Information Theoretic Sensor Management for Multi-Target Tracking with a Single Pan-Tilt-Zoom Camera

Pietro Salvagnini[a], Federico Pernici[b], Marco Cristani[c],
Giuseppe Lisanti[b], Iacopo Masi[b], Alberto Del Bimbo[b], Vittorio Murino[a,c]
[a]Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Via Morego, Genova, Italy
[b]Media Integration and Communication Center, University of Florence, Viale Morgagni, Florence, Italy
[c]University of Verona, Strada le Grazie, Verona, Italy
[a]name.surname@iit.it, [b]name.surname@dsi.unifi.it, [c]name.surname@univr.it

Abstract

Automatic multiple target tracking with 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 PTZ camera management framework which lies on information theoretic principles: at each time step, the next camera pose (pan, tilt, focal length) is chosen, according to a policy which ensures maximum information gain. The formulation takes into account occlusions, physical extension of targets, realistic pedestrian detectors and the mechanical constraints of the camera. Convincing comparative results on synthetic data, realistic simulations and the implementation on a real video surveillance camera validate the effectiveness of the proposed method.

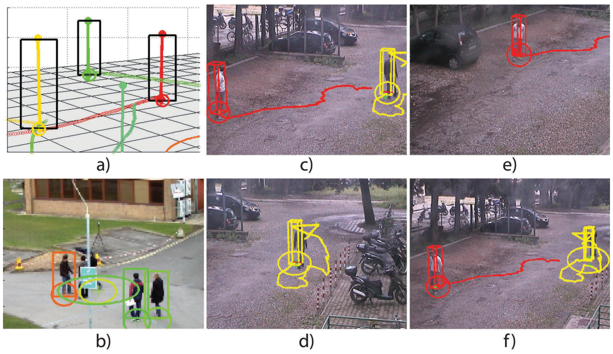

Figure 1. (a) Synthetic scenario; (b) realistic simulation; (c-f) implementation on a real video surveillance camera (best in colors).

1. Introduction

The goal of wide area monitoring has opened new issues in the tracking domain. For example, abnormal behavior detection at a distance demands both trajectory analysis and a proper image resolution to finely recognize human gestures. Hence, a few approaches use pan-tilt-zoom (PTZ) cameras to alternate between large and narrow fields of view [7]. This paper presents a method to automatically select the pose of a single camera (i.e. focal length and pan/tilt angles) in a multi-target tracking scenario.

To date, PTZ sensor management methods for tracking have been approached in a simplified way, discarding many important aspects, such as potential tracking errors, the presence of occlusions, the influence of object detector failures and the varying number of targets in the scene, besides the limited displacement of the camera, due to mechanical constraints. Here, we address such problem in a more realistic and principled way. We suppose a tracking strategy, where an extended Kalman filter works on top of a pedestrian detection module, like in [12]: in this context, our system controls the camera parameters on the basis of targetsâĂŹ trajectories estimation on the ground plane, accounting at the same time possible detection errors, occlusion effects and a variable number of subjects in the scene. All these factors are embedded into a solid theoretical framework: in essence, we want to maximize the information gain produced by an observation on the state of the system, by selecting the camera action which is most informative among those physically plausible. Great attention has been paid to the experimental trials, reaching an optimal compromise between repeatability and realism. Synthetic and real simulations test the performances against occlusions, tracking and detection errors. We conclude that the introduction of the detector performance and the occlusion estimation in the information theoretic based camera management definitively improves the effectiveness of the camera in tracking multiple targets. Finally, a real-time experiment witnesses the suitability of our framework to real situations, and shows how the proposed method, thanks to the heterogeneous aspects taken into account, implicitly produces an effective behavior of the PTZ device. Fig. 1 shows a few frames from the 3 sets of experiments.

The paper continues as follows. Related works are in Sec. 2. The information theoretic formulation is reported in Sec. 3, while our contributions are detailed in Sec. 4. Experiments are presented in Sec. 5, and conclusions in Sec. 6.

2. Related work

PTZ cameras have been studied from different points of view in the past years. Due to the nature of the device, geometric properties related to the varying focal length and positioning[5], and customized calibration techniques [21] have been deeply analyzed.

These sensors are particularly useful for video surveillance, thanks to their capability of both monitoring a wide area and providing high resolution imagery [15]. The main drawback when developing algorithms for PTZ cameras is that it is not possible to work offline with recorded videos since each frame depends on the way the camera moves, [17]. To deal with this problem, in [16] a completely simulated environment is created through computer graphics tools and different strategies for camera to target assignments are proposed and compared. Such strategies are mainly hand-crafted, and require precise information on the targets' position from other sensors. Managing PTZ cameras in a network is, in fact, a typical challenge. A recent novel solution appears in [10], where the authors propose a game theoretic approach for camera to target assignment.

Principled information theoretic frameworks for controlling pose and focal length of active cameras are introduced in [7, 8] and [20], exploiting the concept of information gain for single object tracking. Later on, multi-target tracking is addressed in [18]: notably, here target positions are evaluated on the image plane only. Multiple zooming cameras which give a 3D representation of target positions are considered in [9] for the single target, and for the multi-target scenario [19]. Both in [18, 19] the evaluation method is simplistic, since it assumes no errors in the detections and data association.

A complete formulation of the planning under uncertainty, mainly from a robotics perspective, can be found in [2], where Markov decision processes (MDPs) are employed; such framework explicitly models the temporal evolution of the targets' states and designs a policy for the actions selection based on a reward function.

Our contribution improves substantially the state of the art: we are actually furnishing a principled framework for a single PTZ camera which simultaneously manages heterogeneous challenges, critical for a real implementation. We propose to use a Markov decision process, using an information theoretic reward which copes with different issues in an original fashion. Summarizing, it deals with full-3D positions and takes into account people height, contrarily to [18]; the modeling of the detection noise enriches the "ideal" formulation of [18, 19], at the same computational cost of [18]. The management of the occlusions has no prior in the related literature, and taking into account the mechanical constraints is necessary for an effective implementation on a real camera. Finally, instead of relying on ad-hoc metrics, [18, 19] we exploit standard tracking evaluation protocols and figure of merits, easing future comparisons.

3. MDP with Information Gain Reward

The approach builds upon a multi-target tracking method, and is based on a Markov decision process.

3.1. Basic algorithm for multi-target tracking

Following [18], we design a multi-target tracking system, instantiating one extended Kalman filter (EKF) for each detected target. A pedestrian detector [4] is run at each frame, and the Hungarian algorithm [13] associates each observation to the corresponding filter, initializing a new filter in the case of unassociated observations. Each target position on the ground plane $\mathbf{s}_t = [x^s_t, y^s_t]$ is modeled as a Markov process, while the estimation of the target state in the filter is $\mathbf{x}_t$, containing its location on the ground plane and its speed: $\mathbf{x}_t = [x^w_t, y^w_t, \dot{x}^w_t, \dot{y}^w_t]^\top$. The observation $\mathbf{o}_t = [u_t, v_t]^\top$, *i.e.*, the target location on the image plane in pixels, only depends on the current state and on the action $\mathbf{a}_t$, that is selected in the set of the L possible actions $\mathcal{A}$, and describes the camera pose through the discrete finite parameter vector $\mathbf{a} = (\phi, \theta, \mathrm{f}) \in \mathcal{A}$ (the pan ϕ, tilt θ and focal length f respectively). Formally, we have:

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

where $f(\cdot)$ and $g(\cdot)$ are the dynamical and the observation model, respectively, and $\mathbf{U}$ and $\mathbf{V}$ are the related covariance matrices for the model and observation noise, respectively. The observation function $g(\cdot)$ is the homography from the ground plane to the image plane that obviously depends on the camera parameters defined in the action $\mathbf{a}_t$.

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 $\mathbf{x}^-$ and $\mathbf{P}^-$ 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

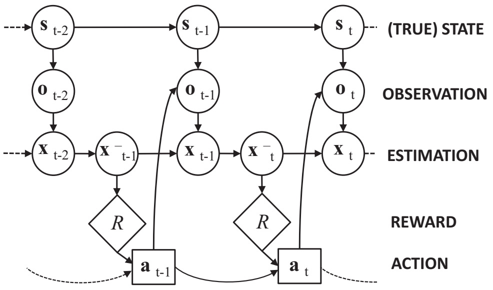

Figure 2. Graphical representation of our approach.

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 $\mathbf{C_x}(\mathbf{a}_t)$; in fact observing a target with an higher magnification will produce a smaller covariance $\mathbf{P}_t^+$ [8].

Eqs. 2 can be seen as modeling the transition probabilities in the MDP (see Fig. 2). To complete the MDP model, we need the reward function $\mathbf{R}(\mathbf{x}_t^-, \mathbf{a}_t)$, 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.

3.2. Information gain formulation: notes

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 [7], we can write:

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

where $H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t)$ is the conditional entropy[1]. Thus, we want to minimize:

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

where we split the domain of integration for $p(\mathbf{o}_t|\mathbf{a}_t)$: Ω_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 only depends on its covariance[2] and Eqs. 2 provide $\mathbf{P}_t^+$ in the case $\mathbf{a}_t$ allows to get the observation for the target, and $\mathbf{P}_t^-$ otherwise. For more details, see [7].

In conclusion, 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 $\mathbf{a}_t$. Extending to K independent targets amounts to sum up the information gains I_k for each target k.

4. Realistic modeling of scene observation from a PTZ camera

The formulation of $\alpha(\mathbf{a}_t)$ in [7] is limited, ignoring many aspects of a real scenario . In the next sections we will redefine it, presenting two versions with four main contributions: in the first version, we introduce (a) the visibility constraint, accounting for the physical dimension of the target in the limited camera FoV, and (b) a realistic detection modeling, maintaining a Gaussian distribution that can be numerically integrated in an efficient way, independently for each target; in the second version, we introduce (c) an additional term that models the occlusions between the targets; such term requires to consider the relative positions between the targets and can be computed only through sampling. Finally, both versions take into account (d) the mechanical limits on the camera motion. Dealing with the variability of the number of targets is managed through the patrolling term as in [18].

4.1. Modeling visibility and detection factors

Introducing the visibility constraint requires to define properly the set Ω_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 updated by considering this new variable:

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

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 $\mathbf{o}_t^- =$

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

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

$g(\mathbf{x}_t^-, \mathbf{a}_t)$, since the actual observation $\mathbf{o}_t$ is yet not available when selecting the actual action $\mathbf{a}_t$[3]. Then, 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:

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

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 Ω_t 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 Eq. 2, we have $\mathbf{o}_t \sim \mathcal{N}(\mathbf{o}_t^-, \Sigma_{\mathbf{o}_t})$, where:

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

Visibility domain. Ω_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 [19], 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 parameters 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 $\mathtt{W}_{\mathbf{a}_t}$, as in [3]. The set Ω_t is then defined as:

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

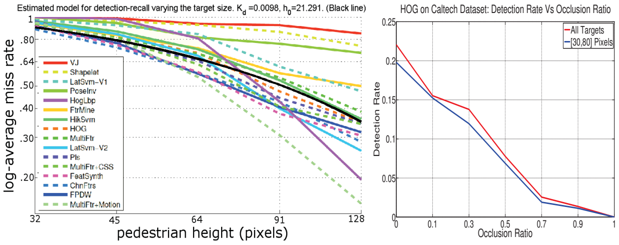

Figure 3. *Left*: Black curve is the function that we used to model the pedestrian detection recall as the target size varies. *Right*: HOG pedestrian detector performance for targets with different occlusion ratios.

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:

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

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

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

In conclusion, the y coordinate for the $\mathbf{e}_t$ is linearly obtained from the y 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 the fact that the performance of pedestrian detectors depends on the height $\mathbf{r}$ of the target (in pixels). We estimate such 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 Caltech pedestrian dataset reported in [11]. Fig. 3 (*left*) shows a comparison between the function we use and the ones reported in [11].

The target height $\mathbf{r}_t = |y_t^{\mathbf{e}} - y_t^{\mathbf{o}}|$ can be computed as a function of the observation $y_t^{\mathbf{o}}$ and the camera position $\mathbf{a}_t$, exploiting the above homology as in Eq. 12. Linearizing the homology around the expected observation $\mathbf{o}_t^-$, we obtain an exponential function, linear in $\mathbf{o}_t$. Therefore we can write:

$$p(\mathbf{d}_t{=}1|\mathbf{o}_t, \mathbf{a}_t) = 1 - e^{-K_d(\mathbf{D}y_t^{\mathbf{o}} + d - \mathbf{r}_0)}, \tag{13}$$

where the matrix $\mathbf{D}$ and vector d 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. 13, gives another Gaussian distribution. Thus the weight $\alpha(\mathbf{a}_t)$ in Eq. 8 can be numerically computed as bounded integration of a Gaussian distribution.

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

4.2. Handling Occlusions

Occlusions represent a serious problem for the selection of the action. Actually, when a target is occluded, it is not seen by the camera, failing to get the information gain, that in the case of no occlusion one would expect.

Without any information on possible occluding obstacles in the field of view, we can only keep into account inter-occlusions among targets in the scene. To this aim, we introduce a term that infers the percentage of area occluded in the next frame, resembling the depth-sorting method of [14]. In practice, we build a binary occlusion mask which indicates the occluded pixels.

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

$$\begin{aligned} H(\mathbf{x}_t|\mathbf{o}_t, \mathbf{d}_t, \mathbf{c}_t, \mathbf{a}_t) &= \int_{\neg\Omega} p(\mathbf{o}_t|\mathbf{a}_t) d\mathbf{o}_t H(\mathbf{x}_t^-) + \\ &+ \int_{\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) \int p(\mathbf{d}_t{=}0|\mathbf{o}_t, \mathbf{a}_t, \mathbf{c}_t) p(\mathbf{c}_t|\mathbf{o}_t, \mathbf{a}_t) d\mathbf{c}_t d\mathbf{o}_t H(\mathbf{x}_t^-) + \\ &+ \int_{\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) \int p(\mathbf{d}_t{=}1|\mathbf{o}_t, \mathbf{a}_t, \mathbf{c}_t) p(\mathbf{c}_t|\mathbf{o}_t, \mathbf{a}_t) d\mathbf{c}_t d\mathbf{o}_t H(\mathbf{x}_t^+) = \\ &= (1 - \alpha(\mathbf{a}_t)) H(\mathbf{x}_t^-) + \alpha(\mathbf{a}_t) H(\mathbf{x}_t^+) \end{aligned} \tag{14}$$

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

$$\alpha(\mathbf{a}_t) = \int_{\Omega_t} p(\mathbf{o}_t|\mathbf{a}_t) \int p(\mathbf{d}_t{=}1|\mathbf{c}_t, \mathbf{o}_t, \mathbf{a}_t) p(\mathbf{c}_t|\mathbf{o}_t, \mathbf{a}_t) d\mathbf{c}_t d\mathbf{o}_t, \tag{15}$$

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

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: $p(\mathbf{d}_t|\mathbf{c}_t, \mathbf{o}_t, \mathbf{a}_t) = p(\mathbf{d}_t|\mathbf{o}_t, \mathbf{a}_t) p(\mathbf{d}_t|\mathbf{c}_t)$, where the first factor has been computed in Sec. 4.1. To estimate $p(\mathbf{d}_t|\mathbf{c}_t)$, i.e., the effect of the occlusion on the detection performance, we use again the Caltech pedestrian dataset [11], obtaining the plots shown in Fig.3(*right*). We choose to approximate this relation as linear: $p(\mathbf{d}_t{=}1|\mathbf{c}_t) = 1 - \mathbf{c}_t$.

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 $^{\neg k}$, $\{k\} \cup \{\neg k\} = \mathcal{K}$), so we need to expand it as:

$$p(\mathbf{c}_t^k|\mathbf{o}_t^k, \mathbf{a}_t) = \int p(\mathbf{c}_t^k|\mathbf{o}_t^{\mathcal{K}}, \mathbf{a}_t) p(\mathbf{o}_t^{\mathcal{K}}|\mathbf{a}_t) d\mathbf{o}^{\neg k} \tag{16}$$

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

$$p(\mathbf{c}_t^k|\mathbf{o}_t^{\mathcal{K}}, \mathbf{a}_t) = \delta(\mathbf{c}_t^k - \bar{\mathbf{c}}_t^k), \quad \bar{\mathbf{c}}_t^k = \frac{\int \delta(\mathbf{x}_t^k \underset{u}{<} \mathbf{x}_t^{\neg k}|\mathbf{a}_t) du}{\int \delta(\mathbf{x}_t^k|\mathbf{a}_t) du}, \tag{17}$$

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, and sample from $p(\mathbf{x}_t^{-,1}, \ldots, \mathbf{x}_t^{-,k}, \ldots, \mathbf{x}_t^{-,K}) = \prod_{k=1}^K p(\mathbf{x}_t^{-,k})$ getting M sets of possible positions for all the targets $\left\{\tilde{\mathbf{x}}_{t,j}^{-,1}, \ldots, \tilde{\mathbf{x}}_{t,j}^{-,k}, \ldots, \tilde{\mathbf{x}}_{t,j}^{-,K}\right\}_{j=1\ldots M}$. Then, given a candidate action $\mathbf{a}_t$, the corresponding weight $\alpha(\mathbf{a}_t)$ is estimated as follow: the related sets of observation predictions are computed $\{\tilde{\mathbf{o}}_{t,j}^1, \ldots, \tilde{\mathbf{o}}_{t,j}^k, \ldots, \tilde{\mathbf{o}}_{t,j}^K\}_{j=1\ldots M}$ according to the model of Eq. 2; each of this set j is used to evaluate the inner integral in Eq. 15:

$$\tilde{\mathbf{d}}_{t,j}^k = \int p(\mathbf{d}_t{=}1|\mathbf{c}_t^k, \tilde{\mathbf{o}}_{t,j}^k, \mathbf{a}_t) p(\mathbf{c}_t^k|\tilde{\mathbf{o}}_{t,j}^k, \mathbf{a}_t) d\mathbf{c}_t^k \tag{18}$$

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. 15 with a summation over the samples:

$$\alpha^k(\mathbf{a}_t) = \frac{1}{M} \sum_{j=1}^M \tilde{\mathbf{d}}_{t,j}^k. \tag{19}$$

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

4.3. Modeling the camera mechanics: action set reduction

In this case, 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

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

$\Delta\phi$ 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 magnification factor, that can be easily obtained combining the expected system frame rate and the camera specifications.

4.4. Patrolling term for new target detection

To take into account possible new targets occurring in the scene, the PTZ has to randomly patrol, looking for new evidence. To model this factor we get inspiration from [18], where an additional information gain $I_p(\mathbf{b}_t|\mathbf{a}_t)$ 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 and on exploration of the scene we obtain:

$$I_t = \sum_{k=1}^{K} I(\mathbf{x}_t^k; \mathbf{o}_t^k|\mathbf{a}_t) + \beta I_p(\mathbf{b}_t|\mathbf{a}_t) \qquad (20)$$

where β 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. Eq.8 and Eq.15 characterize the two proposed versions, the first more efficient and the second one more complex, which also takes into account the occlusions.

5. Experiments

Experimenting PTZ tracking solutions is a classic problem in computer vision: the current protocols span between being quantitative and perfectly repeatable with a low realism, and considering real scenarios, where each test is qualitative and cannot be repeated. 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.

To ease future comparisons, we adopt standard multi-target tracking metrics: the MOTA (the higher the better) which tells how reliable the tracks are: false detections, missed targets and identity switches penalize it; the MOTP (the lower the better) [1], 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 [18]: the bigger a target appears on the screen, the more information could be extracted for higher level tasks (recognition, re-identification etc.). Important parameters for appreciating our performance are the number of detections on the whole sequence (# Dets) and the average zoom value for the camera (Zoom).

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 1. Synth. data, IDEAL detector: comparison among standard strategies and the information theoretic strategy, with and without the sampling (M=100) to cope wiht 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 2. Synth. data, REALISTIC detector: comparison among standard strategies and the information theoretic strategy, with and without the sampling (M=100) to cope wiht 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

5.1. Synthetic Experiments

The synthetic scenario consists in a 15×15m area, with 7 targets following random trajectories mimicking human motion, Fig. 1 (a). The targets are always in the scene, thus the exploration term I_p in Eq. 20 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 in tracks. The action set has 4 steps for the zoom, 7 for pan angle and 10 for tilt angle (*i.e.*, 280 different actions). To model the mechanic the camera can move by a maximum displacement of 2 steps for the angles and 1 for the zoom. We exploit the occlusion term of Eq 20, comparing the Gaussian integral solution of Eq. 8, namely ’intg’, with the sampling strategy of Eq. 19, namely ’smpl’.

A first session considers a perfect detector, whose performance do not decay for smaller targets, but still worsens when the target gets occluded; results are in Tab. 1. The following observations can be made: (1) both the ’intg’ and ’smpl’ approaches outperform the competing PTZ strategies ’patrol’, ’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 create more reliable tracks (higher MOTA) and better localization (lower MOTP); (4) the sampling approach, which prevents the camera from observing targets which may be occluded, outperforms the ’intg’ approach. Note that using an ideal detector and the ’intg’ version, which discards the occlusions, our method slightly differs from the approach in [18]. In fact in this case ours 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 [18].

In the second session, we consider Eq. 13, 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. 2, leading to considerations similar to the previous test. In addition, the presence of a realistic detector brings in general to worse MOTA and MOTP scores; 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 respect: 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.

5.2. Realistic Experiment

As compromise between repeatability and realism, we consider here the PETS 2009 (S2-L1-View1) benchmark, Fig 1b , 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 sets is made of 140 different actions, 4 steps for the zoom, 7 for pan angle and 5 for tilt angle. The mechanical constraint on the camera is implemented as for the synthetic experiments. The sequence is 795 frames long; we subsample it every 2 frames. Globally, there are 19 different targets, for a total of 2322 true detections.

In a first test, whose results are in Table 3, 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.20), testing two different values for β. Considerations: (1) the sampling strategy gives better results for both values of β, 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 β encourages to focus on the targets already presents (i.e., lower MOTP) instead of capturing new items. The best value for β 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 the term in Eq. 13 in the implementation. Results are in Tab. 4 and in general are dramatically lower than in Tab. 3 because of the many false positives and missed detections from the HOG detector. The improvement of the ’smpl’ method with respect to the competitor strategies is evident considering MOTA, this is due to term in Eq. 13 that pushes the camera to increase the zoom with respect to the previous case of the ideal detector (2.7 vs 1.87). 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, much less than for the ’smpl’ case. The whole framework has been implemented in MATLAB and it works at 10 fps for the ’intg’ formulation and 0.3 fps for the ’smpl’. However it is easily parallelizable both in the sampling stage and in evaluating Eq. 5 for the various actions.

Table 3. PETS dataset. Ideal detector. Integral solution vs sampling with occlusion management.

Strategy	’fix’	’patrol’	’rnd’	$\beta = 9$		$\beta = 1$	
				intg	smpl	intg	smpl
MOTA	80.6%	50.7%	30.6%	75.2%	76.5%	64.8%	**81.1%**
MOTP [m]	0.20	0.29	0.39	0.22	0.22	0.18	**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

Table 4. PETS dataset. HOG detector. Integral solution vs sampling with occlusion management.

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

5.3. Real Trials

We also apply our system to 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 [6]. 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, and the whole system works online at about 15 fps for the tracker and 3 fps for the action selection. Some frames (video: http://youtu.be/QvRa8_d2vs8) are shown in Fig. 4 with a detailed description. The method we presented 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.

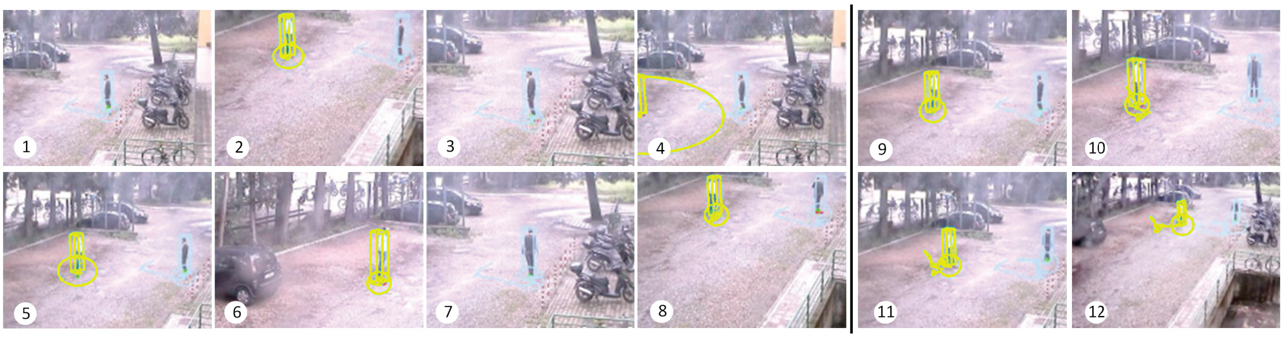

Figure 4. An illustration of camera management with two targets. Targets are marked with their 3D bounding box, the covariance spread of the filter estimate is given by the ellipse. The camera chooses the position automatically, according to the reward function defined in the paper. The resulting behavior 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).

6. Conclusions

In this paper we propose a novel solution to sensor management for multiple target tracking using a PTZ camera. The approach is built upon a information theoretic framework, enriched in a way to mimic real working PTZ conditions, that have never been jointly considered so far. In specific, the formulation takes into account occlusions, physical extension of targets, realistic pedestrian detectors and the mechanical constraints of the camera.

We analyze the characteristics and demonstrate the effectiveness of our approach through synthetic experiments, realistic simulations and an implementation on a real video surveillance camera.

References

[1] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *JIVP*, 2008:1, 2008.

[2] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *JAIR*, 11(1):94, 1999.

[3] A. Criminisi, I. D. Reid, and A. Zisserman. Single view metrology. *IJCV*, 40(2):123–148, 2000.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[5] L. de Agapito, E. Hayman, and I. Reid. Self-calibration of rotating and zooming cameras. *International Journal of Computer Vision*, 45(2):107–127, 2001.

[6] A. Del Bimbo, G. Lisanti, I. Masi, and F. Pernici. Continuous recovery for real time pan tilt zoom localization and mapping. In *AVSS*, pages 160 –165, 30 2011-sept. 2 2011.

[7] J. Denzler and C. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *PAMI*, 24(2):145–157, 2002.

[8] J. Denzler, M. Zobel, and H. Niemann. Information theoretic focal length selection for real-time active 3d object tracking. In *ICCV*, pages 400–407. IEEE, 2003.

[9] B. Deutsch, S. Wenhardt, and H. Niemann. Multi-step multi-camera view planning for real-time visual object tracking. In *Pattern Recognition*, pages 536–545. Springer, 2006.

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

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 99:1–1, 2011.

[12] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR*, volume 1, pages 260–267. IEEE, 2006.

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

[14] O. Lanz. Approximate bayesian multibody tracking. *PAMI*, 28(9):1436, 2006.

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

[16] F. Qureshi and D. Terzopoulos. Planning ahead for ptz camera assignment and handoff. In *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*, pages 1–8. IEEE, 2009.

[17] P. Salvagnini, M. Cristani, A. Del Bue, and V. Murino. An experimental framework for evaluating PTZ tracking algorithms. *Computer Vision Systems*, pages 81–90, 2011.

[18] E. Sommerlade and I. Reid. Information-theoretic active scene exploration. In *CVPR*, pages 1–7. IEEE, 2008.

[19] E. Sommerlade and I. Reid. Probabilistic surveillance with multiple active cameras. In *ICRA*, pages 440–445. IEEE, 2010.

[20] B. Tordoff and D. Murray. A method of reactive zoom control from uncertainty in tracking. *CVIU*, 105(2):131–144, 2007.

[21] Z. Wu and R. Radke. Keeping a pan-tilt-zoom camera calibrated. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 99(PrePrints):1, 2012.