metrics for multi-target tracking evaluation.

- We firstly show how to task a single PTZ camera according to a sophisticated sensor management strategy to support multiple object tracking in a 3D world coordinate frame [22], and demonstrate it working online in a real scenario.

## 4. MDP with Information Gain Reward

### 4.1. Baseline method for multi object tracking

Similarly to [22] and [38], our baseline multi object tracking uses Extended Kalman filter (EKF) for each initialized target. Pedestrian detection [51] is used to extract object observations, and the Hungarian algorithm [52] is applied to associate each observation to the corresponding EKF-filter and to initialize a new filter in the case of unassociated observations.

At time $t$, the real object state, $\mathbf{s}_t$, and its estimation, $\mathbf{x}_t$, include its location in world coordinates and its speed: $\mathbf{s}_t = [x^w_{s,t}, y^w_{s,t}, \dot{x}^w_{s,t}, \dot{y}^w_{s,t}]^\top$, $\mathbf{x}_t = [x^w_t, y^w_t, \dot{x}^w_t, \dot{y}^w_t]^\top$. The observation $\mathbf{o}_t = [u_t, v_t]^\top$, *i.e.*, the target location on the image plane, only depends on the current state and on the action $\mathbf{a}_t$, that is selected from the finite set $\mathcal{A}$ which comprises $L$ different possible actions each of which corresponds to a particular PTZ camera pose $\mathbf{a} = (\phi, \theta, \mathrm{f}) \in \mathcal{A}$ (the pan $\phi$, tilt $\theta$ and focal length f respectively). Formally, we have:

$$\begin{aligned} \mathbf{s}_t &= f(\mathbf{s}_{t-1}) + m_t, \quad m_t \sim \mathcal{N}(0, \mathbf{U}), \\ \mathbf{o}_t &= g(\mathbf{s}_t, \mathbf{a}_t) + n_t, \quad n_t \sim \mathcal{N}(0, \mathbf{V}), \end{aligned} \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ are the motion model and the observation model, respectively, $m_t$ and $n_t$ are the process and the measurement noise with $\mathbf{U}$ and $\mathbf{V}$ their respective covariance matrices. In particular, the function $g(\cdot)$ represents the homography from the world plane to the image plane parameterized by the actions defined in the set $\mathbf{a}_t$.

Let $\mathbf{x}^-_t$ be the predicted state estimate at time $t$, i.e. before having made the observation at $t$, while $\mathbf{x}^+_t$ incorporates the observation. The final estimate for the state at time $t$, $\mathbf{x}_t$, is either $\mathbf{x}^+_t$ or $\mathbf{x}^-_t$, depending whether the target is observed or not (*e.g.*, when the camera is not pointing at it, or the detector misses it). $\mathbf{P}^-_t$,$\mathbf{P}^+_t$ and $\mathbf{P}_t$ are the covariance matrices for $\mathbf{x}^-_t$,$\mathbf{x}^+_t$ and $\mathbf{x}_t$, respectively. If the target is not observed, only $\mathbf{x}^-_t$ and $\mathbf{P}^-_t$ are considered. The EKF equations are then:

$$\begin{aligned} \mathbf{x}^-_t &= \mathbf{F}\mathbf{x}_{t-1}, \\ \mathbf{P}^-_t &= \mathbf{F}^\top \mathbf{P}_{t-1}\mathbf{F} + \mathbf{U}, \\ \mathbf{K}_t &= \mathbf{P}^-_t \mathbf{C}_\mathbf{x}(\mathbf{a}_t)(\mathbf{C}^\top_\mathbf{x}(\mathbf{a}_t)\mathbf{P}^-_t \mathbf{C}_\mathbf{x}(\mathbf{a}_t) + \mathbf{V})^{-1}, \\ \mathbf{x}^+_t &= \mathbf{x}^-_t + \mathbf{K}_t(\mathbf{o}_t - g(\mathbf{x}^-_t, \mathbf{a}_t)), \\ \mathbf{P}^+_t &= (\mathbf{I} - \mathbf{K}_t \mathbf{C}^\top_\mathbf{x}(\mathbf{a}_t))\mathbf{P}^-_t, \end{aligned} \tag{2}$$

where $\mathbf{C}_\mathbf{x}(\mathbf{a}_t) = \nabla_\mathbf{x} g(\mathbf{x}, \mathbf{a}_t)|_{\mathbf{x}=\mathbf{x}^-_t}$ is the linearized homography $g$ evaluated in $\mathbf{x}^-_t$ and $\mathbf{F}$ is the 4×4 matrix that models the system dynamics. Importantly, $\mathbf{C}_\mathbf{x}(\mathbf{a}_t)$ depends on the action, so that diverse camera poses lead to different observation matrices, and different estimations for $\mathbf{x}^+_t$ and $\mathbf{P}^+_t$. It is worth to highlight that also the zoom modifies the linearized projection matrix $\mathbf{C}_\mathbf{x}(\mathbf{a}_t)$; in fact observing a target with an higher magnification will produce a smaller covariance $\mathbf{P}^+_t$ [32].

Eqs. 2 can be seen as modeling the transition probabilities in the MDP (see Fig. 3). To complete the MDP model, we need the reward function $R(\mathbf{x}^-_t, \mathbf{a}_t)$, which tells how informative is a given action $\mathbf{a}_t$ performed in the state $\mathbf{x}^-_t$. Notably, the reward must depend on $\mathbf{x}^-_t$ (not on $\mathbf{x}^+_t$), since we want to select the action *before* performing the observation. Given the reward function, at each time step we can evaluate its value for all the possible actions $\mathbf{a}_t \in \mathcal{A}$, choosing the one which gives the maximal reward.

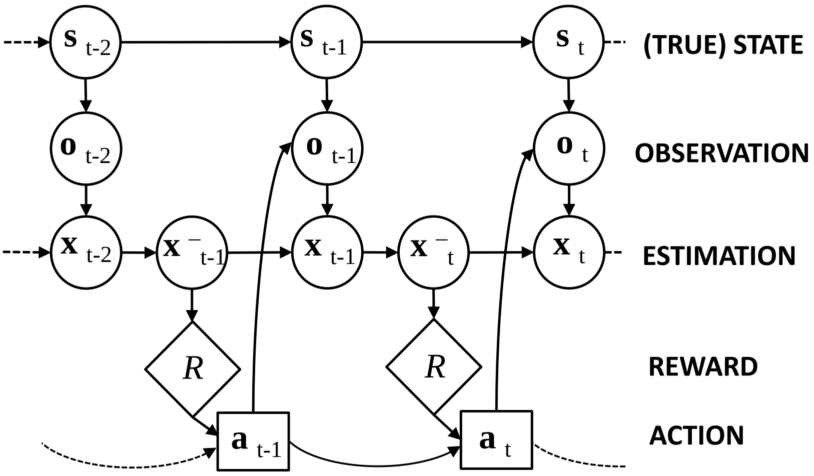

Figure 3: Graphical representation of our approach.

### *4.2. Information gain formulation*

In designing the reward function $R(\mathbf{x}_t^-, \mathbf{a}_t)$ we directly relate it to the expected information gain $I(\mathbf{x}_t; \mathbf{o}_t|\mathbf{a}_t)$ between the state $\mathbf{x}_t$ and the observation $\mathbf{o}_t$, for a given action. In practice, it expresses the amount of information shared between state and observation. Adopting the same formulation of [53], we can write:

$$\begin{aligned} \mathbf{a}_t^\star &= \arg\max_{\mathbf{a}_t} R(\mathbf{x}_t^-; \mathbf{a}_t) = \arg\max_{\mathbf{a}_t} I(\mathbf{x}_t; \mathbf{o}_t|\mathbf{a}_t) = \\ &= \arg\max_{\mathbf{a}_t} H(\mathbf{x}_t^-) - H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t) = \arg\min_{\mathbf{a}_t} H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t), \end{aligned} \tag{3}$$

where $H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t)$ is the conditional entropy[5]. Thus, we want to minimize:

$$\begin{aligned} &H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t) = \\ &= -\int p(\mathbf{o}_t|\mathbf{a}_t) \int p(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t) \log\left(p(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t)\right) d\mathbf{x}_t d\mathbf{o}_t = \\ &= \int_{\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) d\mathbf{o}_t H(\mathbf{x}_t^+) + \int_{\neg\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) d\mathbf{o}_t H(\mathbf{x}_t^-) = \\ &= \alpha_t(\mathbf{a}_t) H(\mathbf{x}_t^+) + (1 - \alpha_t(\mathbf{a}_t)) H(\mathbf{x}_t^-), \end{aligned} \tag{4}$$

where we split the domain of integration for $p(\mathbf{o}_t|\mathbf{a}_t)$. $\Omega_t$ is the set of points in which the target is visible, $\neg\Omega_t$ is the set where it is not visible, i.e., it is out of the camera field of view (FoV), is occluded, or is too small to be detected. Assuming the distribution for $\mathbf{x}_t$ as Gaussian and being the system in Eqs. 2 linear, we can derive the entropy $H(\mathbf{x}_t^+)$ directly from the EKF equations. In fact, the entropy of a Gaussian distribution only depends on its covariance[6] and Eqs. 2 provide $\mathbf{P}_t^+$ if $\mathbf{a}_t$ allows to get the observation for the target, and $\mathbf{P}_t^-$ otherwise. For more details, see [53].

[5] The conditional entropy for two random variables $x$ and $y$ is defined as $H(x|y) = -\int\int p(x, y) \log p(x|y) dx dy$.

[6] The entropy of a Gaussian distributed random vector $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ is: $H(\mathbf{x}) = \frac{n}{2} + \frac{1}{2}\log((2\pi)^n||\mathbf{\Sigma}||)$.

In other words, to ensure maximal expected information gain $I(\mathbf{x}_t; \mathbf{o}_t|\mathbf{a}_t)$ we need only to consider how the term $\alpha(\mathbf{a}_t)$ varies for different actions $\mathbf{a}_t$. Intuitively, such term estimates the probability that at the next step a target will be observed by the camera, as a function of the pose of the camera itself. Extending to $K$ independent targets correspond to sum up the information gains $I_k$ for each target $k$.

## 5. Modeling real world scenarios

As analyzed in the previous section, the formulation of $\alpha(\mathbf{a}_t)$ in [53] is limited, since it neglects aspects of real world scenarios. We model (a) the visibility constraint, accounting for the physical dimension of the target in the current field of view; (b) a realistic person detector whose performance varies according to the occlusion ratio and the imaged object size; (c) the occlusions between the targets that considers the relative positions between the imaged objects (evaluated through sampling); and (d) the mechanical speed limits of camera motion. The variability of the number of targets is managed through the patrolling term as in [38].

### *5.1. Modeling visibility and detection factors*

Introducing the visibility constraint requires to define properly the set $\Omega_t$ in Eq. 4, while introducing the estimation of the detector performance implies to modify $p(\mathbf{o}_t|\mathbf{a}_t)$. Let $\mathbf{d}_t$ be a binary variable which is 1 if the target is found by the detector and 0 otherwise. In practice, $\mathbf{d}_t$ tells us whether the Kalman filter will be updated with a new observation or only the information from the previous prediction will be considered. Hence, Eq. 4 can be modified by considering this new variable:

$$\begin{aligned} H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{d}_t, \mathbf{a}_t) = -\int\int p(\mathbf{o}_t, \mathbf{d}_t|\mathbf{a}_t) \\ \int p(\mathbf{x}_t|\mathbf{o}_t, \mathbf{d}_t, \mathbf{a}_t) \log\left(p(\mathbf{x}_t|\mathbf{o}_t, \mathbf{d}_t, \mathbf{a}_t)\right) d\mathbf{x}_t d\mathbf{o}_t d\mathbf{d}_t. \end{aligned} \tag{5}$$

Let us start by analyzing $p(\mathbf{o}_t, \mathbf{d}_t|\mathbf{a}_t)$ and introducing some assumptions. First, $p(\mathbf{o}_t|\mathbf{a}_t) = p(\mathbf{o}_t^-|\mathbf{a}_t)$ (where $\mathbf{o}_t^- = g(\mathbf{x}_t^-, \mathbf{a}_t)$ ) since the actual observation $\mathbf{o}_t$ is yet not available when selecting the actual action $\mathbf{a}_t$[7]. In this way, we assume that the expected positions of the targets on the image plane only depend on the prediction of the state and the action. Second, we assume that the visibility of a target only depends on its position on the image plane, being unaware of obstacles or other occluders in the scene. Therefore, the term $p(\mathbf{o}_t, \mathbf{d}_t|\mathbf{a}_t)$ in Eq. 5 factorizes as:

$$p(\mathbf{o}_t, \mathbf{d}_t|\mathbf{a}_t) = p(\mathbf{o}_t|\mathbf{a}_t)p(\mathbf{d}_t|\mathbf{o}_t, \mathbf{a}_t). \tag{6}$$

Being $\mathbf{d}_t$ binary, Eq. 5 may be rearranged as:

$$\begin{aligned} H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{d}_t, \mathbf{a}_t) = &\int_{\neg\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) d\mathbf{o}_t H(\mathbf{x}_t^-) + \\ &+ \int_{\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) p(\mathbf{d}_t{=}0|\mathbf{o}_t, \mathbf{a}_t) d\mathbf{o}_t H(\mathbf{x}_t^-) + \\ &+ \int_{\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) p(\mathbf{d}_t{=}1|\mathbf{o}_t, \mathbf{a}_t) d\mathbf{o}_t H(\mathbf{x}_t^+) = \\ =&(1-\alpha(\mathbf{a}_t))H(\mathbf{x}_t^-) + \alpha(\mathbf{a}_t)H(\mathbf{x}_t^+), \end{aligned} \tag{7}$$

[7] In the remaining, for the sake of clarity, we omit the apex $^-$ from $\mathbf{o}_t^-$, if not otherwise specified.

Table 2: Comparing the miss rate as a function of target size between HOG pedestrian detector and the parametric estimation proposed in this paper. See Fig. 4 for the corresponding plots.

| target height [pixel] | 32 | 45 | 64 | 91 | 128 |
|---|---|---|---|---|---|
| HOG (dashed orange) | 0.928 | 0.8272 | 0.696 | 0.473 | 0.350 |
| estimation (solid black ) | 0.900 | 0.792 | 0.657 | 0.503 | 0.350 |

where we also suppose that a detection is possible only if the observation is visible in the image. In conclusion, we just need to compute for any possible action $\mathbf{a}_t$ the weight $\alpha(\mathbf{a}_t)$:

$$\alpha(\mathbf{a}_t) = \int_{\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t)p(\mathbf{d}_t{=}1|\mathbf{o}_t, \mathbf{a}_t)d\mathbf{o}_t. \tag{8}$$

Now, to preserve the Gaussian distribution and therefore the efficient integration for the weight $\alpha(\mathbf{a}_t)$, the two pdfs in Eq. 8 and the integration domain $\Omega$ are defined as follows.

*Observation Distribution.* $p(\mathbf{o}_t|\mathbf{a}_t)$ is the predicted distribution of the observation. Based on the prediction of the state from Eqs. 2, we have $\mathbf{o}_t \sim \mathcal{N}(\mathbf{o}_t^-, \Sigma_{\mathbf{o}_t})$, where:

$$\mathbf{o}_t^- = C_\mathbf{x}(\mathbf{a}_t)\mathbf{x}_t^-, \quad \Sigma_{\mathbf{o}_t} = C_\mathbf{x}(\mathbf{a}_t)P^- C_\mathbf{x}^\top(\mathbf{a}_t) + \mathbf{V}. \tag{9}$$

*Visibility Probability.* $\Omega_t$ is the set of possible observations $\{\mathbf{o}_t\}$ for which the target is fully visible in the camera field of view, considering the limited size of the image plane $\mathcal{S} \subset \mathbb{R}^2$. In defining such set, we originally extend the work in [39], and consider the spatial dimension of the targets, assuming that objects are almost vertical on the ground plane and that their projected height is known for at least one target. Since we know the extrinsic calibration parameter for the camera, we can estimate the head position $\mathbf{e}_t(\mathbf{o}_t)$ on the image plane for a target whose feet are in $\mathbf{o}_t$, through the homology $\mathtt{W}_{\mathbf{a}_\mathrm{t}}$, as in [54]. The set $\Omega_t$ is then defined as:

$$\mathbf{o}_t \in \Omega_t \quad \Leftrightarrow \quad \mathbf{o}_t \in \mathcal{S} \wedge \mathbf{e}_t(\mathbf{o}_t) \in \mathcal{S}. \tag{10}$$

To integrate $p(\mathbf{o}_t|\mathbf{a}_t)$ on the set of points defined above we linearize the homology through the Jacobian $\mathtt{J}_{\mathbf{a}_t} = \nabla_{\mathbf{o}_t}\mathtt{W}_{\mathbf{a}_t}|_{\mathbf{o}_t=\mathbf{o}_t^-}$ of $\mathtt{W}_{\mathbf{a}_t}$ around $\mathbf{o}_t^-$. Therefore:

$$\mathbf{e}_t \approx \overline{\mathbf{e}}_t + \mathtt{J}_{\mathbf{a}_t}(\mathbf{o}_t - \mathbf{o}_t^-), \quad \overline{\mathbf{e}}_t = \mathtt{W}_{\mathbf{a}_\mathrm{t}}(\mathbf{o}_t^-). \tag{11}$$

Assuming that people are vertical in the scene, and that the image plane $y$-axis is vertical, we can discard the horizontal component getting:

$$y_t^\mathbf{e} = y_t^{\overline{\mathbf{e}}} + \mathtt{J}_{\mathbf{a}_{t2,2}}(y_t^\mathbf{o} - y_t^{\mathbf{o}^-}), \qquad x_t^\mathbf{e} = x_t^\mathbf{o}. \tag{12}$$

In conclusion, the $y$ coordinate for the head $\mathbf{e}_t$ is linearly obtained from the $y$ coordinate of $\mathbf{o}_t$, thus the integration on the image plane is still equivalent to integrating over a rectangle whose sides are parallel to the $x$-$y$ axis.

*Detection probability.* $p(\mathbf{d}_t{=}1|\mathbf{o}_t, \mathbf{a}_t)$ is the probability that a target will actually be detected given its position in the image plane. In practice, we consider that the performance of any pedestrian detectors depends on the height $\mathbf{r}_t$ of the target on the image plane (in pixels). We estimate such

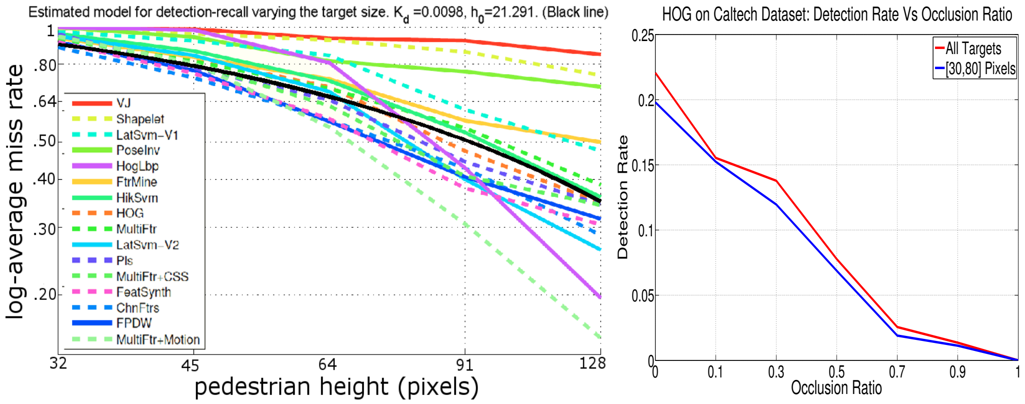

Figure 4: *Left*: Black curve is the function that we used to model the pedestrian detection recall as the target size varies (original plot from [55]). *Right*: HOG pedestrian detector performance for targets with different occlusion ratios. The performance are obtained from the Caltech Pedestrian Dataset using the HOG pedestrian detector implemented in OpenCV. The plot shows that the dependency of the pedestrian detector performance, as a function of the occluded area, can be approximated as linear.

a relation with the function $p(\mathbf{d}_t = 1) = 1 - e^{-K_d(\mathbf{r}-\mathbf{r}_0)}\mathbf{1}(\mathbf{r} - \mathbf{r}_0)$. The two parameters $K_d = 0.0098$ and $\mathbf{r}_0 = 21.29$ are extrapolated from the performance of HOG pedestrian detector on the Caltech Pedestrian Dataset, reported in [55][8]. More details are given in Tab. 2 where we report the miss rate values as a function of target size for the HOG pedestrian detector and compare them with the parametric estimation used. Fig. 4 (*left*) shows the miss rate values for all the other methods reported in [55].

The target height $\mathbf{r}_t = |y_t^{\mathbf{e}} - y_t^{\mathbf{o}}| = (y_t^{\mathbf{o}} - y_t^{\mathbf{e}})$ can be computed as a function of the observation $y_t^{\mathbf{o}}$ and the camera position $\mathbf{a}_t$, exploiting the homology:

$$\begin{aligned}\mathbf{r}_t = &\begin{bmatrix} 0 & 1 \end{bmatrix}\begin{bmatrix} \mathbf{o}_t - \mathbf{e}_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix}\left[(I_{2\times2} - \mathtt{J}_{\mathbf{a}_t})\mathbf{o}_t - \mathtt{W}_{\mathbf{a}_t}(\mathbf{o}_t^-) - \mathtt{J}_{\mathbf{a}_t}\mathbf{o}_t^-\right] = \\ &= \begin{bmatrix} 0 & 1 \end{bmatrix}\left[\mathtt{T}_t\mathbf{o}_t\right] + \begin{bmatrix} 0 & 1 \end{bmatrix}\left[-\mathtt{W}_{\mathbf{a}_t}(\mathbf{o}_t^-) - \mathtt{J}_{\mathbf{a}_t}\mathbf{o}_t^-\right] = \mathtt{T}_t\mathbf{o}_t + \mathbf{t}_t.\end{aligned} \tag{13}$$

Linearizing the homology around the expected observation $\mathbf{o}_t^-$ give us the exponential function:

$$p(\mathbf{d}_t{=}1|\mathbf{o}_t, \mathbf{a}_t) = 1 - e^{-K_d(\mathtt{T}_t\mathbf{o}_t + \mathbf{t}_t - \mathbf{r}_0)}, \tag{14}$$

where the matrix $\mathtt{T}_t$ and $\mathbf{t}_t$ are constants depending on the linearized homology, see Eq. 12. The product of the Gaussian distribution $p(\mathbf{o}_t|\mathbf{a}_t)$, Eq. 9, and the exponential function in $\mathbf{o}_t$, Eq. 14, gives another Gaussian distribution:

[8] Since in our implementations we use the HOG pedestrian detector, we estimate the parameters $K_d$ and $\mathbf{r}_0$ for the performance of that detector. The same procedure can be applied to any other detector for which the miss-rate as a function of the target size is given.

$$
\begin{aligned}
&p(\mathbf{o}_t|\mathbf{a}_t)p(\mathbf{d}_t = 1|\mathbf{o}_t, \mathbf{a}_t) = \\
=&\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \left(1 - \exp\left(\mathbf{t}_t - \mathbf{r}_0\right) \exp\left(-K_d \mathrm{T}_t \mathbf{o}_t\right)\right) \exp\left(-\frac{1}{2}(\mathbf{o}_t - \mu)^\top \Sigma^{-1} (\mathbf{o}_t - \mu)\right) = \\
=&\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \left(\exp -\frac{1}{2}(\mathbf{o}_t - \mu)^\top \Sigma^{-1} (\mathbf{o}_t - \mu) + \right. \\
&\left. - \exp\left(\mathbf{t}_t - \mathbf{r}_0\right) \exp -\frac{1}{2}(\mathbf{o}_t^\top \Sigma^{-1} \Sigma \mathrm{T}_t K_d + \mathrm{T}_t^\top \Sigma \Sigma^{-1} \mathbf{o}_t K_d + (\mathbf{o}_t - \mu)^\top \Sigma^{-1} (\mathbf{o}_t - \mu))\right) = \\
=&\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{o}_t - \mu)^\top \Sigma^{-1} (\mathbf{o}_t - \mu)\right) + \\
-&\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(\mathbf{t}_t - \mathbf{r}_0 + \frac{K_d^2}{2} \mathrm{T}_t^\top \Sigma \mathrm{T}_t\right) \exp\left(-\frac{1}{2}((\mathbf{o}_t - (\mu - \Sigma \mathrm{T}_t K_d))^\top \Sigma^{-1} (\mathbf{o}_t - (\mu - \Sigma \mathrm{T}_t K_d))\right).
\end{aligned}
\tag{15}
$$

Thus the weight $\alpha(\mathbf{a}_t)$ in Eq. 8 can be numerically computed as bounded integration of a Gaussian distribution, and the boundary are modified to require that the minimum target height is $\mathbf{r}_0$.

At this point we have introduced two factors that increase the realism and completeness of the proposed model, while maintaining a low computational cost for the reward function. In the next sections the management of occlusions among targets will be introduced to further reduce the difference between the expected information gain and the real information gain obtained from the camera.

### 5.2. Occlusions Handling

Occlusions represent a serious problem for the selection of the action due to a wrong estimation of the information gain for a target. In fact, being occlusions not modeled in the above formulation, even an occluded target would bring a contribution to the expected information gain, which will not correspond to real information gain obtained after the action is performed. As analyzed in [55] the larger the occluded area for a target the more probable that the detection algorithm will fail. Without any information on possible occluding obstacles in the field of view, we can only keep into account inter-occlusions among targets. To this aim, we introduce a term that estimates the ratio of area of a person occluded in the frame, resembling the depth-sorting method of [56]. In practice, we build a binary occlusion mask which indicates the occluded pixels for each target. From now on we slightly modify the notation, introducing an index for each target $k \in \mathcal{K}$ (with $|\mathcal{K}| = K$), since we will have to consider also the dependencies among two or more targets.

Formally, let $\mathbf{c}_t^k \in [0, 1]$ be the ratio of the bounding box of the target which is visible at time $t$, we can estimate the relation between the probability of detecting the target and its associated $\mathbf{c}_t$ by injecting this variable in Eq. 5:

$$
\begin{aligned}
&H(\mathbf{x}_t^k|\mathbf{o}_t^k, \mathbf{d}_t^k, \mathbf{c}_t^k, \mathbf{a}_t) = \int_{\neg\Omega_t} p(\mathbf{o}_t^k|\mathbf{a}_t) d\mathbf{o}_t H(\mathbf{x}_t^{-,k}) + \\
&+ \int_{\Omega_t} p(\mathbf{o}_t^k|\mathbf{a}_t) \int p(\mathbf{d}_t^k{=}0|\mathbf{o}_t^k, \mathbf{a}_t, \mathbf{c}_t^k) p(\mathbf{c}_t^k|\mathbf{o}_t^k, \mathbf{a}_t) d\mathbf{c}_t^k d\mathbf{o}_t^k H(\mathbf{x}_t^-) + \\
&+ \int_{\Omega_t} p(\mathbf{o}_t^k|\mathbf{a}_t) \int p(\mathbf{d}_t^k{=}1|\mathbf{o}_t^k, \mathbf{a}_t, \mathbf{c}_t^k) p(\mathbf{c}_t^k|\mathbf{o}_t^k, \mathbf{a}_t) d\mathbf{c}_t^k d\mathbf{o}_t^k H(\mathbf{x}_t^+) = \\
&= (1 - \alpha(\mathbf{a}_t)) H(\mathbf{x}_t^{-,k}) + \alpha(\mathbf{a}_t) H(\mathbf{x}_t^{+,k}).
\end{aligned}
\tag{16}
$$

As for the previous case, we just need to compute for any possible action $\mathbf{a}_t$ a modified version of the weight $\alpha(\mathbf{a}_t)$:

$$\alpha^k(\mathbf{a}_t){=}\int_{\Omega_t} p(\mathbf{o}_t^k|\mathbf{a}_t)\int p(\mathbf{d}_t^k{=}1|\mathbf{c}_t^k, \mathbf{o}_t^k, \mathbf{a}_t)p(\mathbf{c}_t^k|\mathbf{o}_t^k, \mathbf{a}_t)d\mathbf{c}_t^k d\mathbf{o}_t^k, \tag{17}$$

which requires to define $p(\mathbf{d}_t^k{=}1|\mathbf{c}_t^k, \mathbf{o}_t^k, \mathbf{a}_t)$ and $p(\mathbf{c}_t^k|\mathbf{o}_t^k, \mathbf{a}_t)$.

*Detection Probability with Occlusion Term.* We assume that the effect of the occlusion ratio and the target size on the detection performance are independent. This leads to the following factorization: $p(\mathbf{d}_t^k|\mathbf{c}_t^k, \mathbf{o}_t^k, \mathbf{a}_t) = p(\mathbf{d}_t^k|\mathbf{o}_t^k, \mathbf{a}_t)p(\mathbf{d}_t^k|\mathbf{c}_t^k)$, where the first factor has been computed in Sec. 5.1. To estimate $p(\mathbf{d}_t^k|\mathbf{c}_t^k)$, i.e., the effect of the occlusion on the detection performance, we use again the Caltech Pedestrian Dataset [55], obtaining the plots shown in Fig. 4(*right*). We choose to approximate this relation as linear: $p(\mathbf{d}_t^k{=}1|\mathbf{c}_t^k) = \mathbf{c}_t^k$.

*Computing Occlusion Ratio for each Target.* $p(\mathbf{c}_t^k|\mathbf{o}_t^k, \mathbf{a}_t)$ estimates the distribution of the occlusion ratio, given the observation for the target $k$ and the camera position. This term also depends on the position of the other targets in the scene (collectively indexed by $^{\neg k}$), so we need to expand it as:

$$p(\mathbf{c}_t^k|\mathbf{o}_t^k, \mathbf{a}_t) = \int p(\mathbf{c}_t^k|\mathbf{o}_t^{\mathcal{K}}, \mathbf{a}_t)p(\mathbf{o}_t^{\mathcal{K}}|\mathbf{a}_t)d\mathbf{o}^{\neg k}. \tag{18}$$

The term $p(\mathbf{c}_t^k|\mathbf{o}^{\mathcal{K}}, \mathbf{a}_t)$ expresses the visibility probability given by the ratio of visible versus occluded pixels:

$$p(\mathbf{c}_t^k|\mathbf{o}^{\mathcal{K}}, \mathbf{a}_t) = \delta(\mathbf{c}_t^k - \overline{\mathbf{c}}^k), \qquad \overline{\mathbf{c}}_t^k = \frac{\int \delta(\mathbf{x}_t^k \underset{u}{<} \mathbf{x}_t^{\neg k}|\mathbf{a}_t)du}{\int \delta(\mathbf{x}_t^k|\mathbf{a}_t)du}, \tag{19}$$

where $\delta(\mathbf{x}_t^k \underset{u}{<} \mathbf{x}_t^{\neg k}|\mathbf{a}_t)$ is a binary mask that takes value 1 if at pixel $u$ a part of target $k$ is observed, and 0 otherwise. The other term $\int \delta(\mathbf{x}_t^k|\mathbf{a}_t)$ measures the whole target area.

The main limitation of this formulation is that it is not possible anymore to compute the information gain for each target independently, since the relative position among targets is considered when estimating occlusion, and it is also not possible to compute the $p(\mathbf{c}_t^k)$ in closed form.

Therefore, at each possible camera pose we apply a Monte Carlo approach sampling from $p(\mathbf{x}_t^{-,1}, \ldots, \mathbf{x}_t^{-,k}, \ldots, \mathbf{x}_t^{-,K}) = \prod_{k=1}^K p(\mathbf{x}_t^{-,k})$, $M$ sets of positions $\left\{\tilde{\mathbf{x}}_{t,j}^{-,1}, \ldots, \tilde{\mathbf{x}}_{t,j}^{-,k}, \ldots, \tilde{\mathbf{x}}_{t,j}^{-,K}\right\}_{j=1\ldots M}$ for all the targets. Then, for a candidate action $\mathbf{a}_t$, the corresponding weight $\alpha^k(\mathbf{a}_t)$ is estimated from the related sets of observation predictions, $\left\{\tilde{\mathbf{o}}_{t,j}^1, \ldots, \tilde{\mathbf{o}}_{t,j}^k, \ldots, \tilde{\mathbf{o}}_{t,j}^K\right\}_{j=1\ldots M}$, computed according to the model of Eq. 2. Each of this set $j$ is used to evaluate the inner integral in Eq. 17:

$$\tilde{\mathbf{d}}_{t,j}^k = \int p(\mathbf{d}_t^k{=}1|\mathbf{c}_t^k, \tilde{\mathbf{o}}_{t,j}^k, \mathbf{a}_t)p(\mathbf{c}_t^k|\tilde{\mathbf{o}}_{t,j}^k, \mathbf{a}_t)d\mathbf{c}_t^k, \tag{20}$$

providing the detection probability of the target $k$ in the sample $j$. The final $\alpha^k(\mathbf{a}_t)$ for the target $k$ is therefore computed replacing the integral in Eq. 17 with a summation over the samples:

$$\alpha^k(\mathbf{a}_t) = \frac{1}{M}\sum_{j=1}^{M} \tilde{\mathbf{d}}_{t,j}^k. \tag{21}$$

The conditional entropy for that target is then computed according to Eq. 7. The sum of the contribution of each target provides the information gain for all the targets.

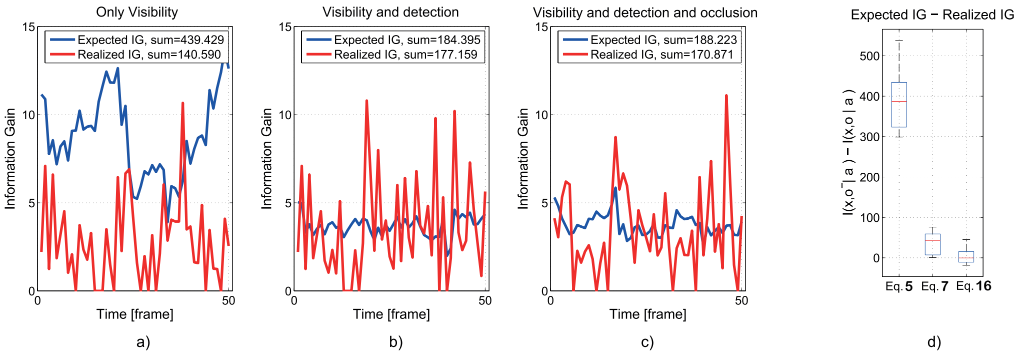

Figure 5: The comparison between the expected information gain for the most promising action, $I(\mathbf{x}_t, \mathbf{o}_t^-|\mathbf{a}_t^\star)$, used as reward function, Eq. 3, and the actual information gain that is obtained with the observation from the selected action $I(\mathbf{x}_t, \mathbf{o}_t|\mathbf{a}_t^\star)$. This test is performed on the synthetic experiments, see Sec. 9, varying the complexity of the reward function. *a)* comparison between the expected and realized information gain using the reward function that only keeps into account the visibility criterion, Eq. 5; *b)* same as before but obtained by keeping into account the performance from the detector, Eq. 7; *c)* keeps into account also the occlusions among particles, Eq. 16, with $M = 100$. *d)* shows the differences for the 3 case on a statistic of 12 runs of 50 frames each. Note that the difference decreases as the model becomes more accurate and is close to 0 for the last formulation of case c).

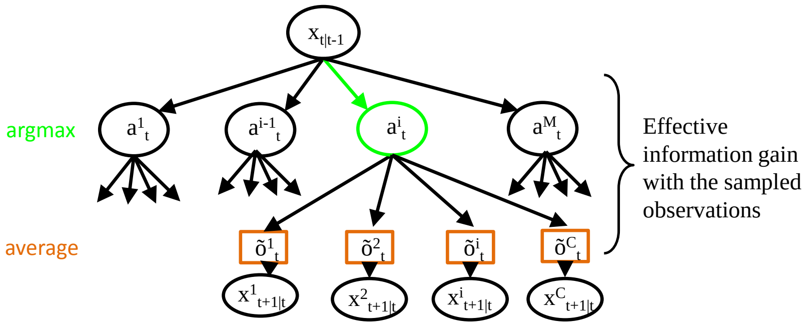

Figure 6: Graph of the algorithm for computing the expected information gain with the occlusions estimation, Eq. 16.

### 5.3. Modeling the camera mechanics: action set reduction

We want to model the mechanical constraints that define the set of positions reachable from the current pose, in a given time interval, of a real PTZ camera. Given the set of all the possible camera actions $\mathcal{A}$ and the previous action $\mathbf{a}_{t-1} = (\phi_{t-1}, \theta_{t-1}, \mathrm{f}_{t-1})$, an action $(\phi, \theta, \mathrm{f}) \in \mathcal{A}$ also belongs to the set of actions $\mathcal{A}_t$, reachable at the next time $t$, if:

$$|\phi - \phi_{t-1}| \leq \Delta\phi \wedge |\theta - \theta_{t-1}| \leq \Delta\theta \wedge |\mathrm{f} - \mathrm{f}_{t-1}| \leq \Delta\mathrm{f}, \tag{22}$$

where $\Delta\phi$ and $\Delta\theta$ are the maximum displacement allowed in the unit of time for the pan and tilt angles and $\Delta\mathrm{f}$ is the maximum variation in the zoom, that can be easily obtained by combining the expected system frame rate and the camera specifications.

### 5.4. Patrolling term for new target detection

To take into account for new targets occurring in the scene, the PTZ has to randomly patrol, looking for new evidence. To model this factor we get inspiration from [38], where an additional term $I_p(\mathbf{b}_t|\mathbf{a}_t)$ related to the patrolling around the scene is defined. Such factor estimates the information gain that could be obtained performing an action $\mathbf{a}_t$ due to the detection of a new target $\mathbf{b}_t$.

When combining the information gain on target position uncertainty with the patrolling term we obtain:

$$I_t(\mathbf{a}_t) = \sum_{k}^{N} I(\mathbf{x}_t^k; \mathbf{o}_t^k|\mathbf{a}_t) + \beta I_p(\mathbf{b}_t|\mathbf{a}_t), \tag{23}$$

where $\beta$ is the weight that mixes the two quantities.

With this last element we complete the definition of the MDP process formed by the EKF equations plus the reward function. In particular, Eq. 8 and Eq. 17 characterize the two proposed versions, the first more efficient and the second one more complex, which also takes into account the occlusions. Alg. 1 shows the pseudo-code for the version with the occlusions handling.

**Algorithm 1:** Algorithm for the myopic approach

```
Input: number of samples $M$, action set $\mathcal{A}_t$, generative-model G (Eq. 1), state $\mathbf{x}_{t-1}$
Output: selected action $\mathbf{a}^\star$
for each target do
    generate $M$ samples $\{\hat{\mathbf{x}}_t\}$ from G
end
for each $\mathbf{a}$ in $\mathcal{A}_t$ do
    for each target do
        get observation set from samples $\tilde{\mathbf{o}}^k = g(\tilde{\mathbf{x}}_t^k, \mathbf{a})$
        compute visibility term $\alpha^k(\mathbf{a}_t)$, Eq. 17
        compute the conditional entropy gain, Eq. 7
    end
    Get whole information gain, $I_t(\mathbf{a})$, Eq. 23
end
Return $\mathbf{a}^\star = \arg\max_{\mathbf{a}\in\mathcal{A}_t}\{I_t(\mathbf{a})\}$
```

## 6. Non-myopic approach

The solution proposed so far is myopic, i.e. the action to be performed at the next step is selected only considering the current system state and the prediction for the next time step. Better results could be achieved if we design a non-myopic approach, where the reward function to be maximized considers more than one step in the future. A non-myopic approach would outperform the myopic one when there are terms which are time-variant: a visibility map on the scene (trees, houses or other occlusion that could prevent the tracker from working properly), occlusions between targets that move close to each others, a target that is leaving the field of view or a target that is going far away from the camera where it will be no longer visible. Indeed, these are all examples of a realistic scenario, that we want to take into account. On the other hand, reasoning on a longer temporal horizon requires a precise modeling of the target future behavior in order to produce a reliable prediction of the targets trajectories.

### 6.1. Look-ahead algorithm

To solve the non-myopic approach we use a sampling strategy inspired by [48], that allows approximate computation on a MDP with very large or infinite dimensionality of the state space.

**Algorithm 2:** Algorithm for the non-myopic approach

```
Function: computeQ($\hbar$,$M$,$\mathcal{A}$,$\gamma$,$G$,$\mathbf{x}_{t-1}$)
Input: horizon $\hbar$, number of samples $M$, action set $\mathcal{A}$, discount-factor $\gamma$, generative-model G
(Eq. 1), state $\mathbf{x}_{t-1}$
Output: rewards $(\hat{Q}^{\star}_{\hbar}(\mathbf{x}_t, \mathbf{a}_1), \ldots, \hat{Q}^{\star}_{\hbar}(\mathbf{x}_t, \mathbf{a}_K))$
if $\hbar = 0$ then
    Return $(0, \ldots, 0)$
else
    For each target generate $M$ samples $\{\hat{\mathbf{x}}_t\}$ from G
    For each $\mathbf{a}$ in $\mathcal{A}$ get $z = f(\hat{\mathbf{x}}_t, \mathbf{a})$
    Compute $\hat{Q}^{\star}_{\hbar}(\hat{\mathbf{x}}_t, \mathbf{a}) = I(\hat{\mathbf{x}}_t, z|\mathbf{a})$
    Take the $r_p L$ actions with highest $\hat{Q}^{\star}_{\hbar}(\hat{\mathbf{x}}_t, \mathbf{a})$, for them get
    $\hat{Q}^{\star}_{\hbar}(\hat{\mathbf{x}}_t, \mathbf{a}) = I(\hat{\mathbf{x}}_t, z; \mathbf{a}) + \gamma \frac{1}{M} \sum_{\hat{\mathbf{x}}_t}$ computeV($\hbar - 1, M, \mathcal{A}, \gamma, G, \hat{\mathbf{x}}_t$)
    Return $(\hat{Q}^{\star}_{\hbar}(\hat{\mathbf{x}}_t, \mathbf{a}_1), \ldots, \hat{Q}^{\star}_{\hbar}(\hat{\mathbf{x}}_t, \mathbf{a}_K))$
end
Function: computeV($\hbar$,$M$,$\mathcal{A}$,$\gamma$,$G$,$\mathbf{x}_t$)
Input: horizon $\hbar$, number of samples $M$, action set $\mathcal{A}$, discount-factor $\gamma$, generative-model G
(Eq. 1), state $\mathbf{x}_t$
Output: value $V(\mathbf{x}_t, \hbar)$
Let
$(\hat{Q}^{\star}_{\hbar}(\mathbf{x}_t, \mathbf{a}_1), \ldots, \hat{Q}^{\star}_{\hbar}(\mathbf{x}_t, \mathbf{a}_K))$ =computeQ ($\hbar$,M,$\mathcal{A}$,$\gamma$,G,$\mathbf{x}_t$)
Return $\max_{\mathbf{a} \in \{\mathbf{a}_1, \ldots, \mathbf{a}_K\}} \{\hat{Q}^{\star}_{\hbar}(\mathbf{x}_t, \mathbf{a})\}$
Function: SelectAction($\hbar$,$M$,$\mathcal{A}$,$\gamma$,$G$,$\mathbf{x}_{t-1}$)
Input: horizon $\hbar$, number of samples $M$, action set $\mathcal{A}$, discount-factor $\gamma$, generative-model G
(Eq. 1), state $\mathbf{x}_{t-1}$
Output: action $\mathbf{a}^{\star}$
Let $(\hat{Q}^{\star}_{\hbar}(\mathbf{x}_t, \mathbf{a}_1), \ldots, \hat{Q}^{\star t}_{\hbar}(\mathbf{x}_t, \mathbf{a}_K))$ =computeQ ($\hbar$,$M$,$\mathcal{A}$,$\gamma$,G,$\mathbf{x}_{t-1}$)
Return $\mathbf{a}^{\star} = \arg\max_{\mathbf{a} \in \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}} \{\hat{Q}^{\star}_{\hbar}(\mathbf{x}_t, \mathbf{a})\}$
```

The main idea is to estimate the future dynamic of the model by sampling the future observations $z_t$ from the $p(\mathbf{o}_t|\mathbf{x}_t^-, \mathbf{a})$. The sampled observation $z_t$ are used to update the state $\hat{\mathbf{x}}_t$ according to the $p(\mathbf{x}_t|\mathbf{o}_t)$. The prediction and sampling iteration are repeated iteratively in future steps and at each step the information gain is computed. The global information gain is computed as a discounted sum of the current and future steps.

Let $L$ be the maximum number of possible actions that are reachable from any action $\mathbf{a}_t$ and $M$ the number of observations to sample at each time for each action. Exploiting the notation of [48], we define the selection of the best action at time $t$ over the finite horizon $\hbar$ as:

$$\mathbf{a}_t^\star = \arg\max_{\mathbf{a}_t} Q(\mathbf{x}_t, \mathbf{a}_t, \hbar), \tag{24}$$

with:

$$\begin{aligned}
Q(\mathbf{a}_t, \mathbf{x}_{t-1}, \hbar) &= \sum_{z_t} (I(\mathbf{x}_t; z_t|\mathbf{a}_t) + \gamma V(\hat{\mathbf{x}}_t, \hbar - 1)),\\
V(\hat{\mathbf{x}}_{t+1}, \hbar) &= \max_{\mathbf{a}_{t+1}} Q(\mathbf{a}_{t+1}, \hat{\mathbf{x}}_{t+1}, \hbar),\\
Q(\mathbf{a}_t, \mathbf{x}_{t-1}, 1) &= I(\mathbf{x}_t; \mathbf{o}_t|\mathbf{a}),
\end{aligned} \tag{25}$$

where $\gamma$ is the discount factor parameter. Such parameter balances the contribution of the information gain expected at the next step and the information gain expected on later steps. A complete recursive description of the algorithm is given in Alg. 2 while Fig. 7 gives a graphical representation of the procedure. At the final step, in the leaves of the tree we can either compute the information gain considering also the occlusion through the sampling procedure or compute the closed form reward, that discards the occlusion effects.

We add the pruning parameter $r_p$ that allows to reduce the size of the tree to be explored, discarding the least promising actions at the current step. The computational cost is $\mathcal{O}(L \cdot K)$ for the myopic case and $\mathcal{O}((K \cdot L)^{\hbar} \cdot (M \cdot r_p)^{\hbar - 1})$ for the non-myopic. The pruning factor $r_p$ is essential when considering a system with many possible actions, we set it to $r_p = 0.1$ in our experiments.

### *6.2. Summarizing samples for efficiency*

The non-myopic approach should be extremely effective in case of occlusions that are considered in the look-ahead procedure via sampling. The number of samples that can be used in our approach is really limited due to exponential growth of the tree in the number of samples. In fact, each new sampled observation generates a state for the tracking algorithm that must be propagated in the future. On the other hand, a small number of samples gives a very rough idea on the expected state of the targets, in particular whether it would be visible or not.

We would like to better predict the expected information gain, keeping a reduced size of the tree, i.e. keeping the same computational complexity as if $M = 1$. To achieve this we first sample $M$ observations for each action and compute the expected information gain, but then these data are merged, resulting in a single updated value for the filter state. This state $\bar{\mathbf{x}} = \frac{1}{M}\sum_{i=1}^{M} \tilde{\mathbf{x}}_j$ is the average of the states corresponding to each observation. In this way, we do not have to generate a subtree for each of the $M$ sample, but a single subtree for each action, as shown in Fig. 7. Its variance $\bar{\mathbf{P}}$ is computed as the weighted geometrical mean between matrix to preserve the same information gain, [58]:

$$\bar{\mathbf{P}} = (\mathbf{P}^-)^{\frac{1}{2}}((\mathbf{P}^-)^{-\frac{1}{2}}\mathbf{P}^+(\mathbf{P}^-)^{-\frac{1}{2}})^{\alpha}(\mathbf{P}^-)^{\frac{1}{2}}. \tag{26}$$

such definition ensures that if $\mathbf{P}^+$ and $\mathbf{P}^-$ are symmetric and definite positive, also $\bar{\mathbf{P}}$ is. Moreover, the entropy $\bar{H}$ associated to a Gaussian distribution with variance $\bar{\mathbf{P}}$ is exactly the weighted

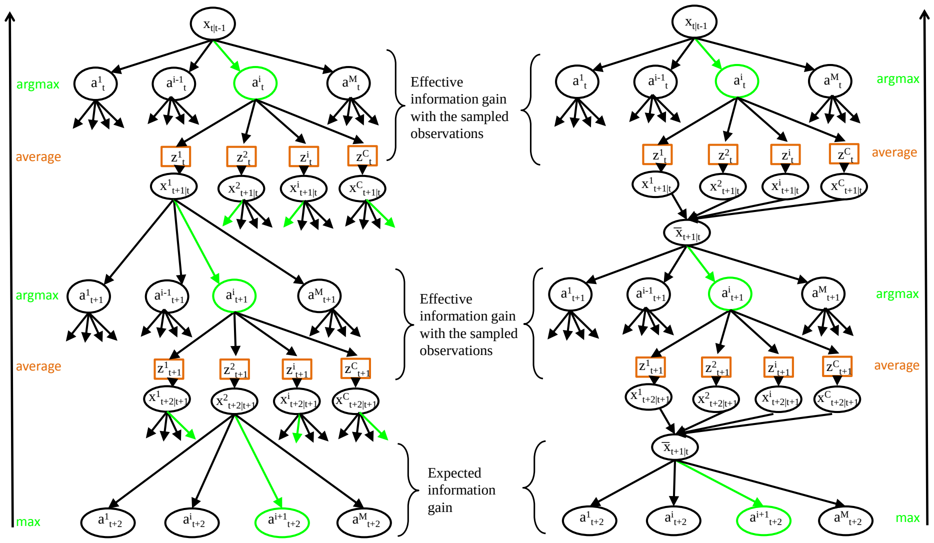

Figure 7: Scheme representing the selection of the best action through the look-ahead algorithm. *Left* Look-ahead optimization according to the original method from [48] applied in [57]; *Right:* proposed approximation which *summarizes* the samples at each horizon.

arithmetic mean of the entropy associated to $\mathbf{P}^{+}$ and $\mathbf{P}^{-}$, with weight $\alpha$. This property guarantees that the entropy obtained averaging over the samples is the same as the entropy associated to this new state $\bar{\mathbf{x}}$:

$$\begin{aligned} H(\bar{\mathbf{x}}) =& \frac{1}{M}\sum_{j=1}^{M} H(\tilde{\mathbf{x}}_j) = \frac{1}{M}\sum_{j=1}^{M} \tilde{\mathbf{d}}_j H(\tilde{\mathbf{x}}_j^{+}) + \\ &+ \frac{1}{M}\sum_{j=1}^{M} (1-\tilde{\mathbf{d}}_j) H(\tilde{\mathbf{x}}_j^{-}) = \alpha H(\tilde{\mathbf{x}}^{+}) + (1-\alpha) H(\tilde{\mathbf{x}}^{-}), \end{aligned} \tag{27}$$

where $\alpha = \frac{1}{M}\sum_{j=1}^{M} \tilde{\mathbf{d}}_j$. By applying this procedure we propagate only one subtree common to all the $M$ observations, thus the computational complexity reduces to $\mathcal{O}((K \cdot L)^{\hbar} \cdot r_p^{\hbar-1} \cdot M)$, which is linear in the number of samples instead of exponential. The summarization causes the loss of the multiple modes, represented by the samples, which are not propagated to future steps.

## 7. Extension to multiple cameras

The method presented in the previous sections describes an algorithm for efficiently and automatically managing a single PTZ camera in a standard multi-target tracking scenario. The extension to a network of multiple PTZ cameras may be obtained in several ways. For example [39] proposes a sequential Kalman filter to combine, in the update stage, the observations of the same target as seen from different cameras. This approach can be applied also to the information gain formulation we proposed. Consider a set $C$ of $N_c$ cameras which successfully observe a target $\mathbf{x}_t$ at time $t$, the covariance resulting from the successive observation according the sequential Kalman filter is the product:

$$\mathbf{P}_t^+ = \left( \prod_{c \in C} (\mathbf{I} - \mathbf{K}_t^c \mathbf{C}_t^c) \right) \mathbf{P}_t^- . \tag{28}$$

As observed in [39] this equation depends on the order in which each camera is considered, mainly because it does not take into account the fact that a camera could miss a target, i.e $\alpha(\mathbf{a}_t^c) < 1$ in Eq. 4. All the possible combinations are exponential in the number of cameras and therefore not suitable for an online application. Hence, we approximate the updated covariance for a single camera with $\alpha(\mathbf{a}_t^c)$, from the weighted average of Eq. 26:

$$\overline{\mathbf{P}}_t^c = (\mathbf{P}^{c,-})^{\frac{1}{2}} ((\mathbf{P}^{c,-})^{-\frac{1}{2}} (\mathbf{I} - \mathbf{K}_t^c \mathbf{C}_t^c) \mathbf{P}_t^{c,-} (\mathbf{P}^{c,-})^{-\frac{1}{2}})^{\alpha(\mathbf{a}_t^c)} (\mathbf{P}^{c,-})^{\frac{1}{2}} . \tag{29}$$

Then, the information gain for a single target across multiple cameras is obtained by successively iterating such computation for each camera, substituting $\mathbf{P}_t^{c,-} = \overline{\mathbf{P}}_t^{c-1}$.

The formulation is similar to the one proposed in [39] except for the fact that Eq. 29 replaces the weighted sum. As introduced in Sec. 6, Eq. 29 guarantees that the entropy related to a gaussian distribution with covariance $\overline{\mathbf{P}}_t^c$ is the weighted sum of the entropy due to covariances $\mathbf{P}_t^{c,-}$ and $\mathbf{P}_t^{c,+}$. This modification does not affect the computational cost, compared to [39]. Indeed, it is exponential in the number of cameras, $O(((K \cdot L)^{\hbar} \cdot r_p^{\hbar-1} \cdot M)^{N_c})$, since all different combinations of actions for the different cameras need to be evaluated.

## 8. Implementation and Evaluation Details

Experimenting PTZ tracking solutions is a classic problem in computer vision [59]. The current protocols span between being quantitative and perfectly repeatable with a low realism [26, 39], and considering real scenarios, where each test is qualitative and cannot be repeated [13, 60]. Here, we consider both the cases, providing a synthetic and a realistic experimental benchmark, which are quantitative and repeatable, concluding with a real experiment where our approach has been implemented in a real-time surveillance platform.

A direct comparison with the methods similar to ours [12, 13, 40], described in Sec. 2, is not possible without changing their own network architecture (number of cameras and the way they communicate), or the experimental protocol (data and ground truth information used by previous works). One of the main features of our system is that it integrates both the multi-target tracking module and the camera management on a single PTZ device, differently from [12] which requires a set of fixed cameras to track the targets and then drive the PTZ cameras towards the target of interest. Both [13, 40] perform live experiments on their own video-surveillance network, hence it is not possible to perform new tests on the same sequences.

We tried to compare our method with the solution proposed in [38], which is the only one focusing on information theoretic management for a single PTZ camera. In particular, this is obtained by simplifying our method (without the detector performance, the occlusion management and the look-ahead optimization) in order to obtain an implementation as close as possible to their method, apart for the fact that we are tracking the targets in the ground plane, instead of the image plane as they do. Results in Sec. 9, 10 demonstrates that the contributions proposed in this paper allow to outperform [38].

### *8.1. Experimental Setup*

We have performed two kind of experiments, synthetic and realistic, for both myopic and non-myopic strategies in order to quantitatively and qualitatively asses the performance of our approach. For the myopic method we have also performed a real experiment with an off-the shelf IP PTZ camera (Sony SNC-RZ30P).

The synthetic scenario consists in a $15 \times 15$m area, with 7 targets following random trajectories mimicking human motion, Fig. 2 (a). The targets are always in the scene, thus the exploration term $I_p$ in Eq. 23 is not considered. We run 12 different sequences, each 50 frames long, with diverse target trajectories, and compute the final scores averaging the per-sequence results. In each sequence, we manage 350 target instances, which have to be detected and associated to tracks. The action set has 4 steps for the zoom, 7 for pan angle and 10 for tilt angle (*i.e.*, $L$=280 different actions). To model the mechanical constraints the camera can move by a maximum displacement of 2 steps for the angles and 1 for the zoom.

For the realistic experiments, as compromise between repeatability and realism, we consider here the PETS 2009 (S2-L1-View1) benchmark, Fig. 2b, where intrinsic calibration matrix $\mathtt{K}_c$ and the extrinsic calibration information are provided. For reproducing the PTZ zoom 1, we reduce the 576×768 resolution to 120×160. The homography $\mathtt{G}_{ptz}$ for this virtual camera to the 3D plane is $\mathtt{G}_{ptz} = \mathtt{K}_c \mathtt{R}_{ptz} \mathtt{K}_{ptz}^{-1}$, where $\mathtt{K}_{ptz}$ and $\mathtt{R}_{ptz}$ are the intrinsic and the rotation PTZ matrices (defined empirically). The original extrinsic calibration data allow to map the ground plane to the original sequence image plane and then, through the $\mathtt{H}_{ptz}$, to the virtual PTZ image plane. The action set is made of 140 different actions, 7 for pan angle, 5 for tilt angle and 4 steps for the zoom. The mechanical constraints of the camera are implemented as in the previous case. The sequence is 795 frames longa and we sub-sample it every 2 frames. Globally, there are 19 different targets, for a total of 2322 true detections.

The whole framework has been implemented in MATLAB and it works at 10 fps for the Gaussian integral solution of Eq. 8 and 0.3 fps for the sampling strategy of Eq. 21. However, it is easily parallelizable both in the sampling stage and in evaluating Eq. 5 for the various actions. Indeed, the evaluation of the expected information gain for each action can be done independently. With $\Delta\phi = 2$, $\Delta\theta = 2$ and $\Delta \mathrm{f} = 1$ the number of reachable actions varies between 18 (at the corners of the grid) and 75. If enough parallel threads are available this can lead to a speedup from 18x to 75x when computing the information gain for each action.

### *8.2. Evaluation Metrics*

To ease future comparisons, we adopt standard multi-target tracking metrics: the Multiple Object Tracking Accuracy (MOTA, the higher the better) which tells how reliable the tracks are and the Multiple Object Tracking Precision (MOTP, the lower the better) [61], which measures the error in localizing the tracked targets on the ground plane. In addition, we calculate the average height of targets as detected in the image, analogously to [38]: the bigger a target appears on the screen, the more information could be extracted for higher level tasks (recognition, re-identification etc.). Other important parameters for appreciating the performance from different strategies are the number of detections on the whole sequence and the average zoom value for the camera.

We use three comparative control strategies: ‘fix’, keeping the camera fixed at the lowest zoom (1x), ‘patrol’, scanning the field of regard according to a preset sequence, ‘random’, performing actions randomly chosen from the set $\mathcal{A}$.

Table 3: Synthetic data, ideal detector: comparison among standard strategies and the information theoretic strategy, with and without the sampling (M=100) to cope with occlusions.

| Strategy | ‘fix’ | ‘patrol’ | ‘rnd’ | *MDP* | |
|---|---|---|---|---|---|
| | | | | ‘intg’ | ‘smpl’ |
| MOTA | 94.6 % | 87.6% | 79.7% | 89.9% | **97.0**% |
| (MOTP [m]) | (0.26 ) | (0.35) | (0.45) | (0.23) | (0.21) |
| Height [pix] | 49.1 | 102.6 | 64.4 | **91.4** | 89.0 |
| # Dets | 278.3 | 75.8 | 55.5 | 186.3 | 214.2 |
| Zoom [x] | 1.00 | 2.54 | 2.53 | 2.05 | 2.00 |

Table 4: Synthetic data, realistic detector: comparison among standard strategies and the information theoretic strategy, with and without the sampling (M=100) to cope with occlusions.

| Strategy | ‘fix’ | ‘patrol’ | ‘rnd’ | *MDP* | |
|---|---|---|---|---|---|
| | | | | ‘intg’ | ‘smpl’ |
| MOTA | 56.6 % | 58.8% | 14.0% | 67.2% | **72.8**% |
| (MOTP [m]) | (0.46) | (0.47) | (0.67) | (0.33) | (0.29) |
| Height [pix] | 57.7 | 106.4 | 84.5 | 121.6 | **122.3** |
| # Dets | 54.5 | 38.3.8 | 15.0 | 80.0 | 89.8 |
| Zoom [x] | 1.00 | 2.54 | 2.46 | 2.64 | 2.61 |

We exploit the occlusion term of Eq. 17, comparing the Gaussian integral solution of Eq. 8, namely ‘intg’ (that represent our implementation of [38]), with the sampling strategy of Eq. 21, namely ‘smpl’.

## 9. Myopic Experiments

### *9.1. Synthetic Experiments for the Myopic Strategy*

Two different experimental sessions are performed on the synthetic scenario, comparing the formulations proposed above for two different types of pedestrian detectors.

A first session considers a perfect detector, whose performance does not decay for smaller targets, but still worsens when the target gets occluded; results are in Tab. 3. The following observations can be made: (1) both the ‘intg’ and ‘smpl’ approaches outperform the competing PTZ strategies ‘patrol’ and ‘rnd’, both in terms of MOTA and MOTP; (2) the ‘fix’ policy is the best among the competitors: actually, it detects a large number of targets (see # Dets), even when they are small, due to the perfect detector; (3) the improvements of our approaches are mainly due to the zooming on the targets (see Zoom [x]), which both creates more reliable tracks (higher MOTA) and better localization (lower MOTP); (4) the sampling approach, that prevents the camera from observing targets which may be occluded, outperforms the ‘intg’ approach.

Note that when using the ideal detector and the ‘intg’ version, which discards the occlusions, our method slightly differs from the approach in [38]. In fact, in this case our approach only considers the mechanical constraints of the camera and the physical extension of the targets as additional elements. Hence, results also show a clear improvement with respect to [38].

In the second session, we consider Eq. 14, substituting the ideal detector with a realistic one, which simulates the HOG performance (i.e. it works worse at small resolutions). Results are in Tab. 4, leading to considerations similar to the previous test. The presence of a realistic detector

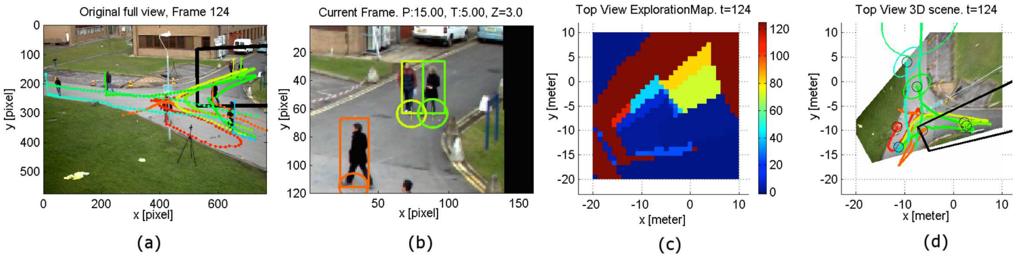

Figure 8: Simulated PTZ on the PETS 2009 dataset. *(a)* Original frame from the dataset, with the tracked targets trajectories and the PTZ field of view (black); *(b)* field of view for the simulated PTZ in the current position (resolution is 160x120 pixels); *(c)* Top view of the exploration map that is used to compute the exploration term of Eq. 23; *(d)* Top view of the warped ground plane where targets are moving, with the estimated trajectories and covariances.

Table 5: PETS dataset, ideal detector: comparison among standard strategies and the information theoretic strategy, with and without the sampling (M=100) to cope with occlusions.

| Strategy | ‘fix’ | ‘patrol’ | ‘rnd’ | $\beta = 9$ | | $\beta = 1$ | |
|---|---|---|---|---|---|---|---|
| | | | | ‘intg’ | ‘smpl’ | ‘intg’ | ‘smpl’ |
| MOTA<br>(MOTP [m]) | 80.6%<br>(0.20) | 50.7%<br>(0.29) | 30.6%<br>(0.39) | 75.2%<br>(0.22) | 76.5%<br>(0.22) | 64.8%<br>(0.18) | **81.1%**<br>(0.17) |
| Height [pix] | 19.9 | 38.5 | 29.3 | 37.5 | **39.2** | 31.8 | 36.4 |
| # Dets | 2160 | 414 | 567 | 998 | 895 | 1524 | 1513 |
| Zoom [x] | 1.00 | 2.00 | 1.55 | 1.97 | 2.08 | 1.63 | 1.87 |

brings in general to worse MOTA and MOTP scores. In addition, the #Dets in the ‘fix’ case decreases dramatically (it cannot zoom to increment the number of detections), and, in general, both the proposed approaches are better in this case. In fact, our strategies know that they need to zoom more (see the Zoom values) to possibly get a detection. Again, the advantage of keeping into account the occlusion term is evident.

### 9.2. *Realistic Experiments on the S2-L1-View1 PETS Sequence for the Myopic Strategy*

A different experimental setup uses a publicly available dataset (PETS2009) and its ground truth for quantitative evaluation. Fig. 8 shows how a PTZ camera is simulated from the original frame, and used to track targets on the ground plane, exploiting the calibration data.

In a first test, whose results are in Tab. 5, we employ the ideal detector, extracting the bounding box from the ground-truth and removing the occluded ones. Since in this sequence people are entering and leaving the scene, we include the exploration term (Eq.23), testing two different values for $\beta$. Considerations: (1) the sampling strategy gives better results for both values of $\beta$, in terms of MOTA, MOTP, and Height; (2) since we have all the detections, MOTA is high also for the fixed strategy. MOTP is higher with our policies, due to the possibility of zooming. (3) reducing $\beta$ encourages to focus on the tracked targets (i.e., lower MOTP) instead of capturing new items. The best value for $\beta$ should be a compromise between tracking accuracy and the capability of capturing novel targets.

In the second test, we introduce a real implementation of the HOG detector, enriching the realism of the simulation, and therefore introduce in the implementation the term in Eq. 14. Results are in Tab. 6 and in general are dramatically lower than those in Tab. 5 because of the many false positives and missed detections from the HOG detector. The improvement of the

Table 6: PETS dataset. HOG detector: comparison among standard strategies and the information theoretic strategy, with and without the sampling (M=100) to cope with occlusions.

| Strategy | 'fix' | 'patrol' | 'rnd' | $\beta = 9$ | | $\beta = 1$ | |
|---|---|---|---|---|---|---|---|
| | | | | 'intg' | 'smpl' | 'intg' | 'smpl' |
| MOTA | 21.6% | 19.0% | 0.0% | 28.3% | 28.3% | 31.6% | **36.4%** |
| (MOTP [m]) | (0.36) | (0.52) | (0.52) | (0.49) | (0.48) | (0.39) | (0.38) |
| Height [pix] | 19.5 | 37.7 | 28.5 | 42.7 | 42.8 | 48.4 | **50.6** |
| # Dets | 1886 | 370 | 581 | 435 | 440 | 714 | 716 |
| Zoom [x] | 1.00 | 2.00 | 1.48 | 2.94 | 2.33 | 2.58 | 2.70 |

Table 7: Synthetic data, ideal detector: comparison among standard strategies and the non-myopic information theoretic strategy with occlusion handling with different horizons $\hbar$, $\gamma = 0.9$, $M = 100$.

| Strategy | 'fix' | 'patrol' | 'rnd' | $\hbar = 1$ | $\hbar = 2$ | $\hbar = 3$ |
|---|---|---|---|---|---|---|
| MOTA | 94.6 % | 87.6% | 79.7% | **97.0**% | 95.6% | 92.1% |
| (MOTP [m]) | (0.26) | (0.36) | (0.45) | (0.21) | (0.21) | (0.21) |
| Height [pix] | 49.1 | 102.6 | 64.4 | 89.0 | 88.6 | **94.7** |
| # Dets | 278.3 | 75.8 | 55.5 | 214.2 | 210.2 | 203.3 |
| Zoom [x] | 1 | 2.54 | 2.46 | 2.07 | 2.11 | 2.23 |

'smpl' method with respect to the competitor strategies is evident considering MOTA, this is due to term in Eq. 14 that pushes the camera to increase the zoom with respect to the previous case of the ideal detector. The MOTP is slightly better for the 'fix' strategy, but this is due to the fact that it is computed only for the targets correctly tracked, that are less than for the 'smpl' case.

### *9.3. Real Trials for the Myopic Strategy*

We also tested our system with a real-time off-the shelf IP PTZ camera, Sony SNC-RZ30P. In order to estimate the calibration parameters of the PTZ camera while moving we use a method similar to [62]. The action set $\mathcal{A}$ is made of 462 actions corresponding to the following grid: 14 values for pan × 11 tilt × 3 zoom. The step between two pan angles is 10.4°, for the tilt is 4.3°, and the zoom values are 1x, 6x and 9x. We set $\beta = 0.667$ and used the 'intg' approach, due to the real-time constraints. The whole system works online at about 15 fps for the tracker and 3 fps for the action selection. Some frames (videos are in supplementary material) are shown in Fig. 9 with a detailed description. The method produces a camera which is able to fully autonomous move in the scene, according to the utility cost, and resulting in a "reasonable" behavior without any supervision from a human operator or other sensors. Effective implementations of computer vision algorithm on PTZ camera are really few, and as far as we know it is the first time a sophisticated algorithm is successfully applied to a stand alone PTZ camera.

## 10. Non-Myopic Experiment

### *10.1. Synthetic Experiments for the Non-Myopic Strategy*

A first session considers a perfect detector, whose performance does not decay for smaller targets, but still worsens as the target gets occluded. Results are in Tab. 7; $\hbar = 1$ indicates the myopic approach, $\hbar = 2, 3$ address the non-myopic strategy, with horizon 2 and 3, respectively.

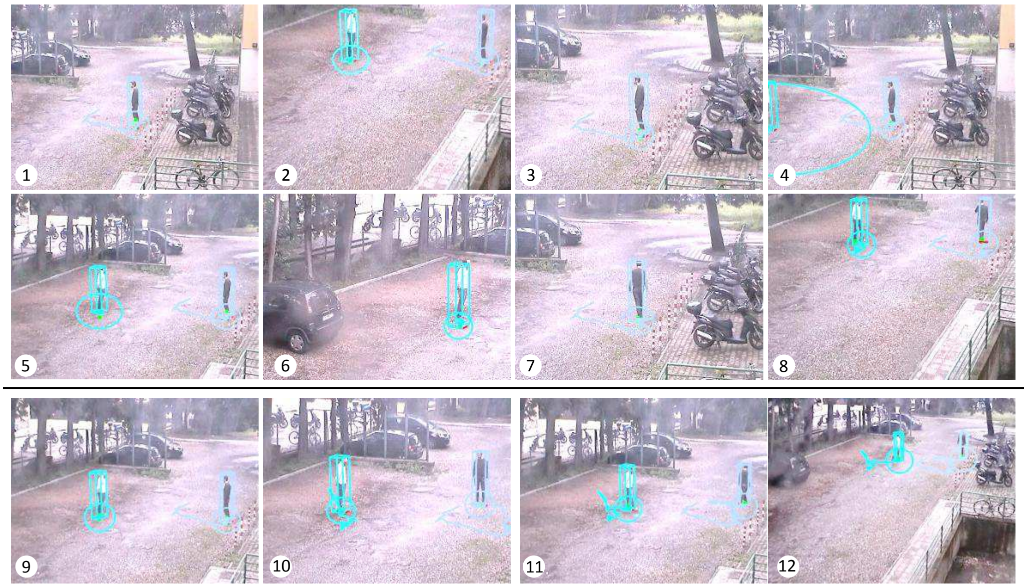

Figure 9: An illustration of camera management with two targets. Targets are marked with their 3D bounding box, the covariance spread of the filter estimate is given by the ellipse. The camera chooses the position automatically, according to the reward function defined in the paper. The resulting behavior produces the following patterns: (1-8) The camera ‘jumps’ between the targets to maximize their localization precision; (9-12) Once the two targets are well localized, the camera widens its field of view to search for novel targets (best in colors).

Many observations can be made: (1) for all the three values of $\hbar$ and in terms of MOTA and MOTP our approach outperforms the other competitors. (2) The ‘fix’ strategy is the best among the competitors: actually, it detects a large number of targets (see #Dets), even when they are small, due to the perfect detector. The improvement of the myopic approach ($\hbar = 1$) is mainly due to the zooming on the targets (see Zoom [x]), which both creates more reliable tracks (higher MOTA) and better localization (lower MOTP). In this scenario the non-myopic approach $\hbar = 2, 3$ does not improve the MOTA metric but provides higher resolution images for the targets. This happens because typically the set of tracked targets is divided on the future steps, hence the camera precisely focuses on few targets at each time. Anyway being the detector ideal, the targets are detected even when they are small in the image, for this reason the MOTA is at is maximum value for $\hbar = 1$. For a deeper understanding of the non-myopic approach, Fig. 10 visualizes the action selection process: the information gain for all the reachable actions is evaluated considering the contribution of each target at each horizon.

In the second session we substitute through Eq. 14 the ideal detector with another one, which simulates the HOG performance. Results are in Tab. 8, leading to considerations similar to the previous test. In addition, the presence of a realistic detector brings in general to lower MOTA and MOTP scores. The #Dets in the ‘fix’ case decreases dramatically (it cannot zoom for taking care of the detector) and in general all the comparative approaches are better in this respect In fact, our strategies know that they need to zoom more (see the Zoom values) to possibly get a detection. From these simulations, it seems that the best horizon is ($\hbar = 2$). This indicates a natural limit of the system in going too far in the future, and this is reasonable. If the target

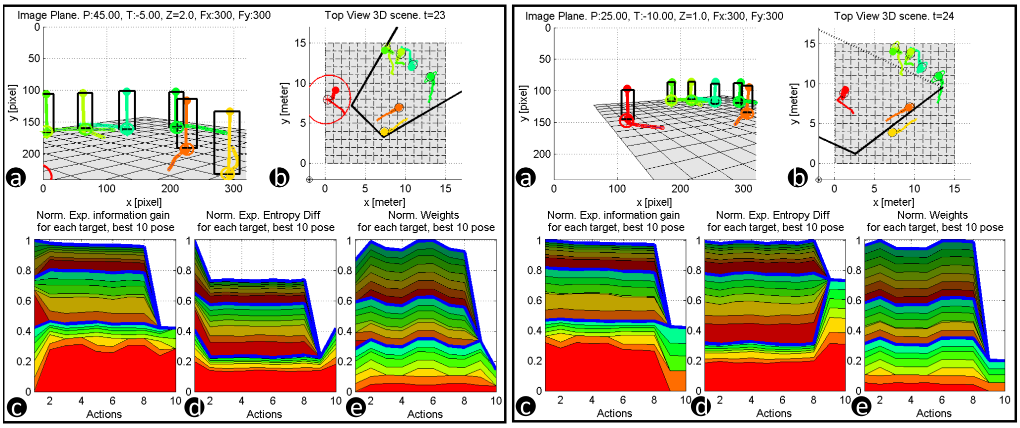

Figure 10: View of the environment for the synthetic experiments in two consecutive frames, $t = 23, 24$. In each of the two frames, some useful data are shown. *a)* Image plane for the PTZ on the synthetic scenario; *b)* Top view of the ground plane where the targets are moving, their movement is limited in a square area of 15×15 meters; *c)* Information gain $I(\mathbf{x}_t, \mathbf{o}_t|\mathbf{a}_t)$ for the 10 best poses (the contribution from the 3 horizon weighted by $\gamma$ is highlighted), for each target at each horizon this term is computed using Eq. 7, the two terms $\alpha(\mathbf{a}_t)$ and $H(\mathbf{x}^+)$ are shown in the other two plots, *d)* and *e)*. This figure shows a case in which the myopic and non-myopic strategy would choose 2 different actions: at $t = 23$ the non-myopic chooses action 1, whereas the myopic would choose action 2, since it cannot predict that waiting to look the red target at $t = 24$ would bring an higher global information gain.

Table 8: Synthetic data, realistic detector: comparison among standard strategies and the non-myopic information theoretic strategy with occlusion handling with different horizons $\hbar$, $\gamma = 0.9$, $M = 100$.

| Strategy | ‘fix’ | ‘patrol’ | ‘rnd’ | $\hbar = 1$ | $\hbar = 2$ | $\hbar = 3$ |
|---|---|---|---|---|---|---|
| MOTA<br>(MOTP [m]) | 56.6%<br>(0.45) | 58.8%<br>(0.46) | 14.0%<br>(0.67) | 72.8%<br>(0.29) | **80.8%**<br>(0.31) | 71.6%<br>(0.32) |
| Height [pix] | 57.7 | 106.4 | 84.4 | 122.3 | 119.5 | **123.1** |
| # Dets | 54.5 | 38.3 | 15.0 | 89.8 | 75.0 | 81.1 |
| Zoom [x] | 1 | 2.54 | 2.50 | 2.62 | 2.66 | 2.71 |

Table 9: PETS dataset, using the detections from GT (except in case they are occluded): comparison among standard strategies and the non-myopic information theoretic strategy with occlusion handling with different horizons $\hbar$, $\gamma = 0.9$, $\beta = 0.4$, $M = 100$.

| Strategy | ‘fix’ | ‘patrol’ | ‘rnd’ | $\hbar = 1$ | $\hbar = 2$ | $\hbar = 3$ |
|---|---|---|---|---|---|---|
| MOTA | 80.6% | 50.7% | 30.6% | 79.7% | **82.1** % | 76.1 % |
| (MOTP [m]) | (0.20) | (0.40) | (0.39) | (0.16) | (0.17) | (0.18) |
| Height [pix] | 19.9 | 38.5 | 29.3 | 34.9 | 35.0 | **36.1** |
| # Dets | 2160 | 414 | 567 | 1631 | 1542 | 1536 |
| Zoom [x] | 1.00 | 2.00 | 1.55 | 1.79 | 1.79 | 1.86 |

Table 10: PETS dataset, HOG detector: comparison among standard strategies and the non-myopic information theoretic strategy with occlusion handling with different horizons $\hbar$, $\gamma = 0.9$, $\beta = 0.4$, $M = 100$.

| Strategy | ‘fix’ | ‘patrol’ | ‘rnd’ | $\hbar = 1$ | $\hbar = 2$ | $\hbar = 3$ |
|---|---|---|---|---|---|---|
| MOTA | 21.6% | 19.0% | -3.6% | 60.5% | 64.1% | **70.5%** |
| MOTP [m] | (0.35) | (0.52) | (0.51) | (0.30) | (0.34) | (0.33) |
| Height [pix] | 19.5 | 37.7 | 28.5 | 37.9 | 34.9 | **38.9** |
| # Dets | 1886 | 370 | 581 | 1292 | 1353 | 1231 |
| Zoom [x] | 1.00 | 2.00 | 1.48 | 2.00 | 1.79 | 2.01 |

abruptly changes directions, when close to the boundaries of the limited area, the EKF engine is not able to predict far in future the probability density for the target’s position. Thus, an action selected relying on such prediction could not be the best.

### *10.2. Realistic Experiments on the S2-L1-View1 PETS Sequence for the Non-Myopic Strategy*

In Tab. 9 we report the results on the PETS sequence using the ground truth as detector. As it can be seen the non-myopic approach obtain slightly better performance than the myopic version in terms of accuracy (MOTA). The best horizon is $\hbar = 2$ since we obtain the highest MOTA value and an high accuracy in the localization. In this first setting the detector is not affected by the target size: all the non occluded targets, even at very low resolution (less than 20 pixels) are correctly detected. For this reason the ‘fix’ strategy has the second best MOTA. Anyway, the zooming capability of the camera helps in the localization. In fact, the MOTP is lower for $\hbar = 1, 2, 3$ and results in a higher average size of targets on the image plane 36.1 pixels for $\hbar = 3$ while the ‘fix’ strategy obtain 19.9 pixels.

In the second test, we introduce a real implementation of the HOG detector, enriching the realism of the simulation. Since we noticed that the HOG detector was performing quite well on this dataset, better than the expected performance (Fig. 4), we did not put the detection term in the estimation ($K_d = +\infty, \mathbf{r}_0 = +\infty$). The results are in Tab. 10. Even in this case, the non-myopic approach does considerable better than the myopic version, as witnessed by the MOTA score. In this case, the best horizon is $\hbar = 3$ since the dynamics of the targets is simpler, with less changes in directions. Therefore, our model (and the EKF dynamics) can predict the future target trajectories with more reliability.

## 11. Conclusions

In this paper, we propose a novel solution to perform sensor management of a single PTZ camera for multiple target tracking. Such solution considers the detector performance at different image resolutions and occlusion ratios. Moreover, it considers the effects of a different camera pose to targets localization. To further improve the tracking performance we apply a non-myopic approach which considers future occlusions among targets in selecting the next actions. We analyze the characteristics and demonstrate the effectiveness of our approach through, synthetic experiments, realistic simulations and effective real-time trials on a real PTZ camera.

## Appendix A. Measurement Accuracy while Zooming

Assuming a constant image measurement error and perfect calibration, world coordinate localization is much more accurate if the backprojection is performed using zoomed views (i.e. long focal length).

We derive analytically the expression that allows to appreciate how zoom affects the measurement equation in the recursive filtering formulation. Measurement uncertainty is mainly oriented along the direction of instantaneous depth because of the panning camera capability. According to this, uncertainty can be quantified assuming a 1D projective camera parametrized by the tilt angle $\theta$ in the instantaneous plane rotating around the vertical camera axis and focal length $f$.

Without loss of generality let's consider that the principal point lies at the image center:

$$\mathtt{K} = \begin{bmatrix} f & 0 \\ 0 & 1 \end{bmatrix}. \tag{A.1}$$

We further have:

$$\mathtt{R} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}. \tag{A.2}$$

The 1-D camera projection matrix $\mathtt{P} = \mathtt{K}[\mathtt{R}|\mathbf{t}]$ results in:

$$\mathtt{P} = \begin{bmatrix} f\cos(\theta) & -f\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & -d \end{bmatrix}, \tag{A.3}$$

where $\mathbf{t} = [0 - d]^\top$ and $d$ is the camera distance with respect to the scene plane. Being the scene plane $Z = 0$, the 1D homography from world to image can be computed from Eq. (A.3):

$$\mathtt{H} = \begin{bmatrix} f\cos(\theta) & 0 \\ \sin(\theta) & -d \end{bmatrix}. \tag{A.4}$$

The inverse:

$$\mathtt{H}^{-1} = \begin{bmatrix} \frac{1}{f\cos(\theta)} & 0 \\ \frac{\sin(\theta)}{f\cos(\theta)d} & -d^{-1} \end{bmatrix}, \tag{A.5}$$

can be used to compute the back-projected uncertainty of a given noise uncertainty $\epsilon$ assumed in the camera image sensor and compute its backprojection $\delta$. Without loss of generality and for the sake of simplicity, let's define:

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} \epsilon \\ 1 \end{bmatrix}, \tag{A.6}$$

their corresponding backprojected points:

$$\mathbf{x}'_1 = \mathtt{H}^{-1}\mathbf{x}_1 = \begin{bmatrix} 0 \\ -d^{-1} \end{bmatrix}, \tag{A.7}$$

$$\mathbf{x}'_2 = \mathtt{H}^{-1}\mathbf{x}_2 = \begin{bmatrix} \frac{\epsilon}{f\cos(\theta)} \\ \frac{\sin(\theta)\epsilon}{f\cos(\theta)d} - d^{-1} \end{bmatrix}, \tag{A.8}$$

are used to compute:

$$\delta = \mathbf{x}'_2 - \mathbf{x}'_1 = \frac{\epsilon\, d}{\sin(\theta)\,\epsilon - f\cos(\theta)}. \tag{A.9}$$

Given $\epsilon$, $\theta$ and $d$, the value of $\delta$ can be increased by increasing the focal length (i.e. performing zoom-in).

## References

[1] Q. Cai, J. Aggarwal, Tracking human motion in structured environments using a distributed-camera system, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (11) (1999) 1241 –1247. doi:10.1109/34.809119.

[2] V. Kettnaker, R. Zabih, Bayesian multi-camera surveillance, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, 1999, pp. 263–259.

[3] L. Lee, R. Romano, G. Stein, Monitoring activities from multiple video streams: establishing a common coordinate frame, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 758 –767. doi:10.1109/34.868678.

[4] R. Collins, A. Lipton, H. Fujiyoshi, T. Kanade, Algorithms for cooperative multisensor surveillance, Proceedings of the IEEE 89 (10) (2001) 1456 –1477. doi:10.1109/5.959341.

[5] Y. Sheikh, M. Shah, Trajectory association across multiple airborne cameras, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 361 –367. doi:10.1109/TPAMI.2007.70750.

[6] B. Leibe, K. Schindler, N. Cornelis, L. Van Gool, Coupled object detection and tracking from static cameras and moving vehicles, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (10) (2008) 1683 –1698. doi:10.1109/TPAMI.2008.170.

[7] A. Ess, B. Leibe, K. Schindler, L. J. V. Gool, Robust multiperson tracking from a mobile platform, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1831–1846.

[8] W. Choi, S. Savarese, Multiple target tracking in world coordinate with single, minimally calibrated camera, in: European Conference on Computer Vision, Springer, 2010, pp. 553–567.

[9] C. Wojek, S. Roth, K. Schindler, B. Schiele, Monocular 3D scene modeling and inference: Understanding multi-object traffic scenes, in: European Conference on Computer Vision, Springer, 2010, pp. 467–481.

[10] M. J. Swain, M. A. Stricker, Promising directions in active vision, International Journal of Computer Vision 11 (2) (1993) 109–126.

[11] S. Intille, J. Davis, A. Bobick, Real-time closed-world tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 697 –703. doi:10.1109/CVPR.1997.609402.

[12] N. Krahnstoever, T. Yu, S.-N. Lim, K. Patwardhan, P. Tu, et al., Collaborative real-time control of active cameras in large scale surveillance systems, in: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2, 2008, pp. 1–12.

[13] C. Ding, B. Song, A. Morye, J. A. Farrell, A. K. Roy-Chowdhury, Collaborative sensing in a distributed ptz camera network, IEEE Transactions on Image Processing, 21 (7) (2012) 3282–3295.

[14] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, D. G. Lowe, A boosted particle filter: multitarget detection and tracking, in: European Conference on Computer Vision, Springer, 2004, pp. 28–39.

[15] Y. Cai, N. de Freitas, J. Little, Robust visual tracking for multiple targets, in: European Conference on Computer Vision, 2006, pp. 107–118.

[16] A. Del Bimbo, G. Lisanti, I. Masi, F. Pernici, Device-tagged feature-based localization and mapping of wide areas with a ptz camera, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010, pp. 39–44.

[17] J. Civera, A. J. Davison, J. A. Magallon, J. M. M. Montiel, Drift-free real-time sequential mosaicing, International Journal of Computer Vision 81 (2) (2009) 128–137.

[18] R. Hess, A. Fern, Improved video registration using non-distinctive local image features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1 –8. doi:10.1109/CVPR.2007.382989.

[19] S. Lovegrove, A. J. Davison, Real-time spherical mosaicing using whole image alignment, in: European Conference on Computer Vision, Springer, 2010, pp. 73–86.

[20] A. Mittal, D. Huttenlocher, Scene modeling for wide area surveillance and image synthesis, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2000, pp. 160 –167. doi:10.1109/CVPR.2000.854767.

[21] E. Hayman, J.-O. Eklundh, Statistical background subtraction for a mobile observer, in: IEEE International Conference on Computer Vision, 2003, pp. 67 –74 vol.1. doi:10.1109/ICCV.2003.1238315.

[22] G. Lisanti, I. Masi, F. Pernici, A. D. Bimbo, Continuous localization and mapping of a pan tilt zoom camera for wide area trackingarXiv:1401.6606v1.

[23] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

[24] C. Boutilier, T. Dean, S. Hanks, Decision-theoretic planning: Structural assumptions and computational leverage, Journal of Artificial Intelligence Research 11 (1999) 1–94.

[25] P. Salvagnini, F. Pernici, M. Cristani, G. Lisanti, I. Masi, A. D. Bimbo, V. Murino, Information theoretic sensor management for multi-target tracking with a single pan-tilt-zoom camera, in: IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 893–900. doi:10.1109/WACV.2014.6836009.

[26] F. Qureshi, D. Terzopoulos, Planning ahead for ptz camera assignment and handoff, in: Third ACM/IEEE International Distributed Smart Cameras. Conference on, 2009, pp. 1–8.

[27] T. Matsuyama, N. Ukita, Real-time multitarget tracking by a cooperative distributed vision system, Proceedings of the IEEE 90 (7) (2002) 1136–1150.

[28] A. W. Senior, A. Hampapur, M. Lu, Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration, in: IEEE Workshop on Applications on Computer Vision, Vol. 1, IEEE, 2005, pp. 433–438.

[29] C. J. Costello, C. P. Diehl, A. Banerjee, H. Fisher, Scheduling an active camera to observe people, in: ACM International workshop on Video surveillance & sensor networks, 2004, pp. 39–45.

[30] C. Micheloni, B. Rinner, G. L. Foresti, Video analysis in pan-tilt-zoom camera networks, Signal Processing Magazine, IEEE 27 (5) (2010) 78–90.

[31] B. Song, C. Ding, A. T. Kamal, J. A. Farrell, A. K. Roy-Chowdhury, Distributed camera networks, Signal Processing Magazine, IEEE 28 (3) (2011) 20–31.

[32] J. Denzler, M. Zobel, H. Niemann, Information theoretic focal length selection for real-time active 3d object tracking, in: IEEE International Conference on Computer Vision, 2003, pp. 400–407.

[33] B. Deutsch, M. Zobel, J. Denzler, H. Niemann, Multi-step entropy based sensor control for visual object tracking, in: Pattern Recognition, Springer, 2004, pp. 359–366.

[34] B. Tordoff, D. Murray, A method of reactive zoom control from uncertainty in tracking, Computer Vision and Image Understanding 105 (2) (2007) 131–144.

[35] J. A. Fayman, O. Sudarsky, E. Rivlin, M. Rudzsky, Zoom tracking and its applications, Machine Vision and Applications 13 (1) (2001) 25–37.

[36] E. Hayman, T. Thorhallson, D. W. Murray, Zoom-invariant tracking using points and lines in affine views. an application of the affine multifocal tensors, in: IEEE International Conference on Computer Vision, Vol. 1, 1999, pp. 269–277.

[37] B. Tordoff, D. Murray, Reactive control of zoom while fixating using perspective and affine cameras, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (1) (2004) 98–112.

[38] E. Sommerlade, I. Reid, Information-theoretic active scene exploration, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[39] E. Sommerlade, I. Reid, Probabilistic surveillance with multiple active cameras, in: IEEE International Conference on Robotics and Automation, 2010, pp. 440–445.

[40] C.-M. Huang, L.-C. Fu, Multitarget visual tracking based effective surveillance with cooperation of multiple active cameras, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41 (1) (2011) 234–247.

[41] X. Zhou, R. Collins, T. Kanade, P. Metes., A master-slave system to acquire biometric imagery of humans at a distance., ACM SIGMM 2003 Workshop on Video Surveillance (2003) 113–120.

[42] A. Del Bimbo, F. Dini, G. Lisanti, F. Pernici, Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks, Computer Vision and Image Understanding 114 (6) (2010) 611 – 623.

[43] D. P. Bertsekas, D. A. Castanon, Rollout algorithms for stochastic scheduling problems, Journal of Heuristics 5 (1) (1999) 89–108.

[44] G. Tesauro, G. R. Galperin, On-line policy improvement using monte-carlo search, in: Advances in Neural Information Processing Systems 1996, Vol. 96, 1996, pp. 1068–1074.

[45] R. Bellman, Dynamic Programming, Princeton Landmarks in Mathematics, Princeton University Press, 2010.

[46] C. Browne, E. Powley, D. Whitehouse, S. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, S. Colton, A survey of monte carlo tree search methods, IEEE Transactions on Computational Intelligence and AI in Games 4 (2012) 1–43. doi:10.1109/TCIAIG.2012.2186810.

[47] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, Vol. 1, Cambridge Univ Press, 1998.

[48] M. Kearns, Y. Mansour, A. Ng, A sparse sampling algorithm for near-optimal planning in large markov decision processes, Machine Learning 49 (2) (2002) 193–208.

[49] C. Kreucher, K. Kastella, A. O. Hero Iii, Sensor management using an active sensing approach, Signal Processing 85 (3) (2005) 607–624.

[50] K. Kitani, B. Ziebart, J. Bagnell, M. Hebert, Activity forecasting, in: European Conference on Computer Vision, Springer, 2012, pp. 201–214.

[51] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, IEEE, 2005, pp. 886–893.

[52] H. Kuhn, The hungarian method for the assignment problem, Naval research logistics quarterly 2 (1-2) (1955) 83–97.

[53] J. Denzler, C. Brown, Information theoretic sensor data selection for active object recognition and state estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2) (2002) 145–157.

[54] A. Criminisi, I. D. Reid, A. Zisserman, Single view metrology, International Journal of Computer Vision 40 (2) (2000) 123–148.

[55] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (4) (2012) 743–761.

[56] O. Lanz, Approximate bayesian multibody tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (9) (2006) 1436–1449.

[57] C. Kreucher, A. Hero, K. Kastella, D. Chang, Efficient methods of non-myopic sensor management for multitarget tracking, in: IEEE Conference on Decision and Control, Vol. 1, 2004, pp. 722–727. doi:10.1109/CDC.2004.1428735.

[58] M. Pálfia, Weighted matrix means and symmetrization procedures, Linear Algebra and its Applications 438 (4) (2013) 1746–1768.

[59] P. Salvagnini, M. Cristani, A. Del Bue, V. Murino, An experimental framework for evaluating PTZ tracking algorithms, in: Internation Conference on Computer Vision Systems, Springer, 2011, pp. 81–90.

[60] C. Piciarelli, C. Micheloni, G. Foresti, Ptz camera network reconfiguration, in: ACM/IEEE International Conference on Distributed Smart Cameras., 2009, pp. 1–7. doi:10.1109/ICDSC.2009.5289419.

[61] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the CLEAR MOT metrics, Journal on Image and Video Processing.

[62] A. Del Bimbo, G. Lisanti, I. Masi, F. Pernici, Continuous recovery for real time pan tilt zoom localization and mapping, in: IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2011, pp. 160 –165. doi:10.1109/AVSS.2011.6027312.