A Decade of Networked Intelligent Video Surveillance

Terrance E. Boult[†‡], R.C. Johnson[†], Tracy Pietre[†], R. Woodworth[‡], Tao Zhang[†]
† Vision and Security Technology (VAST) Lab,
University of Colorado at Colorado Springs
1420 Austin Bluffs Parkway. Colorado Springs CO 80933
719 262 3900 lastname@vast.uccs.edu

‡ Securics Inc
www.securics.com
719 387 8660
lastname@securics.com

ABSTRACT
This paper reviews nearly a decade of work on multi-camera sensor networks combining multiple omni-directional imaging sensors, traditional stationary cameras and pan-tilt sensors. It reviews significant issues, design constraints and accomplishments from the DARPA VSAM project and the commercial systems based on that early work. With commercial intelligent camera networks deployed with hundreds of sensors, we review the key components in effective "distributed video surveillance," then discuss the major open issues, including hardware-accelerated algorithms needed for increasing resolution while reducing power, and the issues of mobile surveillance. We briefly review our recent results in these challenging areas.

1. INTRODUCTION
Intelligent networked video surveillance is a well-established commercial technology. While video-based research has been developing for more than two decades, significant advances in video-surveillance began in the mid-90s, including the DARPA Video Surveillance and Monitoring (VSAM) project in the US and ESPRIT funded efforts in the EU. In the VSAM effort, the main outdoor demonstration included a dozen different cameras with distributed processing and network communications integrating algorithms from teams including static and PTZ cameras [6], Airborne Cameras [10] and omni-directional cameras [2]. The technology and the key investigators from each of these three teams became the core of commercial video-surveillance products from ObjectVideo, Sarnoff/Pyramid Vision, Guardian Solutions and specialized systems from RemoteReality.

Visual Surveillance is a broad area and no amount of review in a workshop paper will cover it adequately, so we will not try. Good reviews of the state-of-the-art at the turn of the century surveillance systems can be found in a special issue of IEEE PAMI from August 2001 and the *Proceeding of the IEEE*, October 2001 and more recent review in Image and Video Computing July 2004. Recent work can be found in many venues with concentrations in regularly held IEEE Workshops on Visual Surveillance (VS), Advanced Video Surveillance Systems (AVSS) and Performance Evaluation of Tracking Systems (PETS).

Intelligent video surveillance is a systems level problem with 6 major components:

1. Sensor Architecture
2. Low-level detection/processing algorithms
3. Hardware/Computation architecture
4. Software/Communication architecture
5. User-Interface
6. Higher-level algorithms for combining data and filtering out uninteresting events

One could write thesis on any one of these, but this paper is a high-level review of the first four components in terms of constraints, system design issues and open issues. This paper draws on nearly a decade of networked intelligent video surveillance systems development by Dr. Boult[1] and students and introduces the ongoing efforts of the team at the Vision and Security Technology Laboratory at the University of Colorado at Colorado Springs.

Dr. Boult's first efforts on multi-camera networked video surveillance include both omni-directional and other traditional video sensors, as part of the initial DARPA VSAM effort [8], and on the ONR MURI program. Dr. Boult continued to lead the development of systems that then moved through initial field-testing for the DARPA Small Unit Operations program to multi-week field testing at several Army bases as part of the Army SmartSensorWeb program, eventually leading to wireless

ACM Workshop on Distributed Camera Network. October 2006.

[1] This work started while Dr. Boult was at Lehigh University and has been funded over the years by multiple contracts from DARPA, ONR, ARMY and SBIRS with Remote Reality and Securics Inc. It also includes his experience while the founding CTO of Guardian Solutions and founder/CEO of Securics Inc. The views expressed are the authors, not necessarily those of the funding agencies, current or past employers.

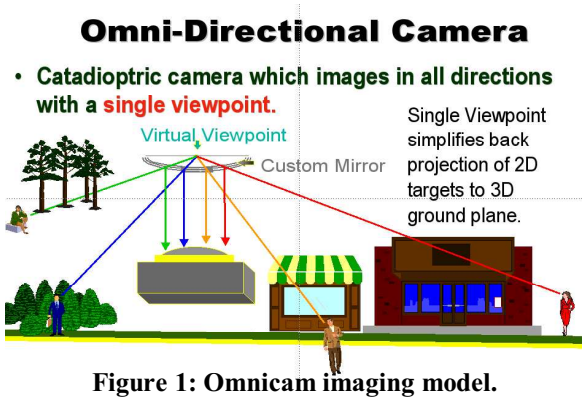

Figure 1: Omnicam imaging model.

geo-spatially enabled commercial systems developed for Guardian Solutions and Remote Reality [3].

In each area of this review, we cover the limitations and introduce the ongoing SEE PORT effort, which tackles the much harder problem of surveillance and tracking from moving vessels and our ongoing FIINDER effort, which addresses low-power networked detection systems with hardware accelerated mega-pixel sensors.

2. SENSOR ARCHITECTURE AND LOW-LEVEL DETECTION

The first issue that must be addressed in an intelligent sensor system is the selection of the sensor(s) – if you cannot sense the target, no amount of post processing is going to find it. The issues here include selection of the imaging technology and the lens system. For the basic sensor technology, the choices are visible sensors (CCD/CMOS), an intensified low-light sensor, or a thermal or LWIR sensor. For each technology there is also the potential choice of the resolution of the sensor, with "analog" sensors supporting CIF (320x240), NTSC/QCIF (640x480), and with digital sensors supporting these plus 1, 3 or 5 mega-pixel resolutions at 8, 10 or 12 bits per pixel. A major issue in these decisions is the cost of the sensors, (e.g. a LWIR sensor costs $30-$80K), as well as the lighting. Since the surveillance is often most important at night, a true cost comparison must include the costs of lighting along with the sensors' resolutions, (e.g. installation of a single wide-area lighting system may cost $50K-$70K per pole).

Lens choices impact both the field of view and range to target. For a traditional lens, the tradeoff is well-known. For the catadioptric omni-directional sensors with which we have worked, it is less obvious. As seen in Figure 1, these sensors use a mirror to collect the scene light and the camera points at the mirror. If the mirror is pointed down, (or up) the system has a hemispherical, or more field-of-view, seeing all around the camera. The resulting sensor has non-uniform resolution. The non-intuitive part is that the resolution is maximum along the horizon, where targets are most distant. If the mirror is imaged at the center in a QCIF omni-image, there will be a 480 pixel mirror radius and about 1500 pixels on the horizon, or about 4.2 pixels per degree resolution. In comparison, a standard camera has 4.2 pixels per degree when using a 150 degree lens, which means it would require 3 such cameras to watch the horizon. Though omni-directional sensors cost more, this resolution/FOV tradeoff is the reason that omni-cameras are frequent components in Dr. Boult's surveillance work [4].

Tightly coupled to the sensor architecture is low-level detection – if it cannot be detected, no amount of higher-level architecture will help. To be viable commercial video surveillance systems, the systems need to reliably and robustly handle small and non-distinctive targets from great distances. The need for detection of small targets at a distance is a conflict of security concerns versus cost. Distance translates to response time – the goal of security is not only to record events but also to respond to them while they are transpiring. Therefore, it is necessary to detect events far enough in advance to respond. While one could increase standoff distance by increasing the focal length of the imaging system, this results in a narrowing FOV and reduction of the overall imaged area, which means that protecting a reasonable area requires numerous cameras, proving to be generally cost prohibitive. Figure 2 shows the impact of the minimum size detection target on the number of sensors needed to cover the staging area of an airfield. While minimum target size is clearly a function of the algorithms, it is also clear that sensor resolution impacts this as well; for a fixed minimum target size, a mega-pixel sensor covers considerably more area than a NTSC or CIF sensor.

In a real system, end users would investigate each alarm,

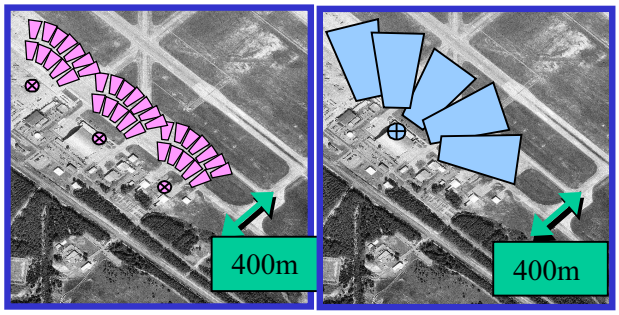

Figure 2: Impact of minimum target size on number of sensors required to secure an area. Left assumes minum target size of 48 pixels (6x8 pixel human target) and requires 33 sensors to cover the staging area. Right diagram is assuming a minimum size of 12 pixels (3x4 pixel human target) and requires only 5 sensors. If false alarms accumulate independently, the impact of number of sensors on overall system false alarms is obvious.

and in many of the current government deployment 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 in multiple programs for a less than 5% miss-detection on very distant targets (1-2km). Again, the overall minimum target size has a significant impact on system false alarm rates, though for almost any algorithm, the FA/MD rates are strongly impacted by the choice of minimum target size. Having a formal model that allows one to make such tradeoffs is therefore an important requirement for a video surveillance algorithm.

With each NTSC video containing 10^{20} potential target regions per camera per hour, achieving acceptable False Alarm/Misdetection Rates place very strong demands on the low-level processing of the system. In [4,2] we investigated formal methods for analyzing FA and MD rates for this type of problem. These papers are, to our knowledge, still the only work to formally model pixel group aspects that allow meaningful FA/MD for a video surveillance system. These papers analyzed the grouping that 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 or brush produce real changes that are "significant," but not interesting motion. To address this class of nuisance alarms, we added saliency models, similar in spirit to those used by [10]. However, even with salience models, birds, bugs and other animals are a significant nuisance issue. If one is using only 6-12 pixels on target, then distinguishing a crawling human from a dear or other large animal is quite difficult.

This then brings us back to the sensor architecture. An effective low-level detection can detect/track small targets, but assessment and identification need more resolution, which suggests a multi-camera system where the detected targets are then handed-off to a Pan-Tilt-Zoom (PTZ) sensor. As is well-demonstrated in the VSAM effort, an effective way to address this issue of identification would require the cameras to be calibrated and use geo-spatial coordinates, passed through the network, to pass control from one sensor to another.

With a PTZ in a lower security setting, the system can also use a temporal stop-and-stare approach to trade probability, or time to detection, against cost. Instead of 5 sensors, the left side of Figure 2 could be 5 stops on a PTZ tour. It becomes most interesting when the "PTZ" in question becomes a self-contained wireless smart camera like that shown on the right. This sensor is the core of the Guardian Solutions Threat Watch system, commercially released in 2004, that combines an embedded vision processor with 25x zoom rugged PTZ, IR illumination, 802.11B networking. Sold with special tripods to allow 4 units, plus a laptop, to coordinate to protect sensitive cargo for the military or provide a portable electronic-fence capability wherever needed.

While sensors and low-level processing are probably the most advanced components of intelligent video systems, there are three important open issues: no-illumination sensors, moving sensors, and increasing sensor resolution.

Visible or NIR sensors still require illumination, so an important issue is moving to thermal and intensified systems. While the equipment is more expensive, the reduced infrastructure and power costs often tip the balance.

A larger issue is that of low-level detection, tracking and control when the sensors themselves are moving. This makes background subtraction impractical without specialized hardware such as that developed by Sarnoff. Using COTS hardware, we are addressing this in our SEE-PORT effort using both visible and LWIR 360° detection and tracking from moving ships. Because we could not use standard background subtraction, we needed a new approach to detect targets. Based on local saliency, we developed a new algorithm to determine dual thresholds, called symmetric subtraction, see [9], and are now developing new window-variance based detectors for improved sensitivity in more complex wave clutter. Another approach is using cascaded Haar-based classifiers to directly "recognize" particular classes of targets.

The SEE-PORT effort also includes a PTZ for acquiring images with sufficient resolution for assessment identification, automated identification of targets that have been previously detected, and an architecture for integration with other sensors, (e.g. onboard sonar and onshore radar and cameras). While PTZ slew-to-cue has been demonstrated by many groups, and our earlier versions are in commercial products of RemoteReality and Guardian Solutions, the problem for SEE-PORT is more difficult because of the potential latency involved with a moving observer. By the time the PTZ gets to the target, it may no longer be where we thought it was, or conversely, we may no longer be located at our previous position. Furthermore, with many targets we must develop a schedule for when the PTZ should move from target to target, and hence may need to predict where the target will be even when we can no longer see it.

To address this problem, we use predictive motion models fed into a multi-target priority-scheduling algorithm. An extended Kalman filter, with target disambiguation,

including image-based properties of the targets, predicts target position with uncertainty. Using the confidence data and predicted location, we developed a program to schedule the slew of a PTZ to integrate high confidence targets that could be a threat and targets whose uncertainty is becoming too large. Because both the omni-camera and PTZ will be moving/rocking on a ship, the prediction requires an ego-motion model for the vehicle motion. We then combine the target motion model, sensor motion model and latency models for all the computations and communications to predict where the target will be.

3. HARDWARE/COMPUTATION

Real-time video processing is a very computationally intensive task and, if the system is not designed well, can swamp most communication networks.

The early VSAM systems at CMU mostly used tower-type 500Mhz PCs and Unix workstations, computing CIF data at around 10-15fps. While Sarnoff used custom hardware boards, which eventually lead to their Arcadia chips, Dr. Boult's efforts were computing 30fps at 640x480 using an embedded 233Mhz system.

As algorithms improve from those early versions, they have required considerably more computing; ObjectVideo uses dual 3Ghz processors for 4 QCIF channels and Guardian Solutions uses a 2.4Ghz for 4 QCIF channels. More significantly, as higher-resolution sensors are starting to be used, they demand considerably more powerful approaches.

High-end PC-class processors are still the dominant forms of computing intelligent video surveillance. ObjectVideo recently introduced "ObjectVideo on board" versions using a TI DaVinci DSP to provide their core computation at 15fps on CIF video. Their goal is reduced cost and reduced system size, not increased performance nor the support for larger sensors.

Moving to higher-resolution sensors is, however, critical to improving overall system performance. This is probably the most significant open issue-- how to accelerate the detection/tracking techniques using hardware accelerations such as FPGA and/or local DSPs. As we move to programs using 3, 5 and soon, 16Mpixel sensors, we are focusing on FPGAs to address this.

In FIINDER (FPGA-enhanced Intensified Image Network Detectors with Embedded Recognition), we have been using the Elphel 333 network cameras with a Spartan 3, an Etrax processor, 64M of memory and a 3Mpixel sensor. We have designed a version of our Cascaded Haar-wavelet which, in simulations, requires an average of 6ms (worst case is, however, 300ms), and will be porting it to the new 5Mpixel version of the camera this fall. This approach is not simply taking previous video surveillance techniques and adapting them to an FPGA; it is looking at what we can do well in hardware and developing new approaches to exploit that ability.

4. SOFTWARE/COMMUNICATION

At the core of any distributed system is its software and communication architecture. We break our discussion off into the issues of the network protocol and the overall software architecture.

4.1 Network Protocols

Due to the potentially massive amounts of video data and the need for this to be real-time, design must address some means of communicating target information and cannot simply use standard streaming video protocols. Dr. Boult was part of the team to define the original VSAM communication protocol, [7].

The VSAM protocol represents key target properties as well as image data. It was sufficient for the dozen or so sensors used in the VSAM, but had limitations that prohibited its use in larger systems, including using a single central coordination node and a fixed packet structure. Dr. Boult enhanced that protocol as he developed architectures for wireless video [3] to support hundreds of nodes. The most critical extension for scalability was adaptive bandwidth control. The overall Scalable Network architecture is shown in Figure 3. The SPM (Sensor-Processing Modules) do the actual video processing and detect/track targets. They then send these into the distributed architecture with a description of the target, including its geo-location and a localized image chip of an area around each target. The AGM (Archive Gateway Module) provides traffic routing, reliable multi-cast support, archiving to support replay on lightweight nodes and provides traffic bandwidth adaptation. The operator control units (OCU) are a display and control user interface.

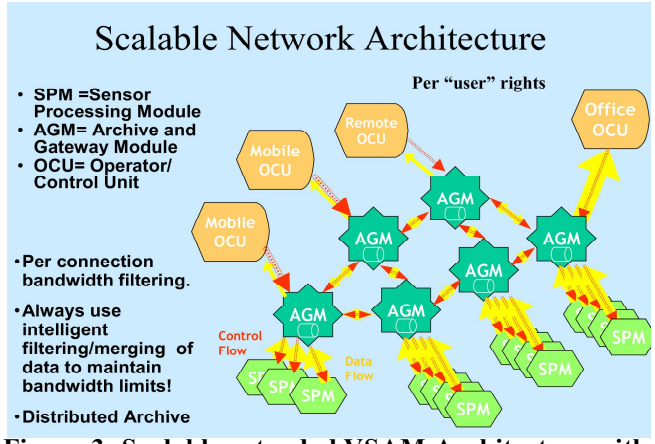

Figure 3: Scalable extended VSAM Architecture with distributed adaptive bandwidth control

Using this extended protocol has allowed large deployments on limited bandwidth, which Dr. Boult has used to install a sensor network with over 100 sensors including 88 camera covering 2Km of a major US port, all using a single 802.11B channel.

In a similar manner, using our own reliable protocols in addition to low-level best-effort transmissions, has allowed us to achieve an order of magnitude better throughput when sending images on Zigbee networks.

4.2 SOAP-based interfaces

While the extended VSAM architecture provides an effective transport and scalability, it does not address sensor discovery or larger management issues. For a collection of reasons, DOD has been moving their overall network architectures toward a Service Oriented Architecture, requiring multiple programs to follow those guidelines. For two of our ongoing ONR projects, we have moved to a SOAP-based protocol. For the SEE-PORT project (see Figure 4), this provides for a scalable and flexible way for other sensors to be integrated and multiple "users" to view the data and potentially control the PTZs.

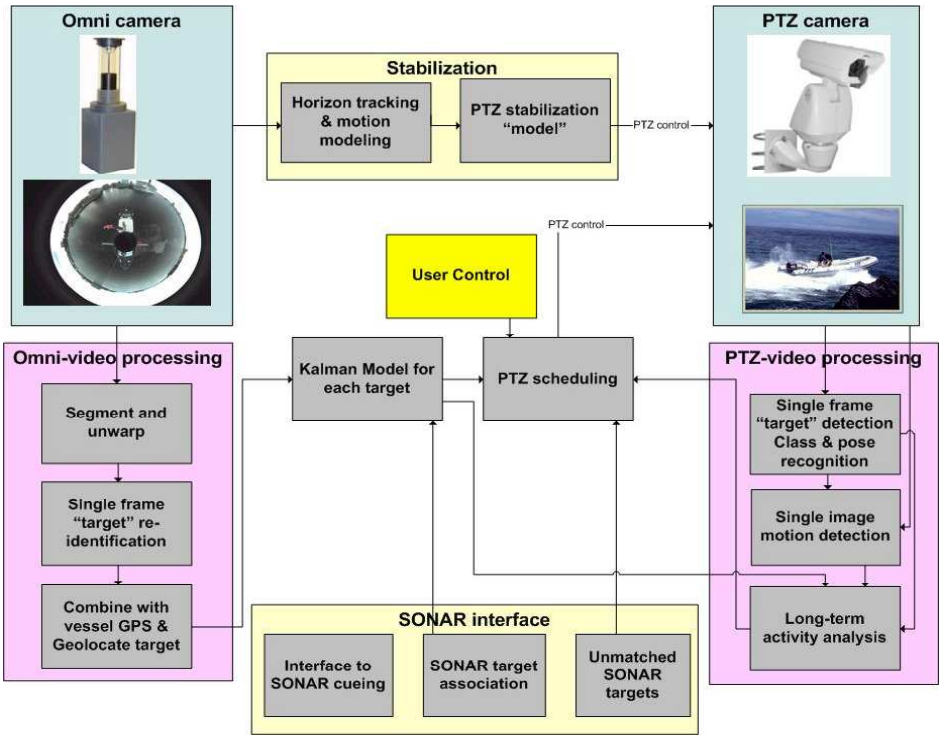

Figure 4: SEE-PORT architecture. Each Colored box represents an interdependent embedded computer. Both visible and thermal Omni-Cameras will be used and eventually 4 omni-cameras and 4-6 PTZ will be used on larger vessels. Current Development is using 1.2Mpixel visible and 640x480 LWIR but designing algorithms for 16Mpixel visible sensor.

5. CONCLUSIONS

Networked distributed video surveillance has moved from an academic research area to major commercial efforts with installations often involving hundreds of sensors. This paper reviewed some of the core lower-level issues and discussed open area for continued research. The major areas of high-level algorithms and user-interfaces are also critical areas for research.

6. REFERENCES

[1] T.Boult, A.Erkin, P.Lewis, R.Micheals, C.Power, C.Qian, and W.Yin, ``Frame-rate multi-body tracking for surveillance," in Proc. of the DARPA IUW, 1998.

[2] T.E. Boult, R.Micheals, X.Gao, and M. Eckmann, ``Into the woods: Visual surveillance of non-cooperative and camouflaged targets in complex outdoor settings," *The Proceeding of* the IEEE, vol. 89, pp. 1382--1402, Oct 2001

[3] T. Boult. Geo-spatial Active Visual Surveillance on Wireless Networks. Proc. of IEEE Applied Imagery Pattern Recognition (AIPR) Workshop, October 2003

[4] T. Boult, X. Gao, R. Micheals, and M. Eckmann. Omni-directional Visual Surveillance. Special issue of Image and Vision Computing on Surveillance, Elsevier House, 2004.

[5] X. Gao, T. Boult, F. Coetzee, and V. Ramesh, ``Error analysis of background adaptation," in Proc IEEE Conf on Computer Vision and Pattern Recognition, June 2000.

[6] 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.

[7] A Lipton, T. Boult, and Y. Lee, "Video surveillance and monitoring communication specification document 98-2.2," tech. rep., CMU, Sept. 1998. Available at www.vast.uccs.edu/~tboult/vsam_protocol_98_22.ps.gz

[8] S.K. Nayar and T. Boult. Omnidirectional vision systems: PI report. In Proceedings of the 1997 DARPA Image Understanding Workshop, pp 55-62, May 1997.

[9] T. Petrie and T. Boult, "Separable Features as a Basis for Adaptive Multi-Thresholding and Segmentation", Second Biotechnology and Bioinformatics Symposium, 2005.

[10] L.Wixson, ``Detecting salient motion by accumulating directionally-consistent flow," IEEE Tran. PAMI pp.774-781, Aug. 2000