Information-theoretic Active Scene Exploration

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

Abstract

Studies support the need for high resolution imagery to identify persons in surveillance videos[13]. However, the use of telephoto lenses sacrifices a wider field of view and thereby increases the uncertainty of other, possibly more interesting events in the scene. Using zoom lenses offers the possibility of enjoying the benefits of both wide field of view and high resolution, but not simultaneously. We approach this problem of balancing these finite imaging resources – or of exploration vs exploitation – using an information-theoretic approach. We argue that the camera parameters – pan, tilt and zoom – should be set to maximise information gain, or equivalently minimising conditional entropy of the scene model, comprised of multiple targets and a yet unobserved one. The information content of the former is supplied directly by the uncertainties computed using a Kalman Filter tracker, while the latter is modelled using a "background" Poisson process whose parameters are learned from extended scene observations; together these yield an entropy for the scene. We support our argument with quantitative and qualitative analyses in simulated and real-world 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 [13] 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 obviates the restriction to a single camera 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 all, or a subset of, 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, which 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,

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

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*. [14] used a first come, first serve rule in their simulator, based on evaluations of network routing algorithms [4]. Contrary to our work, they 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*. [10] 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 [7, 14, 15].

Probabilistic reasoning for camera zoom control is used by Tordoff *et al*. and Denzler *et al*. [6, 18] 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 [11, 12, 17]. We would like to stress that this paper is neither about a new tracking or scene activity analysis method. Instead, our concern is how these results can be used for camera scheduling. Davis *et al*. [5] 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.

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*. [6]. 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

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

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

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

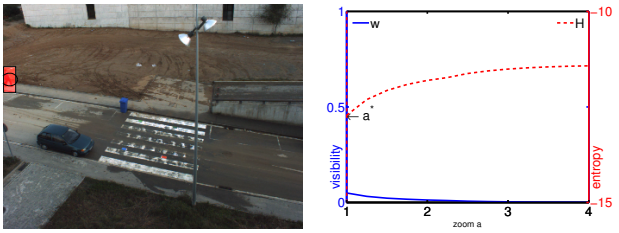
Since the entropies of the state estimate $H(\hat{\mathbf{x}}_t^+)$, $H(\hat{\mathbf{x}}_t^-)$ do not depend on the actual observation $\mathbf{o}$, this integral can be simplified to

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

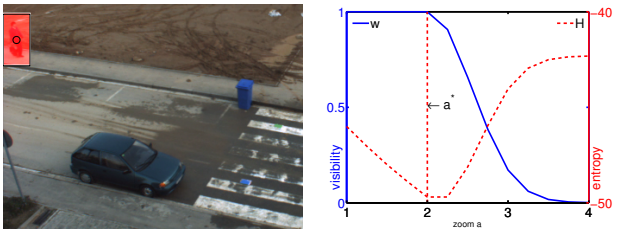
The factor $w(\mathbf{a}_t) = \int_v p(\mathbf{o}_t|\mathbf{a}_t)\, d\mathbf{o}_t$ expresses the probability of making an observation of the object in the image. We simplify the evaluation of this integral by making use of the error function for the integral over the Gaussian distribution of the position,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

We give an example of the resulting behaviour on a sequence from the HERMES Outdoor dataset, camera 1[8]. Figure 1 shows the temporal development of the zoom selection process, with images corresponding to the resulting zoom settings on the left. The right column shows the probability of making an observation, w, and the expected entropy, H, over a given zoom range (1x to 4x). Frame 442 (1a) shows the full view of the scene right after initialisation of a Kalman filter on a newly detected target, and the 1σ covariance ellipse of the location. Due to this high initial uncertainty and its proximity to the edge of the field of view, the probability of making an observation is low, but highest for the smallest zoom setting. In this case it is "1x", as indicated by the label $\mathbf{a}^*$ in the figure. For this frame, zoom selection is governed by the chance of making a visibility. Five frames later, at frame 447 (1b), the position is more certain since the target has been reliably tracked for a short time, as shown by the smaller covariance ellipse. The confidence in making an observation has increased, to a constant maximum of one up to the zoom value of 2x, where it slowly drops. There, the entropy also rises. The minimum of the entropy effectively defines the highest zoom setting which does not risk a loss of the target. A similar behaviour is shown in the last row (1c), which portrays frame 458 of the sequence. Here, the camera started panning to follow the object.

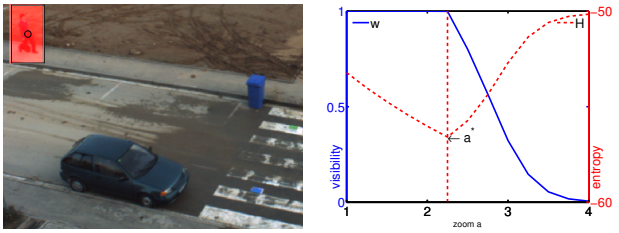
The development of the entropy for a single target and its visibility is shown for the first 20 frames after the initial detection in figure 2.



(a) Frame 442



(b) Frame 447



(c) Frame 458

Figure 1. Visibility term w and entropy $H(\mathbf{x}|\mathbf{o})$ (see equation 5) 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.

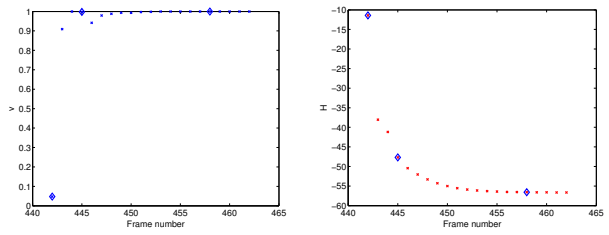


Figure 2. Visibility term w (left) and entropy $H(\mathbf{x}|\mathbf{o})$ (right) for consecutive frames in the HERMES sequence. The frames 442, 447 and 458, detailed in figure 1, have been highlighted.

3.2. Poisson process for unobserved events

Imagine a street scene where pedestrians show up regularly, but unpredictably. The street has a lot of houses and doors - nearly everywhere a person can appear. The absolute times of two appearances are independent random variables - the number of appearances before an occurrence is independent of the number of the following ones.

In the following discussion, an appearance of an object

is modelled by a homogeneous Poisson process for every location at which it can show up. The waiting time T until the next appearance of an object at location $\mathbf{y}$ thus has an exponential distribution with the appearance rate $\lambda(\mathbf{y})$. The probability of no appearance after having waited for time t is

$$p(T > t, \mathbf{y}) = e^{-\lambda(\mathbf{y})t} \tag{6}$$

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

$$p(T < (t - t_0(\mathbf{y})), \mathbf{y}) = p(h_t, \mathbf{y}) = 1 - e^{-\lambda(\mathbf{y})(t-t_0(\mathbf{y}))} \tag{7}$$

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

$$\begin{aligned} p(h_t|\mathcal{F}_t) &= 1 - \prod_{\mathbf{y}\in\mathcal{F}_t} 1 - p(h_t, \mathbf{y}) && (8)\\ &= 1 - \prod_{\mathbf{y}\in\mathcal{F}_t} e^{-\lambda(\mathbf{y})(t-t_0(\mathbf{y}))} && (9)\\ &= 1 - \exp(-\sum_{\mathbf{y}\in\mathcal{F}_t} \lambda(\mathbf{y})(t - t_0(\mathbf{y}))) && (10) \end{aligned}$$

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_u):

$$H = H_u(\mathbf{x}_{k+1,t}) + \sum_{k=1}^{|T|} H_{tracked}(\mathbf{x}_{k,t}) \tag{11}$$

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

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 $\mathcal{F}$:

$$\begin{aligned} H_u(\mathbf{x}_{k+1,t}) &= p(h_t|\mathcal{F}) H_{tracked}(\mathbf{x}_{k+1,t_0}) \\ &+ (1 - p(h_t|\mathcal{F}))\hat{H}_u \end{aligned} \tag{12}$$

The entropy $H_{tracked}(\mathbf{x}_{k+1,t_0})$ is the entropy of the object after instantiation of a new tracker, whereas $\hat{H}_u$ 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 practice, it is set to be the logarithm of ten times the covariance of the uninitialised tracker.

Let $p_{k,t} = p(h_t|\mathcal{F}_t)$ be the probability of an observation of an object k at time t and correspondingly $H_{tr,k,t} = H_{tracked}(\mathbf{x}_{k,t}|\mathbf{o}_{k,t}, \mathcal{F}_t)$ be the entropy of a tracked object. The relative entropy to the next time-step is thus

$$\begin{aligned} H_{t+1} - H_t &= \sum_{k=1}^{|T|} (H_{tr,k,t+1} - H_{tr,k,t}) \\ &+ p_{k+1,t+1} H_{tr,k+1,t_0} \\ &+ (p_{k+1,t} - p_{k+1,t+1})\hat{H}_u && (13)\\ &\approx \sum_{k=1}^{|T|} (H_{tr,k,t+1} - H_{tr,k,t}) \\ &+ p_{k+1,t+1}(H_{tr,k+1,t_0} - \hat{H}_u) && (14) \end{aligned}$$

Here, we approximate equation 13 by assuming that at current time t the probability $p_{k+1,t}$ of finding a new object $k+1$ in the current field of view is 0. This effectively disregards missed detections, but can be mitigated by overestimating the appearance rate. 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 $H_{tr,k,t+1}$.

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 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

$$\mathbf{a}_t^* = \arg\min_{\mathbf{a}_t} \Delta H \tag{15}$$

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

$$\mathbf{a}_t^* = \arg\min_{\mathbf{a}_t, \Omega} \sum_{\Omega} \Delta H \tag{16}$$

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. Learning scene activity

In most scenes there are areas where fewer 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. Whereas these areas could be specified by user input, such as entry and exit points, in this work we learn these entry points from longtime observations by modelling the appearance rate $\lambda(\mathbf{y})$ for every scene point $\mathbf{y}$. 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 in figure 3 for both viewpoints in all sequences of the EC-funded CAVIAR "shopping mall" data set [1].

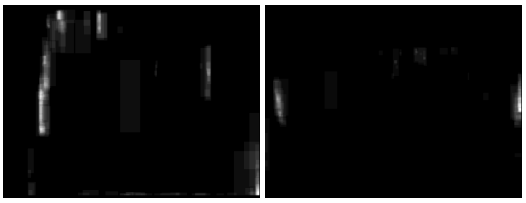


Figure 3. Appearance in corridor and frontal view of the CAVIAR "shopping mall" data set.

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 CAVIAR test case scenarios 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 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.

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

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 [19]. 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 partial trajectories 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, i.e. all contained targets, with a zoom setting of 2 would result in an overall coverage of 2. The reasoning behind this metric is that higher resolution benefits following tasks, such as identification or action recognition, a claim supported by recent studies[13].

4.2. Experiments

We made two experiments, one with a constant appearance rate for every pixel and a full overview over the scene, and one where the appearance rate has been determined a priori from the data set (see figure 3), and the camera is required to explore the scene because it can never observe it as a whole. 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. In the latter case, 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 furthermore evaluated standard rule based scheduling methods, i.e. random selection of targets and the first come, first serve rule (FCFS) as used by [14].

For the first experiment we assume a minimum zoom setting which allows to observe the whole scene. Such a setup

simulates a 'virtual zoom' camera, which simply downsamples or resizes 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 minute, every 12 and full second, or 5 appearances per second, respectively.

The number of false positives, shown in figure 4, shows that with entropy based scheduling methods more tracks are not assigned to any of the ground truth tracks, i.e. yield a higher fragmentation of tracks. This is effectively the number of targets which needed to be initialised again due to a longer focus onto other targets. The false negatives in figure 4 shows how the number of completely missed target goes down if the appearance rate is high enough. For both of these figures, the absolute number of targets in all sequences is 324. While the methods barely differ in the latency (figure 5, left), the advantage of the methods presented here can be seen in figure 5, where the entropy based scheduling methods result in a better coverage of the targets.

λ/s	Subset	All	1
1/60	14	14	12
5/60	14	15	13
1	13	19	12
5	14	16	9

(a) FP. FCFS: 11, Random: 15

λ/s	Subset	All	1
1/60	1	6	5
5/60	1	6	5
1	0	0	6
5	0	0	5

(b) FN. FCFS: 4, random: 3

Figure 4. False positives (FP): reacquired targets and False Negatives (FN): missed targets, both out of a total of 324.

λ/s	Subset	All	1
1/60	4.94	4.78	4.83
5/60	5.02	4.88	4.84
1	5.11	4.93	4.87
5	5.03	4.96	4.80

(a) FCFS: 4.90, Random: 5.11

λ/s	Subset	All	1
1/60	1.69	2.18	1.65
5/60	1.62	2.11	1.61
1	1.59	1.65	1.58
5	1.58	1.53	1.56

(b) FCFS: 1.13, Random: 1.38

Figure 5. Latency in frames (left) and Observation area relative to ground truth (right) for constant appearance rates.

The second experiment compared the algorithms with a minimum zoom value of 2 and a maximum of 4, which requires an exploration, or scanning, of the scene. The camera settings continuously have to balance the reduction in uncertainty for a few targets with the risk of missing a target in the area currently not observed. The result is shown in figure 6. 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. 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 local appearance rates. The points of high activity are more often visited than the less active areas of the scene.

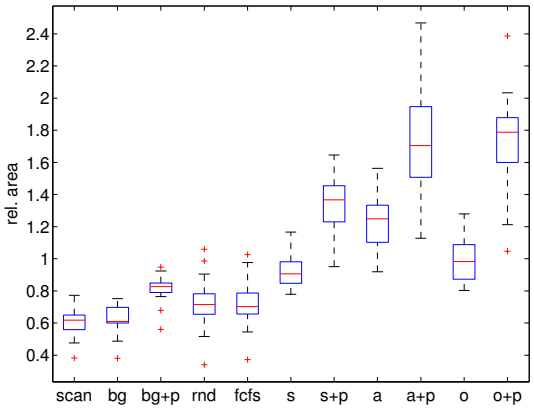


Figure 6. Second experiment. Box plot of area covered by different scheduling methods on CAVIAR data set. Measured area is relative to unzoomed ground truth. See text for explanation of labels.

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*. [6], in which 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 to this criterion. The chance of an appearance of a target is modelled by local Poisson processes; the chance of making an observation thus rises with the time passed since the last observation of this location. This acts as a counterbalance to the zoom-in behaviour, and yields behaviour in which a target is tracked while its surrounding area is maximally covered by observations.

The zoom onto a current target is discouraged 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.

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 $\mathbf{y}$ is not correct. This dependency can be approximated by finding typical trajectories in a scene (e.g. [11]). 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 approaching each other. 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 suboptimal 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[16].

Acknowledgements

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

References

[1] A. D. Bagdanov, A. D. Bimbo, and F. 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.

[2] Y. Bar-Shalom and T. E. Fortmann. *Tracking and data association*, volume 179 of *Mathematics in Science and Engineering*. Academic Press Professional, Inc., San Diego, CA, USA, 1987.

[3] F. Bashir and F. Porikli. Collaborative tracking of objects in EPTZ cameras. In C. W. Chen, D. Schonfeld, and J. Luo, editors, *Visual Communications and Image Processing 2007*, volume 6508. SPIE, 2007.

[4] C. J. Costello, C. P. Diehl, A. Banerjee, and H. 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.

[5] 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.

[6] J. Denzler, M. Zobel, and H. 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.

[7] 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.

[8] J. Gonzàlez, X. F. Roca, and J. J. Villanueva. Hermes: A research project on human sequence evaluation. In *Computational Vision and Medical Image Processing (VipIMAGE'2007)*, Porto, Portugal, 2006. IEEE.

[9] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Meissner, G. Bradski, P. Baumstarck, S. Chung, and A. 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.

[10] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. 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.

[11] N. Johnson and D. 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.

[12] 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.

[13] 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.

[14] F. Z. Qureshi and D. Terzopoulos. Surveillance in virtual reality: System design and multi-camera control. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[15] D. Rother, K. A. Patwardhan, and G. Sapiro. What can casual walkers tell us about a 3D scene? In *11th IEEE International Conference on Computer Vision*. IEEE Computer Society, 2007.

[16] N. Roy and C. 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.

[17] 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.

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

[19] F. 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.