Distributed Interactive Video Arrays for Event Capture and Enhanced Situational Awareness

Mohan M. Trivedi, Tarak L. Gandhi, and Kohsia S. Huang,
University of California, San Diego

A new video surveillance approach employing a large-scale cluster of video sensors demonstrates the promise of multicamera arrays for homeland security.

Video surveillance activity has dramatically increased over the past three years. Earlier work dealt mostly with single stationary cameras, but the recent trend is toward active multicamera systems. Such systems offer several advantages over single camera systems—multiple overlapping views for obtaining 3D information and handling occlusions, multiple nonoverlapping cameras for covering wide areas, and active pan-tilt-zoom (PTZ) cameras for observing object details.

Research interests have thus migrated from simple static image-based analysis to video-based dynamic monitoring and analysis.[1,2] Researchers have made strides addressing illumination, background, color, and perspective invariance issues.[1,2] They can also better track and analyze deformable shapes associated with moving human bodies and moving cameras[3] and have improved activity analysis and control of multicamera systems.[2,3] Our own research deals with a distributed array of cameras that offer wide area monitoring and scene analysis at multiple levels of abstraction. However, installing multiple sensors introduces new system design issues and challenges. We need handoff schemes for passing tracked objects between sensors and clusters, methods for determining the best view given the scene's context, and sensor-fusion algorithms to best exploit a given sensor or sensor modality's strengths.

To address these issues, we've developed a multicamera video surveillance approach, called Distributed Interactive Video Array. The DIVA framework provides multiple levels of semantically meaningful information ("situational" awareness) to match the needs of multiple remote observers. A large-scale, cluster of video streams lets us observe a remote scene, and using automatic focus-of-attention and event-driven servoing (motorized control of camera PTZ) captures desired events at appropriate resolutions and perspectives. We've designed DIVA-based systems that can track and identify vehicles and people, monitor perimeters and bridges, and analyze activities. Deployment of select DIVA modules at Super Bowl XXXVII and on roadways and a bridge in San Diego has proven the value of computer vision techniques in homeland security (see the "Computer Vision's Role in US Homeland Security" sidebar).

Framework and functionalities

Single-perspective-camera-based systems limit the quantity and quality of data available from the viewable environment. Furthermore, systems that use a single dedicated processor to analyze and record data can't distribute the processing, select from an array of available sensors, or access real-time or archived data at multiple remote sites.

DIVA, on the other hand, supports *distributed video networks* that distribute sensors over a wide area for complete coverage (see figure 1). It also offers *televiewing* capabilities—all sources of information are available through a TCP/IP connection to the distributed computers. In addition, DIVA includes *active camera systems* that exploit redundant sensing by having one or more central monitors select the camera with the best view of a given area (focus of attention) in response to an event.

1541-1672/05/$20.00 © 2005 IEEE
Published by the IEEE Computer Society

Computer Vision's Role in US Homeland Security

Computer vision is a key AI research area. From the 1970s to the 1990s, computer vision proved its practical value in a wide range of application domains including medical diagnostics, semiconductor manufacturing, automatic target recognition and smart weapons, remote sensing, and various environmental applications.

It thus wasn't surprising that the first set of post-9/11 *requests for proposal* by the Combating Terrorism Technology Office of the Technical Support Working Group, managed by the US Secretary of Defense, included several computer-vision-related research topics. The RFPs solicited projects that planned to develop rapid prototypes in less than two years.[1] They also urgently sought new concepts and systems for

- remote monitoring of real- or near-real-time movements of forces and resources—in particular, networked autonomous systems that provide a fused picture of the environment and movements;
- locating faces in video images containing one or more human faces, with special interest in "natural environments" with unconstrained lighting and pose angles;
- identifying faces in video images under unconstrained lighting and pose conditions with potential for real-time applications;
- systems for tracking a single person through multiple sequential video images or through multiple cameras in uncontrolled lighting environments;
- terrorist behavior and action prediction technology to assist the analysis and identification of patterns, trends, and models of behavior of terrorist groups and individuals, including visualization and display tools for understanding the relationships between people, events, and behavior patterns; and
- physical security support to protect personnel, equipment, and facilities against terrorist activities.

The US Department of Homeland Security also recognized the importance of the computer vision field, with one of its first set of RFPs issued in April 2004 titled "Automated Scene Understanding." Many other US government agencies, including the National Research Council, encouraged realignment of research agendas and programs to support homeland security applications.[2,3] For example, the National Science Foundation sponsored a number of workshops to identify and encourage research in cyberinfrastructure and sensor network fields.[4] Computer vision was once again identified as an important topic in an NSF report that highlighted the need for developing "ubiquitous vision" with networked and cooperative arrays of cameras (see www.calit2.net/news/2003/3-17_NSF.html).

References

1. Under Secretary of Defense for Acquisition, Technology and Logistics, Technical Support Working Group, Broad Agency Announcement 02-Q-4655, Oct. 2001.
2. Committee on Science and Technology for Countering Terrorism, National Research Council, *Making the Nation Safer: The Role of Science and Technology in Countering Terrorism*, National Academies Press, 2002.
3. H. Chen, F. Wang, and D. Zeng, "Intelligence and Security Informatics for Homeland Security: Information, Communication, and Transportation," *IEEE Trans. Transportation Systems*, vol. 5, no. 4, 2004, pp. 329–341.
4. S. Mehrotra et al., "Project Rescue: Challenges in Responding to the Unexpected," *Proc. SPIE*, vol. 5304, 2003, pp. 179–192.

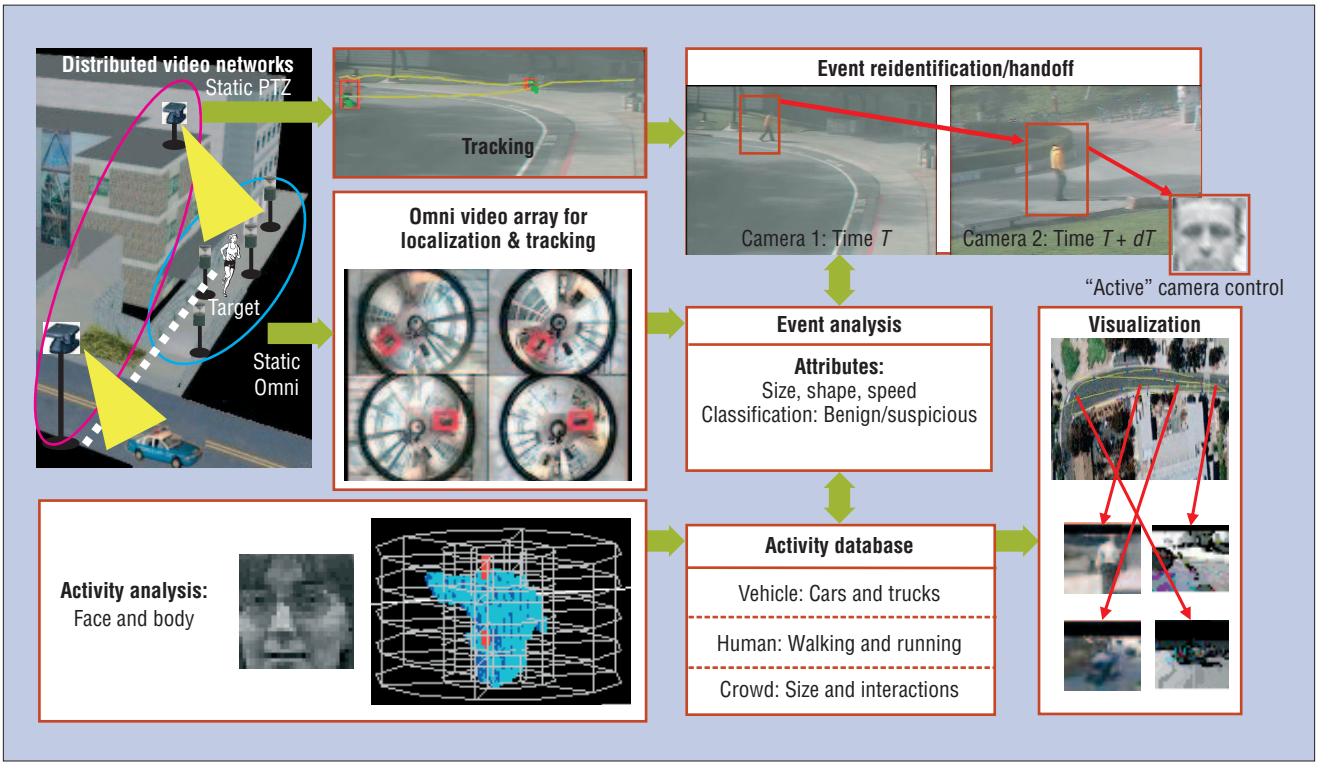


Figure 1. Active video capture and analysis for multilevel situational awareness using Distributed Interactive Video Array (DIVA).

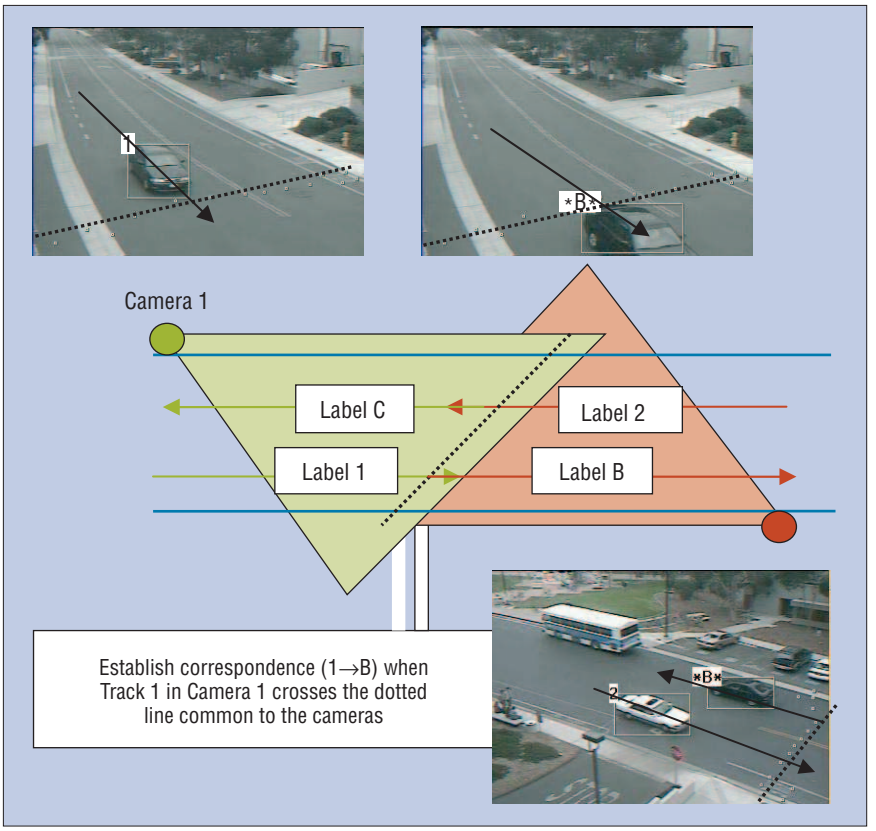


Figure 2. Vehicle tracking and handoff between cameras with partially overlapping fields of view.

Another feature is *multiple object tracking and handoff* (reidentification) using several cameras. DIVA creates a model of the environment with which it can interact to detect, segment, and track objects in a scene between viewpoints. It can then perform *3D localization* to locate the objects in 3D world coordinates.

Finally, DIVA offers *multisensor integration*, exploiting information from rectilinear CCD cameras, omnidirectional cameras, and infrared cameras in an integrated and effective way.

To realize the fusion for integrated situational awareness, we've also developed the Networked Sensor Tapestry framework for multilevel semantic integration.[4] The NeST server takes inputs from visual analysis modules for both indoor and outdoor scenarios through the network and archives the data in a database along with appropriate timestamps. NeST assures the tracked person's privacy using a set of programmable plug-in *privacy filters* operating on incoming sensor data. The filters either prevent access to the data or remove any personally identifiable information. We specify the privacy filters using a *privacy grammar* that can connect multiple low-level data filters and features to create arbitrary data-dependent privacy definitions.[4] The Context Visualization Environment tool provides users with a 3D virtual reality interface for ongoing activities. CoVE also lets users replay previous records of surveillance spaces for investigative purposes.

Observing roads, bridges, and perimeters

Homeland security must protect transportation infrastructures[5] from terrorist attacks and natural disasters as well as from continuous degradation caused by heavy traffic and elements of nature. Bridges are critical in such infrastructures, and monitoring them requires both seismic sensors and cameras. Multimodal sensory systems characterize important patterns associated with structural movements and dynamic loads from vehicular traffic. For security purposes, it would be useful if such systems could also identify and track the same vehicle in different cameras spread over a wide area.

Multicamera vehicle tracking

To extract moving vehicles from a video sequence, we need to be able to identify where changes occur in a video scene. We can accomplish this using background subtraction, a commonly used and computationally inexpensive method that generates a background image using several frames of video.[6] We can then subtract that background image from the current video image to separate moving foreground objects.

The vehicle-tracking system processes the resulting image to extract *blobs* and then identifies vehicles based on blobs that satisfy certain size, area, and density constraints. To robustly track vehicles over multiple frames, the system associates existing tracks with appropriate blobs. It combines measured blob positions with track parameters using a Kalman filter to improve accuracy.[6] It then generates new tracks from unassociated blobs and removes tracks that aren't associated with any of the blobs for a certain number of frames.

To seamlessly track vehicles using multiple cameras, the system needs to consistently maintain object identity. When the cameras' fields of view (FOV) are partially overlapping, we have a handoff problem (similar to handoffs in a cellular network); objects leaving one camera must immediately transfer to the other camera. As figure 2 shows, when an object touches any point on a dotted line in one camera, the system checks the corresponding point in the other camera to locate the object and passes the track from the first camera to the second.

When the camera FOVs don't overlap and are separated by a large distance, we have a *reidentification problem*. Reidentification is more difficult than camera handoff because an object in one camera could have several potential matches in the other camera, and we might not always be able to disambiguate all the matches. In such a case, it's more useful to get a few reliable matches than many less reliable matches. Timothy Huang and Stuart Russell developed a probabilistic framework for the vehicle reidentification problem.[7] Based on this framework, we use properties of color, size, and time of transit between the cameras to match vehicles between cameras. Because the proportion of colored vehicles and large vehicles is small, the matches with such vehicles are more likely to be reliable. So, we only select vehicles having sufficiently high color saturation or a size larger than a threshold to avoid false matches. The algorithm for vehicle reidentification is as follows:

- *Vehicle detection:* Detect vehicles in up-

stream and downstream cameras (the first and second camera along the direction of traffic flow) using background subtraction and extract their snapshots.

- *Feature extraction:* For all vehicles in both cameras, use K-means clustering to group colors in vehicle pixels and select the color whose cluster has the largest number of pixels. Estimate the vehicle size based on the pixels of dominant color. Finally, select vehicles meeting the minimal color-saturation and size requirements.
- *Matching:* For each vehicle in the upstream camera, select vehicles in the downstream camera that arrived within a window based on expected transit time. Then compute the weighted distance between the upstream vehicle and all selected downstream vehicles. Assign confidence scores to each match based on the weighted distance. Finally, select matches with a confidence score greater than a threshold (we adjust the threshold to detect as many matches as possible while keeping the number of false matches to a tolerable level).

Figure 3 shows our experiments with vehicle reidentification using a pair of cameras on the Coronado Bridge. We're currently evaluating and improving the algorithm.

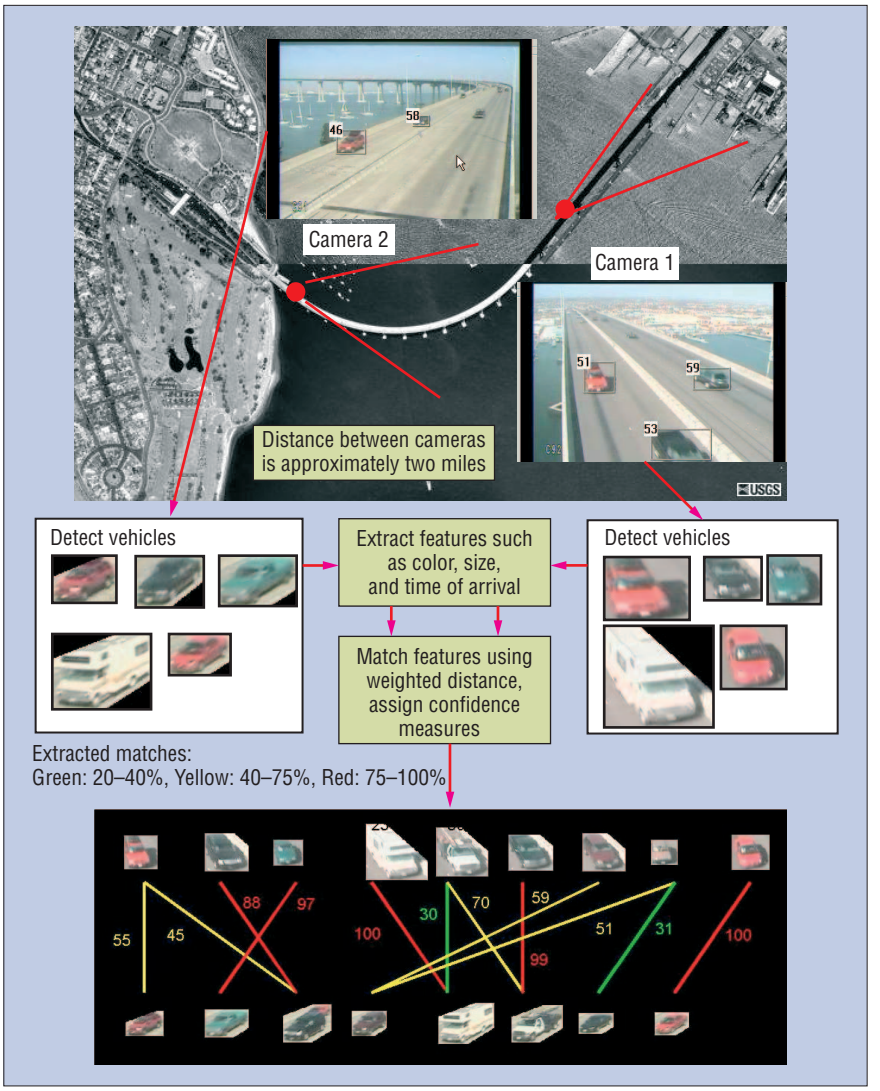


Figure 3. Vehicle reidentification using videos from the Caltrans Traffic Management Center's cameras on the Coronado Bridge. (aerial image courtesy of the US Geological Survey)

Sensor fusion for monitoring infrastructure health

Many civil structures have been instrumented with various types of sensors for monitoring their structural health. Seismic sensors such as strain gauges and accelerometers can provide temporal signatures of vehicles passing over them, which could be used to extract the weight and effect of vehicles on the structure. However, seismic sensors are also sensitive to other natural and artificial phenomena, such as earthquakes, blasts, and external vibrations. Video sensors could help distinguish these phenomena from normal vehicular traffic and give rich information about the vehicles' shape, size, color, velocity, and track history (paths taken).

Figure 4 shows a block diagram of a system under development for monitoring a bridge's structural health by combining information from both seismic and video sensors. The system's vision module processes video streams to detect and track vehicles and extract their image properties. These can be used in conjunction with the responses from seismic sensors to help determine the effect of various types of vehicle loads on a bridge.

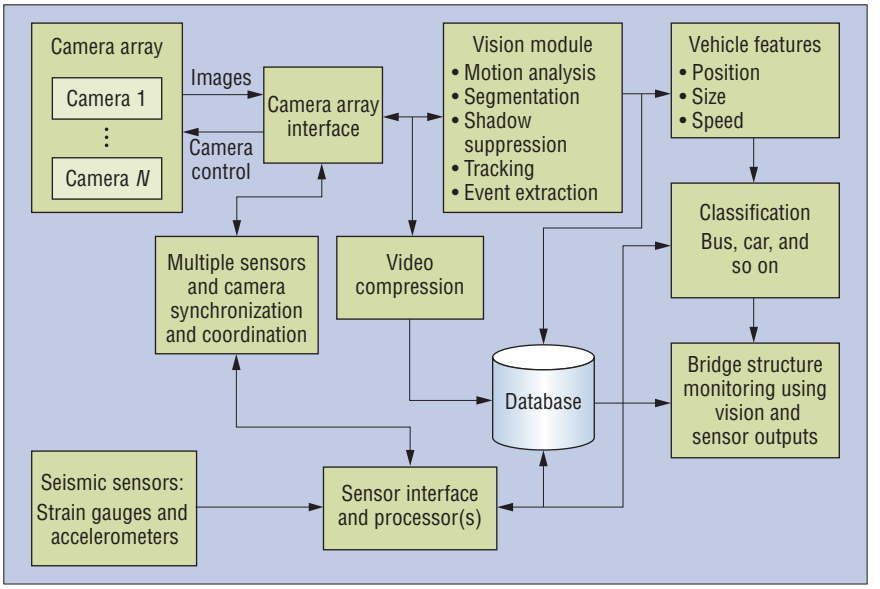


Figure 4. Block diagram for civil-infrastructure monitoring.

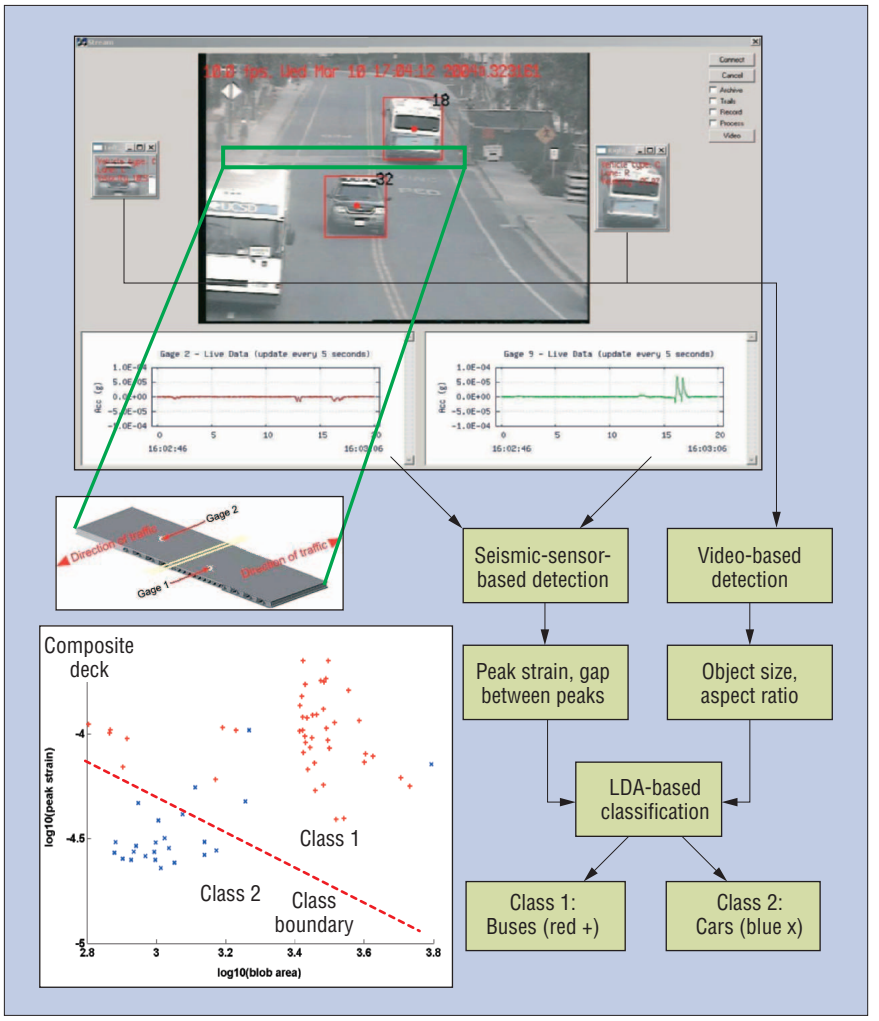


Figure 5. Civil-infrastructure monitoring with camera videos and vibration sensors embedded in the roadway. Vehicles are classified with features from both sensor modalities using linear discriminant analysis.

Figure 5 shows an application for detecting vehicles and extracting their properties, including the vehicle snapshots from video processing and responses from the strain gauges for both directions of traffic. To distinguish between buses and cars, we recorded the

- larger of the two peak responses (corresponding to each wheel base) that the strain gauge recorded when detecting the vehicle,
- time interval between two peak responses,
- vehicle blob area obtained from video-based detection, and
- vehicle blob's aspect ratio (height to width).

Each property is larger for buses than for cars.

We combined these features using Fisher linear discriminant analysis[8] to find an optimal linear combination of the logarithms (for scale invariance) of these properties that maximizes the variation between the classes and minimizes the variation within each class. The boundary between the two classes is a hyperplane obtained by thresholding this linear discriminant using Bayesian error criterion to minimize the number of classification errors.[8]

Perimeter sentry with active camera control

For continuous monitoring of wide areas, it isn't always practical to have a person continuously view the video to identify suspicious activities. It helps to have a system that can automatically extract and summarize interesting events. An important application of such a system is a perimeter sentry, which guards a preconfigurable monitoring zone or virtual fence. Background subtraction detects moving objects such as people and vehicles, which are tracked over frames.[6] Any track that breaches the virtual fence, such as a person passing in that zone, triggers an alarm. In addition, such an event can also initiate active control of other cameras in the array. For example, using the location of the monitoring zone, a system could make a PTZ camera point to the zone to obtain finer event details.

Figure 6 shows an application of the perimeter sentry. A car entering the protected zone at the garage triggers another camera that zooms in toward the event and captures a high-resolution video sequence. A face detection module then captures the intruder's face. In another scenario, a stalled vehicle triggers another camera to zoom in on details such as the vehicle's license plate.

Deployment

We've successfully deployed and tested the systems just described on the UCSD campus and on other sites. In addition to the vehicle reidentification experiment we did on the Coronado Bridge, we've performed vehicle tracking and traffic parameter estimation with cameras overlooking the Interstate 5 freeway passing through the campus. We also demonstrated perimeter sentry near the Coronado Bridge, successfully detecting intrusions in the security zone, and have deployed multimodal vehicle data extraction with video and seismic sensors on a campus road with significant traffic. We're working on deploying a similar system on a bridge over the I-5 freeway. We've also demonstrated multicamera handoff with cameras on a campus street.

In addition, we successfully deployed several DIVA modules at Super Bowl XXXVII (see figure 7).[9] We mounted a high-resolution thermal camera near a riverbed beside the stadium to detect humans and animals in visually cluttered scenes on a 24-hour basis. We used traffic-flow analysis to monitor the peripheral traffic on a nearby road. We also installed an omnicamera in downtown San Diego to simultaneously monitor traffic conditions using a digital televiewer (a software interface unwarping omnivideo to perspective video on customizable PTZ settings) and estimate the crowd size. These surveillance nodes were remotely linked to and controlled by the perimeter sentry command center in Sea Port Village (in downtown San Diego), for the city authorities, police, and first responders.

Tracking people, capturing events, and analyzing activities

In contrast to the outdoor applications we've described, indoor DIVA systems use multiple types of cameras with highly overlapped FOVs for versatile human-related event and activity analysis. The objectives for such systems involve developing sensor networks that derive multilevel awareness of human activity and identity. Figure 8a shows a DIVA system for deriving such multilevel semantic description of activities in a room. The system includes a video analysis level that processes camera array videos for person segmentation. With both omni and rectilinear PTZ video arrays, the system can obtain a multiresolution representation of human activities.[10] Next, the localization level detects people and tracks them continuously. PTZ array also captures human faces.[10] Then, the gesture analysis and identification levels derive higher semantic details for human gesture and identity. The integration and visualization level derives the spatial-temporal co-occurrence of the events for high-level activity awareness to focus on certain humans. This level also archives and visualizes the events in real time as well as replays them for offline investigative purposes.[4]

Figure 6. Active camera control. DIVA can define a security zone for intrusion detection, and an event can activate another camera to capture a close-up view.

Real-time 3D person tracking

For real-time indoor 3D person tracking, the omni array videos first undergo pixel-level processing that segments the human silhouettes by background subtraction with shadow elimination (see figure 8). The system measures the horizontal locations and heights of people from the silhouettes by triangulation with the calibrated omnivideo array.[10] It then associates these 3D measurements of humans with the existing tracks and decides track initialization and termination according to time constraints. Finally, it updates Kalman track filters with the new measurements to output the estimated and predicted track locations.

After tracking the person, the system determines a focus of attention to capture the face

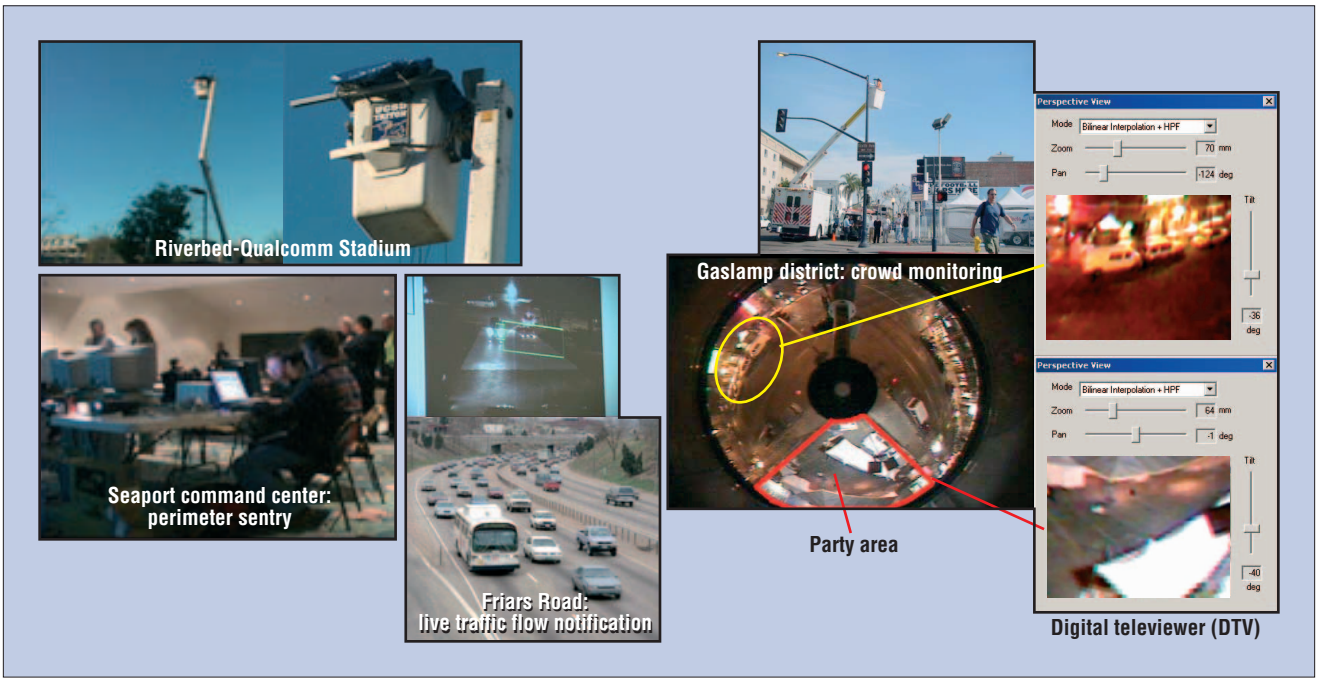


Figure 7. The DIVA security network deployment in Super Bowl XXXVII at San Diego in January 2003.

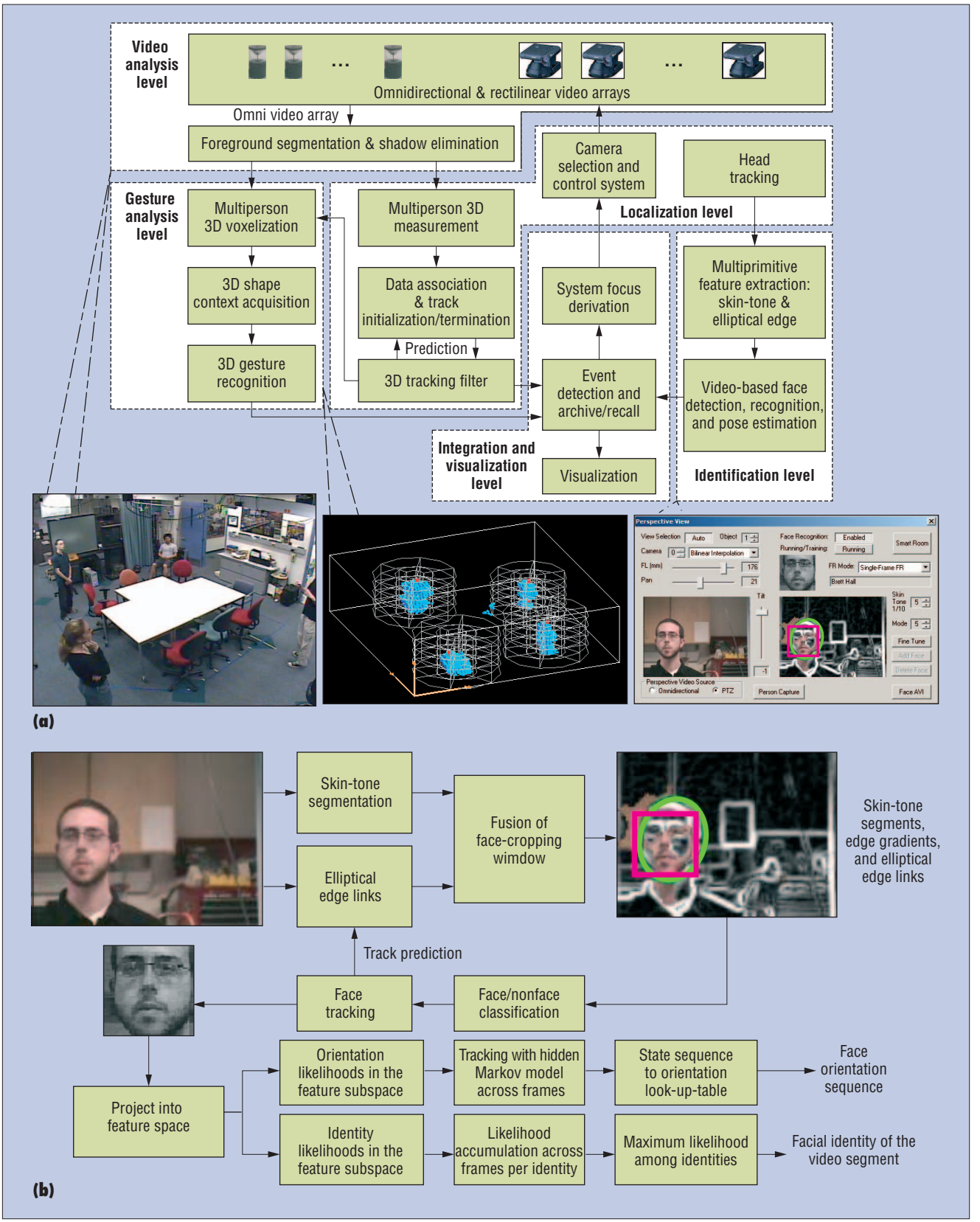


Figure 8. (a) An architecture of the indoor DIVA system with multilevel visual context abstraction. (b) A flow chart for video-based face capture, pose estimation, and recognition.

(see figure 8a). Based on the head location, the system can choose a nearby PTZ camera to focus on the subject's face.[10] It can then detect skin-tone regions and elliptical edges for face contour from the image to find plausible faces (see figure 8b). It verifies the candidates using a face classifier and updates the face tracks. Using this real-time scheme, we can robustly detect a face under challenging environmental conditions.

Face orientation estimation is useful in assessing a subject's focus of attention and intent. As figure 8b shows, the system first projects the face video frames into a facial feature subspace and then computes the likelihood scores of a face frame associated with various face orientation clusters. It tracks these orientation likelihoods across frames using a hidden Markov model (HMM), whose state sequence is equivalent to the final face orientation sequence. For face recognition,[10] it trains clusters of different identities in the feature subspace and accumulates the identity likelihoods across frames in a video segment to make the final decision. Our experiments have shown that these novel video-based face analysis algorithms surpass single frame-based methods in reliability due to information accumulation over time.

DIVA can also capture human activities using 3D human body gesture analysis. As figure 8a shows, the system can reconstruct voxels (volume elements—3D analog of pixels) of human subjects from the array omnivideos. Then we can form a cylindrical, 3D-shape-context descriptor to each subject to capture the body configurations. A vocabulary of HMMs model the dynamics of the 3D body configurations or gestures. Given a gesture sequence, the 3D-shape-context histograms are vector quantized and the index sequence goes to the HMM vocabulary to decide the final gesture by maximum likelihood. With this scheme, we can robustly perform gesture recognition even with noisy and low-resolution human body voxelization.

Integrated situational awareness

We've deployed the real-time 3D tracker in a 6.7 × 6.6 m room with four omni cameras, each of which captures a 640 × 480 pixel video. This lets us obtain tracking accuracy of approximately 20 cm for five people simultaneously.[10] For a person entering or exiting the room, access zones are defined and displayed in the CoVE interface as shown in figure 9a. The tracker sends data to the NeST server to monitor the zones and to archive data over long periods. This lets us accumulate passing counts of the zones with the track indices.

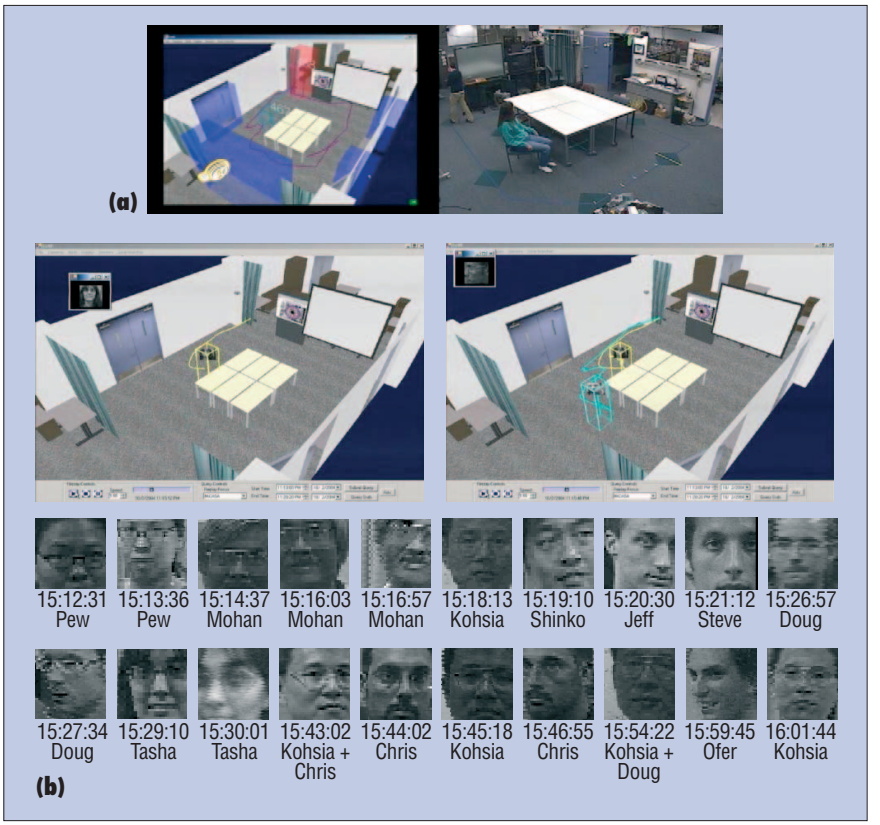


Figure 9. (a) Long-term zone watch. Zone count changes as the human track passes through it. (b) Results of automatic tracking-based face capture and archiving during approximately 50 minutes. In multiperson cases, the subjects are captured one at a time.

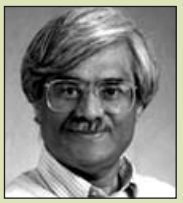
As shown in the dialog box in figure 8a, a PTZ camera is driven by the 3D tracker to capture the human face upon entrance. The face is detected in approximately 15 frames per second and identified by the system.[10] Also, the face image is attached to the human bounding box in CoVE as shown in figure 9b. Long-term face archive is shown in figure 9b. The 3D tracker monitors the room continuously and archives the entering people automatically with a timestamp. The tracker is suitable for visual surveillance and forensic support applications. As an attentive scenario, multiple people can be sequentially scanned using the PTZ camera closest to each person. When a person enters or exits, the system would reset the scanning order.

Currently, the gesture recognition and video-based face orientation and recognition modules that involve HMM are implemented in Matlab. Although running offline, they give very promising accuracies.[10] System situational awareness would be enhanced once their C++ implementations are available.

Computer vision will play a significant role in enhancing personal safety and protecting infrastructure and properties within national borders. Remote monitoring of transportation facilities and public spaces as well as automatic notification systems triggered by potentially dangerous events can certainly incorporate vision systems as essential components. However, such applications pose major challenges to the existing and commercially available systems, mainly due to strict requirements of very high detection rates and almost zero false alarm rates, robustness to environmental variations, distributed and almost ubiquitous coverage, and real-time or near-real-time performance.

Fortunately, systems similar to those we've discussed promise to provide solutions that address specific "threats" encountered in protecting critical infrastructures, national landmarks, or public spaces. For instance, in the event of natural or man-made disaster, a DIVA-type system can provide an exact visual and seismic damage assessment. If a protected site such as an airport runway, port facility, or military base is breached, a DIVA-based system

The Authors



Mohan M. Trivedi is a professor of electrical and computer engineering and the founding director of the Computer Vision and Robotics Research Laboratory at the University of California, San Diego. His research interests include intelligent systems, computer vision, intelligent (smart) environments, intelligent vehicles and transportation systems, and human-machine interfaces. He received his PhD in electrical engineering from Utah State University. He has received the Distinguished Alumnus Award from Utah State University, the Pioneer Award (Technical Activities), and the Meritorious Service Award from the IEEE Computer Society. Contact him at the Computer Vision and Robotics Research Lab., Univ. of California, San Diego, 9500 Gilman Dr. 0434, La Jolla, CA 92093; mtrivedi@ucsd.edu.

Tarak L. Gandhi is a postdoctoral researcher at the Computer Vision and Robotics Research Laboratory at the University of California, San Diego. His research interests include computer vision, motion analysis, image processing, robotics, target detection, and pattern recognition. He is working on projects involving intelligent driver assistance, motion-based event detection, traffic-flow analysis, and structural health monitoring of bridges. He received his PhD in computer science and engineering, specializing in computer vision, from Pennsylvania State University. Contact him at the Computer Vision and Robotics Research Lab., Univ. of California, San Diego, 9500 Gilman Dr. 0434, La Jolla, CA 92093; tgandhi@ucsd.edu.

Kohsia S. Huang is a postdoctoral researcher at the Computer Vision and Robotics Research Laboratory at the University of California, San Diego. His research interests include multimodal intelligent environments, computer vision, machine learning, and signal processing. He received his PhD in electrical engineering from the University of California, San Diego. He is a member of the IEEE. Contact him at the Computer Vision and Robotics Research Lab., Univ. of California, San Diego, 9500 Gilman Dr. 0434, La Jolla, CA 92093; khuang@ucsd.edu.

can identify the point of breach, take close-up video images of the event, track the vehicle or person responsible for the breach, and warn the appropriate authorities. ■

Acknowledgments

We thank the reviewers for their constructive comments and the following research sponsors: the Technical Support Working Group of the US Department of Defense, NSF Information Technology Research Grant for Structural Health Monitoring, and NSF Information Technology Research-sponsored Rescue Project. UC Discovery Grants supported development of various research test beds.

References

1. *Proc. ACM 2nd Int'l Workshop Video Surveillance & Sensor Networks*, ACM Press, 2004.
2. *Proc. IEEE Conf. Advanced Video and Signal Based Surveillance*, IEEE Press, 2003.
3. *Multimedia Systems*, special issue on visual surveillance, vol. 10, no. 2, 2004, pp. 116–180.
4. D. Fidaleo, R.E. Schumacher, and M.M. Trivedi, "Visual Contextualization and Activity Monitoring for Networked Telepresence," *Proc. ACM 2nd Int'l. Workshop Effective Telepresence*, ACM Press, 2004, pp. 31–39.
5. H. Chen, F. Wang, and D. Zeng, "Intelligence and Security Informatics for Homeland Security: Information, Communication, and Transportation," *IEEE Trans. Transportation Systems*, vol. 5, no. 4, 2004, pp. 329–341.
6. D.A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice-Hall, 2003.
7. T. Huang and S. Russell, "Object Identification: A Bayesian Analysis with Application to Traffic Surveillance," *Artificial Intelligence*, vol. 103, nos. 1–2 , 1998, pp. 77–93.
8. R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, *2nd ed.,* Wiley InterScience, 2000.
9. D. Ramsey, "Researchers Work with Public Agencies to Enhance Super Bowl Security," California Institute for Telecommunications and Information Technology, 2003, www.calit2.net/news/2003/2-4_superbowl.html.
10. M.M. Trivedi, K.S. Huang, and I. Mikić, "Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces," *IEEE Trans. Systems, Man and Cybernetics, Part A*, vol. 35, no. 1, 2005, pp. 145–163.