Introduction to the Special Section on Video Surveillance

Robert T. Collins, Alan J. Lipton, and Takeo Kanade, *Fellow, IEEE*

AUTOMATED video surveillance addresses real-time observation of people and vehicles within a busy environment, leading to a description of their actions and interactions. The technical issues include moving object detection and tracking, object classification, human motion analysis, and activity understanding, touching on many of the core topics of computer vision, pattern analysis, and aritificial intelligence. Video surveillance has spawned large research projects in the United States, Europe, and Japan, and has been the topic of several international conferences and workshops in recent years.

There are immediate needs for automated surveillance systems in commercial, law enforcement, and military applications. Mounting video cameras is cheap, but finding available human resources to observe the output is expensive. Although surveillance cameras are already prevalent in banks, stores, and parking lots, video data currently is used only "after the fact" as a forensic tool, thus losing its primary benefit as an active, real-time medium. What is needed is continuous 24-hour monitoring of surveillance video to alert security officers to a burglary in progress or to a suspicious individual loitering in the parking lot, while there is still time to prevent the crime. In addition to the obvious security applications, video surveillance technology has been proposed to measure traffic flow, detect accidents on highways, monitor pedestrian congestion in public spaces, compile consumer demographics in shopping malls and amusement parks, log routine maintainence tasks at nuclear facilities, and count endangered species. The numerous military applications include patrolling national borders, measuring the flow of refugees in troubled areas, monitoring peace treaties, and providing secure perimeters around bases and embassies.

The 11 papers in this special section illustrate topics and techniques at the forefront of video surveillance research. These papers can be loosely organized into three categories.

Detection and tracking involves real-time extraction of moving objects from video and continuous tracking over time to form persistent object trajectories. C. Stauffer and W.E.L. Grimson introduce unsupervised statistical learning techniques to cluster object trajectories produced by adaptive background subtraction into descriptions of normal scene activity. Viewpoint-specific trajectory descriptions from multiple cameras are combined into a common scene coordinate system using a calibration technique described by L. Lee, R. Romano, and G. Stein, who automatically determine the relative exterior orientation of overlapping camera views by observing a sparse set of moving objects on flat terrain. Two papers address the accumulation of noisy motion evidence over time. R. Pless, T. Brodský, and Y. Aloimonos detect and track small objects in aerial video sequences by first compensating for the self-motion of the aircraft, then accumulating residual normal flow to acquire evidence of independent object motion. L. Wixson notes that motion in the image does not always signify purposeful travel by an independently moving object (examples of such "motion clutter" are wind-blown tree branches and sun reflections off rippling water) and devises a flow-based salience measure to highlight objects that tend to move in a consistent direction over time.

Human motion analysis is concerned with detecting periodic motion signifying a human gait and acquiring descriptions of human body pose over time. R. Cutler and L.S. Davis plot an object's self-similarity across all pairs of frames to form distinctive patterns that classify bipedal, quadripedal, and rigid object motion. Y. Ricquebourg and P. Bouthemy track apparent contours in XT slices of an XYT sequence volume to robustly delineate and track articulated human body structure. I. Haritaoglu, D. Harwood, and L.S. Davis present *W4*, a surveillance system specialized to the task of looking at people. The *W4* system can locate people and segment their body parts, build simple appearance models for tracking, disambiguate between and separately track multiple individuals in a group, and detect carried objects such as boxes and backpacks.

Activity analysis deals with parsing temporal sequences of object observations to produce high-level descriptions of agent actions and multiagent interactions. In our opinion, this will be the most important area of future research in video surveillance. N.M. Oliver, B. Rosario, and A.P. Pentland introduce Coupled Hidden Markov Models (CHMMs) to detect and classify interactions consisting of two interleaved agent action streams and present a training method based on synthetic agents to address the problem of parameter estimation from limited real-world training examples. M. Brand and V. Kettnaker present an entropy-minimization approach to estimating HMM topology and

- *R.T. Collins and T. Kanade are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA E-mail:{rcollins, tk}@cs.cmu.edu.*
- *A.J. Lipton is with DiamondBack Vision, Inc., Washington DC. E-mail: ajl@dbvision.net.*

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112011.

parameter values, thereby simultaneously clustering video sequences into events and creating classifiers to detect those events in the future. Y.A. Ivanov and A.F. Bobick recognize gestures and multiobject interactions from noisy, low-level tracking data by parsing a stochastic context-free grammar (SCFG) that defines multiple events that can be occuring simultaneously in the scene. T. Wada and T. Matsuyama present a hypothesize-and-test approach to recognizing multiple object behaviors directly from video sequences using a Nondeterministic Finite Automaton (NFA) that allows all feasible interpretation states to be simultanously active. They also introduce a colored-token propagation mechanism to keep track of the partial interpretations being assembled for different objects over time and present extensions to handle multiple simultaneous video streams.

Several of these papers represent work funded under the recent DARPA Video Surveillance and Monitoring (VSAM) research program. Carnegie Mellon University was chosen to lead this effort by of developing an end-to-end testbed system that integrates a wide range of advanced surveillance techniques: real-time moving object detection and tracking from stationary and moving camera platforms, recognition of generic object classes (e.g., human, sedan, truck) and specific object types (e.g., campus police car, FedEx van), object pose estimation with respect to a geospatial site model, active camera control and multicamera cooperative tracking, human gait analysis, recognition of simple multiagent activities, real-time data dissemination, data logging, and dynamic scene visualization. We invite the reader to visit the VSAM web page at http://www.cs.cmu.edu/~vsam/ for more information.

Discussions of video surveillance research with nonpractioners invariably lead to comments about Big Brother. Although this is obviously not the goal of current video surveillance research, the concern is reasonable. In 1998, the NYC Surveillance Camera Project run by the New York Civil Liberties Union documented nearly 2,500 surveillance cameras viewing public spaces within Manhattan. The vast majority are privately owned cameras installed outside businesses and apartment complexes, with no mechanism to correlate information between them. However, it would not be infeasible for a sufficiently well-funded government to install a network of thousands of cameras capable of tracking individual citizens as they walk through the city. As the two research paths of video surveillance and biometric identification begin to merge, this scenario becomes even more troubling. Is the promise of never being mugged worth the loss of privacy implied by always being watched? These larger societal questions stray outside the scope of this technical journal, but now is a good time to begin to specify what data should be collected, how long it should be stored, and who has access, so that an ethical framework will be in place to guide the development and application of the powerful technology that will soon be available.

Robert T. Collins
Alan. J. Lipton
Takeo Kanade

Robert T. Collins received the PhD degree in computer science in 1993 from the University of Massachusetts at Amherst for work on scene reconstruction using stochastic projective geometry. He is a member of the Research Faculty at the Robotics Institute of Carnegie Mellon University (CMU). From 1992 to 1996, he was technical director of the DARPA RADIUS project at the University of Massachusetts, culminating in the ASCENDER system for populating 3D site models from multiple, oblique aerial views. From 1996 to 1999, Dr. Collins was technical codirector of the DARPA Video Surveillance and Monitoring (VSAM) project at CMU. This project developed real-time, automated video understanding algorithms that guide a network of active video sensors to monitor the activities of people and vehicles in a complex scene. Dr. Collins has published for more than a decade on topics in video surveillance, 3D site modeling, multiimage stereo, projective geometry, and knowledge-based scene understanding.

Alan J. Lipton received the PhD degree in electrical and computer systems engineering from Monash University, Melbourne, Australia in 1996. For his thesis, he studied the problem of mobile robot navigation by natural landmark recognition using on-board vision sensing. He is a senior scientist at DiamondBack Vision, Inc., an internet startup company based in Washington, D.C. From 1997 through 2000, he served on the faculty of CMU's Robotics Institute. During his time at CMU, Dr. Lipton was a project comanager of DARPA's Video Surveillance and Monitoring (VSAM) project. On this project, Dr. Lipton developed algorithms for detection and tracking of people and vehicles from video streams, integration and fusion of video data, user interfaces for vision system networks, and intelligent sensor control.

Takeo Kanade received the BE degree in electrical engineering from Kyoto University in 1968, the ME degree in 1970, and the PhD degree in 1973. He is the U.A. and Helen Whitaker Professor of Computer Science and Robotics and director of the Robotics Institute at Carnegie Mellon University. He has made widely known technical contributions in multiple areas of computer vision, robotics, and sensor design. At CMU, he has led many major projects on vision and robotics sponsored by NSF, DARPA, NASA, DOE, and NIMH. He is a member of the National Academy of Engineering and a fellow of the IEEE, ACM, and AAAI, respectively. He has received several awards, including the Joseph F. Engelberger Award, the JARA Award, the Yokogawa Prize, the Hip Society, Otto AuFranc Award, and the Marr Prize. He is founding chief editor of the *International Journal of Computer Vision.*