Geo-spatial Active Visual Surveillance on Wireless Networks

Terrance E. Boult
Univ. of Colorado at Colorado Springs and Guardian Solutions, Inc

Keywords: Detection, Tracking, video surveillance, video motion detection Geo-spatial, ESRI, wireless video

Abstract

This paper reviews some of the history of automated visual surveillance, from the second and third generation VMD days of the early 1990s, to the current state of the art. It discusses the inherent limitations that resulted in an nearly negligible "increase" in performance throughout the 1990s and still exist in commercially available systems. Then we review an approach that overcomes these limitations – active visual surveillance with geo-spatial rules.

Active visual surveillance uses data from computer controlled Pan/Tilt/Zoom (PTZ) units combined with state of the art video detection and tracking to, in a cost effective manner, provide active assessment of potential targets. This active assessment allows an increase in the number of pixels on target and provides a secondary viewpoint for data fusion, while still allowing coverage of a very large surveillance area. This active approach and multi-sensor fusion, not a new concept, was developed as part of the DARPA Video Surveillance and Monitoring (VSAM) program in the late 90's. While we have continued to expand upon it since that time, there has been no commercial video surveillance, before Guardian Solutions, that provided these important abilities.

The core ideas in this paper address limitations of the original VSAM designs, briefly introducing our enhancements including geo-spatial rules for wide area multi-sensor fusion, and key design issues to allow us to support wireless networks.

1. Background & Previous Research

Visual Surveillance is a broad area and no amount of review in this paper will cover it adequately. In addition to the papers cited herein, a good review of many state-of-the-art visual surveillance systems can be found in a special issue of the *Proceeding of the IEEE* (Oct. 2001) as well as recent IEEE Workshops on Visual Surveillance and Workshops on Performance Evaluation of Tracking Systems.

In [1], Ringer and Hoover of Sandia National Lab present a detailed evaluation of the (then) commercially available exterior digital Video Motion Detectors (VMDs). These systems used specialized hardware, some were boards in a PC, other were standalone units, to allow real-time processing. This study was extremely well done, analyzing 13 different commercial systems in a controlled outdoor study (on a clean dirt background within a double fenced area). The stated detection criterion of the evaluation was 90% probability of detection (pd) at 95% confidence. Another requirement was an average of 10 or fewer false/nuisance alarms in 24 hours on a day with few clouds, bright sunny and calm weather. Only 6 of the 13 systems achieved the stated goals. In more challenging lighting conditions (still on dirt background), they developed 24 hours of "test tape" to test systems. (This set of tapes is still available and a good place to start for static VMD type system evaluation.) On the test tapes, the 6 systems that "passed" the clear-day tests averaged over 50 nuisance alarms. Even with the testing on simple dirt backgrounds, their final conclusions were that "VMDs in general, when used in an outdoor environment, are susceptible to nuisance alarms from environmental effects ... all had some problems rejecting nuisance alarms". Applying them in even more complex environments, such as grass, water, woods, where the backgrounds are themselves moving, would have been even more problematic.

As the Sandia study was ongoing, DARPA as developing its plans for the Video Surveillance and Monitoring (VSAM) program. That program, which funded research in video surveillance in the late 90's, sought to move beyond single camera VMD, to networks of video sensors, [2]. Most of the research was focused at higher level analysis like classifying activities as uncommon[3], seeking particular patterns of (indoor) activity[4] and reasoning about human movements [5, 6, 7, 8]. Unfortunately, most of the work in VSAM did not stress the detection capabilities – it was done in good lighting with color cameras and moderate size targets (approximately .1% to 1% of the image

(between 300 and 3000 pixels on target), with those doing human modeling often having the target represent 10% of the image. Furthermore, most of the groups exploited color (which cannot be used at night or in thermal video) to simplify detection and tracking.

The main efforts with a significant focus on low-level detection were the work of Sarnoff [9], which addressed detection/tracking from a moving plane and [10] which addressed lighting independent background subtraction (though it was not tested on complex outdoor scenes). The VSAM work at Maryland [11, 12] included non-parametric models for background subtraction and low-level people tracking, but all the examples were color imagery with simple lighting and large targets. Finally, our work, [13], addressed detection/tracking in omni-directional video and included analysis in very challenging situations including snipers.

There have been a few projects which have explicitly addressed lighting, which is a major issue outdoors, with a few important projects in Europe. Again there is too much to site in a workshop paper, but a few important works include PASSWORDS project, [15] uses an illumination change compensation algorithm to allow it to work in outdoor settings, and Riddler, et.al., [16], which uses Kalman filtering for adaptive background estimation which takes into account changing illumination so as not to mistake lighting as objects of interest. A similar approach is used in [17]. In [18], an approach was explored that used local order statistics to detect significant lighting changes. While each of these approaches was moderately effective for dramatic changes, none of them work well for a fast moving localized cloud shadow, and none of them discussed use at night, were moving "lights", illuminating static scene elements, are the often the only visible sign of an intruder.

To be viable systems, however, automated video surveillance needs to be able to work at night (maybe its most critical time), with small and non-distinctive targets that are as far away as possible. (Distance translates to response time – the goal is not just to record events but to respond to them while they are happening). None of the aforementioned systems discudded on nocturnal video, [17] and our own work has seriously addressed gray-scale data, the only type available at night (low-light or thermal).

For many current US government projects, the requested goal is to produce less than 3 false alarms (FA) per day. For these military applications, undetected targets could be, literally, deadly, so the miss detection (MD) rates also need to be low, with stated goals ¡ 5distant targets (1-2km). With each NTSC video containing 10^{20} potential target regions per camera per day achieving these performance goals place very strong demands on the low-level processing of the system. In [19, 20, 21], we investigated FA and MD rates for this type of problem. These papers analyzed the grouping algorithm that has allowed us to address the "signal-level" FA and impacts of random noise. However, they did not address nuisance alarm (NA) rates, where lighting, water, grass or trees produce real changes that are not "significant" motion.

How then does one reduce the impact of FA and NA? One approach, which we are pursuing, is active surveillance with geo-spatial models. This approach combines sensor-fusion with active sensor control and calibration information to allow the system to use multiple sensors to analyze an event. Thus providing added information that can significantly reduce the FA and NA rates.

2. Active Geo-spatial surveillance

One of the issues addressed in the VSAM project was the importance of situational awareness and the role of geo-spatial information in providing that awareness for a large facility.

We content that, while geo-spatial information has much more to offer than just "situational awareness", we consider it the key to scalability and robustness of a video surveillance system, see 1. How to calibrate cameras and do ray intersection is well known for regular cameras, and for omni-directional cameras we discussed the geometric calibration and back-projection in [22].

For scalability we need to provide coverage with fewer cameras. One aspect of this is to detect targets at greater ranges, but then the targets will be too small in the image for assessment. By using the geo-spatial information, the camera that detected the target can "handoff" to another sensor with greater optical magnification for assessment. For high-end cameras, e.g. a thermal camera with 300mm lens which might cost $250,000, making efficient use of that camera for assessment is important to scalability. A similar issue arises with having fixed cameras watching choke points – using active PTZ control, it can then "hand-off" a target to a PTZ when that target begins leaving the choke region. With active tracking, the PTZ can then follow the target through a large open area. The geo-spatial sensor-to-sensor hand-off was demonstrated as part of the VSAM project, with hand-offs from the Lehigh OmniCamera to the CMU PTZ systems and multiple CMU fixed to PTZ hand-offs.

For robustness, 3D provides significant advantages that were not exploited in the VSAM project. In addition to the back-projected position, the calibrated cameras and geo-spatial information allows computation of the targets approximate size, speed and heading in meaningful world coordinates. In most perspective images for surveillance, the variations in "image" size of targets makes it difficult to define a "pixel" size that is meaningful. E.g. the feet of the subject behind the truck (¡ 1sq.ft of target) in the foreground of figure 3, take up almost as much area as the backhoe out-

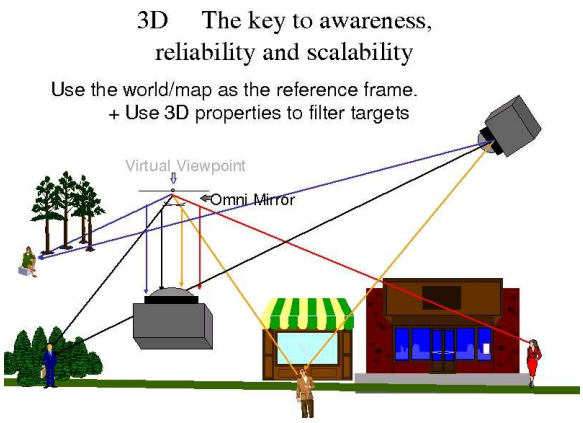

Figure 1. 3d is the key to scalability and robustness. From a single calibrated camera, one can intersect rays with a digital terrain model. The simplest model, a local plane, is sufficient for many settings. Back-project of the ray from the camera to the ground plane produces target location. It also produces a distance which can be used to scale the pixels on the target to approximate size in square feet (or meters). Using true 3D positions it also allows computing speed in mph and heading (degrees from north).

Figure 2. An example of a situational awareness display with an 8 camera system showing a fixed to PTZ hand-off. Without the map, it is difficult to understand that the two images with targets (boxed regions upper left and lower right) are showing the same thing. On the map are shown the FOVs of the different cameras and the recent locations of the target (shown as question marks, ????, since it did not know the type.). This this knowledge it is clear there are two cameras with overlapping FOVs that are looking at the same target. In this case the fixed camera was loosing the target and "handed off" to the PTZ to continue following it.

lined in black in the far part of the scene (¿ 50sq.ft of target). A small motion of the brush in the front of the scene would generate a motion target much larger than the backhoe. But by using 3D information, one can filter the detect objects in a more meaningful size. Our approach, implemented in GuardianWatch, is to use an image-map such as figure 4, where each region has associated XML rules describing the targets of interest (i.e. what sets of an alarm rather than just what moves).

Guardian represents its rules in two different forms. The first is as ESRI shape files (ESRI is a trademark of ESRI incorporated, www.esri.com). The ESRI shape file format allows sharing geo-spatial information (we can back-project the coordinates of the regions) in a form that quickly is becoming the de-facto standard for government GIS. (Internally it is really just a DB3 database with special column attributes). For simpler interfaces with other tools, we also use an XML format with an associated image.

A partial XML rule-set might look like:

```
<rule_set>
 <sensor_ruleset id="0x01 0x01">
  <image> manatee-alarm-0.ppm </image>
  <rule name="Parking exit ">
   <alarm_id> 230-230 </alarm_id>
   <threatcon> 0-5 </threatcon>
   <days_of_week> sun-fri </days_of_week>
   <start_time> 0:00 </start_time>
   <duration> 24:00 </duration>
   <size> 5-25 sqft </size>
   <speed> .1-5 mph </speed>
   <angle> 240-270 degrees </angle>
   <gps_poly> .. (saveing space) </gps_poly>
  </rule>
  <!-- ship boarding sat 10am-4pm,
     handle it separately, rules not shown -->
  <!-- We ignore people on saturdays..
   too busy with passengers, crew -->
  <rule name="People in parking area ">
   <alarm_id> 230-230 </alarm_id>
   <threatcon> 0-5 </threatcon>
   <days_of_week> sun-fri </days_of_week>
   <start_time> 0:00 </start_time>
   <duration> 10:00 </duration>
   <size> 2-5 sqft </size>
   <speed> 0-5 mph </speed>
   <angle> 0-360 degrees </angle>
   <!-- handoff syntax is
       camid dlat dlon dalt zoom_fov_degrees -->
   <handoff> 0x8 0 0 0 5 </handoff>
   <gps_poly> .. (saveing space) </gps_poly>
   </rule>
   <!-- ... rest rules -->
<rule_set>
```

The move to geo-spatial representations allows us to not only have rules for targets based on their position in the image, but our patent-pending approach allows us to define

Figure 3. As is well known, distant targets appear smaller than near-by targets. The feet of the person behind the truck have more pixels than the backhoe in the distance. The perspective effect makes image-based rules difficult to apply in filtering nuisance alarms.

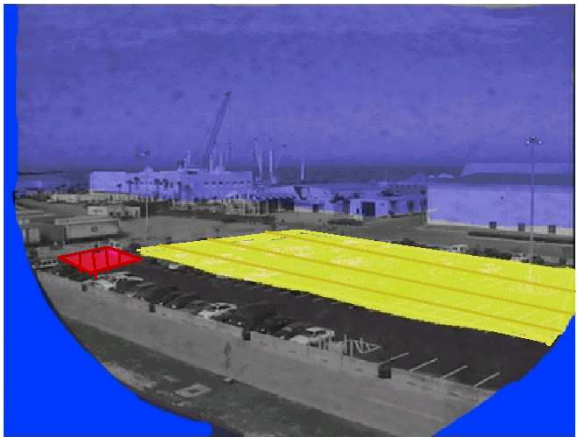

Figure 4. An alarm map associates rules with pixels in the image. The red (small vertically striped region on left) has rules that activated only for cars leaving (not entering) off-hours. The yellow (lighter region in parking lot) has rules that alarm for people-sized targets anytime. The yellow region hands-off to a PTZ so a guard can assess what the human is doing. The blue regions (sky and lower corners) are ignored. The remaining regions will detect/track targets but do not set off alarms.

rules in a map-based geo-spatial nature and then use it to filter targets for any cameras that are looking at the area. This allows rules to apply to PTZ cameras as well as fixed cameras. As the PTZ follows a target it will have better estimates of size and speed and can have more certainty about it.

While the VSAM project investigated some uses of geo-spatial information for situational awareness and camera coordination, the biggest advantage of 3D, filtering nuisance went unstudied. Using geo-spatial information allows the video surveillance system to ignore a wide range of "nuisance" alarms (e.g. most lighting, many birds, blowing trash), that would otherwise render the systems unusable. While some of these can be ignored using 2D information, the 3D information makes it far simpler for the end user to understand. Combining 3D analysis with the sophisticated detection/tracking from [20] has allowed us to detect and track small targets on very complex moving backgrounds, such as a zodiac on water in figure 6, with 30 pixels on target. Zodiac was 700ft away in 2ft waves using a 320x240 thermal sensor with 50mm lens. Guardian Solutions commercialized this technology and now has multiple commercial clients who are using automated video surveillance on a 24/7/365 basis.

3. Wireless video surveillance

Our approach to video surveillance has been, from the beginning, intended for distributed processing. With the data demands of real-time video processing, it simply makes sense to push the processing as close to the camera as possible. The VSAM project also embarrassed the "networked" approach. The original VSAM communication protocol, [23], provided rudimentary means for communication of the necessary data, but it was not particularly efficient. Once the video surveillance system has the ability to decide salient motion and use rules to decide what is important, it has the inherent ability to filter not just "alarms", but also the vast streams of video data it has analyzed.

With moderate analysis as that described above and a significantly extended network protocol, we can support video surveillance on lower-bandwidth networks. The GuardianWatch software has implemented a basic adaptive bandwidth control with priority filtering. Using the "significance" of the motion along with the alarm rules, the system can decide the priority of each video item.

To support efficient adaptive bandwidth management, the system uses a sprite-based representation.

The system represents each moving target separately (as a jpeg chip) on a reference background (a full size jpeg). The detection system then queues up the video chips, their geo-spatial position, shape descriptions etc, and the system can decide what it can afford to send. The "video" encoding is, in spirit, similar to sprites in Mpeg4 (which almost no one supports), but we drop all the inter-frame motion coding both to make it easy to dynamically drop data, and so writing display code is easier. (The approach/protocol supports differential motion coding but we have not found it

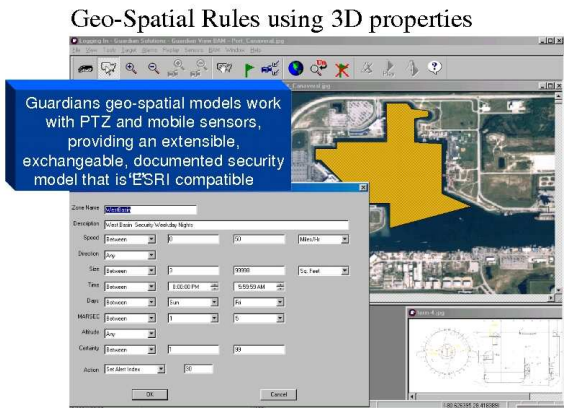

Figure 5. An example showing a geo-spatial alarm zone and the GUI interface for defining the alarm rule associated with that polygon. The geo-spatial rules can apply to any camera watching the area, including PTZs that follow the target into the area.

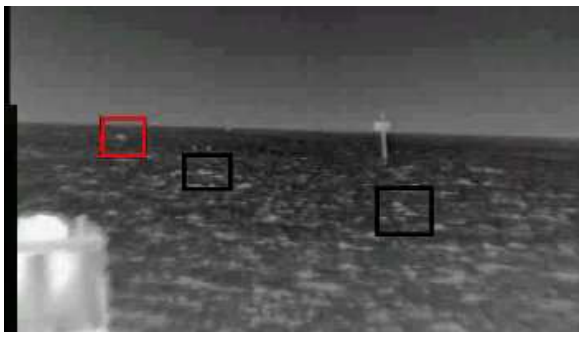

Figure 6. Zodiac tracking example from thermal video. Zodiac is detected as a confident target (red/light gray box). Two other boxes (black) show hypothesis of potential targets. Using 3D information, the waves never become confident.

necessary). If there is minimal bandwidth (e.g the 8 camera hand-off example above was monitored over a 33Kbps dialup link), then the reference background images are sent very infrequently (often ¡ 1 per min) and significant chips are sent more frequently. In the zodiac example, more and more of the hypothesis chips would be pruned as the bandwidth was reduced. At the same time, the target position and other (very small) descriptors are passed around so we can maintain situational awareness.

The system's processors/communication nodes forms a "tree", see figure 7, with communication filtering as data moves up the tree. If the user is mobile and near one

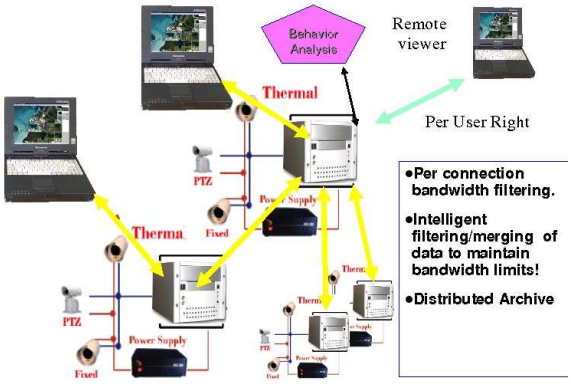

Figure 7. The communication architecture for GuardianWatch. Each computer processes video, stores video and geo-spatial data, and then filters data based on both a priority allocated bandwidth and the currently available bandwidth. The geo-spatial data is analyzed by the Behavioral Analysis node which implements the overall geo-spatial rule engine. The system supports encryption and per user/node access rights.

of the clusters of cameras, e.g. the guard truck pulls up near a terminal in the port, they have more available bandwidth and get more of the video chips and faster reference frames. If the local first responders (e.g. police) are examining an alarm via their Internet connection, the data might be bouncing off a commercial "Internet" satellite with much lower bandwidth. The first responders will still get a good situational awareness but it will be far from "full motion video". However rather than degrading all the data uniformly, the "chips" of potential targets will have much higher quality than the less significant data. If the communication links between the processors are wireless, as they are at many of Guardian Solutions commercial and military clients, then the "storage" of archival video also needs to be distributed. The extended VSAM protocol supports the necessary "DVR" features as well. In August of 2003, an 88 camera installation of Guardian Solutions automated video surveillance system went "online" at Port Canaveral, the largest Cruise ship port in the country. The system communicates the processed results from these cameras back to the central monitoring and the mobile monitoring stations using COTS 802.11 technology. Two guards can manage this large set of cameras because they are not watching them unless something significant is happening.

4. Conclusion & Future work

Studies the early 90's pointed to the problem of nuisance alarms, yet little of the research since that time has really

focused on addressing this issue. While many researchers have been trying to "recognize" complex human activities, the ability to ignore simple nuisances such as birds, lighting and trash, have been largely ignored. This paper discussed how to use geo-spatial rules and filtering using 3D propertie to reduce or reject the nuisance alarms. It also briefly discussed how advanced "filtering" can produce data that supports more efficient situational awareness of video-based systems on wireless networks.

Even with 3D filtering, there are many situations where nuisances still arise. More advanced algorithms that fuse the data from multiple complimentary sensors can reduce these even farther. Already our system can use its "confidence" to filter its alarms, so a fixed sensor could detect a target, but uncertain about its properties, it can hand-off to another sensor to automatically assess the target. But adding multiple sensors could reduce the nuisance alarm rate even farther, e.g. with good lighting a thermal and visible sensor would see a human target with similar size, but would see "trash" very differently. The issue of nuisance from large animals (deer) is more problematic (especially in woods where target shape cannot be used because of the frequency of occlusion). But sensitive acoustic/seismic sensors that could monitor the "footsteps" might be able to distinguish them. Our future work includes multiple issues in multi-sensor integration and cueing.

References

[1] C. Ringler and C. Hoover, "Evaluation of commercially available exterior digital vmds," tech. rep., SANDIA, Sept. 1998. SAND94-2875 UC-706. (Printed June 1995).

[2] T. Kanade, R. Collins, A. Lipton, P. Burt, and L. Wixson, "Advances in cooperative multi-sensor video surveillance," in *Proc. of the DARPA IUW*, pp. 3–24, 1998.

[3] W. Grimson, C. Stauffer, R. Romano, and L. Lee, "Using adaptive tracking to classify and monitor activities in a site," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22–29, 1998.

[4] B. Flinchbaugh and T. Olson, "Autonomous video surveillance," in *25th AIPR Workshop: Emerging Applications of Computer Vision*, May 1996. See also DARPA IUW May 1997.

[5] J. Davis and A. Bobick, "The representation and recognition of human movements using temporal templates.," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 928–934, 1997.

[6] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfi nder: Real-time tracking of the human body," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[7] A. Lipton, H. Fuijiyoshi, and R. Patil, "Moving target detection and classifi cation from real-time video," in *Proc. of the IEEE Workshop on Applications of Computer Vision*, 1998.

[8] I. Haritaoglu, D. Harwood, and L. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, pp. 809–830, August 2000.

[9] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, pp. 774–781, August 2000.

[10] Y. Ivanov and A. Bobick, "Fast lighting independent background subtraction," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 199–207, 1998.

[11] I. Haritaoglu, D. Harwood, and L. Davis, "W^4s: A real-time system for detecting and tracking people in 2.5D," in *Computer Vision—ECCV*, 1998.

[12] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *FRAME-RATE Workshop*, IEEE, 1999. Eletronic (only) proceedings at www.eecs.lehigh.edu/FRAME.

[13] T. Boult, C. Qian, W. Yin, A. Erkin, P. Lewis, C. Power, and R. Micheals, "Applications of omnidirectional imaging: Multi-body tracking and remote reality," in *Proc. of the IEEE Workshop on Computer Vision Applications*, Oct. 1998.

[14] T.E.Boult, R.Micheals, X.Gao, P.Lewis, C.Power, W.Yin, and A.Erkan, "Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets," in *Second IEEE International Workshop on Visual Surveillance*, pp. 48–55, IEEE, 1999.

[15] M. Bogaert, N. Chleq, P. Cornez, C. Regazzoni, A. Teschioni, and M. Thonnat, "The passwords project," in *ICIP*, pp. 1112–1115, IEEE, 1996.

[16] C. Riddler, O. Munkelt, and H. Kirchner, "Adaptive background estimation and foreground detection using kalman fi ltering," in *ICRAM*, pp. 193–199, 1995.

[17] G. Foresti, "Object detection and tracking in time-varying and badly illuminated outdoor environments," *SPIE Journal of Optical Engineering*, vol. 37, no. 9, pp. 2550–2564, 1998.

[18] B. Xie, V.Ramesh, and T. Boult, "Sudden illumination change detection using order consistency," in *Workshop on Statistical Methods in Video Processing (in conjunction with ECCV2002)*, 2002.

[19] X. Gao, T. Boult, F. Coetzee, and V. Ramesh, "Error analysis of background adaption," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[20] T.E.Boult, R.Micheals, X.Gao, and M. Eckmann, "Into the woods: Visual surveillance of noncooperative and camouflaged targes in complex outdoor settings," *Proceeding of the IEEE*, vol. 89, pp. 1382–1402, October 2001.

[21] X. Gao, V. Ramesh, and T. Boult, "Statistical characterization of morphological operator sequences," in *Computer Vision—ECCV*, May 2002.

[22] T.E.Boult, X.Gao, , R.Micheals, and M.Eckmann, "Omnidirectional visual surveillance," *Image and Vision Computing*, 2004. to appear.

[23] A. Lipton, T. Boult, and Y. Lee, "Video surveillance and monitoring communication specifi cation document 98-2.2," tech. rep., CMU, Sept. 1998. `http://www.cs.cmu.edu/~vsam/Documents` as `vsam_protocol_98_22.ps.gz`.