Information-theoretic Decision Making for Exploration of Dynamic Scenes

Eric Sommerlade, Ian Reid

Department of Engineering Science, University of Oxford
OX1 3PJ, Oxford, UK
{eric,ian}@robots.ox.ac.uk

Abstract. This work presents an information-theoretic approach to controlling attention in video sequences. Our work balances the information gain from following an object of interest with the information potentially gained from shifting the focus of interest, thus exploring the scene. We propose that the camera parameters of a single camera – pan, tilt and zoom – should be set to maximise expected information gain to reduce the uncertainty in the scene model, which is equivalent to minimising the expected conditional entropy. The scene model comprises multiple targets and a yet unobserved one, where the increase in information content of the former is supplied directly by using a Kalman Filter tracker. The latter is modelled using a "background" Poisson process, and its parameters are learned from extended scene observations; We deliberately model the scene activity with a frequentist approach to emphasise the differences among the scheduling functions presented. We support our argument with quantitative and qualitative analyses in simulated environments, demonstrating that this approach yields sensible exploration behaviours in which the camera alternates between obtaining close-up views of the targets while paying attention to the background, especially to areas of known high activity.

1 Introduction

Object detection and tracking of objects in video data are crucial elements for further reasoning in modern vision-based systems. In the context of video surveillance, a high coverage of supervised area is desired to maximise the number of object detections, which are then used for further processing, e.g. identification or classification. These tasks usually require or profit from a higher resolution [16] that usually cannot be obtained from cameras that serve to overview the scene. Since the cost of installation and resulting amount of video data to be transferred, stored and observed prohibits naïve addition of cameras, an alternative solution is to use cameras with a pan/tilt/zoom (PTZ) functionality, which explore the area in a sensible fashion and focus onto occurrences of interest to surveillance. Not only are these devices readily available commercially, but also does the restriction to a single camera obviate the need to relate several cameras to each other spatially. Unfortunately, the exploration of the scene conflicts with a close-up inspection of objects of interest. Zooming into a part of the scene decreases the field of view of the camera, and areas with possibly interesting behaviour are not covered any longer. Furthermore, in active zoom control a balance has to be struck between the

maximum attainable zoom onto an object and the risk of losing lock.
These problems are directly addressed in our work, which presents a new method to schedule a single active camera by making use of a probabilistic framework. We address three issues, firstly how to explore the scene to search for new, yet undetected actors, secondly how to decide which of the detected actors to observe more closely and finally how far to zoom onto the chosen target, minimising the risk of losing track. Specifically, we use the information–theoretic concept of entropy to measure the uncertainty of each object in the scene and to compare the utility of pan/tilt/zoom settings for decreasing these uncertainties. We use an activity map to incorporate scene specific actor behaviour. This map keeps track of the rate actors appear in this area of the scene, and is modelled by a Poisson process for each location. The probability of making a new detection is obtained from the locations which are missed when a set of parameters is chosen. This acts as a counterbalance for the zoom onto the actors. The best parameters are the ones which maximally reduce the uncertainty of a subset of or all actors and minimise the chance of an undetected appearance of a new actor.

We compare our method with standard approaches using recent metrics and a new one. The latter measures the increase in area of observations when using a given scheduling algorithm. For repeatability, we test the scheduling policies on a common video data set with available ground truth data, and on a new sequence which has been preprocessed with a background detection algorithm. Both tests yield confirming results.

2 Related work

Camera scheduling has been addressed in some recent work by the vision community. Qureshi *et al.*[17] used a first come, first serve rule in their simulator, based on evaluations of network routing algorithms [5]. Contrary to our work, these use a supervisor camera to make wide area observations and to coordinate PTZ control. This kind of master–slave–configuration is also used by Bagdanov *et al.*[1], who considers scheduling as a dynamic discrete optimisation problem. All works address the camera assignment problem, i.e. more persons to be observed than cameras available, but not the zoom selection. All authors use synthetic data to run evaluations for control of one or several PTZ cameras.
Hampapur *et al.*. [11] uses hand crafted rules to assign active cameras to actors, and chooses the zoom setting via geometric reasoning. The system uses multiple calibrated supervisor cameras for 3D tracking, and incorporates a head detector to focus the zoomed view onto the face of persons. A disadvantage of supervisor cameras is the need for a mapping of image contents to the active camera, which has to be obtained from restricted camera placements, movements or temporally extended observations [18,17,8].

Probabilistic reasoning for camera zoom control is used by Tordoff *et al.*and Denzler *et al.*[21,7] to minimise the chance of losing the target while maximising zoom level at the same time. Whereas both works address only one target, the latter makes use of a stereo platform to track in 3D.

We are aware that the concept of scene activity has been studied intensely [13,20,14,4]. We would like to stress that this paper is neither about a new tracking or scene activ-

ity analysis method. Instead, our concern is how these results can be used for camera scheduling. Davis *et al.*[6] use detected motion from randomly chosen pan/tilt settings to learn a map which is then used to select future camera parameters. The authors propose several methods to navigate through the learned map, but all goal locations are chosen randomly by assuming the map entries stem from an unnormalised probability distribution. Another kind of activity map can be found in Gould *et al.*[9]. Here, a sophisticated perceptual model is learned and used to drive the focus of attention. Objects are classified in a close-up view which is selected from a wide angle view having a high chance of containing classifiable objects. The actual distribution is given by a previously trained Bayes network.

Sequential camera scheduling in a decision theoretic framework has also been used to generate a series of fixation points. Gu *et al.*[10] obtain these with a set of thresholds to accommodate for the heuristic modelling of the costs and rewards obtained in from directing the eye fixation point. Similar to our work, Paletta and Fritz' [15], model these gains with entropy as a measure of local information content and use reinforcement learning to obtain a general policy which needs shorter sequences of fixations to reduce uncertainty than selecting these points in a random fashion.

3 Active Zooming vs Exploration

One goal of our system is to track objects and obtain images at a high resolution to aid in processing steps, e.g. identification or classification. For this, we desire minimal uncertainty in the location of the objects. At the same time, the zoom is bounded by the uncertainty of the object's motion, as well as its spatial extent.

We make use of the same optimality criterion for the selection of the camera parameters as Denzler *et al.*[7]. However, while Denzler's work is specifically concerned with optimising tracking accuracy, we are seeking balance between this and the possibility of acquiring new targets.

The criterion is as follows: Before making an observation at time t, we choose the best parameter $\mathbf{a}_t$ for the observation. The parameter $\mathbf{a}_t$ summarises the different settings for the observation process, i.e. pan, tilt and zoom. Among all choices, this parameter will maximally reduce the expected uncertainty in a given probability distribution of the true state $\mathbf{x}_t$. Applying the chosen parameter yields an observation $\mathbf{o}_t$ which is finally used to update the distribution $p(\mathbf{x}_t)$.

A natural measure for the uncertainty is the expected conditional entropy

$$H_{\mathbf{a}_t}(\mathbf{x}_t|\mathbf{o}_t) = -\iint p(\mathbf{x}_t, \mathbf{o}_t|\mathbf{a}_t) \log(p(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t))\, d\mathbf{x}_t d\mathbf{o}_t \tag{1}$$

Note that this measure is independent of the next observation.

The best parameter $\mathbf{a}$ is then found by minimisation of the conditional entropy:

$$\mathbf{a}_t^* = \arg\min_{\mathbf{a}_t} H_{\mathbf{a}_t}(\mathbf{x}_t|\mathbf{o}_t) \tag{2}$$

In the following two sections we will first summarise the results of Denzler *et al.*, then introduce our approach for scene exploration.

3.1 Object Tracking

We assume independence of the objects in the scene, and assign each detection a Kalman filter. Our observation model is the position and bounding box of the object in the 2-d image plane. The state model for each of the tracked targets is the position, velocity and extent in each coordinate. The motion model is a simple constant-velocity target for the position and velocity [2], whereas the width and height are assumed to be constant.

The differential entropy of such a Gaussian distributed state vector $\mathbf{x}$ with covariance matrix $\mathbf{P}$ is

$$H(\mathbf{x}) = 3 + \frac{1}{2}\log((2\pi)^6|\mathbf{P}|). \tag{3}$$

We use the notation $\hat{\mathbf{x}}_t^+$ for a state which has been updated with the latest observation $\mathbf{o}_t$, and $\hat{\mathbf{x}}_t^-$ the state which has been predicted by the Kalman filter, but not updated because no observation was made. The analogous notation is used for the covariance matrices, $\hat{\mathbf{P}}_t^+$ and $\hat{\mathbf{P}}_t^-$, respectively.

The conditional entropy in equation 1 integrates over the domain of all observations. This domain can be split into the area inside (v) and outside ($\neg v$) the image. The integral then splits into a part where the target is visible, and a part where it is not visible:

$$\begin{aligned} H_{\mathbf{a}_t}(\mathbf{x}_t|\mathbf{o}_t) = &\int_v p(\mathbf{o}_t|\mathbf{a}_t)H(\hat{\mathbf{x}}_t^+)\,d\mathbf{o}_t \\ &+ \int_{\neg v} p(\mathbf{o}_t|\mathbf{a}_t)H(\hat{\mathbf{x}}_t^-)\,d\mathbf{o}_t \end{aligned} \tag{4}$$

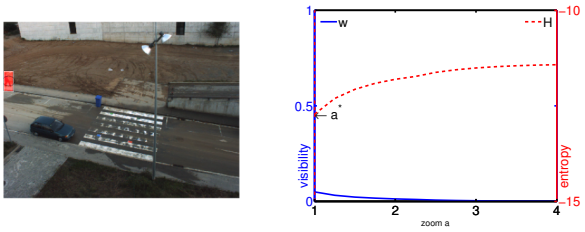
Since in the Kalman filter case the entropies of the state estimate $H(\hat{\mathbf{x}}_t^+)$, $H(\hat{\mathbf{x}}_t^-)$ do not depend on the actual observation $\mathbf{o}$, this integral can be simplified to

$$\begin{aligned} H_{\mathbf{a}_t}(\mathbf{x}_t|\mathbf{o}_t) = &\, w(\mathbf{a}_t)H(\hat{\mathbf{x}}_t^+) \\ &+ (1 - w(\mathbf{a}_t))H(\hat{\mathbf{x}}_t^-) \end{aligned} \tag{5}$$

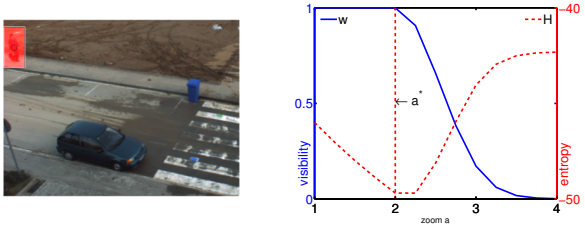
The factor $w(\mathbf{a}_t) = \int_v p(\mathbf{o}_t|\mathbf{a}_t)\,d\mathbf{o}_t$ expresses the probability of making an observation of the object in the image. We simplify the evaluation of this integral by making use of the error function for the integral over the Gaussian distribution of the position and the axis alignment of the observation. We weight this result by the expected observed area of the bounding box.

The development of the entropy H and the probability of making an observation, w, are shown in figure 1. Here, the first frame (a) shows the 1σ - covariance ellipse of the location right after initialisation on a newly detected target. Due to the high initial uncertainty of the location, the probability of making an observation is highest when not zooming in. The fifth frame (b) shows the decreased covariance ellipse, and that the confidence in the making an observation in the next frame rises. The camera zooms in, but is limited by the visibility of the bounding box of the target. If the camera zoomed in too far, the bounding box would be cropped. In the 16th frame (c), the camera zooms in further and starts panning to follow the object.

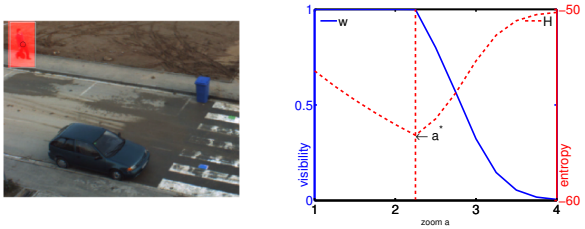
The development of the entropy for a single target and its visibility is shown in figure 2 for the first 20 frames after the detection of a target in the sequence.



(a) Frame 442



(b) Frame 447



(c) Frame 458

Fig. 1. Visibility term w and entropy H for given levels of zoom for frames 442, 447 and 458 of the HERMES Outdoor sequence, camera 1. (a) after the initialisation of a Kalman filter on a new object. (b) The covariance gets smaller, and the confidence in the visibility rises. The camera zooms in. (c) The camera pans to follow the object.

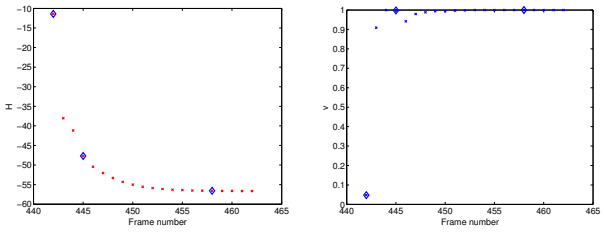


Fig. 2. Entropy H (left) and visibility term w (right) for consecutive frames in the HERMES sequence. The frames 442, 447 and 458, detailed in figure 1, have been highlighted.

3.2 Poisson process for unobserved events

In many man-made environments, people show up regularly, but unpredictably. The absolute times of two appearances of persons in a scene are independent random variables, i.e. the number of appearances before an occurrence is independent of the number of the following ones. The rate at which persons appear is dependent on the location in the scene. People usually use doors or enter the scene along a typical pathway, and less often take shortcuts.

In this discussion, for every point in the scene, the appearance of objects is modelled by a homogeneous Poisson process. The waiting time T until the next appearance of an object at location $\mathbf{x}$ thus has an exponential distribution with the appearance rate $\lambda(\mathbf{x})$. The probability of no appearance after having waited for time t is

$$p(T > t, \mathbf{x}) = e^{-\lambda(\mathbf{x})t}$$

The chance of an activity (one or more appearances) h_t at location $\mathbf{x}$ since the last observation $t_0(\mathbf{x})$ is thus

$$p(T < (t - t_0(\mathbf{x})), \mathbf{x}) = p(h_t, \mathbf{x}) = 1 - e^{-\lambda(\mathbf{x})(t - t_0(\mathbf{x}))}$$

Assuming that all probabilities of appearance at locations $\mathbf{x}$ are independent, the probability of making an observation in a given view $\mathcal{F}_t$ is:

$$\begin{aligned} p(h_t|\mathcal{F}_t) &= 1 - \prod_{\mathbf{x}\in\mathcal{F}_t} 1 - p(h_t, \mathbf{x}) \\ &= 1 - \prod_{\mathbf{x}\in\mathcal{F}_t} e^{-\lambda(\mathbf{x})(t - t_0(\mathbf{x}))} \\ &= 1 - \exp(-\sum_{\mathbf{x}\in\mathcal{F}_t} \lambda(\mathbf{x})(t - t_0(\mathbf{x}))) \end{aligned}$$

An object is detected by the system if it is in the field of view $\mathcal{F}$ of the camera. Once an object is detected, the object is tracked by a Kalman filter. If the object leaves the scene (out of the maximum field of view of the camera, or beyond a certain region of interest in the scene), the Kalman filter is stopped and the object is not considered any more.

We seek to reduce the uncertainty in our scene model as much as possible at each time step. As in the previous section, we take the entropy as a natural measure of uncertainty. Under the assumption of independence of objects, this entropy reduces to the sum of the entropy terms of all objects, which comprises of the set T of tracked objects and one which has potentially appeared and remained undetected:

$$H = \sum_{k=1}^{|T|} H_{tracked}(\mathbf{y}_{k,t}) + H_{untracked}(\mathbf{y}_{k+1,t})$$

Whereas we denote the state of object k at time t as $\mathbf{y}_{k,t}$.

In this formulation, only one previously undetected object can appear. The entropy of such an undetected object depends on the probability $p(h_t|\mathcal{F})$ of an appearance at

time t in field of view $\mathcal{F}$:

$$H_{untracked}(\mathbf{y}_{k+1,t}) = p(h_t|\mathcal{F})H_{tracked}(\mathbf{y}_{k+1,t_0}) + (1 - p(h_t|\mathcal{F}))H_u$$

The entropy $H_{tracked}(\mathbf{y}_{k+1,t_0})$ is the entropy of the object after instantiation of a new tracker, whereas H_u is a constant equal to the uncertainty in the state of the undetected object. It can be interpreted as the entropy of a uniform distribution of the object being in the now unsupervised areas or not having appeared yet - in practise, it is set to be the logarithm of ten times the covariance of the uninitialised tracker.

Let $p_{k,t} = p(h_t|\mathcal{F}_t)$ the probability of an observation of an object k at time t and correspondingly $H_{tr,k,t} = H_{tracked}(\mathbf{y}_{k,t}|\mathbf{o}_{k,t}, \mathcal{F}_t)$ be the entropy of a tracked object. Disregarding missed detections, we assume that at the current time t the probability $p_{k+1,t}$ of finding a new object $k+1$ in the current field of view is 0, because we would have found them in the previous observation otherwise. The relative entropy to the next time-step is thus

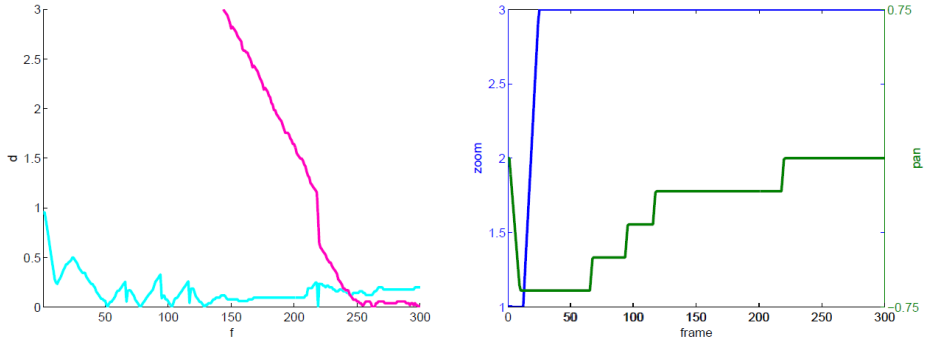
$$\begin{aligned} H_{t+1} - H_t &= \sum_{k=1}^{|T|}(H_{tr,k,t+1} - H_{tr,k,t}) \\ &\quad + p_{k+1,t+1}H_{tr,k+1,t_0} \\ &\quad + (1 - p_{k+1,t+1})H_u - (1 - p_{k+1,t})H_u \\ \\ &= \sum_{k=1}^{|T|}(H_{tr,k,t+1} - H_{tr,k,t}) \\ &\quad + p_{k+1,t+1}(H_{tr,k+1,t_0} - H_u) \end{aligned}$$

If all known targets are sufficiently well tracked, a view $\mathcal{F}_t$ is chosen which reduces the entropy by tracking a new object, weighted by the chance of its appearance. This view can exclude other, already tracked objects - the uncertainty in their position rises, increasing their entropy $H_{tr,k,t+1}$.

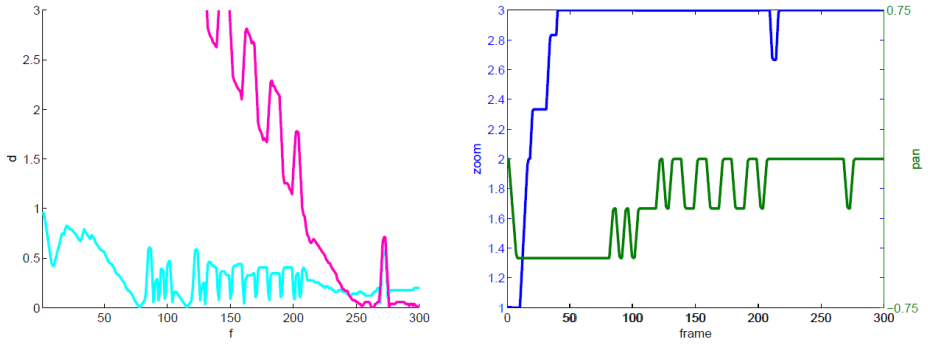
Figure 3 shows plots of the current zoom setting per frame and the horizontal distance of actors to the camera's focus normalised to the current width of the visible region. Any distance below 1 indicates a visible object. The figure shows the effect of the Poisson process on the detection of the second actor in the HERMES sequences. If appearance is not modelled, the camera simply follows the object by adjusting the pan setting accordingly. In the latter case, the camera scans the area of the currently tracked object while there is sufficient certainty about its next position. This local scanning behaviour results in an earlier detection of the second object.

3.3 Multi- vs single target tracking

As outlined in the previous section, the requirement to optimally observe all targets results in minimising the entropy of all targets at the same time. The behaviour will be fairly predictable, mainly a concentration on the detected targets and limited exploration



(a) Appearance not modelled



(b) One appearance per minute, 25fps

Fig. 3. Left: Horizontal distance of actors to the camera's focus, right: zoom and pan setting per frame. (b) shows the earlier detection of an actor due to the Poisson process.

of the scene. It might be more sensible to direct attention on a new target first, then move to the old one to confirm its position, and thenceforth supervise both of them at the same time.

To address the concept of novelty a new target introduces, or the importance of a target which has not been under scrutiny for a longer period of time, the best action is the minimum relative entropy to the next action selection step

$$\mathbf{a}_t^* = \arg\min_{\mathbf{a}_t} \Delta H$$

Whereas $\Delta H = H_t(\mathbf{x}_t|\mathbf{o}_t, \mathbf{a}_t) - H_{t-1}(\mathbf{x}_{t-1}|\mathbf{o}_{t-1}, \mathbf{a}_{t-1})$ is the entropy that is reduced in one time step. Furthermore, the choice of the objects to be tracked is considered. Instead of focusing onto a single target, any subset Ω of the currently tracked targets might yield the best decrease of uncertainty

$$\mathbf{a}_t^* = \arg\min_{\mathbf{a}_t, \Omega} \sum_{\Omega} \Delta H$$

This gives us three different choices of policies. The minimum joint entropy of all targets is simply the sum of all single entropies. The minimum relative entropy chooses the target which yields the greatest overall gain in information. The third choice is the extension of the latter to a subset of targets.

3.4 Modelling scene activity

To emphasise the differences from policies for scheduling we deliberately choose a simple approach to scene activity. More elaborate approaches could be used, e.g. Boiman and Irani's approach to saliency in video [4], or the visual attention proposed by Itti *et al.*[12]. For our approach we use the number of detections in a given image area, since in most scenes there are areas where less events of interest will occur, e.g. the appearances and disappearances of pedestrians are limited by walls, or parts of a camera's view can be blocked. We learn these entry points from longtime observations by modelling the appearance rate $\lambda(\mathbf{x})$ for every scene point $\mathbf{x}$. The appearance rate is trivially obtained as the average over all detected appearances for each pixel. An example of such an appearance map is shown for all CAVIAR sequences in figure 4. We then use this map as input to the Poisson process described in section 3.2.

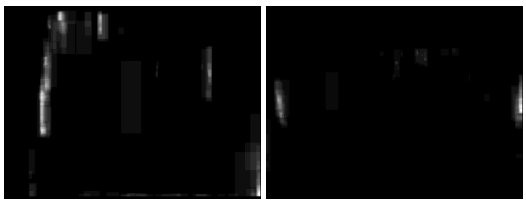


Fig. 4. Appearance in corridor and frontal view of the Caviar data set.

4 Evaluation

Evaluating scheduling algorithms on live video data is difficult. For a fair comparison, each algorithm should run on the same input, which is difficult to obtain with human actors. Pre-recorded video can be used for evaluation if the resolution is high enough to support a "virtual zoom" approach, where the image is down-sampled and cropped to a desired field of view. A high resolution is required if object detectors or trackers are to be run on the down-sampled image.

The approach used in this paper is simulation based on ground-truth data. We use the annotations supplied with the EC-funded CAVIAR test case scenarios [1] and add Gaussian noise of 1 pixel to the labelled bounding boxes. This also removes other sources of error in the evaluation, e.g. from detection, tracking and and data association. Each

[1] EC Funded CAVIAR project/IST 2001 37540, found at URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

detection is assigned a Kalman filter, which is used to obtain the uncertainty of the tracking as described in section 3.1. The Kalman filter uses an observation noise of 1 pixel, and process noise of 0.05 units (both for 1σ). A track is lost if for more than 10 frames no observation has been made, the target leaves the maximum field of view, or the expected measurement does not overlap with the actual measurement.

4.1 Metrics

The metrics we use are the latency, the fragmentation of a track, and the overall coverage of all tracks compared to the ground truth.

The latency is a measure for the delay of the detection of a target in the scene, e.g. when the camera is currently zoomed onto another target.

A track is a list of continuous observations, either from ground truth, or tracked by the Kalman filter. The average spatial coverage of a track is the relative overlap of the ground truth and observed bounding boxes, as introduced in [22]. This metric is less than one if the camera is not constantly observing the object. For example, if an object is seen only during half of the time it resides in the scene, its average spatial coverage would be 0.5.

The fragmentation of tracks into several system tracks due to tracking loss is measured by the number of false positives (FP) and false negatives (FN) of the track association. A track is considered a false positive if the average spatial coverage is above a threshold (here: 0.25), but the temporal overlap is too small (here: 0.16) . A track is considered a false negative if it is overlapping either spatially or temporally below given thresholds, and a true positive if it fulfils both criteria.

The overall coverage is the relative increase of object area due to zooming. This metric measures the average observed area, relative to the ground truth value. Successfully observing the whole scene with a zoom setting of 2 would result in an overall coverage of 2.

4.2 Experiments

We made two experiments, one with a constant appearance rate for every pixel, and one where the appearance rate has been determined a priori from the data set (see figure 4). The experiments evaluated the performance of the entropy minimisation scheduling for single, all and a subset of targets with a maximum number of 3 targets. If a further target was within the bounding box spanned by the bounding number of targets in the subset, it is added to the evaluation.

We finally compared the algorithms with standard rule based scheduling methods, i.e. random selection of targets and the first come, first serve rule (FCFS) as used by [17].

The first experiment assumes a minimum zoom setting which allows to observe the whole scene. Such devices are called 'virtual zoom' cameras, who simply downsample or resize image regions from a sensor with a higher resolution. Another example for such an input is high definition video, which can be processed much faster by restricting analysis to the relevant parts of the image[3]. In this experiment, we set the maximum

zoom to 3. The average activity of the scene pixels λ has been chosen as one appearance every second, 5 seconds, every minute or 5 minutes, respectively.

As can be seen from figure 5(a), the entropy based scheduling methods result in more tracks which are not assigned to any of the ground truth tracks, i.e. a higher fragmentation of tracks. The advantage of the methods presented is shown in figure 5(c), the entropy based scheduling methods result in a better coverage of the targets. The methods barely differ in the latency (5(b)).

λ	Subset	Sum	Single
1/60	4.3	4.3	3.7
5/60	4.3	4.6	4.0
1	4.0	5.9	3.7
5	4.3	4.9	2.8

(a) False positives (FP) in percent of the ground truth tracks(324). FCFS: 3.4, random: 4.3

λ	Subset	Sum	Single
1/60	4.9	4.8	4.8
5/60	5.0	4.9	4.8
1	5.1	4.9	4.9
5	5.0	5.0	4.8

(b) Latency in frames (at 25fps). FCFS latency: 4.9, random: 5.1

λ	Subset	Sum	Single
1/60	1.7	2.2	1.7
5/60	1.6	2.1	1.6
1	1.6	1.7	1.6
5	1.6	1.5	1.6

(c) Observation area relative to ground truth. FCFS: 1.1, Random: 1.4

The second experiment compared the different scheduling methods for a minimum zoom value of 2 and a maximum of 4. The result is shown in figure 5. In addition to the standard methods of scanning ('scan'), random target selection ('rnd') and first–come–first–serve ('fcfs'), we added a background–only policy ('bg'), which results from searching for a new target only, without taking any tracking based utility into account. This last method, as well as the methods described in section 3.3 – all ('a'), single('s'), subset ('o') – have been evaluated with and without local scene activity (label augmented with '+p' in the latter case). The poor performance of scanning, background-only, random and FCFS is easily explained. We made two experiments, one with a constant appearance rate for every pixel, and one where the appearance rate has been determined a priori from the data set (see 4). The first two methods simply scan the parameters in a more or less sensible fashion. They do not react to detected targets at all. FCFS profits from the tie-breaking rule of observing the oldest target next, but both FCFS and the random rule fixate onto an object only for a fixed time, not considering the state of the objects already visited or the duration the rest of the scene has been without observation. Apparent is the increase of the overall coverage when using the learned appearance rates. The points of high activity are more often visited than the less active areas of the scene.

5 Conclusion and Future Work

This paper presented a method of scene exploration combined with zoom control. We extended the information theoretic framework pioneered by Denzler *et al.*[7]. Here, the choice of zoom, pan and tilt settings is driven by the maximal expected decrease in uncertainty augmented by the likelihood of making an observation. To control the exploration of the scene, we added the uncertainty of a potential, yet unobserved target

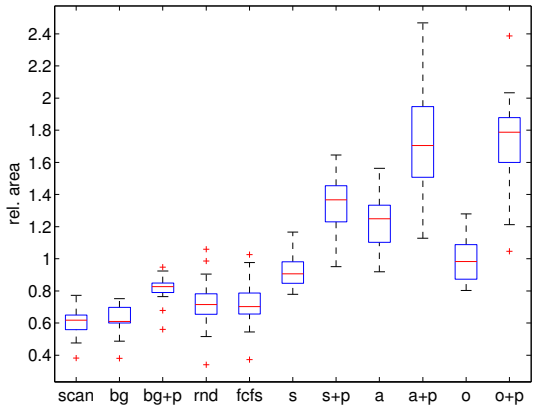


Fig. 5. Box plot of area covered by different scheduling methods on CAVIAR data set. Measured area is relative to unzoomed ground truth. See text for explanation of labels.

to this criterion. The chance of an appearance of a target is modelled by local Poisson processes, the probability of making an observation thus rises with the time not having observed this location. This acts as a counterbalance to the zoom-in behaviour, the target is being tracked while its surrounding area is maximally covered by observations. The zoom onto the current target is disfavoured once the expected decrease in uncertainty is higher for a new, potential target which has not yet been detected. We extended this reasoning to multiple targets. Here, the potential acquisition of a new target must provide more information than a subset of targets which can be observed simultaneously.

We evaluated the performance of this scheduling policy with respect to existing and new metrics. These were in particular the analysis of latency of the target detection, the increase of observed area, and the number of missed targets. The results on two publicly available data sets show that the camera control works well for two different camera systems: virtual zoom cameras and traditional pan-tilt units.

However, several shortcomings of the current method will be focus of our attention in future work. The assumption of independence of the random processes governing appearance at location $\mathbf{x}$ is not correct. This dependency can be approximated by finding typical trajectories in a scene e.g. [13]. Furthermore, the simplifying assumption that the targets are independent leads to difficulties when targets are overlapping. Our future research aims to address this issue by including the dependency into the entropy framework, thus reacting accordingly when the objects are getting closer.

Also the scene activity approach in section 3.4 lends itself to a temporal extension – often appearance rates vary with time, for example, people flock before the box office opens, appear at a restaurant at given times because they have common meal times, or cars gridlock during rush hours. This requires not only a spatially, but also temporally varying appearance rate, i.e. a non-homogeneous Poisson process. Reliably learning

such behaviour requires scene observation over a longer period of time and will be addressed in our future work.

Lastly, we do not incorporate any movement cost into the camera parameter selection process. A change of zoom by one motor step is considered equally fast as a pan and tilt across the whole field of view. This can lead to abrupt behaviour and sub-optimal paths when the parameter selection process is myopic, i.e. a greedy, one step look-ahead. We are therefore looking into methods to efficiently solve for multi-step plans to select the camera parameters [19], or to use reinforcement learning in a dynamic context, extending the [15] into the temporal domain.

Acknowledgements

The authors gratefully acknowledge support by EC grant IST-027110 for the HERMES project in the EU sixth framework programme.

References

1. Andrew D. Bagdanov, Alberto D. Bimbo, and Federico Pernici. Acquisition of high-resolution images through on-line saccade sequence planning. In *VSSN '05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 121–130, New York, NY, USA, 2005. ACM. II
2. Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and data association*, volume 179 of *Mathematics in Science and Engineering*. Academic Press Professional, Inc., San Diego, CA, USA, 1987. IV
3. Faisal Bashir and Fatih Porikli. Collaborative tracking of objects in EPTZ cameras. In Chang W. Chen, Dan Schonfeld, and Jiebo Luo, editors, *Visual Communications and Image Processing 2007*, volume 6508. SPIE, 2007. X
4. Oren Boiman and Michal Irani. Detecting irregularities in images and in video. In *10th IEEE International Conference on Computer Vision (ICCV 2005), Beijing, China*, 2005. II, IX
5. Cash J. Costello, Christopher P. Diehl, Amit Banerjee, and Hesky Fisher. Scheduling an active camera to observe people. In *VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*, pages 39–45, New York, NY, USA, 2004. ACM Press. II
6. J. Davis, A. Morison, and D. Woods. An adaptive focus-of-attention model for video surveillance and monitoring. *Machine Vision and Applications*, 18(1):41–64, February 2007. II
7. Joachim Denzler, Matthias Zobel, and Heinrich Niemann. Information theoretic focal length selection for real-time active 3-d object tracking. In *9th IEEE International Conference on Computer Vision*, pages 400–407. IEEE Computer Society, 2003. II, III, XI
8. U. M. Erdem and S. Sclaroff. Look there! predicting where to look for motion in an active camera network. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2005)*, pages 105–110, 2005. II
9. Stephen Gould, Joakim Arfvidsson, Adrian Kaehler, Benjamin Sapp, Marius Meissner, Gary Bradski, Paul Baumstarck, Sukwon Chung, and Andrew Y. Ng. Peripheral-foveal vision for real-time object recognition and tracking in video. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007. III
10. Erdan Gu, Jingbin Wang, and Norman I. Badler. Generating sequence of eye fixations using decision-theoretic attention model. In *Proceedings of WAPCV 2007*, volume 4840 of *LNAI*, pages 277–292. Springer, 2007. III

11. Arun Hampapur, Sharat Pankanti, Andrew Senior, Ying-Li Tian, Lisa Brown, and Ruud Bolle. Face cataloger: Multi-scale imaging for relating identity to location. In *AVSS '03: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, Washington, DC, USA, 2003. IEEE Computer Society. II
12. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. IX
13. Neil Johnson and David Hogg. Learning the distribution of object trajectories for event recognition. In *BMVC '95: Proceedings of the 6th British conference on Machine vision (Vol. 2)*, pages 583–592, Surrey, UK, UK, 1995. BMVA Press. II, XII
14. D. Makris and T. Ellis. Automatic learning of an activity-based semantic scene model. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 183–188, 2003. II
15. Lucas Paletta and Gerald Fritz. Reinforcement learning for decision making in sequential visual attention. In *Proceedings of WAPCV 2007*, volume 4840 of *LNAI*, pages 277–292. Springer, 2007. III, XIII
16. P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. Technical Report NISTIR 7408, National Institute of Standards and Technology, 2007. I
17. Faisal Z. Qureshi and Demetri Terzopoulos. Surveillance in virtual reality: System design and multi-camera control. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. II, X
18. Diego Rother, Kedar A. Patwardhan, and Guillermo Sapiro. What can casual walkers tell us about a 3D scene? In *11th IEEE International Conference on Computer Vision*. IEEE Computer Society, 2007. II
19. Nicholas Roy and Caleb Earnest. Dynamic action spaces for information gain maximization in search and exploration. In *Proceedings of the American Control Conference (ACC 2006)*, Minneapolis, USA, 2006. IEEE. XIII
20. C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000. II
21. Ben Tordoff and David Murray. Resolution vs. tracking error: zoom as a gain controller. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Madison, Wisconsin*. IEEE Computer Society Press, 2003. II
22. Fei Yin, D. Makris, and S. A. Velastin. Performance evaluation of object tracking algorithms. In *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2007), Rio de Janeiro, Brazil*, October 2007. X