Network Video Image Processing for Security, Surveillance, and Situational Awareness

Abhijit Mahalanobis, Jamie Cannon, S. Robert
Stanfill, Robert Muise
Lockheed Martin, MFC
MP 450
5600 Sand Lake Road
Orlando, FL 32819

Mubarak Shah
School of Computer Science
University of Central Florida
Orlando, FL 32826

Abstract

Lockheed Martin and the University of Central Florida (UCF) are jointly investigating the use of a network of COTS video cameras and computers for a variety of security and surveillance operations. The detection and tracking of humans as well as vehicles is of interest. The three main novel aspects of the work presented in this paper are i) the integration of automatic target detection and recognition techniques with tracking ii) the handover and seamless tracking of objects across a network, and iii) the development of real-time communication and messaging protocols using COTS networking components. The approach leverages the previously developed KNIGHT human detection and tracking system developed at UCF, and Lockheed Martin's automatic target detection and recognition (ATD/R) algorithms. The work presented in this paper builds on these capabilities for surveillance using stationary sensors, with the goal of subsequently addressing the problem of moving platforms.

1. Introduction

In recent years, the need for distributed surveillance systems has emerged in applications ranging from homeland defense to modern network centric warfare concepts. Various DoD efforts are on-going to develop and deploy such capabilities in the near future. The Army's Future Combat Systems (FCS) program is an example of how network centric concepts are changing the battlefield. Military operations in urban terrain (MOUT) scenarios also call for distributed information gathering and processing capabilities. Video cameras abound in civilian life wherever security is of interest (e.g. in commerce, transportation, education, entertainment and so forth). The viability of distributed security and surveillance capabilities is enabled by the advent of low-cost cameras, computers, and networking technology (both wired and wireless). Although the component technologies and the infrastructure for such systems already exists today, the challenge is in developing algorithms that work across multiple platforms, and addressing the bandwidth and communication issues.

The goal of this effort is to take advantage of multiple cameras (with overlapping or non-overlapping fields of view) in order to monitor activity over a large area. The system must be able to handle both stationary and moving objects. While motion analysis can be used to detect vehicles and humans when they are moving, the ATD/R capability is required for detecting and initiation tracks when they are stationary, and recognizing the detected objects. The system must be able to detect, track and handed over moving objects between cameras in real-time. For seamless operation across platforms, this requires the

position of the target in the next field of view (FOV) to be predicted. Based on an analysis of the location of the detections, and registration between the camera views, it becomes possible to depict the positions of the objects and their movements with respect to a site map, thus providing a global composite view of events. This can serve as a powerful monitoring tool by providing situational awareness over the site of engagement.

The rest of the paper is organized as follows. Section 2 is an overview of the ATD/R process, and describes how targets are detected, and multiple views are brought together in a combined view of the world. Section 3 is an overview of the KNIGHT human detection and tracking system [1] developed at UCF. We describe in this section how the process is augmented to track across multiple FOVs. The integration of the ATD/R and tracking system and the highlights of the networking and communication process are discussed in Section 4, along with a description of initial results obtained. Finally, Section 5 provides a summary of the work performed to date, along with a discussion of the challenges and directions for future work.

2. Target Detection in Multiple Views

We first discuss the approach for detecting stationary vehicular objects (interchangeably referred to as targets). Various target detection and recognition methods may be used depending on the sensor type, range to target, resolution and other key driving parameters. The objective here is not to build a better ATD/R capability, but to extend the algorithms to work across multiple platforms. For convenience, we use the *maximum average correlation height* (MACH) Correlation Filtering approach [2] for target detection and classification. The basic concept of operation using correlation filters is shown in Figure 1.

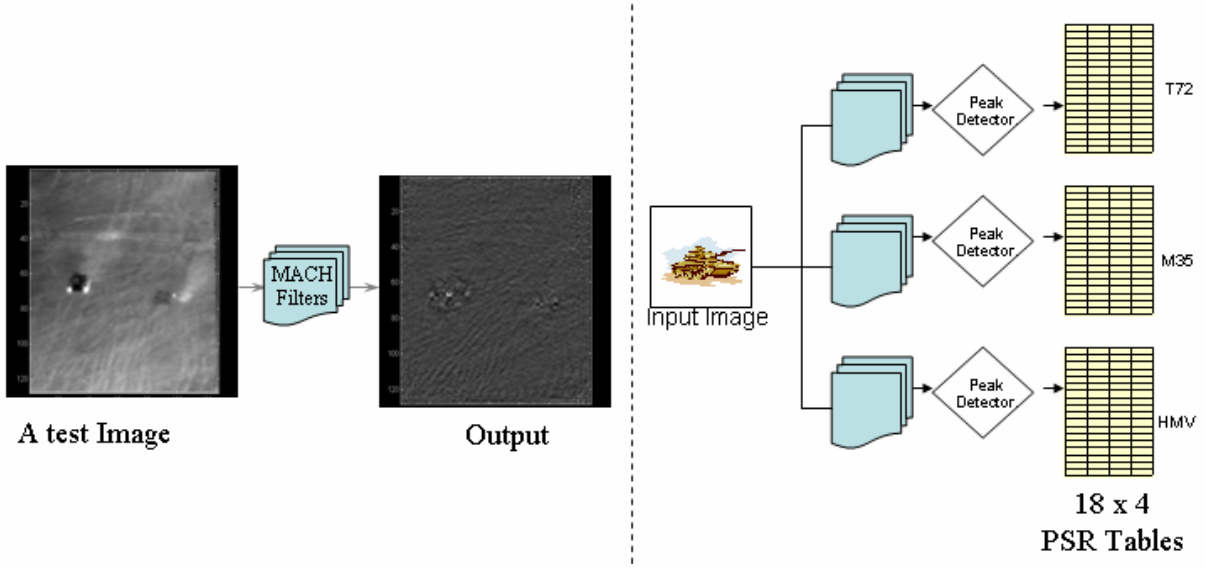

Figure 1: An input image is processed by a bank of correlation filters to detect and identify targets. The filter with the highest PSR determines the class of the object, and the position of the correlation peak indicates its location.

Essentially, the input test image is processed by a bank of linear correlation filters that are optimized to respond to the presence of a target by producing a peak at the corresponding location in the output image (also known as correlation plane). Since correlation is a shift-invariant operation, the position of the peak always represents the location of the target, even when it is moving. Each filter is synthesized using

representative training images to exhibit distortion tolerance over a limited range of orientations and signature variations. Thus multiple filters are required for every class to accommodate all possible distortions. For example in Figure 1, there are 72 correlation filters for every target to cover 18 aspect bins (each 22.5 degrees wide) and 4 different signature types (for thermal images these conditions may be hot, cold, day and night signatures). A metric known as *peak to sidelobe ratio* (PSR) is used to measure the strength of the correlation peaks. The class of the target is declared to be the same as the filter which yields the highest PSR value.

Figure 2 illustrates the concept of networking multiple "nodes", each with a camera, processor and on-board ATD/R capability. The outputs of each node is received at a central "command and control" point where the information is combined. For now, we assume that the sensors and the platforms on which they reside are stationary. Since bandwidth is limited, each node only reports the ATD/R results including pixel position of the detections. Using knowledge of the camera geometry, the location of the target in the sensor view can be converted to a common reference frame, and represented as a point on a site map or an aerial view of the region obtained using an overhead asset. This is further illustrated in Figure 3 where the image in the top window serves as the "site map" and the smaller windows at the bottom represent the three sensor views. The target is detected and recognized in each sensor view, and its pixel position is reported. This data is collected at a central computer where the target coordinates are converted to a common reference frame and fused to depict the location of the target (represented in Figure 3 by the red square) on the site map or overhead image. When the target moves in the sensor views, the site map is update in real-time so that the target position can be seen moving on the site map.

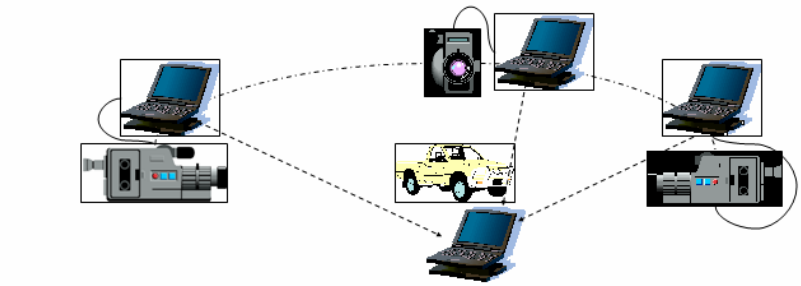

Figure 2: Multiple platforms (sensors) are networked to a central computer where ATD/R information is combined into a common reference frame.

The main advantage of the process illustrated in Figure 3 is that an updated site map that depicts the combined information from multiple sources can be a valuable tool for situational awareness. While individual sensors have only a limited view of the world and may not be able to see around buildings and other obstructions, the combined information can be "dialed-up" by any of the nodes. This allows the local platforms to benefit from the information observed by others in the network. It also allows the command and control center to have a cohesive picture of the battlefield based on multiple observations.

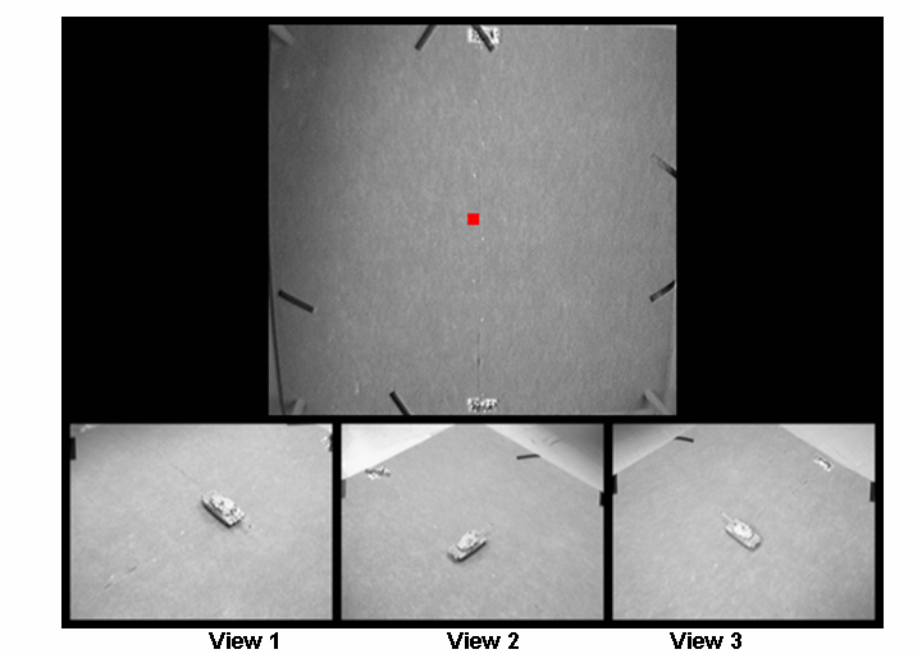

Figure 3: The views from the 3 separate nodes (shown at the bottom) are processed locally by an ATD/R, and the position of the target is reported. This data is collected at a central computer where the target coordinates are converted to a common reference frame and fused to depict the location of the target on a site map (or overhead image).

3. Target hand-over and tracking across multiple platforms

Target tracking is an integral and important part of a surveillance system. We first provide a brief overview of the KNIGHT tracking system [1] designed for single camera systems, and then describe its extension to tracking across multiple FOVs. KNIGHT is a 'smart' surveillance system that detects important changes, events, and activities using computer vision techniques, flags significant events, and presents a summary in terms of key frames and textual description of activities to a monitoring officer for final analysis and response decision. The system is robust to illumination changes and weather conditions. KNIGHT has been installed at four locations in the downtown Orlando area which has Orange Avenue as its primary street, and is currently being field tested. The system employs single camera, and works in real time. KNIGHT consists of four main modules, three of which are shown in Figure 4: object detection and shadow removal, tracking object classification and activity detection.

Specifically, we view tracking as a region correspondence problem where performance is affected by noisy background subtraction, change in the size of regions, occlusion and entry/exit of objects. For these reasons traditional approaches cannot be directly applied to tracking humans. To achieve correct correspondence, we have developed a solution based on linear velocity, size and distance constraints. Furthermore, most of the surveillance systems do not tackle the problems in tracking caused by shadows. To address this issue, we employ a shadow detection approach based on similarity of background and shadow regions.

In addition to tracking moving objects, we believe that motion based classification helps to reduce the reliance on the spatial primitives of the objects and offers a robust but computationally inexpensive way

to perform classification. We have devised a solution to this problem using temporal templates. Temporal templates are used for classification of moving objects. A temporal template is a static vector image in which the value at each point is a function of motion properties at the corresponding spatial location in the image sequence. Motion History and Motion Energy images are examples of temporal templates, proposed by Bobick and Davis [3]. Motion History image is a binary image with a value of one at every pixel where motion occurred. In Motion History image pixel intensity is a function of temporal history i.e. pixels where motion occurred recently will have higher values as compared to other pixels. These images were used for activity detection. We have defined a specific Recurrent Motion template to detect repeated motion. Different types of objects yield very different Recurrent Motion Images (RMI's) and therefore can easily be classified into different categories on the basis of their RMI. We have used the RMIs for object classification and also for detecting carried objects.

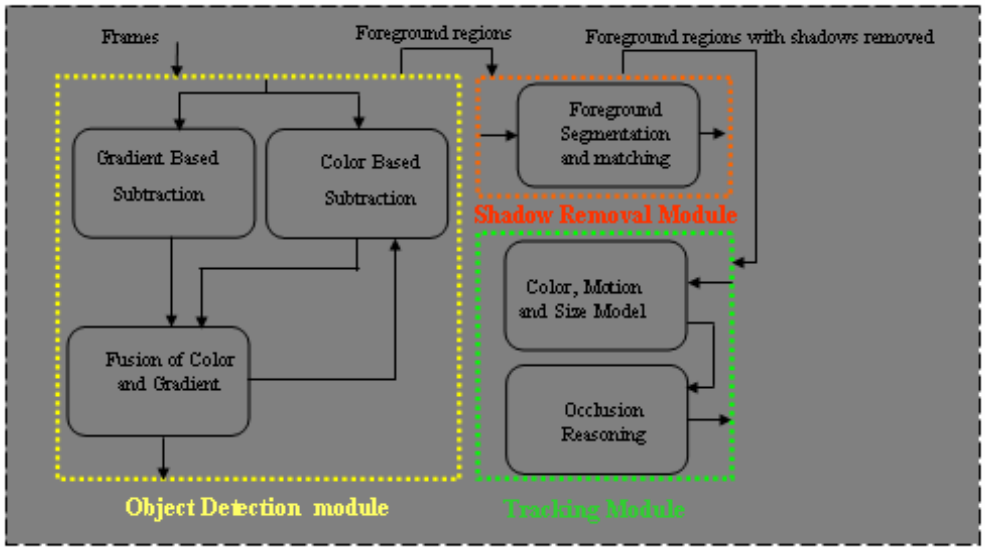

Figure 4: A overview of the KNIGHT motion based detection and tracking system for humans and vehicles. Additional details can be found at http://www.cs.ucf.edu/~vision/projects/Knight/Knight.html

Tracking across multiple fields of view

To track objects successfully in multiple cameras, one needs to establish correspondence between objects detected and tracked in each camera. Our system is able to discover spatial relationships between the camera FOVs and use this information to correspond between different perspective views of the same person. We employ a novel approach of finding the limits of FOV of a camera as visible in the other cameras that is very fast compared to conventional camera calibration based approaches. Using this information, when a person is seen in one camera, we are able to predict all the other cameras in which this person will be visible. Moreover, we apply the FOV constraint to disambiguate between possible candidates for correspondence.

When tracking is initiated, there is no information provided about the FOV lines of the cameras. The system can, however, find this information by observing motion in the environment, as illustrated in Figure 5. Whenever there is an object entering or exiting one camera, it actually lies on the projection of the FOV line of this camera in all other ones in which it is visible. Suppose that there is only one

target. Then, when it enters the FOV of a new camera, we find one constraint on the associated line. Two such constraints will define the line, and all constraints after that can be used in a least squares formulation. In an earlier paper [4], it was demonstrated that the initialization of FOV lines by one person walking in the environment for about 40 seconds was sufficient to initialize the lines. These lines were then used to resolve the correspondence problem between cameras. However it is not always possible to have only one target moving in the scene. When multiple targets are in the scene and if one crosses the edge of FOV, all targets in other cameras are picked as being candidates for the projection of FOV line. Since the false candidates are randomly spread on both sides of the line whereas the correct candidates are clustered on a single line, correct correspondences will yield a line in a single orientation, but the wrong correspondences will yield lines in scattered orientations. We can then use Hough transform to find the best line in this case. This method needs more points for a reliable estimate of the lines and therefore takes longer time to set up correctly. Additional constraints derived from categorization of objects and their motion may be used to reduce the number of false correspondences, thus reducing the time it requires to establish the lines.

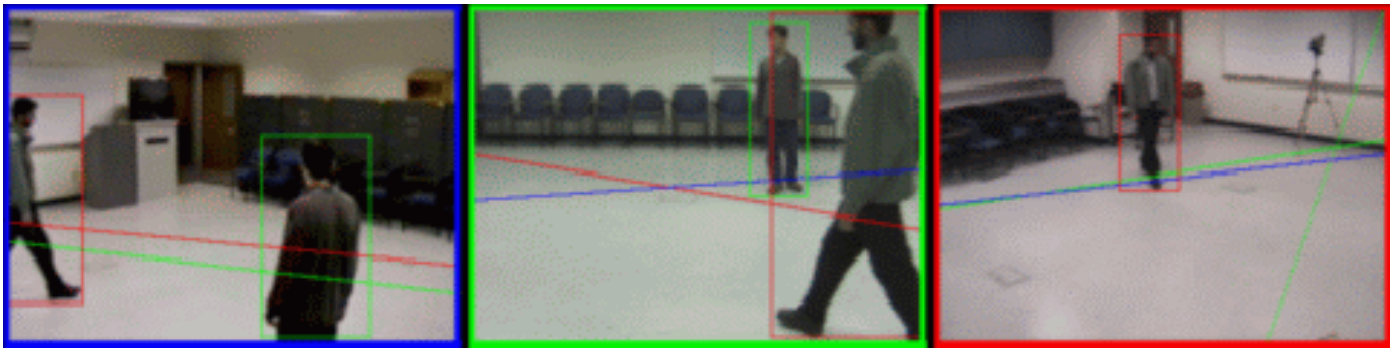

Figure 5: The automatic calibration of three separate cameras with overlapping fields of view (FOVs) is shown. The FOV boundary lines are established by observing where moving objects visible in one camera simultaneously appear at the edge of another camera's view. Places where this occur represent points on the boundaries of FOVs of other cameras that are visible in the current view.

4. Network Integration of Tracking and ATD/R

The ability to detect, track and recognize objects across a network has been demonstrated across both wired and wireless networks. The concept of a wireless peer-to-peer ad-hoc network is shown in Figure 6. This system was built and tested using laptop PCs, each equipped with a Synchrotech^{TM} adapter and a wireless card from MeshLAN^{TM}. We also tested operations on a commercially available 802.11b wireless hub. A socket based communication over TCP/IP was used to network three PCs that acted as "clients" and a fourth one as the "server". Network architecture is traditionally split into layers starting at the top application layer and going progressively down towards the hardware. The Transmission Control Protocol (**TCP**) forms the Transport layer and beneath it the Internet Protocol (**IP**) forms the Network layer. The Transport layer looks after assembling whole messages from individual packets whatever route they may take and the Network layer looks after getting individual packets across the network. If data packets are lost then TCP automatically attempts to retry the operation. It uses a simple acknowledgement interchange to ensure this. Within TCP/IP, the two communicating programs (server and client), allocate sockets and then connection is initiated by the client program. The server continually listens for connect requests and then chooses to accept a connection from them. This client-server model is an appropriate

scheme for the distributed network as there are many clients making connection requests for information from one place (the server).

Figure 6: Example of a ad-hoc Peer-to-Peer network.

The particulars of the interactions are as follows. The KNIGHT tracking system executes locally at each of the clients and the local tracking data is sent to the server. The server ensures that the tracked entities from the clients are de-conflicted and properly associated, and assigned global labels as described in Section 3. It is also essential to synchronize the frames processed at the clients so that the proper temporal correspondence can be made. The global labels are then received back at the clients and used for consistent labeling and display purposes. At startup, the server is in a "training mode" to establish the FOV boundaries based on the entry and exit of moving objects across the different FOVs. Thereafter, the main purpose of the server is to generate and return consistent labels for the tracked objects.

Figure 7: ATR – Tracker interactions occur only at each client. The results of the ATR including class label and confidence are sent to the server.

The ATR-Tracker interactions occur only at each client, as shown in Figure 7. When a moving object is detected[1], a 64 x 128 region of the image containing the tracked object is fed to the ATR for classification, and the result is used for generating the *target call* (class label) and *confidence* associated with that object, which is then sent to the server. When a new object enters the FOV of a client the *target call* sent by the ATR is used as the label. If however the object is already in track (i.e. it corresponds to an existing object) it gets the *target call* with the highest *confidence* (including those from previous classification results) is assigned to the object as its label.

Figure 8 illustrates the interaction of the ATR and tracker across a network using models for a "Tank" and a "Mini", a relatively smaller vehicle. The pictures represent snapshots of actual events that occurred during a real-time test and demonstration of the algorithms. As these objects move from right to left across the three FOVs, the ATR labels are correctly established and handed over across the clients via the server. The color of the box containing the target is set to green if it is recognized to be the Tank, blue when it is the Mini, and red if it cannot be recognized. In this instance, the Tank is visible and correctly recognized in the left and middle camera views. The Mini is visible in all three views, but is too close to the edge in the left camera view to be recognized. It is however recognized correctly in the middle and right camera views. It should be also noted that the tracking labels P122 and P123 are consistently assigned by the server to the Tank and Mini across all three views.

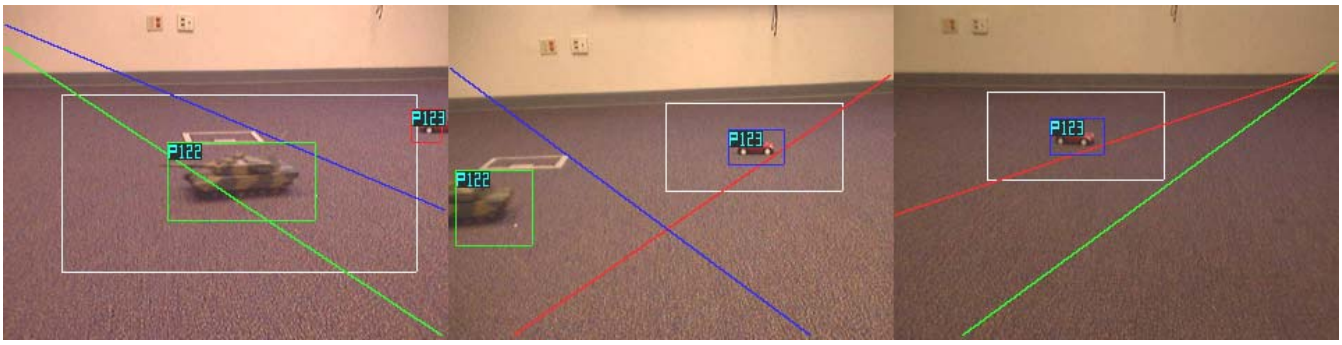

Figure 8: Snapshots from realtime demonstration show the detection, tracking, classification and handover of targets across a network of computers.

5. Summary and Future work

We have shown that several existing components such as COTS computers and networking technology, video trackers and ATD/R algorithms can be brought together to address the need for wide area surveillance in a distributed processing environment. The baseline video tracking system (KNIGHT) has been extended to work across multiple platforms, achieving motion based target detection and handover between multiple FOVs in real-time. The tracking system is able to automatically establish where the camera FOVs intersect, and use this information to generate consistent labeling of objects across the network. This process was further augmented using a correlation based ATR algorithm to classify the tracked objects and assign unique labels. The interactions between the ATR and tracking algorithms were defined, and the algorithms was shown to work across a network of three client computers and a server using the TCP/IP protocol.

[1] For now we use the motion based detection to cue the ATR. The ability to detect stationary targets using the ATD/R and initiate tracks on them will be incorporated in future versions.

The current system works on stationary platforms with fixed mounted cameras. Our goal is to extend the capability to moving platforms, especially unmanned air-vehicles. For simplicity, we envision that initially video data will be wirelessly transmitted to receiving computers on the ground where the processing will take place. In the future, it may be advantageous to process imagery aboard the platform and transmit only the salient results across the network. The greater challenge is to solve the FOV registration and the relative calibration between cameras for the moving platform scenario. While we seek a purely image based solution to this problem, we will also explore the potential benefits of using GPS and other information about the platforms and their positions relative to one another.

In the future, we anticipate that the ability to register the field of views of cameras on moving platforms may potentially lead to novel simplification of the guidance and control required to coordinate the relative behavior of the platforms. There are also new evolving paradigms for collaborative target recognition [5] that require specific configuration of the platforms around the targets. Such algorithms heavily leverage the infrastructure outlined in this paper which has the ability to automatically calibrate multiple moving FOVs to associate and track objects across them.

6. References

1. Omar Javed and Mubrark Shah, "Tracking And Object Classification For Automated Surveillance", The seventh European Conference on Computer Vision (ECCV 2002), Copenhagen, May 2002.

2. Book Chapter, “Correlation Pattern Recognition: An optimum Approach,” in Image Recognition and Classification: Algorithms, Systems, and Applications, Marcel Dekker, New York, Editor: Bahram Javidi, pp. 295-321, 2002.

3. A .Bobick and J. Davis, “The Recognition of Human Movements Using temporalTemplates", *Transactions of IEEE PAMI*, Vol 23, No. 3, March 2001.

4. Sohaib Khan, Omar Javed, Mubarak Shah, "Tracking in Uncalibrated Cameras with Overlapping Field of View", Performance Evaluation of Tracking and Surveillance PETS 2001, (with CVPR 2001), Kauai, Hawaii, Dec 9th, 2001

5. Abhijit Mahalanobis, Alan J. Van Nevel, “A collaborative network of correlation filters for object recognition**,”** Proc. SPIE Vol. 5202, p. 219-226, Optical Information Systems; Bahram Javidi, Demetri Psaltis; Eds., Nov 2003