Saccade planning with an active zooming camera

Alberto Del Bimbo Federico Pernici
Dipartimento di Sistemi e Informatica – Università di Firenze
Via Santa Marta 3, I-50139 Firenze, Italy
Email: {delbimbo,pernici}@dsi.unifi.it

***Abstract*—This paper considers the problem of designing an active observer to plan a sequence of decisions regarding what target to look at, through a foveal-sensing action. We propose a framework in which a pan/tilt/zoom (PTZ) camera schedules saccades in order to acquire high resolution images (at least one) of as many moving targets as possible before they leave the scene. An intelligent choice of the order of sensing the targets can significantly reduce the total dead-time wasted by the active camera and, consequently, its cycle time. The grabbed images provide meaningful identification imagery of distant targets which are not recognizable in a wide angle view. We cast the whole problem as a particular kind of dynamic discrete optimization. In particular, we will show that the problem can be solved by modelling the attentional gaze control as a novel on-line Dynamic Vehicle Routing Problem (DVRP) with deadlines. Moreover we also show how multi-view geometry can be used for evaluating the cost of high resolution image sensing with a PTZ camera.**
Congestion analysis experiments are reported proving the effectiveness of the solution in acquiring high resolution images of a large number of moving targets in a wide area. The evaluation was conducted with a simulation using a dual camera system in a master-slave configuration. Camera performances are also empirically tested in order to validate how the manufacturer's specification deviates from our model using an off-the-shelf PTZ camera.

I. Introduction

Equipping machines with even a limited version of our own visual abilities is proving a remarkable task. The image formation process of the eye alone can be emulated using a lens and CCD array; it is argued as a result of the 3D to 2D mapping of the real world to the visual retina that the major challenge for for machine vision is perception.
Making the biological comparison, it is widely reported that the human visual faculty is not a single comprehensive processing unit, but a series of small task-specific processors whose output can be combined. Image stabilization (fixation) occurs at a level below the brains main visual processors, having direct connections from the ear and retina to the eye muscles. Eye and head movements interact with the visual process, allowing maximum resolution to be focused on specific areas of the scene. In humans this is achieved by either repeated saccade-fixate cycles or by smooth motion tracking. So in reality, we do not scan a scene in raster fashion, our visual attention tends to jump from one point to another. These jumps are called saccade. Yarbus [1] demonstrated that the saccadic patterns depend on the visual scene as well as the cognitive task to be performed. The conclusion is that we do not see, we look [2]. In this paper the focus is visual attention according to task at hand and the scene content.
The lack of works addressing task-driven visual processing is mainly motivated by the fact that its studying seems, as a first sight, too specialized, non-generic, or bordering on hackery. But active vision demands such processes; it is founded in the idea of specialized processing for specialized tasks. Most of the active vision literature is limited to studying low-level subconscious reflexes. One wonders whether truly active and purposeful vision systems will be realized. In other words, while active tracking and visual attention was researched in the past years, purposeful zooming is (and probably will remain) a largely unexplored area in active vision [3]. Basically sensing was not a major issue for computer vision as for example was perception. However despite this for the particular task of object recognition notably works are reported in the literature [4] [5].
dire qui che active recognition is particular Our work is motivated by the goal of reproducing the ability of humans to recognize a person in a crowd of moving people for surveillance purposes. In humans, the process of recognizing a person and that of moving the eyes are served by almost two distinct subcortical brain areas: one specialized for recognizing faces and one specialized for making decisions on whom look at next. The eye acts as a foveal sensor that allows high resolution only at the point of interest, avoiding the cost of uniform high resolution. Indeed during a scan-path in a moving crowd of walking people it is normal to backtrack to a previous observed person thinking "oh that's my friend". This because the gaze planning task does not directly depend on the face recognition task. Visual attention in this particular task is more affected by the target position, the predicted time in exiting the scene and the effort made in moving the head and the eyes from one direction to another. In fact during a saccade, the redirection is so rapid that the gaze lasts only a tenth of a millisecond. During that time the few images obtained are typically blurred because of the fast camera motion. As far as the deployment in sophistication in visual analysis is concerned, saccades are dead times. So our brain avoids doing large redirection of the gaze while undertaking this task, trying to minimize that dead time.
A direct application of that behavior of the human visual system can be applied in Visual Surveillance. Automated surveillance can be a powerful tool in deterrence of crime, but most of the solutions and implementations proposed so far are unnecessarily poor in evidential quality. In this sense, remote identification of targets is and will be an important mandatory capability for modern automated surveillance systems. In particular, recognizing a person or a car license plate requires that high resolution views must be taken before they leave the scene. Using a large number of static or active cameras that operate cooperatively is an expensive and impractical solution. One way to cope with this problem is to make better use of the capabilities of the sensor.
We argue that one active pan/tilt/zoom (i.e. a foveal sensor)

camera (the slave camera) together with a wide angle camera (the master camera) and a good strategy for visiting the targets can be used instead. The fixed camera is used to monitor the scene estimating where targets are in the surveilled area. The active camera then follows each target to produce high resolution images. In this configuration, we show that the visual signal from the master camera provides the necessary information to plan the saccades sequence. Moreover, the introduction of an appropriate scheduling policy allows to maximize the number of targets that can be identified from the high resolution images collected. Indeed, this is achieved by continuously gazing at the most appropriate targets, where the appropriateness strongly depends on the task considered. In fact, tasks may have conflicting requirements, as in the case where different tasks would direct the fovea to a different point in the scene. For systems with multiple behaviors, this scheduling problem becomes increasingly paramount.

The key contributions of this part of the thesis are: (1) We propose a novel formulation for the remote target identification problem in terms of saccadic gaze planning. (2) We give a general framework in which an active camera can be modelled. (3) The use of uncalibrated methods makes the proposed framework function in any planar scene. (4) We extend previous approaches on PTZ greedy scheduling proving through simulation that our framework yields better system performance.

II. Related Work

Recent years (especially after 9/11) have seen a continued increase in the need for and use of automatic video surveillance for remote identification problems. The few works addressing this subject do not address the planning problem or do not fully exploit all the information intrinsically present in the structure of the problem. In [6] the problem of deciding which camera should be assigned to which person was addressed and some general approaches are given. It should also be noted that there is no work except [7] on objectively evaluating the performance of multi-camera systems for acquiring high resolution imagery of people. Most results are presented in the form of video examples or a series of screen captures without explicit system performance evaluations. Very little attention is given to the problem of what to do when there are more people in the scene than active cameras available.

Many works in literature uses a master/slave camera system configuration with two [8][7][9][10][11] or more cameras [12][13][14][6][15]. The remote target identification problem is also termed as distant human identification (DHID). In [8], a single person is tracked by the active camera. If multiple people are present in the scene, the person who is closest to the position of the previous tracked individual is chosen. In [7] the authors use greedy scheduling policies taken from the network packet scheduling literature. They are the first to describe the problem formally and propose a solution. In particular, in this work the authors, albeit mentioning that there is a transition cost measured in time to be paid whenever the camera switches from person to person, do not explicitly model this cost in their problem formulation. The consequence is that their analysis wrongly motivates an empirically determined watching time instead of at least a single video frame. Moreover the work uses greedy policies instead of policies with a time horizon. Also in [12] the authors propose a form of collective camera scheduling to solve surveillance tasks such as acquisition of multi-scale images of a moving target. They take into account the camera latency and model the problem as a graph weighted matching. In the paper there are no experimental results and no performance evaluation for the task of acquiring as many multi-scale images of many targets as possible in real time. In [10] another similar approach with a dual camera system was recently proposed in indoor scenes with walking people. No target scheduling was performed, targets are repeatedly zoomed to acquire facial images by a supervised learning approach driven by skin, motion and foreground features detection. In [16] a ceiling mounted panoramic camera provides wide-field plan-view sensing and a narrow-field pan/tilt/zoom camera at head height provides high-resolution facial images. The works in [17][9] concentrate on active tracking. In both works the respective authors propose a simple behavior (a policy) with a finite state machine in order to give some form of continuity when the currently tracked target is changed. In [13] two calibration methods to steer a PTZ camera to follow targets tracked by another camera are proposed. The authors give some criteria of optimization leaving the formal optimization as future research. Though performing coarse registration the methods [13] and [8], generally suffice to bring the target object within a narrow zoomed field of view.

Another body of literature, concerning the mathematical optimization framework, comes from the motion planning literature and in particular from the context of rapid deployment automation. Specifically, those problems related to rearranging parts by a robot in an industrial assembly line setting. A representative work in this context is [18]. In that work the problem is: given n identical parts initially located on a conveyer belt, and a robot arm of capacity k parts, compute the shortest route for the robot arm to grasp and deliver the parts, handling at most at k a time. A PTZ-camera can be interpreted as a robot arm, we will use such analogy in our problem formulation.

The other important work related to our problem is [19], in which the authors study the problem in which a vehicle moves from point to point (customers) in a metric space with constant speed, and at any moment a request for service can arrive at a point in the space. The objective is to maximize the number of served customers. They analyze several policies showing that in such a problem lower bounds on system performance can be obtained analytically. This work is reminiscent of our problem, the main differences are that our customers (targets) are moving and have deadlines. A further important difference is that the nature of our particular vehicle (a PTZ-camera) does not allow us to model the cost of moving from target to target in the euclidean space.

III. Problem Formulation

In this section we formulate and discuss the three main features that characterize this problem: targets motion, arrivals as a continuous process, and deadlines. Once a subset of moving target is selected the correct camera tour can be optimized as a Kinetic Travelling Salesperson Problem (KTSP). The problem of how choosing the best permutation subset from the currently tracked targets is an instance of the Time Dependent Orienteering (TDO) with deadlines.

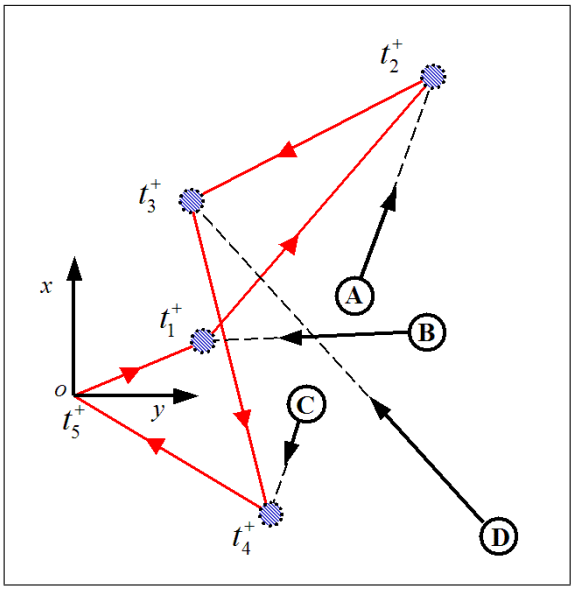


Fig. 1. An instance of Kinetic-TSP with four targets. The shortest-time tour (light line).

A. Kinetic Travelling Salesperson Problem

As cameras can be calibrated with automatic or manual methods such as in [13] it is possible to associate to each point in the plane where targets are moving a vector of PTZ-camera parameters. According to this, at each point in the world plane it is possible to issue camera commands in order to bring a moving target in a close up view by giving to the camera the 3D vector (p, t, z), specifying pan, tilt and zoom values to be applied. In our formulation we model the PTZ-camera as an intercepter with restricted resources (e.g., limited speed in setting its parameter). The dynamics of the targets are assumed known or predictable (i.e., for each target one can specify its location at any time instant). The problem is expressed as that of finding a policy for the PTZ-camera which allows to "visually hit" (with a saccade sequence) as many targets as possible in accordance with the device speed. This allows to cast the problem as a Kinetic Travelling Salesperson problem (KTSP) [20]. In fig.1 are shown four targets A, B, C, D moving on a plane. The shortest-time tour is shown with the respective interception points. At each interception point is also shown the time instants of the sequence when the intercepter visually hits the targets. Formally this problem is formulated as follow:

KTSP : *Given a set $S = \{s_1, s_2, ..., s_n\}$ of moving targets, each s_i moving with known or predictable motion $x_i(t)$, and given an active camera intercepter starting at a given position and having maximum speed $V_{ptz} \geq V_i \; \forall i$, find the shortest-time tour starting (and ending) at the origin, which intercepts all targets. V_i indicates the imaged speed of target i and V_{ptz} indicates the maximum speeds of the pan-tilt-zoom device. The solution is defined as the permutation of the discrete set S that has the shortest travel time.*

It is necessary that the intercepter run faster than the targets. This is not generally a problem even for slower PTZ-cameras. By imagining the PTZ-camera as a robot manipulator with two revolute (pan-tilt) and one prismatic (zoom) joint, it is possible to view the principal axis of the camera as a robot arm which rotates and move forward to reach a point in the space. In such settings, due to the typically high distance at which PTZ-cameras are mounted, the speeds of the virtual end-effector are generally higher than common moving targets such as cars or humans.

B. Time Dependent Orienteering (TDO)

In a typical surveillance application, targets arrive as a continuous process, so that we must collect "demands to observe", plan tours to observe targets, and finally dispatch the PTZ camera. In a such dynamic-stochastic setting there is a lot of interdependency between the state variables describing the system. Moreover, tours must be planned while existing targets move or leave the scene, and/or new targets arrive. Basically the whole problem can be viewed as a global dynamic optimization. Since for such a problem no a-priori solution can be found, an effective approach is to determine a strategy to specify the actions to be taken as a function of the state of the system. In practice, we consider the whole stochastic-dynamic problem as a series of deterministic-static subproblems, with the overall goal of tracking the time progression of the objective function as close as possible. In our problem, targets are assumed to enter the scene at any time from a finite set of locations. The camera must steer its foveal sensor to observe any target before it leaves the scene. Assuming with no loss of generality that the paths of the targets are straight lines and that targets move at constant speeds, the time by which a target must be observed by the camera can be estimated. Moreover, real-time constraints may impose bounds on the total amount of time needed to plan the target observation tour. According to this, given a fixed reference time, KTSP can be reformulated as a Time Dependent Orienteering (TDO) problem [21]. In the classical formulation of the static orienteering problem there is a resource constraint on the length of the tour; the problem solution is the one that maximizes the number of sites visited. The time dependent orienteering problem for a single PTZ-camera can be formulated as follows:

TDO : *Given a set $S = \{s_1, s_2, ..., s_n\}$ of moving targets, each s_i moving with a known or predictable motion $x_i(t)$, the deadline t, and a time-travel function $l : S \times S \times N \longmapsto I\!R^+ \cup \{0\}$ the salesperson's tour to intercept a subset $T = \{s_1, s_2, ..., s_m\}$ of m targets is a sequence of triples: $(s_1, t_1^+, t_1^-), (s_2, t_2^+, t_2^-), ...(s_m, t_m^+, t_m^-)$, such that: for $i \in \{1, 2, ..., m\}$, $t_i^+, t_i^- \in N \cup \{0\}$ with $0 = t_1^+ \leq t_1^- \leq t_2^+ \leq ... \leq t_m^+ \leq t_m^- \leq t$. The subset T is composed by the maximum number of targets interceptable within the time t, imposed by the real-time constraint.*

Orienteering problems are classified as path-orienteering or cycle-orienteering problems depending on whether the network to be induced by the set of pairs of consecutive targets visited is supposed take the form of a path or of a cycle, respectively. The deadline t breaks the dynamic problem into a sequence of static problems. Such a formulation has a great advantage which is computationally helpful. Since there is no polynomial time algorithms to solve the KTSP, it is impossible to solve an instance of the KTSP problem with more than eight or nine targets in a fraction of a second, by the exhaustive search. However even

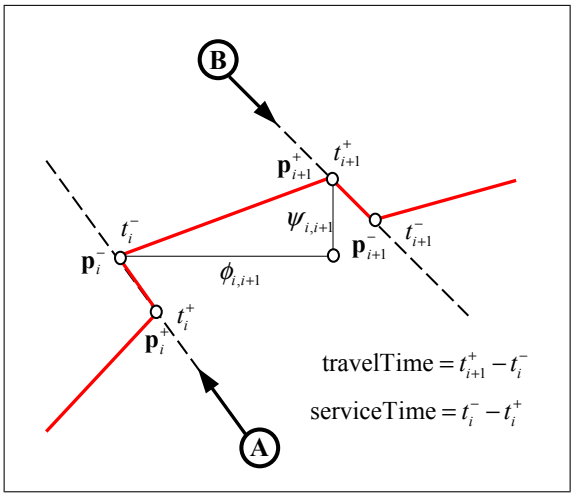


Fig. 2. A symbolic scheme representing a saccade from the target A to the target B. The $\phi_{i,i+1}$, $\psi_{i,i+1}$ are respectively the pan and tilt angles as seen from the slave camera when the camera leaves target A at time t_i^- and intercepts B at time t_{i+1}^+.

if such an algorithm did exist the time needed to switch to all the targets would be so large that novel targets would not be observed due to the time needed to complete the tour. *So, the brute force approach enumerating and evaluating all the subsets permutations perfectly fits with the nature of our dynamic incremental formulation.*

C. Deadlines

Based on the tracking predictions targets are put in a queue, according to their residual time to exit the scene. TDO is instantiated for the first k targets in the queue. If $\mathcal{A}_k$ is the set of the permutations of the subsets of k targets then it can be shown that:

$$|\mathcal{A}_k| = \sum_{i=0}^{k} \frac{k!}{(k-i)!} \tag{1}$$

where $|\mathcal{A}_k|$ is the cardinality of the set $\mathcal{A}_k$. So for example with a queue of $k = 7$ targets we have $|\mathcal{A}_7| = 13700$. In this case the exhaustive enumeration requires 13700 solutions evaluations. As remarked in the previous section, solutions with a large number of scheduled targets would not be practical for an incremental solution, since the time needed to switch to all the targets would be so large that novel targets would not be observed due to the time needed to complete the tour.

The framework is fairly general and more elaborated policies can be estimated by changing optimization cost and/or the sorting used in the queue (priority in the queue can be specified according to some combined quality measure of the imagery of the targets, for example preferring targets moving in certain specified directions). Here we want to maximize the number of targets taken at high resolution. With the deadlines the TDO becomes a constrained combinatorial optimization, where the feasible set can be defined as follow (see the TDO definition in the previous section):

$$t_i^- < t_i^d \quad , \quad \forall i = 1..|T| \tag{2}$$

Where $T \in \mathcal{A}_k$ is an instance of the permutations of the subsets, and t_i^d is the deadline for the target at position i in T. That means the the camera must leaves the target i in T at time t_i^- before the target leaves the scene at time t_i^d.

The TDO solution is calculated by assuming a constant speed for the pan-tilt-zoom camera motors as specified by the manufacturer. There is no need for an exact specification of these speeds, in that they are used only for the prediction of the cost of the saccadic sequences. In order to keep the computation tractable the number of target in the queue k should not be greater than 8 (9 with optimized code). For example on a Pentium IV 2.0 GHz running Matlab, computing and evaluating the permutations of the subsets of 8 targets takes a fraction of a second.

IV. Saccades Planning Geometry

In order to show the advantages of adopting this framework for our research objective, we consider the classic camera system in a master/slave configuration [8][7]. In this configuration a static, wide field of view master camera is used to monitor a wide area and track the moving targets providing the position information to the foveal camera. The foveal camera is used to observe the targets at high resolution. We estimate the interception times of a target for each of the three foveal camera control signals (respectively t_ϕ, t_ψ, t_z for pan, tilt, zoom). Since the effects of the three control signals are independent from each other (i.e. the pan motor operates independently from the tilt motor) the time needed to conclude a saccade is dominated by the largest one. The largest time is taken as the time spent by the foveal camera to observe the target and is taken into account to derive the overall time needed to complete the tour in the TDO formulation.

With reference to fig.2 the estimated t_ϕ, t_ψ, t_z are assumed as the times needed to make the foveal camera gaze at the target at position $i+1$, leaving the target at position i in the sequence $S = \{s_1, ..., s_i, s_{i+1}, ..., s_m\}$ (in fig.2 the targets at position i and $i+1$ are respectively indicated as $\mathbf{A}$ and $\mathbf{B}$). In other words they represent the times needed for changing the pan and tilt angles and zoom respectively by $\phi_{i,i+1}$, $\psi_{i,i+1}$ and $z_{i,i+1}$ (not shown in the figure) in order to intercept the new target at time t_{i+1}^+ while leaving the old target at time t_i^-. The time $t^\star = \max\{t_{\phi_{i,i+1}}, t_{\psi_{i,i+1}}, t_{z_{i,i+1}}\}$ is the travel time needed to change the gaze.

By assuming targets moving on a calibrated plane, these times can be computed, at least in principle, by solving for t from each of the following equations:

$$\phi(t) = \omega_\phi t + \phi_{t_i^-} \qquad \psi(t) = \omega_\psi t + \psi_{t_i^-} \tag{3}$$

Where $\phi(t)$ and $\psi(t)$ are time varying functions, representing the angles between rays from the image points corresponding to the target trajectory w.r.t to a reference ray in the foveal camera. The ω_ϕ and ω_ψ are, respectively, the pan and tilt angular speeds and the angles $\phi_{t_i^-}$ and $\psi_{t_i^-}$ represent the angle positions at time t_i^-. By separately solving the two equations in t we estimate the interception times t_ϕ and t_ψ, needed to intercept the target through pan and tilt camera motion. Each of the above equations is non-linear due to the image formation process. In order to make the TDO problem solvable, a closed

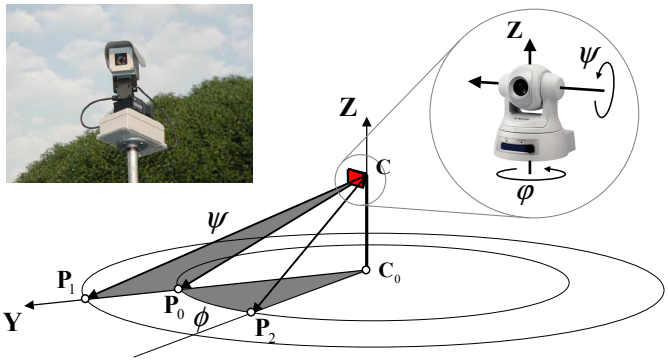


Fig. 3. The geometry of a PTZ camera viewing a world plane in which the pan axis coincides with the normal of the plane. Also shown are the angles ϕ and ψ travelled by the pan-tilt device gazing from the target $\mathbf{P}_1$ to the target $\mathbf{P}_2$.

form solution is obtained by assuming that during the camera interception process, the target motion is negligible. Now the TDO can be solved by exhaustive enumeration without an iterative root finder for the eq.3. With this assumption eq.3 becomes time independent and simplifies:

$$\phi_{t^+_{i+1}} = \omega_\phi t + \phi_{t^-_i} \qquad \psi_{t^+_{i+1}} = \omega_\psi t + \psi_{t^-_i} \tag{4}$$

defining the values for

$$t_{\phi_{i,i+1}} = \frac{\phi_{t^+_{i+1}} - \phi_{t^-_i}}{\omega_\phi} \qquad t_{\psi_{i,i+1}} = \frac{\psi_{t^+_{i+1}} - \psi_{t^-_i}}{\omega_\psi} \tag{5}$$

In order to keep tractable the estimate of the angles of the targets as seen by the slave camera we assume that the PTZ-camera is not mounted oblique w.r.t. the world plane. The camera pan axis it is approximately aligned with the normal of the world plane. This is generally the case when PTZ-cameras are mounted on top of a pole (see fig.3). This means that during continuous panning while keeping a fixed angle for the tilt, the intersection of the optical axis with the 3D plane approximately describes a circle. The principal axis sweeps a cone surface so its intersection with the 3D world plane is in general an ellipse with an eccentricity close to one. In the same sense during continuous tilting while keeping a fixed angle for the pan, the intersection of the optical axis with the 3D plane describes approximately a line. The swept surface is a plane (see fig.3). In such conditions the tilt angle between a reference ray and the ray emanating from the image point corresponding to a target trajectory can be measured once the intrinsic internal camera parameters for the slave camera are known as [22]:

$$cos(\psi) = \frac{\mathbf{x'_1}^T \omega \mathbf{x'_0}}{\sqrt{\mathbf{x'_1}^T \omega \mathbf{x'_1}} \sqrt{\mathbf{x'_0}^T \omega \mathbf{x'_0}}} \tag{6}$$

where $\omega = \mathtt{K}^{-\mathrm{T}}\mathtt{K}^{-1}$ is the image of the absolute conic an imaginary point conic directly related to the internal camera matrix $\mathtt{K}$. While $\mathbf{x'_1}$ and $\mathbf{x'_0}$ (as also shown in fig.4) are, respectively, the projection of the world point $\mathbf{X}_1$ as seen by the master camera and transformed through $\mathtt{H}'$ to the slave camera, and the projection of the point $\mathbf{C}'_0$.

$\mathbf{C}'_0$ is the orthogonal projection of the camera center of the slave camera $\mathbf{C}'$ onto the world plane. By choosing as reference

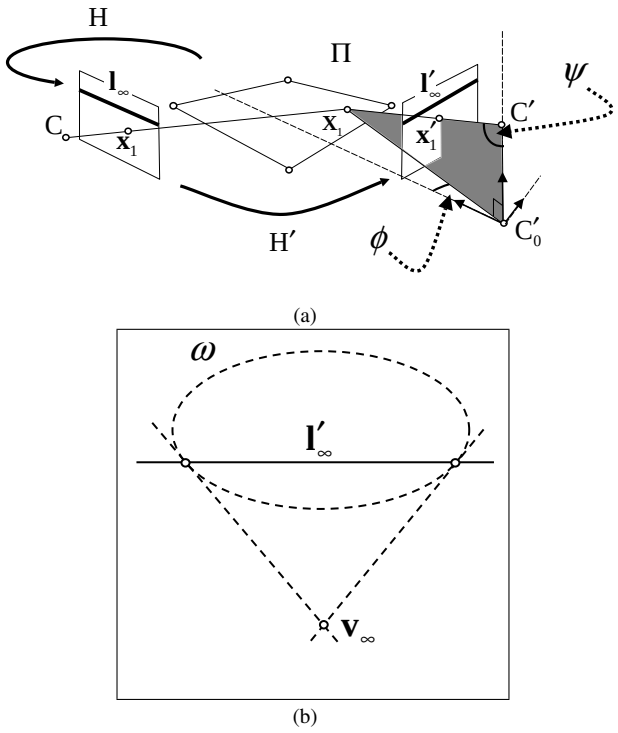


Fig. 4. (a) The geometry used for computing the pan ϕ and tilt ψ angles of a target $\mathbf{X}_1$ as seen from the slave camera $\mathbf{C}'$ in its home position between. (b) Pole-polar relationship between vanishing point $\mathbf{v}_\infty$ of the plane normal and its the vanishing line $\mathbf{l}'_\infty$ used to compute the tilt angle ψ. The IAC is shown dashed to remind that it is a pure imaginary conic.

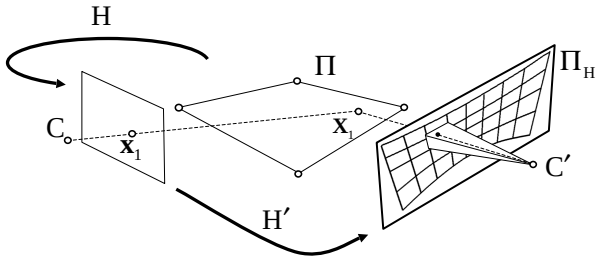


Fig. 5. The slave camera is internally calibrated and the inter-image homography $\mathtt{H}'$ between the master camera $\mathbf{C}$ and the slave camera $\mathbf{C}'$ is computed in its home position (image plane Π_H). We can consider the slave camera as an angle measurement device using the extended image plane composed of the planar image mosaic having Π_H. as a reference plane.

ray to represent tilt angles of the ray passing through $\mathbf{C}'_0$ and $\mathbf{C}_0$ as shown in fig.4, the value of $\mathbf{x'_0}$ can be computed directly using the pole-polar relationship as:

$$\mathbf{x'_0} = \omega^{-1}\mathbf{l}'_\infty \tag{7}$$

Where $\mathbf{l}'_\infty$ is the vanishing line of the plane Π as seen from the slave camera and it can be computed by transferring the vanishing line $\mathbf{l}_\infty$ in the master camera to the slave camera as $\mathbf{l}'_\infty = \mathtt{H}'^{-\mathrm{T}}\mathbf{l}_\infty$. The above formula can be applied because $\mathbf{x'_0}$ coincides with the vanishing point of the directions normal to the plane Π (see [23]). Summarizing, in this configuration the slave camera, in addition to its foveal capability also uses calibration (in its home position) as angle measurement device.

Internal camera parameters necessary for the PTZ-camera can be computed very accurately as recently shown in [24] using the method originally described in [25].

The pan angle of a world point in the plane can be computed directly from the master camera once the world to image homography $\mathtt{H}$ is known and the point $\mathbf{C}'_0$ is measured from the master camera. If that point cannot be measured because it is not visible from the master camera, it can also be computed using the inter-image homography $\mathtt{H}'$. In fact since the slave camera is internally calibrated at its home position, it is possible to obtain its pose and so its camera center w.r.t. the world reference once the world to image homography $\mathtt{H}'_0$ is known from the slave camera. This can be computed as: $\mathtt{H}'_0 = \mathtt{H}'\mathtt{H}$ (see fig.4).

The same approach of eq.4 is followed to obtain the zoom control, once the amount of zoom needed to obtain the desired close-up is calibrated for each point in the world plane. A look-up table using an equispaced grid of points can be used to perform this calibration manually or automatically as shown in [13]. The equation for the estimation of the time needed for changing the zoom to intercept the new target can be written similarly as for pan and tilt:

$$z_{t^+_{i+1}} = v_z t + z_{t^-_i} \tag{8}$$

where v_z is the zooming speed and $z_{t^-_i}$ is the zooming value at time t^-_i, when the target is left and $z_{t^+_{i+1}}$ is the zooming value at time t^+_{i+1} when the next target is intercepted.

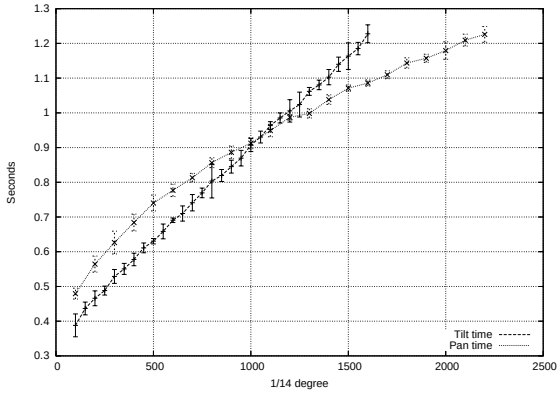
V. Simulation Results

A. Estimating Camera Speeds

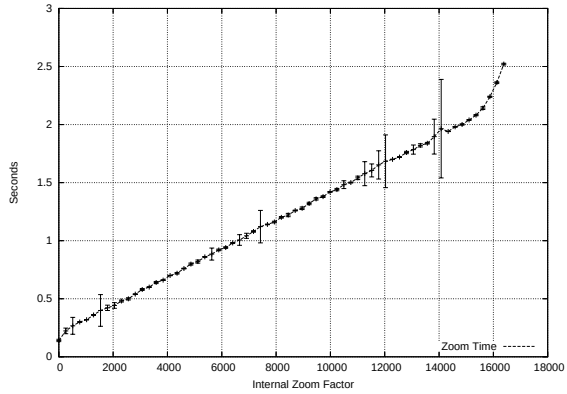
We ran several experiments to empirically estimate the pan/tilt/zoom speeds of our cameras in order to validate the constant velocity kinematic models used in the eq.4 and eq.8. The results of these experiments are shown figure 6. In particular we have conducted several trials and then we have averaged the results in fig.6(a) are shown the pan and tilt speeds while in fig.6(b) are reported the zoom speeds. Worthy of note is the fact that, contrary to manufacturer specification, the cameras do not move at a constant speed. Indeed, there are situations in which either panning or tilting might be the slowest of motions, as indicated by the crossover point of the two curves in figure. When moving such short distances, camera motion is nearly instantaneous and we found that assuming a constant camera velocity when planning a saccade sequence worked just as well as the more complex camera performance model.

B. Congestion Analysis

Evaluating different planning strategies using a video surveillance system installed in a real context is a very complicated task. In fact, while we can easily collect video from a static camera, and use it for target tracking, it is almost impossible to collect all the information needed to plan tours in a master-slave camera configuration with a foveal slave camera. To address these difficulties, we have created a Monte Carlo simulation for evaluating scheduling policies using randomly generated data.



(a)



(b)

Fig. 6. Empirically estimated pan-tilt (a) and zoom (b) times for the Sony SNC-RZ30, averaged over thirty trials.

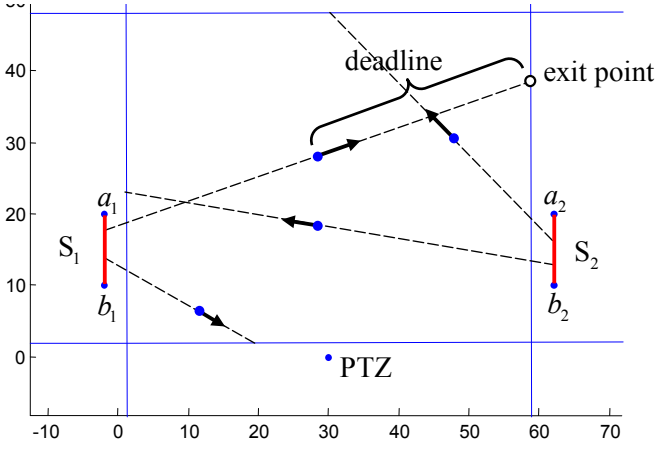


Fig. 7. The simulated surveillance scene. The segment $\mathbf{S_1}$ and $\mathbf{S_2}$ represent sources where targets originates with a given statistics.

But there is also another main reason for using randomly generated data. The use of randomly generated data often enables more in-depth analysis, since the datasets can be constructed in such a way that other issues could be addressed. For example the arrival rate parameter, generally denoted λ, describes the "congestion" of the system. This is basically the only important parameter which is worth of testing in a similar scenario. We stress the importance of this kind of testing: real data testing cannot evaluate the algorithm performance in this context.

We performed a Monte Carlo simulation that permits evaluating the effects of different scheduling policies in a congestion analysis setting. We used in our simulator a particular scene in which our framework could be of invaluable benefit. A large area of approximatively 50x60 meters (half of a soccer field) is monitored with the slave camera placed as shown in fig.7 at position $(30, 0, 10)$. The master camera views the monitored area at a wide angle from above (more suitable for tracking due low occlusion between target). Arrivals of targets are modelled as a Poisson process. The scene is composed of two target sources situated at opposite positions in the area. Targets originate from these two sources S_1 and S_2 from initial positions that are uniformly distributed in given ranges of length 10 meters positioned as shown in fig.7. The starting angles for targets are also distributed uniformly with the range $[-40, 40]$ degrees. Target speeds are generated from a truncated Gaussian with a mean of 3.8 meter/sec and standard deviation of 0.5 meter/sec. (typical of a running person) and are kept constant for the duration of target motion. Targets follow a linear trajectory. This is not a restrictive assumption since each TDO has in this simulation a deadline of $t = 5$ seconds, and the probability of maneuvering for targets with a running-human dynamic in an interval of five seconds is very low. So the overall performance of the system is not generally affected. The deadline t has a role similar to a sampling time for traffic behavior and can be generally tuned depending on the speeds of the targets. In our simulated scene it is quite improbable that a target enters and exits the scene before five seconds are elapsed.

The used scene can represent a continuous flow of people, in a crisis situation. An example is people exiting from a stadium or from the subway stairs. It can be interesting, for crime detection purposes, to acquire as many high resolution images of such running people as possible before they leave the scene.

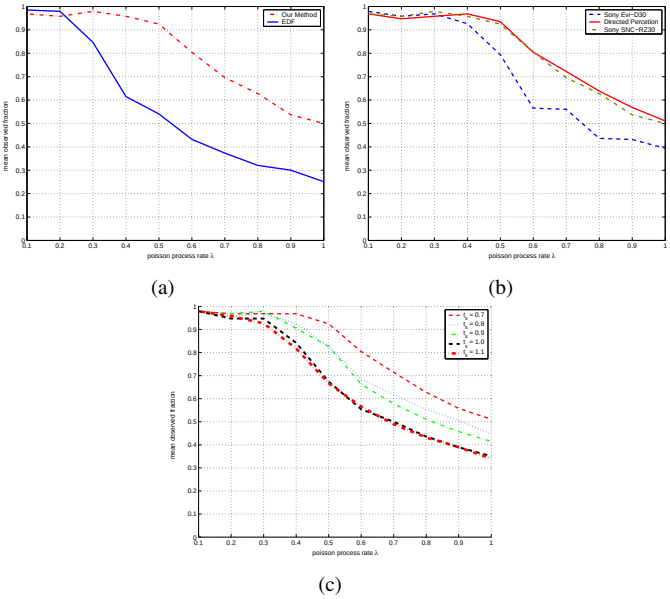
We assume that all targets have the same size in the scene (average humans height) and a specific size is fixed at which the target must be observed by the foveal camera. For pinhole cameras, as the focal length of the camera changes, the pinhole model predicts that the images will scale in direct proportion to the focal length [26]. By assuming a constant speed for the zooming motor and a linear mapping of focal length to zoom it is possible to build a look-up table in the simulator as: $\text{Zoom}[x, y] = M \cdot \text{dist}(\mathbf{C}', \mathbf{X})$ where x and y are the imaged coordinates of the world plane point $\mathbf{X}$ as seen by the master camera, $\mathbf{C}'$ is the camera center of the slave camera and M is the constant factor which depends on the size at which targets are imaged and on the target size in the scene. We want to collect human imagery with an imaged height of approximatively 350 pixels using an image resolution of 720×576. In fig.8, plots indicate the number of targets that are observed by the foveal camera (ordinates) as a function of the arrival rate λ (abscissa) for three different situations. Since there are two sources with the same arrival rate, λ actually refers to half the number of arrivals per second. The size of the queue is six elements which guarantees that the enumeration of all the subsets with their permutations is generated in a fraction of a second (basically a negligible time). Performance is measured by running a scenario in which 500 targets are repeatedly generated one hundred times and the performance metric was estimated by taking the mean. The metric corresponds to the fraction of people observed in the scene. In particular we take the mean (over the experiments) of the number of observed target divided by number of all the targets.

Fig.8(a) shows a comparison of our methods with the earliest deadline first policy studied in [7]; it evident that our policy, using long term planning plus the cost of moving the sensor, outperforms a simple greedy strategy. While there is no need for planning in very modest traffic scenes, traffic monitoring, in large, wide areas would receive an invaluable great advantage of more than 40% by adopting the proposed techniques. Fig.8(b) shows experiments conducted using different speeds

	Pan Speed deg/sec	Tilt Speed deg/sec	Zoom Speed #mag/sec
Sony EVI-D30	80	50	0.6
Sony SNC-RZ30	170	76.6	8.3
Directed Perception	300	300	11.3

TABLE I

OFF THE SHELF PTZ-CAMERAS PERFORMANCE. THE #MAG MEANS MAGNIFICATION FACTOR PER SECOND AND IS CALCULATED DIVIDING THE MAXIMUM OPTICAL ZOOM (FOR EXAMPLE 25X) BY THE ZOOM MOVEMENT TIME FROM WIDE TO TELE (FOR EXAMPLE 2.2 SECONDS).



(a) (b)

(c)

Fig. 8. Policy performance versus arrival rate λ. (a) Our methods and simple earliest deadline first policy. (b) Three different PTZ-camera under test with different pan-tilt-zoom speed. (b) Performance variation at varying service time t_s (the specified time to watch a target).

for PTZ motors typical of off-the-shelf active cameras. Three cameras were selected using their respective performance as indicated by the technical specification (see tab.I). Using this performance values in the simulator produce the plots of fig.8(b). Although the three models are very different in performance, such differences are less evident for the observing task under test. This is mostly caused by the camera position w.r.t. the scene plane; the performance in tilt speed was practically never employed because of the latency of the other controls w.r.t. the imaged motion pattern of targets. The control which delayed most of the saccades, employing the largest setup time, was the zoom control (mostly caused by the scene depth). This explains why the two fastest cameras exhibit similar performances. This type of analysis can be useful for determining the type of cameras and ultimately the cost needed to monitor an area with a multi-camera system.

Fig.8(c) shows the performance degradation w.r.t. the service time (or the watching time) t_s. This time is directly related to the quality of the acquired images and can potentially affect recognition results. The figure also shows that varying t_s does not affects the performances in direct proportion.

VI. Summary

Automated high resolution imaging of targets using PTZ cameras is an important and mandatory capability for modern automated surveillance. In such systems, and especially in the case of wide area surveillance applications, to view multiple moving targets each camera must share observation time. We have presented a solution for planning saccade sequences using a single foveal camera in a master-slave camera system configuration. The system models the attentional gaze planning, with a novel approach combining ideas from Dynamic Vehicle Routing Problem (DVRP) and multiview geometry. Results are presented using a simulator that indicates how many targets are missed as a function of the arrival rate, camera speed parameters and watching time. Results have been derived under realistic assumptions in a challenging scene. We proved that our framework gives good performance in monitoring wide areas with little extra effort with respect to other cumbersome approaches coordinating a large number of cameras doing the same task.

The same principles presented here can also be applied to camera-networks to build large surveillance systems; the framework is open and may be extended easily in several different ways; e.g. a real-time face recognition/detection can be incorporated in the optimization.

One main limitation of the presented method is that it does not take advantage of persistent motion patterns generally present in common scenes, for example an intersection with moving cars. Such knowledge would be of invaluable benefit in cases where targets are following pre-defined paths. Ongoing research will address on-line learning algorithms capable of finding more long-term policies. Moreover, further research can apply supervised machine learning methods to a simulated data set (as generated by our approach) to understand the behaviors of complex saccadic patterns for the task under consideration.

References

[1] A. Yarbus, *Eye Movements and Vision*, Plenum Press, 1967.

[2] R.K. Bajcsy., "Active perception," *Proc of the IEEE*, vol. 76, no. 8, pp. 57–62, 1988.

[3] Michael J. Swain and Markus A. Stricker, "Promising directions in active vision," *International Journal of Computer Vision*, vol. 11, no. 2, pp. 109–126, 1993.

[4] J. Denzler and C.M. Brown., "Information theoretic sensor data selection for active object recognition and state estimation.," *Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 145–157, 2002.

[5] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "Appearance based active object recognition," *Image and Vision Computing*, vol. 18, pp. 715–727, 2000.

[6] S. Stillman, R. Tanawongsuwan, and I. Essa., "A system for trackng and recognizing multiple people with multiple cameras.," *Technical Report GIT-GVU-98-25 Georgia Institute of Technology, Graphics, Visualization, and Usability Center*, 1998.

[7] C. J. Costello, C. P. Diehl, A. Banerjee, and H. Fisher, "Scheduling an active camera to observe people," *Proceedings of the 2nd ACM International Workshop on Video Surveillance and Sensor Networks*, pp. 39–45, 2004.

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

[9] J. Batista, P. Peixoto, and H. Araujo., "Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking.," *In Proceedings of the IEEE Workshop on Visual Surveillance*, pp. 18–25, 1998.

[10] S. J. D. Prince, J. H. Elder, Y. Hou, and M. Sizinstev, "Pre-attentive face detection for foveated wide-field surveillance," *IEEE Workshop on Applications on Computer Vision*, pp. 439–446, 2005.

[11] L. Marchesotti, L. Marcenaro, and C. Regazzoni, "Dual camera system for face detection in unconstrained environments," *ICIP*, vol. 1, pp. 681–684, 2003.

[12] Ser-Nam Lim, Davis L.S, and A. Elgammal, "Scalable image-based multi-camera visual surveillance system," *Proceedings IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 205–212, 2003.

[13] Andrew Senior, Arun Hampapur, and Max Lu, "Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration," *IEEE Workshop on Applications on Computer Vision*, 2005.

[14] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle., "Face cataloger: Multi-scale imaging for relating identity to location.," *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 21–22, 2003.

[15] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1456–1477, 2001.

[16] M. Greiffenhagen, V. Ramesh, D. Comaniciu, and Heinrich Niemann, "Statistical modeling and performance characterization of a real-time dual camera surveillance system," *IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 335–342, 2000.

[17] D. W. Murray, K. J. Bradshaw, P. F. McLauchlan, I. D. Reid, , and P.M. Sharkey, "Driving saccade to pursuit using image motion," *Int. Journal of Computer Vision*, vol. 16, no. 3, pp. 205–228, 1995.

[18] P. Chalasani, R. Motwani, and A. Rao., "Approximation algorithms for robot grasp and delivery," *In Proceedings of the 2nd International Workshop on Algorithmic Foundations of Robotics*, pp. 347–362, 1996.

[19] Dimitris Bertsimas and Garrett Van Ryzin, "A stochastic and dynamic vehicle routing problem in the euclidean plane," *Operations Research*, vol. 39, pp. 601–615, 1991.

[20] C.S. Helvig, G. Robins, and A. Zelikovsky, "The moving-target traveling salesman problem," *Journal of Algorithms*, vol. 49, no. 1, pp. 153–174, 2003.

[21] F. V. Fomin and A. Lingas, "Approximation algorithms for time-dependent orienteering," *Information Processing Letters*, vol. 83, no. 2, pp. 57–62, 2002.

[22] R. I. Hartley and A. Zisserman., "Multiple view geometry in computer vision.," *Cambridge University Press, second edition*, 2004.

[23] C. Colombo, A. Del Bimbo, and F. Pernici, "Metric 3d reconstruction and texture acquisition of surfaces of revolution from a single uncalibrated view," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 99–114, 2005.

[24] S.N. Sinha and M. Pollefeys., "Towards calibrating a pan-tilt-zoom cameras network," *P. Sturm, T. Svoboda, and S. Teller, editors, OMNIVIS*, 2004.

[25] L. De Agapito, R. Hartley, and E. Hayman., "Linear selfcalibration of a rotating and zooming camera.," *In Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 15–21, 1999.

[26] B.J. Tordoff, "Active control of zoom for computer vision," *DPhil thesis, Univ. of Oxford*, 2002.