Coopetitive Visual Surveillance using Model Predictive Control

Vivek K. Singh
School of Computing
National University of Singapore
Republic of Singapore
vivekkum@comp.nus.edu.sg

Pradeep K. Atrey
School of Computing
National University of Singapore
Republic of Singapore
pradeepk@comp.nus.edu.sg

ABSTRACT

Active cooperative sensing with multiple sensors is being actively researched in visual surveillance. However, active cooperative sensing often suffers from the delay in information exchange among the sensors and also from sensor reaction delays. This is because simplistic control strategies like Proportional Integral Differential (PID), that do not employ the look-ahead strategy, often fail to counterbalance these delays at real time. Hence, there is a need for more sophisticated interaction and control mechanisms that can overcome the delay problems. In this paper, we propose a *coopetitive* framework using Model Predictive Control (MPC) which allows the sensors to not only 'compete' as well as 'cooperate' with each other to perform the designated task in the best possible manner but also to dynamically swap their roles and sub-goals rather than just the parameters. MPC is used as a feedback control mechanism to allow sensors to react not only based on past observations but also on possible future events. We demonstrate the utility of our framework in a dual camera surveillance setup with the goal of capturing the high resolution images of intruders in the surveyed rectangular area e.g. an ATM lobby or a museum. The results are promising and clearly establish the efficacy of *coopetition* as an effective form of interaction between sensors and MPC as a superior feedback mechanism than the PID.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]; I.4.m [**Artificial Intelligence**]: Image Processing and Computer Vision—*Miscellaneous*

General Terms

Security

Keywords

Visual Surveillance, Coopetition, Model Predictive Control

VSSN'05, November 11, 2005, Singapore.

1. INTRODUCTION

In a multi-sensor environment, active cooperative sensing has been shown to be better than the passive non-cooperative sensing [2, 3]. By active cooperative sensing, we mean that the sensors in use work with each other to better perform the task and also offer an innate ability to respond to the stimuli. However, the active cooperative sensing often suffers from the delay in information exchange among the sensors and also from the delay in sensors reacting to the situation [1, 14]. This problem is often compounded by the use of simplistic control strategies like Proportional Integral Differential (PID), which do not employ the look-ahead strategy to counterbalance the various delays at real time [7, 12]. Hence, there is a need for a more sophisticated control mechanism that can overcome the delay problems.

In this paper, we propose an active and cooperative multi-sensor framework that uses Model Predictive Control (MPC) as a feedback mechanism to allow sensors to react not only based on past observations but also on possible future events. MPC is a control method that makes explicit use of the process model to predict the system control inputs and the outputs and has been widely used in robotics and chemical engineering fields [1, 7]. To correlate MPC to a real-life example of chess; we play our next move not only based on past opponent moves but also on his/her anticipated future moves.

Our framework introduces the concept of *coopetition* among the sensors. We define *coopetition* as a process in which the sensors 'compete' as well as 'cooperate' with each other to perform the designated task in the best possible manner. In a *coopetitive* environment, the sensors work towards global optimization for the task at hand even if the global optimization requires the sacrifice of certain local optimizations. For example, consider a dual camera surveillance scenario where the system goal is to obtain high resolution images of an intruder entering in an enclosed rectangular area. In this scenario, the camera which is in a better position to capture the images of the intruder should be allowed to track him/her even if it means competing against the peer camera in certain instances e.g. based on size of facial image obtained. Hence in this case even though the cameras are 'competing' in a local context (i.e. in capturing the images of intruder), they are still 'cooperating' towards a common goal in a global context by working together to obtain the best high resolution images of the intruder.

We further extend the definition of co-operation as often described in sensor literature to include the concept of trans-

fer of *roles and sub-goals* rather than just parameters. Thus we believe that sensors should not only pass useful information (as in [17, 9] etc.) to each other but also pass over their entire strategy and role to the other sensor if it helps in achieving better overall performance. Hence, we introduce the idea of *dynamic role swapping* between similar sensors for achieving global optimization. Our definition of overall *coopetitive* interaction can also be understood by looking at the popular card game of bridge in which partners try to outbid each other in an attempt to obtain best possible results for their team. We also notice that partners do not just pass each other parameters but rather give up their role (e.g. as a bidder) if they realize that doing so will lead to over-all best results for the team.

Though our framework is scalable to multiple sensors/ cameras, we use a dual-camera surveillance scenario as a base case to demonstrate its effectiveness and utility. Moreover, though the framework is applicable for any well-defined surveillance setting, we choose the specific system goal to capture the high resolution images of an intruder entering an enclosed single door rectangular premises e.g. an ATM lobby or a museum sub-section. This has been chosen in order to provide quantifiable system performance measures/results. The results are evaluated in terms of the ability of the framework to obtain high resolution frontal facial images of the intruder. The results obtained are compared with the base case scenarios like those of static sensors or *non-coopetitive* interactions and also contrasted with those of similar approaches described in literature.

To summarize, our main contributions in this paper are -

1. Adoption of *coopetitive* interaction approach between sensors in order to achieve overall system-level optimization. We also extend the definition of cooperation between cameras to transfer of roles and sub-goals rather than just parameters. These features help us assign the best-suited vision sensor for the task at hand and achieve best over-all results for surveillance.

2. Introduction of MPC (Model Predictive Control) as a novel means of handling feedback in active sensors. The ability to react to possible future scenarios gives us an effective tool to offset transmission and reaction delays which often affect the performance of multiple visual sensor frameworks.

The organization of the remainder of the paper is as follows. In section 2, we describe the related work done by the research community. Section 3 presents the detailed description of our framework and also provides a theoretical analysis of the various tools and techniques used. In section 4, we describe the results and perform a detailed analysis. Section 5, summarizes the work done and the contributions with a discussion on future work.

2. RELATED WORK

As stated earlier, the main contributions of this paper are - proposal for a new mode of interaction (i.e. *coopetition*) among sensors, and introduction of MPC as a novel feedback mechanism. We describe below the related work keeping these two points in focus.

VSAM (Visual Surveillance and Monitoring) project [2, 3] describes the concept of cooperation among multiple active cameras for tracking objects. It also provides a very good overview of the visual surveillance area, however it does not deal with the specific issues such as *coopetition* and delay counter-action which we are handling.

Barreto group [1] has done some appealing work on feedback control mechanisms for vision based systems. They have highlighted the use of Model Predictive Control mechanism to control the motion of a camera placed in robot head. They also use a Kalman filter for predicting the future positions of the tracked object so as to handle system delays. Papanikolopoulos et al. [11] have also proposed the use of Model Predictive Control and Kalman Filter to track objects for a camera placed in the robot head. Saedan [13], on the other hand has described the use of PID control for visual tracking by robot head camera. However, these works deal with single camera robotic systems and do not cover the interaction between multiple cameras and the complexities arising such as delay and the competition/cooperation issues. We intend to create a combined system which would have an efficient feedback mechanism combined with the ability to handle interactions across multiple cameras.

Recently some interesting dual-camera frameworks have been proposed. Collins group [17] has described a master-slave approach to detect human beings at a distance. In this system, the master camera takes wide panoramic images and the slave camera zooms into the person to obtain his images. Liu et al [9] also describe a similar master-slave approach. Their system has a wide-angle panoramic camera with a PTZ (Pan Tilt Zoom) camera on top.

Anastasio et al [15] have also described the use of a wide-angle camera combined with an active camera to obtain human images. The work by Greiffenhagen et al [6] uses an omni-view camera attached to a PTZ active camera to obtain images of the ROI. However, in all these works, there is no movement of the master camera and both cameras are placed at the same physical position hence reducing the possibilities of obtaining good quality images of the ROI e.g. a change of face direction by the human being will cause these systems to lose out on his facial information. Hence we have adopted a dual camera co-opetitive approach, which would allow interaction between the two PTZ cameras so as to employ the best suited camera to obtain images of the ROI.

Lam et al [8] have described a constantly panning camera system. They forward predict the position of panning camera and the moving object in order to schedule video analysis tasks for future instances of time. However, their work focuses on reducing computational complexity while ours is aimed at obtaining good quality images of the ROI. Our framework also adopts the constantly panning strategy but enhances it's application to facilitate dynamic role swapping.

Looking at the multi-agent robotics literature, Stentz [5] describes cooperation and competition between robots in order to fulfill assigned robotic tasks e.g. Clearing toxic waste etc. The idea described in this paper is clearly interesting, however it is described from a robotic perspective with no correlation to visual surveillance or vision sensor interactions. In our current work we aim to describe the effective use of a combination of cooperation and competition for obtaining best performance in vision-based systems.

A summary of reasonably related work has been shown table 1. It clearly highlights the attributes of visual surveillance which have already been adopted by the research community and also those which have been proposed for the first

Table 1: Summary of different attributes across various related work

Work	Active cameras	Multiple cameras	Camera interaction	Model predictive control	Dynamic role swapping	Continuous panning approach
Collins [2, 3]	Yes	Yes	Cooperation	-	-	-
Barretto [1]	Robot head movement	-	-	Yes	-	-
Papanikolopoulos [11]	Robot head movement	-	-	Yes	-	-
Saedan [13]	Robot head movement	-	-	-(PID)	-	-
Collins [17]	Yes	Dual camera	Cooperation	-	-	-
Liu [9]	Yes	Dual camera	Cooperation	-	-	-
Anastasio [15]	Yes	Dual camera	Cooperation	-	-	-
Greiffenhagen [6]	Yes	Dual camera	Cooperation	-	-	-
Lam [8]	Yes	Yes	Cooperation	-	-	Yes
Proposed framework	**Yes**	**Dual camera**	***Coopetition***	**Yes**	**Yes**	**Yes**

time in this paper. As can be seen in table 1, a significant work has already been done using multiple active cameras. Interaction between cameras is also commonly described, but the *coopetitive* approach of interaction between cameras for global optimization has been introduced for the first time in our proposed framework. MPC has been described from robotics perspective in a couple of works but its prowess for visual surveillance has been highlighted for the first time in this paper. The concept of dynamic role swapping has also been described for the first time in this paper. Lastly, the concept of continuously panning cameras is also fairly new in visual surveillance and this work is the second after [8] to adopt it.

3. PROPOSED FRAMEWORK

We propose a dual camera surveillance framework which adopts *coopetitive* interaction strategy combined with Model Predictive Control for countering system delays. The aim of our surveillance framework is to obtain high quality frontal facial images of intruders entering a single door enclosed rectangular environment e.g. an ATM lobby or a museum subsection. The high quality frontal images can be useful for further automated processing e.g. face recognition etc.

3.1 Overview of the framework

The proposed framework uses two cameras to undertake the tasks of scanning the room for detecting new intruders and focusing (i.e. tracking and zooming) onto their faces. The two cameras have been placed directly opposite each other at the same vertical height. The first camera is placed directly above the important artifact e.g. ATM machine and the second camera is placed facing inwards directly above the door. A diagrammatic representation of the surveyed premises used for experiments in our framework has been shown in figure 1. The two cameras have been represented as $C1$ and $C2$ in the figure. One of the two cameras constantly pans the entire surveyed area to detect new ROI objects and the second camera focuses on the ROI to obtain quality images of ROI. The framework decides on which camera is better suited for the task of focusing and allows it to the undertake this task. The other camera automatically takes over the task of panning the entire surveyed area. The algorithmic approach for the proposed framework has been illustrated in figure 2.

We make the following assumptions -

1. We need to take care of only the most important ROI in the surveyed area. If need be, of course additional cameras can be inserted into the same framework to take care of the 2nd most important ROI, 3rd most important ROI and so on.

2. We assume that we have efficient means available for detecting an ROI. We are currently employing a frontal face detector for ROI detection.

3. We also assume that we have access to a mechanism to decide which camera is giving better images of the ROI. Currently we are using un-zoomed size of detected face as a measure to decide on 'better' images. The camera which captures higher resolution images of the face is considered as 'better' camera.

In our framework, a minimum of two cameras are required so as to provide frontal face images in both directions. Besides this, a minimum of two cameras are also required from an analytical perspective so as to perform two distinct roles of focusing and panning. Further more, we have resisted the temptation of using both cameras to track/focus on the 2 most important ROIs as this can lead to all further intrusions going totally undetected.

The proposed framework translates the PTZ parameters using simple geometrical transformations as the cameras are placed directly opposite each other at a fixed distance in the room at the same vertical height as shown in figure 1.

In the following two subsections, we discuss the two main ingredients of our framework - use of *coopetitive* interaction approach and the use of Model Predictive Control.

3.2 Use of coopetitive interaction approach

We call our framework's interaction approach as *coopetitive* in the sense that the cameras both compete and cooperate for efficient visual surveillance. However, it is important to note that the competition we are dealing with is intra-team i.e. the competing entities still share a common overall goal (e.g. members of same team in a bridge card game trying to outbid each other) as opposed to inter-team (e.g. members of opposite team in a bridge card game) in which case the entities may have opposing system goals.

Initially the cameras compete against each other to undertake the role of focusing onto the ROI. In a single intruder case, this competition is clearly won by the camera towards which the intruder is facing. However, this competition becomes more interesting in the multiple intruder scenario where the winner of competition must be decided

Figure 1: An overview of the surveyed premises

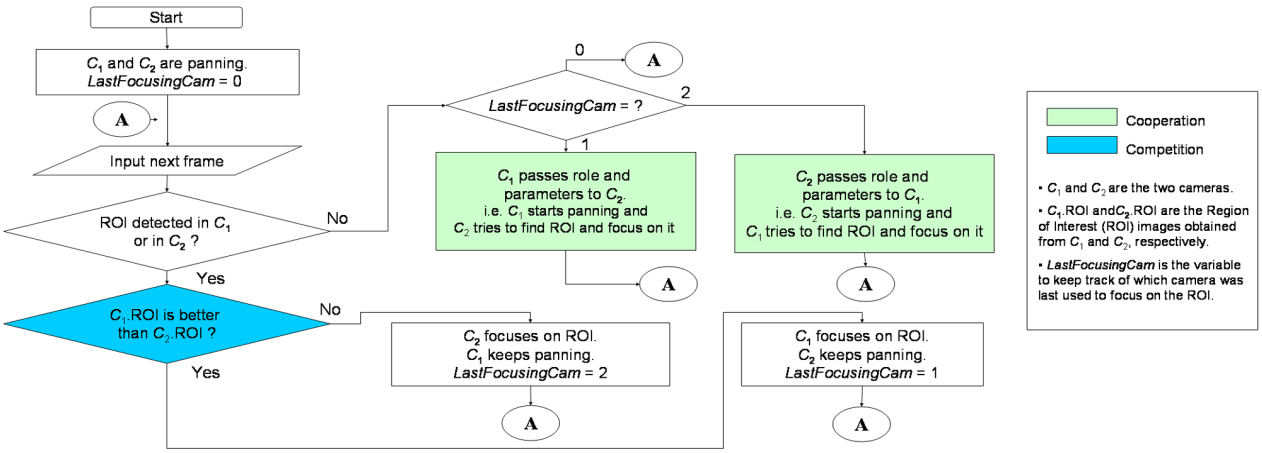

Figure 2: Algorithmic approach for the proposed framework

based on a measure of merit. In fact, our system provides only a generic framework and can support measures of merit of all varieties. It can be a simple decision based on face size in images or a complex function based on highest resolution facial image of person obtained as yet, room parameters, lighting conditions and so on. A measure of discrimination which gives higher priority to new intruder's faces is also plausible. However, for the purpose of our current experiments we use the size of un-zoomed face as a measure of merit to discriminate between sensors.

Let us say, without the loss of generality, camera C_1 wins the initial competition and starts focusing on the ROI. Then camera C_2 starts panning the entire area searching for new ROIs. However, if at a later point of time C_1 can not obtain ROI images anymore e.g. due to change of facial direction by intruder, it would cooperate by passing over its role as well as the information regarding possible location of the ROI to C_2. Hence our cameras both compete and cooperate at different moments of time to allow the best suited sensor to take over the task of obtaining ROI images.

3.3 Use of MPC

3.3.1 Introduction to MPC

Model Predictive Control provides better results than traditional control approaches [7, 12] as it also considers future values before deciding on the optimal control action. Hence, it provides the ability to counter various delays which are commonly encountered in visual surveillance tasks. For example, single frame visual feedback delay in all tracking processes and the transfer and reaction delays between interacting cameras in multi-camera frameworks, both affect the performance of the vision based systems [1, 14].

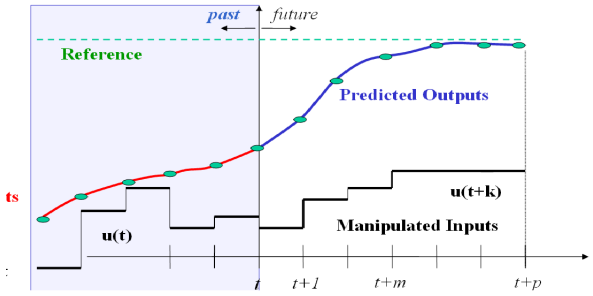

Figure 3: Working of Model Predictive Control

In our proposed framework in particular, we noticed an average transfer and reaction delay of 1.5 seconds when parameters were transferred between the cameras. Hence we use MPC to estimate future trajectories and preemptively react to them in order to neutralize the effects of the various delays. Simply put, the system tries to estimate what would be the *actual* position of the ROI by the time other camera is able to react to the incoming parameters. The system makes use of this information to pre-compensate for the various delays to be encountered. This in turn helps

in improving the quality of surveillance undertaken by the framework.

The working of the Model Predictive Control has been illustrated in figure 3 adapted from [10]. At time t, based on the past and future (estimated) values, the system tries to decide optimal values of manipulated inputs $u(t+k)$. However, only one input $u(t)$ is actually feeded into the system. The same process is repeated at time $(t+1)$ i.e. based on the input and output values till time $(t+1)$, future values of manipulated input and predicted output are decided. Such a process is repeated at the end of each time interval in the duration of interest i.e. till time $(t+p)$.

3.3.2 Our MPC Framework

The MPC framework adopted by us for surveillance has been divided into four parts as shown in figure 4. The *input* to the system is the movement to be made by the camera in order to try to bring the ROI centroid to the center of the image plane. The reference point for the signal is the center of the image plane and the *output* of the system is the actual position of the ROI centroid obtained on the image plane.

The aim of our framework is to obtain the images with the ROI centroid placed at the center of the image plane. In order to achieve this, the framework works as follows. The System Dynamics (Part A) is responsible for converting the system input i.e. control signal in terms of image plane $[x, y]$ coordinates into pan and tilt movement parameters for the camera. Based on this camera movement, we measure the ROI centroid position as the output of the system. However, in MPC, very often we try to predict the input and output data for future instances in order to achieve global optimization. Under such circumstances, we use a state estimation mechanism (Part B) to estimate the future ROI centroid positions. We are using a Kalman Filter approach to estimate the future ROI positions.

Part C refers to the reference point which is the center of the image plane. After obtaining the actual/estimated ROI positions for the duration of interest and comparing it with the Reference value we are able to obtain the error values in terms of $[x, y]$ coordinates. This value is passed to the Optimizer (Part D), which decides on the optimal control signal to be sent at the current instant so as to achieve an overall minimum value for the penalty function. The penalty function is a weighted average of the estimated error and the control effort required. The above mentioned process is repeated at the end of each frame in an effort to bring the ROI image centroid to the center of the image plane.

Now let us look at each of the four parts of our MPC framework mentioned above in more detail.

A: System Dynamics

The system dynamics part includes the conversion of the input control signal in term of $[x, y]$ parameters into pan and tilt angle values which are implemented on the camera. The relation between $[x, y]$ coordinate deviations and the corresponding angles can be found using camera calibration. In our particular framework using Canon VC-C4 cameras with 384 by 288 pixel resolution frames, we found the pixel to angle ratio to be 16 pixels/degree in both horizontal and vertical directions.

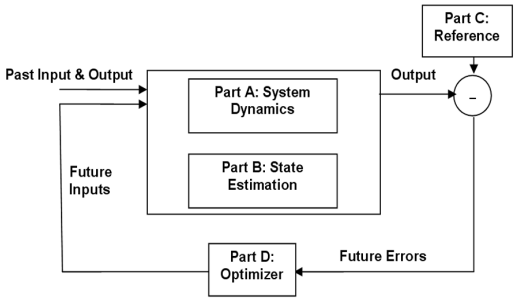

Figure 4: The proposed MPC framework for Visual Surveillance

B: State Estimation

The process of state estimation is required when we need to estimate the ROI positions for future instants of time. We have adopted the widely used [3, 1, 8] Kalman Filter approach for the purpose of state estimation. We have assumed a constant velocity model to undertake state estimation and modelled system noise as the difference between measured and predicted values at current time (t).

Our overall equation to calculate the ROI position for next time instant i.e. $(t+1)$ is -

$$y_o = \begin{bmatrix} p_x \\ p_y \end{bmatrix}_{t+1} = \begin{bmatrix} p_x \\ p_y \end{bmatrix}_t + \begin{bmatrix} T \\ T \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix}_t + G \begin{bmatrix} px_t - px_{t|t-1} \\ py_t - py_{t|t-1} \end{bmatrix} \quad (1)$$

y_o is the optimal estimate,
p_x is the position on x axis,
p_y is the position on y axis,
T is the time duration between frames,
v_x is the x-axis velocity,
v_y is the y-axis velocity,
G is the Kalman gain,
px_t is the x-axis position of ROI as measured at time t,
$px_{t|t-1}$ is the x-axis position of ROI at time t as predicted at time $t-1$,
py_t is the y-axis position of ROI as measured at time t,
$py_{t|t-1}$ is the y-axis position of ROI at time instant t as predicted at time $t-1$.

C: Reference

In our framework, we want the ROI centroid to be imaged at the center of the image plane. This allows high quality facial images to be obtained in the center of the image plane together with contextual information from the non-center portions. Our reference point always remains at the center of image plane [0, 0] assuming that egomotion i.e. the motion of the camera itself, has been compensated for. The compensation for egomotion is handled by the Optimizer section when it makes calculations on appropriate input parameters to be fed into the system.

D: Optimizer

The basic aim of the optimizer is to find out the optimal current input (u), which decreases the ROI tracking error as well as the control effort required. Hence, we want to

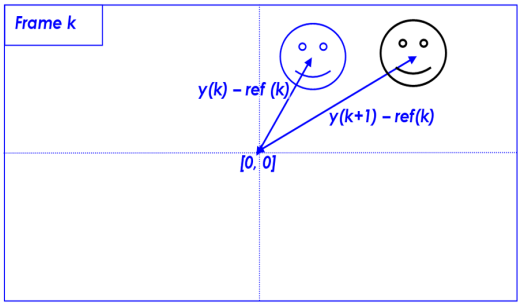

Figure 5: Movement of ROI centroid on the image plane between two consecutive frames

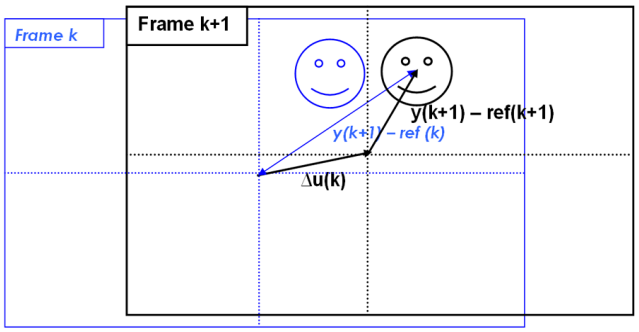

Figure 6: Impact of control input on the reference frame

minimize a penalty term that represents the weighted sum of future errors and control actions. This can be represented in a mathematical form as follows -

$$\min_{\Delta u} \quad Q \sum_{k=N_1}^{N_2} (y(k) - ref(k))^2 + R \sum_{k=N_1}^{N_2} (u(k+1) - u(k))^2 \tag{2}$$

where,
Q, R are the weight factors,
k is the instant of time being considered,
N_1, N_2 are the start and end point of the duration of interest,
$y(k)$ is the output i.e. ROI centroid on the image plane,
$u(k)$ is the control input i.e. movement of camera in terms of image plane coordinates,
$ref(k)$ is the reference signal i.e. image plane center [0,0].

The factors Q and R decide the relative importance given to the reduction of future error and the control action. After a few rounds of experiments and tuning, the values of 0.8 and 0.2 were found to be most appropriate for parameters Q and R respectively. This is due to the fact that in our framework reducing tracking error is much more important than reducing the camera movement.

The $y(k)$ as mentioned in equation 2 is the position of ROI centroid on the image plane and can be obtained for future frames by using Kalman Filter as described in equation 1. The $ref(k)$ represents the center of the image plane [0,0], but it needs to be compensated for the movement of the camera itself. Hence for future instants of time -

$$ref(k) = \Delta u(k) \tag{3}$$

This process of parameter translation across frames has been demonstrated in figure 5 and figure 6. While figure 5 demonstrates the normal error as observed for two consecutive frames in the case of no camera movement between frames, figure 6 demonstrates how these parameters are translated in the next frame in case there is camera movement between frames. Basically the reference itself moves by $\Delta u(k)$, and hence the effective error should be reduced by $\Delta u(k)$ to compensate for the camera movement.

The second term of equation 2 clearly represents the movement of the camera between frames and can be written as $\Delta u(k)$ too. Hence, equation 2 can also be written as -

$$\min_{\Delta u} \quad Q \sum_{k=N_1}^{N_2} (y(k) - \Delta u(k))^2 + R \sum_{k=N_1}^{N_2} (\Delta u(k))^2 \tag{4}$$

To facilitate the solution of this equation we bring it to a standard quadratic form, which is -

$$\min_{x} \ \mathbf{x}' H \mathbf{x} - g' \mathbf{x} \tag{5}$$

For which the solution is described in [10] as $\mathbf{x} = 0.5 g H^{-1}$. We translate our MPC problem of equation 4 to the standard quadratic form given in equation 5. The process of translation and simplification is similar to one described in [10]. The specific mathematical steps have been omitted due to space constraints. After simplification we obtain the final value of input parameters as -

$$\Delta U(k) = \frac{Q^2}{Q^2 + R^2} Y(k)_{n \times 2} \tag{6}$$

where,
$\Delta U(k)$ is a matrix containing optimal values of next n incremental movements to be made by system at time k
$Y(k)$ is a matrix that represents the next n estimated ROI positions at time k,
Q and R are the weight parameters as described earlier.

We use the above equation to obtain the values of $\Delta U(k)$ matrix after each iteration. The first value of this matrix is actually fed into the system and after this $\Delta U(k+1)$ is calculated at the next iteration. The first value of $\Delta U(k+1)$ matrix is fed into the system and such a process is repeated for the entire period of interest.

4. RESULTS AND ANALYSIS

In this paper, we have described an active multi-sensor *coopetitive* interaction framework for visual surveillance. It is fairly intuitive that our framework has gained considerably through the use of active sensors (allowing camera movement for tracking) and multiple sensors (obtaining facial data in both directions).

Before embarking upon detailed results and analysis, it is imperative to realize that there is no direct comparison possible between our proposed system and other systems described in literature e.g. [17], [15], [9] etc. While these systems can capture facial images in only one direction, our

system is capable of handling facial image data in both directions i.e. even when the person traversing the surveyed area starts walking in the opposite direction. Hence our system will perform significantly better in all cases where the person traversing is allowed to change his facial direction. Also, even if we consider just the subset of trajectory where the person faces in only one direction, there is no direct comparison with other systems possible as our system uses just one sensor per direction where as all other systems use two sensors (e.g. one Wide-angle sensor and one PTZ sensor) per direction to undertake a similar tracking task.

To establish the veracity of our proposed framework in performing surveillance tasks, we have conducted comprehensive experiments to compare the *coopetitive* interaction approach and MPC feedback mechanism with their possible alternatives. We compare the *coopetitive* interaction approach with other plausible approaches such as 'only cooperation' and 'only competition'. We also compare the performance of the proposed MPC feedback mechanism with that of the popular PID control using various experiments.

We do not use standard data sets like those described in PETS [4] etc. to evaluate the performance of our surveillance system, as such data sets do not provide any means for estimating the performance of real time surveillance experiments. They provide off-line images to be used for evaluating the performance, which is not possible in our system as it needs to undertake physical panning, tilting and zooming operations in real time in order to capture the ROI. Hence, we use real-time experiments to evaluate the performance of our system.

Experiment 1 helps us in determining which camera interaction approach or feedback mechanism provides us with best ability to obtain images of intruders traversing certain definitive trajectories in an enclosed environment. Experiment 2 helps us in determining which interaction approach or feedback mechanism can detect an intruder most number of times in a given time period even if the intruder is allowed freedom to *choose his own trajectory* and *intentionally try to avoid detection*. This is a significantly bold experiment to undertake considering that none of the other surveillance works have tried to use performance under intentional detection avoidance as a measure of their effectiveness. In Experiment 3, we do a final comparison between MPC and PID feedback mechanisms in terms of their ability to track and obtain high resolution images of the ROI.

For the purpose of all three experiments we have adopted the scenario of an intruder entering an enclosed single door rectangular premises e.g. an ATM lobby or a museum subsection. The specific location used for our experiments was a 4 meter by 6 meter rectangular area. Also, for the purpose of our experiments the volunteer intruders were allowed to face either directly in front or directly opposite with respect to the principal artifact. This was due to the availability of only frontal face detectors. This condition may be relaxed in future if non-frontal face detectors become available.

4.1 Experiment 1: Good quality images in an enclosed environment

A common surveillance goal is to obtain 'good quality' images of an intruder entering an enclosed environment which can be used for further automated processing e.g. face recognition etc. Images containing frontal face data with a minimum resolution of 90 by 90 pixels can be used by automated

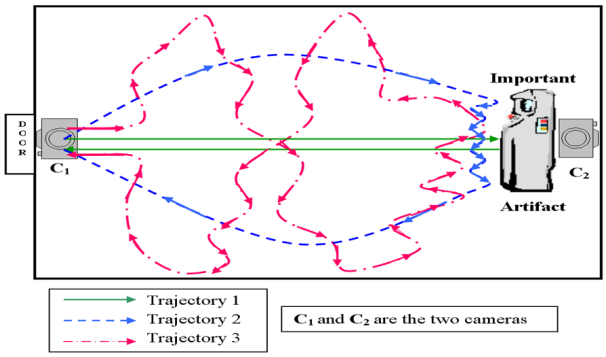

Figure 7: Different trajectories for intruders traversing the enclosed rectangular area

face recognition procedures with an accuracy of around 90% [16], hence we have adopted this as a definition of 'good quality' image in our current experiment. In this experiment we allowed the intruder to take three definitive and representative trajectories to traverse the enclosed environment. The ability of different surveillance frameworks in terms of mode of interaction and feedback mechanism was compared with our proposed framework. The three trajectories which were used for this experiment have been shown in figure 7.

The first trajectory involves only straight line motion and represents the simplest and most common path of traversing the enclosed area. The second trajectory displays the possible path for an intruder who walks in with a slightly curved path, spends sometime in front of the important artifact but keeps on changing the direction of his face. This could be a symptom that he does not want his 'good quality' images to be captured by the surveillance system.

The third trajectory represents a possible path to be taken by a determined adversary who might be aware of the mechanisms of cameras interaction and the delays involved. He may want to change his direction very often and do so when he is moving forward as well as when he is moving sideways. His intention would be to exploit the delays encountered in camera role swapping to avoid getting his good quality images captured.

The idea behind choosing these trajectories was to represent most of the possible intruder movements in enclosed environments. We have covered both horizontal and lateral movements of the intruder. We have also tried to represent various situations ranging from simple straight motion to complex face direction change coupled with horizontal and lateral movement.

The performance of the different possible frameworks was measured quantitatively in terms of percentage of frames in which they were able to obtain good quality images of the intruder for the various trajectories.

For the first part of the experiment, we studied how the mode of camera interaction affects the performance of the surveillance framework. We compared our proposed *coopetitive* approach with the base case approaches of 'only cooperation' and 'only competition'. We define 'only cooperation' as an interaction mechanism in which cameras try to help each other without any measure of merit or discrimination between them. Hence we have adopted an equal time-slicing approach to represent the process of 'only cooperation'. The

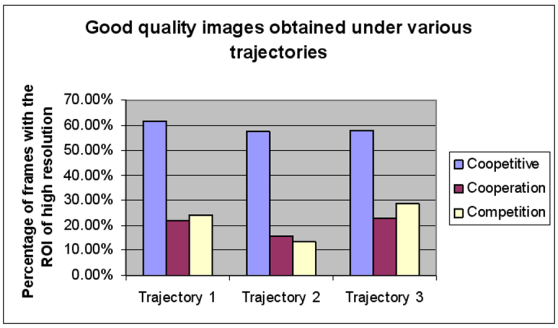

Figure 8: Comparison between different modes of camera interaction

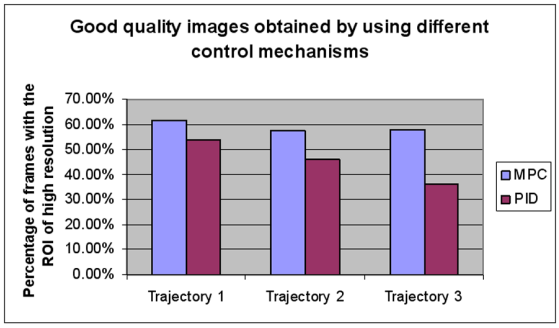

Figure 9: Comparison between different feedback mechanisms

cameras pass each other the parameters and the roles after a fixed interval of time.

The 'only competition' approach has been defined as that in which cameras continuously compete with each other for the task of tracking and focusing on the intruder. There is no cooperation between cameras and hence the first camera does not pass any information to its peer camera in case it can not detect the intruder face anymore.

The results obtained for this experiment have been shown in figure 8. We notice a clear advantage obtained by the use of *coopetitive* interaction over 'only cooperation' or 'only competition' in all three trajectories. For example in trajectory 1, good quality of images were obtained for 62% of the frames using *coopetitive* interaction approach. The 'only cooperation' or 'only competition' interaction approaches on the other hand could obtain good quality images for only 22% and 24% frames respectively.

Another interesting point to note is that on average for the three trajectories, the 'only cooperation' approach obtained good quality images for only 20% of frames which was even lesser than 22% obtained by 'only competition' approach.

In the second part of this experiment, we studied the impact of feedback mechanism, on the ability of the surveillance framework to obtain good quality images of the intruder. The results obtained are shown in figure 9. We noticed a clear advantage by the use of MPC mechanism for feedback. For the first trajectory, MPC feedback framework could capture good quality images for 62% of frames as compared to 54% frames for PID. While this did not represent a very significant performance improvement, the difference became very obvious as the trajectories became more complex. For example, in trajectory 3, MPC showed its scalability and continued to perform reasonably well, capturing good quality images for 58% of frames where as PID mechanism's performance degraded significantly to capturing good quality images for just 36% of frames.

4.2 Experiment 2: Intruder detection

An important measure of a surveillance mechanism's competency is its ability to detect intrusions in the surveyed premises. This intrusion detection should be achievable even if the intruder is intentionally avoiding detection. For the second experiment, we measured the performance of different frameworks in terms of their ability to capture facial images of the intruder even when he chooses his own trajectory and changes his direction as many times as he wants in a deliberate effort to avoid detection by the surveillance system.

In this experiment, we asked the volunteers to enter the surveyed premises for duration of one minute. They were given freedom to *define their own trajectory* and also *when and how often they change their directions*. They were asked to spend at least 5 seconds in front of the important artifact of the surveyed premises. This was to simulate real life scenario in which intruders may try to avoid detection but would still need to spend at least some time in front of the important artifact to undertake any task of significance.

Ten runs of this experiment conducted with the help of volunteers who were post-graduate students from our multimedia research laboratory. The average values for the number of times the intruder was detected were calculated and have been shown in figure 10. We found that our framework which adopts *coopetitive* interaction approach combined with MPC feedback mechanism performed significantly better than other plausible frameworks. On average our proposed framework obtained facial images 17.9 times within one minute which is 27.4 % higher than nearest alternative of using PID feedback mechanism which could detect faces only 13 times within one minute.

The 'only cooperation' approach could capture intruder face only 11.3 times on average which was again poorer than 12.4 of 'only competition' approach. The results obtained from experiments 1 and 2 indicate to us that cooperation undertaken without any measure of merit of discrimination between sensors can actually *reduce* the overall performance of the system rather than increasing it.

4.3 Experiment 3: Tracking error and size of ROI

Experiments 1 and 2 clearly indicate the superiority of *coopetitive* mode of interaction as compared to 'only cooperation' and 'only competition'. On the whole we have realized that amongst the various possible alternatives, our framework which uses MPC combined with *coopetitive* interaction has performed best followed by the PID feedback framework.

In this experiment, we further investigate the performance of our framework as compared to PID feedback framework. We shall evaluate the performance in terms of tracking error and the size of ROI. In an ideal tracking scenario, the ROI centroid should coincide exactly with the center of the image plane. Hence, we define tracking error as the Euclidean

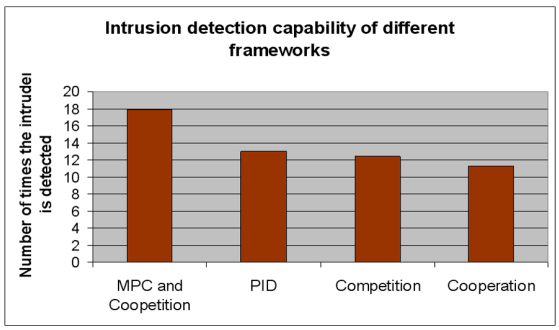

Figure 10: Comparison of different possible frameworks in terms of ability to detect intrusions in enclosed premises

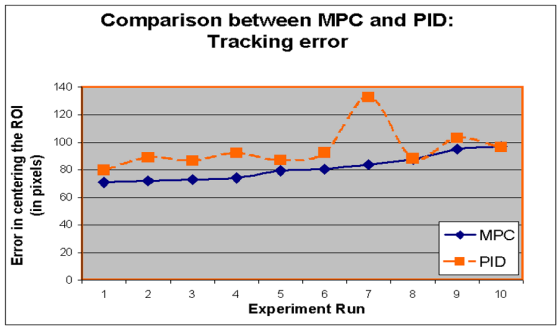

Figure 11: Comparison between MPC and PID in terms of tracking error

distance in pixels between the center of the image plane and the ROI centroid. The size of ROI for our experiment is measured in terms of the size of the frontal face in frames which detect the intruder.

The scenario and settings for experiment 3 were same as those of experiment 2 but now we are focusing on the tracking ability and the ROI capturing abilities of the different frameworks. The average values of tracking error for the ten experiment runs have been shown in figure 11. We notice a clear trend that PID suffers from more error than MPC. The average error for PID was 94.8 pixels as compared to 81.4 pixels for MPC. This represents an error reduction of 16.5% by the use MPC in tracking.

The values for the size of the ROI (frontal facial image) have been summarized in figure 12. Again, we notice that MPC performs significantly better than PID. The average size of the ROI in the images obtained by MPC was 68 by 68 pixels as compared to 59 by 59 pixels for PID. This represents an improvement of around 13% in the size of ROI by the use of MPC as a feedback mechanism. This further convinces us on MPC's superiority as a feedback mechanism when compared to PID which has been adopted in frameworks such as that described in [13].

4.4 Further Discussion

Another important concept which we have introduced in this paper is that of transfer of roles and sub goals rather than just parameters between sensors. Such an approach

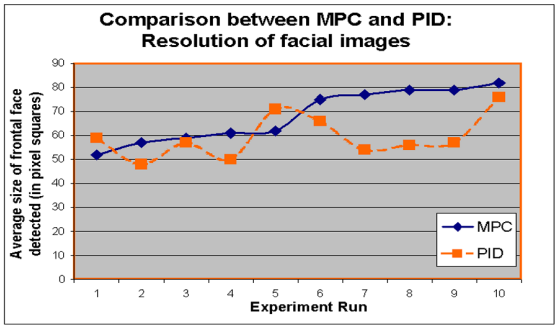

Figure 12: Comparison between MPC and PID in terms of size of ROI in captured frames

can be implemented in practice via dynamic role swapping. This dynamic role swapping allows the best suited sensor at the best suited position to undertake the sensing task for global optimization. This concept of dynamic role swapping also helps to significantly reduce the cost of sensors for obtaining data.

Looking at it analytically, Let us consider a multi-sensor scenario in which sensors with n different roles must be placed at m different positions of vantage. Hence using a simplistic framework we would need $n \times m$ different sensors to undertake a task such as surveillance. Dynamic role swapping on the other hand proposes the use of just m competent sensors placed one at each location which can assume the n different roles at different points of times. This would reduce over-all surveillance cost without any loss of quality. Hence the reduction in number of sensors would take place in the ratio of $1 : n$. This reduction ratio represents a significant cost reduction for many real life situations.

Dynamic role swapping is often possible and must be adopted at all such available opportunities. For example, surveillance frameworks described in [17, 9, 15] use a wide-angle camera and one PTZ camera to obtain surveillance images. However, the use of just 2 cameras in these frameworks provides facial information in only one direction. To obtain facial images in both directions similar to our proposed framework, these frameworks would need 4 cameras as opposed to just 2 in our framework. This indicates a reduction ratio of 1 : 2 in terms of number of cameras used for surveillance task. This has been possible with the use of continuous panning as an alternative to wide angle cameras and the use of competent active cameras which can assume the role of both focusing and continuous panning.

To sum up our discussion for this section, we have proposed an active multi-sensor framework with *coopetitive* interaction for surveillance tasks. Through three comprehensive practical experiments we have established the superiority of MPC as a feedback mechanism for active visual sensing. Through the results of experiments 1 and 2, we have also clearly demonstrated the advantages of using *coopetition* mode of camera interaction as compared to 'only cooperation' or 'only competition'. Finally we have also put forward the analytical reasoning for the improvements created by the transfer of roles between sensors rather than just parameters in various multi-sensor environments.

5. CONCLUSIONS

In this paper we have proposed enhancements for cooperative multi-sensor environments in terms of their mode of interaction and feedback mechanism. We have proposed *coopetitive* approach for interaction between sensors which allows sensors to cooperate based on a merit decided by competition. We have also established the value of MPC as an efficient feedback mechanism that can help to counterbalance various transfer and reaction delays observed in cooperative sensing. The application of all these concepts has been shown by the means of a dual camera surveillance framework used for surveying single door enclosed environments.

From the results obtained we can clearly conclude that for interaction between sensors, *coopetition* i.e. cooperation based on merit performs significantly better than 'only cooperation' or 'only competition' approaches. In fact, we realized that 'only cooperation' without a notion of merit or discrimination between sensors may reduce the effectiveness of the system rather than improving it. We also deduce that MPC performs significantly better than PID as a feedback mechanism for vision sensors. This is by virtue of MPC's capability to consider estimated future values rather than just past data to make its control decisions.

Future work scope in this area remains in creating more precise means for defining and handling cooperation and competition between sensors. More sophisticated means of estimating future states would also be very useful. In terms of the framework itself, better ROI detection mechanisms as well as incorporating different types of ROI are worth exploring so as to achieve further improvements in surveillance results.

Future work would also be undertaken to extend the proposed framework to handle multiple ROIs and multiple sensors which could be visual as well as non-visual e.g. audio sensor, infra-red sensors etc. Such non-visual sensors would allow the framework to handle situations where visual sensing alone might fail e.g. intruder hiding the face or using a face-mask etc.

6. ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our supervisor Dr. Mohan Kankanhalli, who has guided us and provided valuable insights at each step in the course of this work.

7. REFERENCES

[1] J. P. Barreto, J. Batista, P. Peixoto, and H. Araujo. Integrating vision and control to achieve high perfomance active tracking. Technical report, TR-BAR-0202, ISR/DEEC - University of Coimbra, Febraury 2002.

[2] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multi-sensor surveillance. In *Proceedings of the IEEE*, pages 1456–1477, October 2001.

[3] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A system for video surveillance and monitoring. Technical report, CMU-RI-TR-00-12, Robotics Institute, CMU, USA, May 2000.

[4] Datasets. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Breckenridge, USA, January 2005.

[5] M. Dias and A. Stentz. A market approach to multirobot coordination. Technical report, CMU-RI -TR-01-26 - Carnegie Mellon University, August 2001.

[6] M. Greiffenhagen, V. Ramesh, and D. Comaniciu. Statistical modeling and performance characterization of a real-time dual camera surveillance system. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, USA, June 2000.

[7] M. Khadir and J. Ringwood. Linear and nonlinear model predictive control design for a milk pasteurization plant. *Journal of Control and Intelligent Systems*, 31(1), 2003.

[8] K.-Y. Lam and C. K. H. Chiu. Adaptive visual object surveillance with continuously moving panning camera. In *Proceedings of the 2nd ACM International Workshop on Video surveillance and Sensor Networks*, October 2004.

[9] Q. Liu, D. Kimber, J. Foote, L. Wilcox, and J. Boreczky. FLYSPEC: A multi-user video camera system with hybrid human and automatic control. In *ACM International Conference on Multimedia*, New York, USA, December 2002.

[10] M. Morari, J. H. Lee, C. E. Garcia, and D. M. Prett. *Model Predictive Control*. Prentice Hall, Englewood Cliffs, New Jersey, 2003.

[11] N. Papanikolopoulos, P. Khosla, and T. Kanade. Visual tracking of a moving target by a camera mounted on a robot: A combination of control and vision. *IEEE Transactions on Robotics and Automation*, 9(1), 1993.

[12] P. K. Roy, G. Mann, B. C. Hawlader, V. Masek, and S. O. Young. Comparative study of model predictive and decoupled PID controller for a multivariable soil heating process. In *The 14th Annual IEEE Newfoundland Electrical and Computer Engineering Conference*, Newfoundland, Canada, October 2004.

[13] M. Saedan and M. H. A. Jr. 3D vision-based control on an industrial robot. In *Proceedings of IASTED International Conference on Robotics and Applications*, Clearwater, USA, November 2001.

[14] P. Sharkey and D. Murray. Delays versus performance of visually guided systems. In *IEE Proceedings on Control Theory Applications*, pages 436–447, September 1996.

[15] S. Swarup, T. Oezer, S. R. Ray, and T. J. Anastasio. A self-aiming camera based on neurophysical principless. In *Proceedings of The International Joint Conference on Neural Networks*, Portland, USA, July 2003.

[16] J. Wang, C. Zhang, and H. Shum. Face image resolution versus face recognition performance based on two global methods. In *Asian Conference on Computer Vision*, Jeju Island, Korea, January 2004.

[17] X. Zhou, R. Collins, T. Kanade, and P. Metes. A master-slave system to acquire biometric imagery of humans at distance. In *ACM International Workshop on Video Surveillance*, Berkley, USA, November 2003.