Towards On-Line Saccade Planning for High-Resolution Image Sensing

Alberto Del Bimbo [a] Federico Pernici [b]

[a] *Dipartimento di Sistemi e Informatica – Università di Firenze Via Santa Marta 3, I-50139 Firenze, Italy*

[b] *Dipartimento di Sistemi e Informatica – Università di Firenze Via Santa Marta 3, I-50139 Firenze, Italy*

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.

Key words: Active Vision, Travelling Salesperson Problem, Video Surveillance, Sequential Decision Making

1 Introduction

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

Email addresses: delbimbo@dsi.unifi.it, pernici@dsi.unifi.it (Federico Pernici).

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 transition 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 saccade 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 different points in the scene. For systems with multiple behaviors, this scheduling problem becomes increasingly paramount.

The key contributions of the paper 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. We have also discussed in (Del Bimbo and Pernici, 2005) the basic ideas underlying the

approach presented in this paper.

2 Related Work

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 (Stillman et al., 1998) the problem of deciding which camera should be assigned to which person was addressed and some general approaches were given. It should also be noted that there is no work except (Costello et al., 2004) 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 the literature uses a master/slave camera system configuration with two (Zhou et al., 2003) (Costello et al., 2004) (Batista et al., 1998) (Prince et al., 2005) (Marchesotti et al., 2003) or more cameras (Lim et al., 2003) (Senior et al., 2005) (Stillman et al., 1998) (Hampapur et al., 2003) (Collins et al., 2001). The remote target identification problem is also termed as distant human identification (DHID).
In (Zhou et al., 2003), a single person is tracked by the active camera. If multiple people are present in the scene, the person closest to the position of the previous tracked individual is chosen.
In (Costello et al., 2004) 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 method uses greedy policies instead of policies with a time horizon. Also in (Lim et al., 2003) 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 (Prince et al., 2005) 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 (Greiffenhagen et al., 2000) 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 (Murray et al., 1995)(Batista et al., 1998) 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 (Senior et al., 2005) 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 (Senior et al., 2005) and (Zhou et al., 2003), generally suffice to bring the target object within a narrow zoomed field of view.
The other important work related to our problem is (Bertsimas and Van Ryzin, 1991), 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.

3 Saccade Planning as a novel Dynamic Vehicle Routing Problem

Three main features characterize the task under consideration: targets' motion and position, target arrivals, and target deadlines.
The first one is that targets moving farthest from the camera appear to move slower in the image while closer targets appear to move faster. For example gazing from a closer target to a distant one is generally slower mainly as a result of a zoom induced delay (as pan and tilt motions are much faster in comparison). In the case when the targets are at similar distances and opposite directions from the slave camera the gazing depends mainly on the pan and/or tilt. Once a finite number of targets is at hand this trade-off can be exploited by using an appropriate cost for changing the orientation and zoom of the PTZ camera in an instance of the Kinetic Travelling Salesperson Problem (KTSP). This is an extension of the classical Travelling Salesperson Problem (TSP) with moving cities.
The second and third main features are that targets arrive unpredictably so it is impossible to know how many targets will be in the scene at any time and what target to choose to look at especially if solutions have to be computed on-line. The problem of how to choose the best permutation subset from the currently tracked targets is an instance of the Time Dependent Orienteering (TDO) with deadlines.

3.1 Kinetic Travelling Salesperson Problem (KTSP)

As cameras can be calibrated with automatic or manual methods such as in (Senior et al., 2005) 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) [1] , 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" [2] (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) (Helvig et al., 2003). In fig.1(a) 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 (or path) 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) joints and one prismatic (zoom) joint, it is possible to view the principal axis of the camera as a robot arm which rotates and moves forward to reach a point in the space. In such a setting, 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.

3.2 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 such a dynamic-stochastic setting there is a lot of interde-

[1] In general this vector can be thought as the set of the controllable parameters of an active camera. For example (p, t, z, f) where f is the focal length.

[2] capture high-resolution images or videos of as much as possible of any object that passes through a designated area.

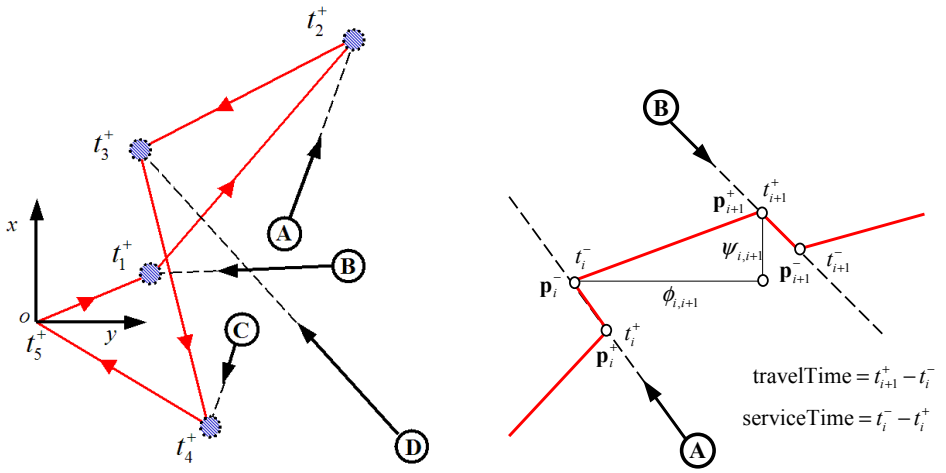
pendency between the state variables describing the system. Moreover, tours must be planned while existing targets move or leave the scene, and/or new targets arrive. Basically the whole problem can be viewed as a global dynamic optimization. Since for such a problem no a-priori solution can be found, 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 (Fomin and Lingas, 2002). 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.*

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 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 exhaustive search approach enumerating and evaluating all the subsets permutations perfectly fits with the nature of our dynamic incremental formulation.*

3.3 Deadlines

Based on the tracking predictions the 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



(a) (b)

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

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 solution evaluations. 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 follows (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 leave the target i in T at time t_i^- before the target leaves the scene at time t_i^d.

4 Master-Slave Camera System 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 (Zhou et al., 2003) (Costello et al., 2004). 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.

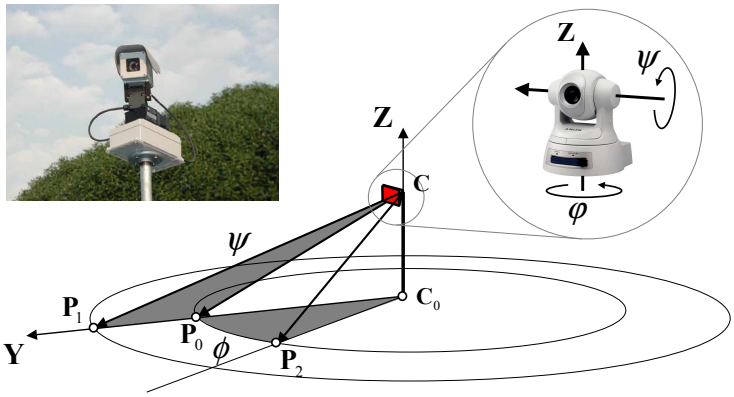


Fig. 2. The geometry of a PTZ camera viewing a world plane in which the pan axis coincides with the normal of the plane. Also shown are the angles ϕ and ψ travelled by the pan-tilt device gazing from the target $\mathbf{P}_1$ to the target $\mathbf{P}_2$.

4.1 Modelling the cost of changing orientation and zoom

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.1(b) 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.1(b) 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 adopting a linear constant speed model for the motors: A closed form solution is obtained by assuming that during the camera interception process, the target motion is negligible. This is basically the same assumption made so far: 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. With this assumption the travel times $t_{\phi_{i,i+1}}$ and $t_{\psi_{i,i+1}}$ can be computed as

$$t_{\phi_{i,i+1}} = \frac{\phi_{t^+_{i+1}} - \phi_{t^-_i}}{\omega_\phi} \qquad t_{\psi_{i,i+1}} = \frac{\psi_{t^+_{i+1}} - \psi_{t^-_i}}{\omega_\psi} \tag{3}$$

where ω_ϕ, ω_ψ are the pan and tilt angular speeds and $\phi_{t_i^-}$, $\psi_{t_i^-}$, $\phi_{t_{i+1}^+}$, $\psi_{t_{i+1}^+}$ represent the angle positions respectively at time t_i^- and t_{i+1}^+.

4.2 Computing the Pan, Tilt Angles and the Zoom

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.2). This means that during continuous panning while keeping a fixed angle for the tilt, the intersection of the optical axis with the plane approximately describes a circle. The principal axis sweeps a cone surface so its intersection with the 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.2). 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 in a home position (i.e. a reference position) for the slave camera are known as (Hartley and Zisserman., 2004):

$$cos(\psi) = \frac{\mathbf{x'}_1^T \omega \mathbf{v}_\infty}{\sqrt{\mathbf{x'}_1^T \omega \mathbf{x'}_1}\sqrt{\mathbf{v}_\infty^T \omega \mathbf{v}_\infty}} \tag{4}$$

where ω is the image of the absolute conic (IAC) which is directly related to the internal camera matrix $\mathtt{K}$ as $\omega = \mathtt{K}^{-T}\mathtt{K}^{-1}$ (see the appendix A for basic definition and properties). We can consider the slave camera as an angle measurement device using the extended image plane composed by the planar image mosaic having Π_H as a reference plane. Indeed fig.3 shows exactly that: the point $\mathtt{x}_1$ on the camera $\mathbf{C}$ is transferred by the homography $\mathtt{H}'$ on the camera $\mathbf{C}'$. The $\mathtt{H}'$ relates the master camera plane to the home position plane Π_H which is the reference view where the calibration is computed (i.e. all the homographies $\mathtt{H}_{ij}$ are related to that view). So it is like transferring the point $\mathtt{x}_1$ onto a camera having the mosaic plane as the image plane and then using eq. 4 there for computing the tilt angle.

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. Given a pinhole camera model the zoom is proportional to the scene depth. The depth can be computed by triangulation given the camera internal parameter and the world to image homography.

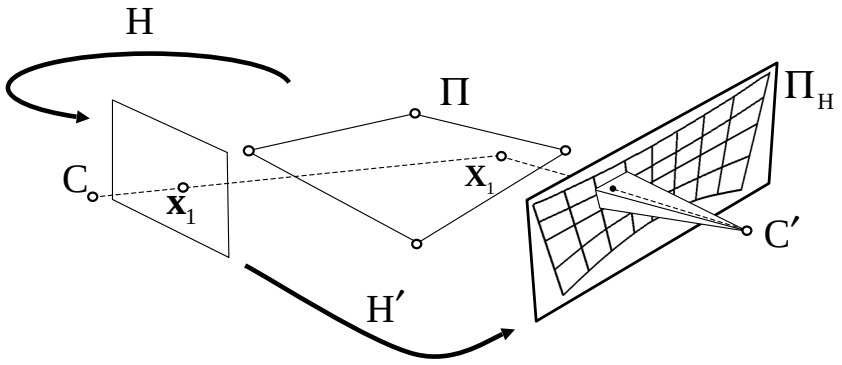


Fig. 3. The slave camera is internally calibrated and the inter-image homography $\mathtt{H}'$ between the master camera $\mathbf{C}$ and the slave camera $\mathbf{C}'$ is computed in its home position (image plane Π_H). We can consider the slave camera as an angle measurement device using the extended image plane composed of the planar image mosaic having Π_H. as a reference plane.

5 Experimental Results

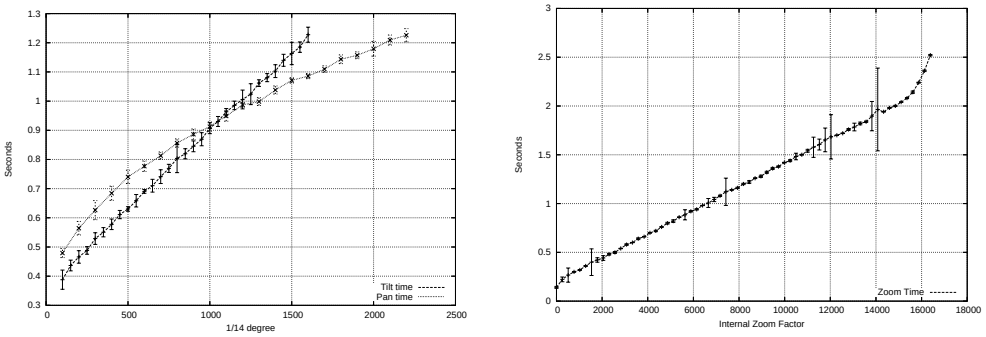
In this section we show two different kind of experiments. The first regards the camera speed kinematic model used and the second regards a statistical performance analysis in a congestion setting. They are not intended to fully evaluate the performance of the saccades planning framework described in the paper with a full real implementation.

5.1 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 eq.3. The results of these experiments are shown figure 4. In particular we have conducted several trials and then we have averaged the results. In fig.4(a) the pan and tilt speeds are shown while in fig.4(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.

5.2 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



(a) (b)

Fig. 4. Empirically estimated pan and tilt times for the Sony SNC-RZ30, averaged over thirty trials.

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. Moreover, a wide range of traffic levels and paths through the scene need to be used. It is also difficult to derive statistics about the performance of the different scheduling policy implementations by observing the system. For example, errors could be due to a mistake performed by the tracking camera rather than a weakness of the scheduling policy. It is not easy to separate the performance of the scheduling policy from that of the overall system.

To address all of these difficulties, we have created a Monte Carlo simulation for evaluating scheduling policies using randomly generated data. 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. For example, errors could be due to a mistake performed by the tracking camera rather than a weakness of the policy. In order to evaluate how good the proposed approach is it, is mandatory to separate the performance of the algorithm from that of the overall system.
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 at position $(30, 0, 10)$. The master camera sees the monitored area at a wide angle from above (more suitable for tracking due to low occlusion between targets). Arrivals of targets are modelled as a Poisson process. The scene is composed of two target sources situated at opposite sides of the area. Targets originate from these two sources from initial positions that are uniformly distributed in given ranges of length 10 meters positioned at opposite sides of the area. 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. Generally when people move in a free space, a minimal distance is generally followed (i. e. so a line path in generally used).

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.

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.5, plots indicate the number of targets that are observed by the foveal camera (ordinates) as a function of the arrival rate λ (abscissa) for two 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 targets divided by number of all the targets.

Fig.5(a) shows a comparison of our methods with the 'earliest deadline first' policy studied in (Costello et al., 2004); 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.5(b) 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 performance in direct proportion.

Fig.5(c) shows experiments conducted using different speeds for PTZ motors typi-

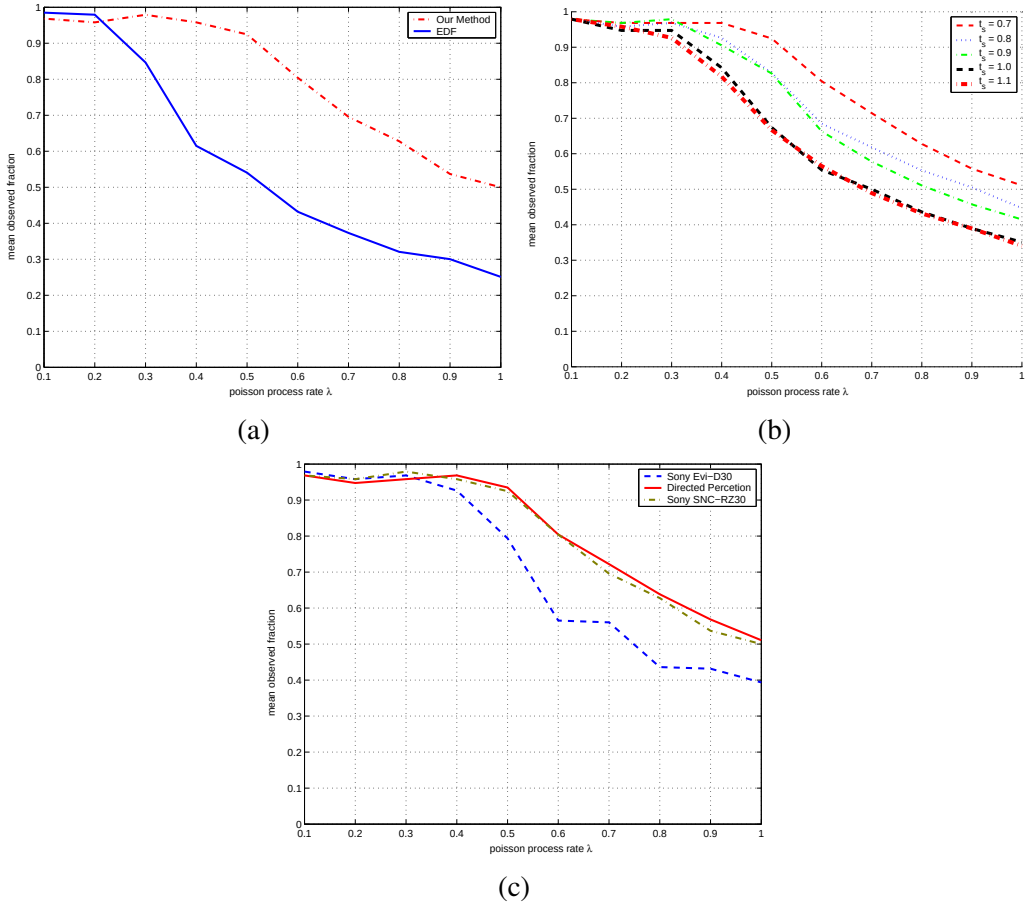


Fig. 5. Policy performance versus arrival rate λ. (a) Our methods and simple earliest deadline first policy. (b) Performance variation at varying service time t_s (the specified time to watch a target). (c) Three different PTZ-camera under test with different pan-tilt-zoom speed.

cal of off-the-shelf active cameras. Three cameras were selected using their respective performance as indicated by the technical specification (see table1). Using performance values in the simulator produces the plots of fig.5(c). 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 performance. This type of analysis can be useful for determining the type of camera and ultimately the cost needed to monitor an area with a multi-camera system.

	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 1
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).

6 Conclusions

Planning saccade sequences is mandatory for making PTZ cameras usable in real environments. In this paper, we have defined a new scheduling policy to acquire high resolution images of as many moving targets as possible before they leave the scene, taking into account the costs of camera movements, and performed a congestion analysis putting the number of target sensed in relationship with the number of moving targets in the camera field. Results have been derived under reasonable assumptions on camera travel times and by performing a Monte Carlo simulation. The whole framework allows the derivation of useful quantitative evaluations that otherwise would be impossible to obtain from real observations. In fact, in a real scenario it is almost impossible to configure experimental test conditions in order to collect all the information needed to plan tours with a PTZ camera in a master slave configuration; and on the other hand, since errors due to target tracking are interwoven with errors due to the scheduling policy, the effects of the scheduling policy cannot be identified reliably. The same approach followed here can also be applied to camera networks for large surveillance systems. The framework can be easily extended to deal with additional functions like object recognition or face detection, or management of targets with different degrees of interest. Future research will address the definition of on-line learning algorithms to obtain camera scheduling with no constraints on the number of moving targets and the size of the temporal window, and the application of the proposed policy to real surveillance environments with inclusion of face detection and high resolution target acquisition.

Appendixes

A Absolute Conic and rotating cameras

The image of the absolute conic is the projection of the absolute conic Ω. This is an imaginary point conic that lies on the plane at infinity Π_∞ in 3D and has the property that it is invariant to similarity transformations of space (Hartley and Zisserman., 2004). The conic relevant to calibration is its projection onto the image

plane, i.e. the image of the absolute conic ω (IAC). The IAC is related to the camera calibration parameter by $\omega = \mathtt{K}^{-\mathrm{T}}\mathtt{K}^{-1}$. The calibration matrix $\mathtt{K}$ may be computed from ω according to the Cholesky decomposition.
One important property of the IAC is that it can be transferred from one image to another through the infinite homography $\mathtt{H}^{\infty}$ as:

$$\omega_i = \mathtt{H}_{ij}^{\infty -T}\omega_j\mathtt{H}_{ij}^{\infty -1} \tag{A.1}$$

Once we have $\mathtt{H}^{\infty}$ the equation above can be used to impose constraints on ω. Points at infinity (like for example vanishing point) are mapped between views by the infinite homography $\mathtt{H}^{\infty}$ and this is independent on translation between views. In particular when there is no translation between the views, the infinite homography $\mathtt{H}^{\infty}$ relates points of any depth. This simplification can be exploited when images are taken with cameras having a common center. The $\mathtt{H}^{\infty}$ coincides with the inter-image homographies, so we have a convenient method of measuring $\mathtt{H}^{\infty}$ directly from images.

B Computing the Slave internal camera parameter

Internal camera parameters necessary for the PTZ-camera can be computed very accurately as recently shown in (Sinha and Pollefeys., 2004) using the method originally described in (De Agapito et al., 1999). When images are taken with cameras all located at the same camera center point in space, camera matrices can be simplified. It is possible to analyze this situation by representing each camera as a 3×3 matrix instead of a general 3×4 camera matrix. A point in the i-th image, represented by a homogeneous 3-vector $\mathbf{X}_i$ corresponds to a ray in space consisting of points of the form $\lambda\mathtt{P}_i^{-1}\mathbf{x}_i$. Points on this ray are mapped into the j-th image to a point $\mathbf{x} = \mathtt{P}_j\mathtt{P}_i^{-1}\mathbf{x}_i$. Denoting the transformation $\mathtt{H}_{ij} = \mathtt{P}_j\mathtt{P}_i^{-1}$ one sees that he i-th and j-th images are related by a projective planar transformation $\mathtt{H}_{ij}$. Clearly this can be estimated by at least four matched points. Each transformation estimated by point correspondences is related to the internal camera parameter as eq.A.1. Once the homographies are known the equation above can be expressed linearly in the terms of ω. If the skew is zero which is usual in modern cameras, there are four unknown in the internal camera parameters: focal length (1 DOF), principal point (2 DOF) and aspect ratio (1 DOF). Four homographies suffices to compute the minimal solution. In fact each equation provides a single constraint on ω. Since we need the internal camera parameter in a specified home position only for computing angles we don't need to use the zoom. This means that the internal camera parameter does not varying while panning and tilting so in eq.A.1 $\omega_i = \omega_j = \omega$ and becomes

$$\omega = \mathtt{H}_{ij}^{-T}\omega\mathtt{H}_{ij}^{-1} \tag{B.1}$$

References

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

Bertsimas, D., Van Ryzin, G., 1991. A stochastic and dynamic vehicle routing problem in the euclidean plane. Operations Research 39, 601–615.

Collins, R., Lipton, A., Fujiyoshi, H., Kanade, T., 2001. Algorithms for cooperative multisensor surveillance. Proceedings of the IEEE 89 (10), 1456–1477.

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

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

Del Bimbo, A., Pernici, F., 2005. Saccade planning with kinetic tsp for distant target identification. In Proceedings of the IEE International Symposium on Imaging for Crime Detection and Prevention (ICDP), The IEE, Savoy Place, London, UK.

Fomin, F. V., Lingas, A., 2002. Approximation algorithms for time-dependent orienteering. Information Processing Letters 83 (2), 57–62.

Greiffenhagen, M., Ramesh, V., Comaniciu, D., Niemann, H., 2000. Statistical modeling and performance characterization of a real-time dual camera surveillance system. IEEE Conf. Computer Vision and Pattern Recognition 2, 335–342.

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

Hartley, R. I., Zisserman., A., 2004. Multiple view geometry in computer vision. Cambridge University Press, second edition.

Helvig, C., Robins, G., Zelikovsky, A., 2003. The moving-target traveling salesman problem. Journal of Algorithms 49 (1), 153–174.

Lim, S.-N., L.S, D., Elgammal, A., 2003. Scalable image-based multi-camera visual surveillance system. Proceedings IEEE Conference on Advanced Video and Signal Based Surveillance, 205–212.

Marchesotti, L., Marcenaro, L., Regazzoni, C., 2003. Dual camera system for face detection in unconstrained environments. ICIP 1, 681–684.

Murray, D. W., Bradshaw, K. J., McLauchlan, P. F., Reid, I. D., , Sharkey, P., 1995. Driving saccade to pursuit using image motion. Int. Journal of Computer Vision 16 (3), 205–228.

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

Senior, A., Hampapur, A., Lu, M., 2005. Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration. IEEE Workshop on Applications on Computer Vision.

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

Stillman, S., Tanawongsuwan, R., Essa., I., 1998. 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.

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