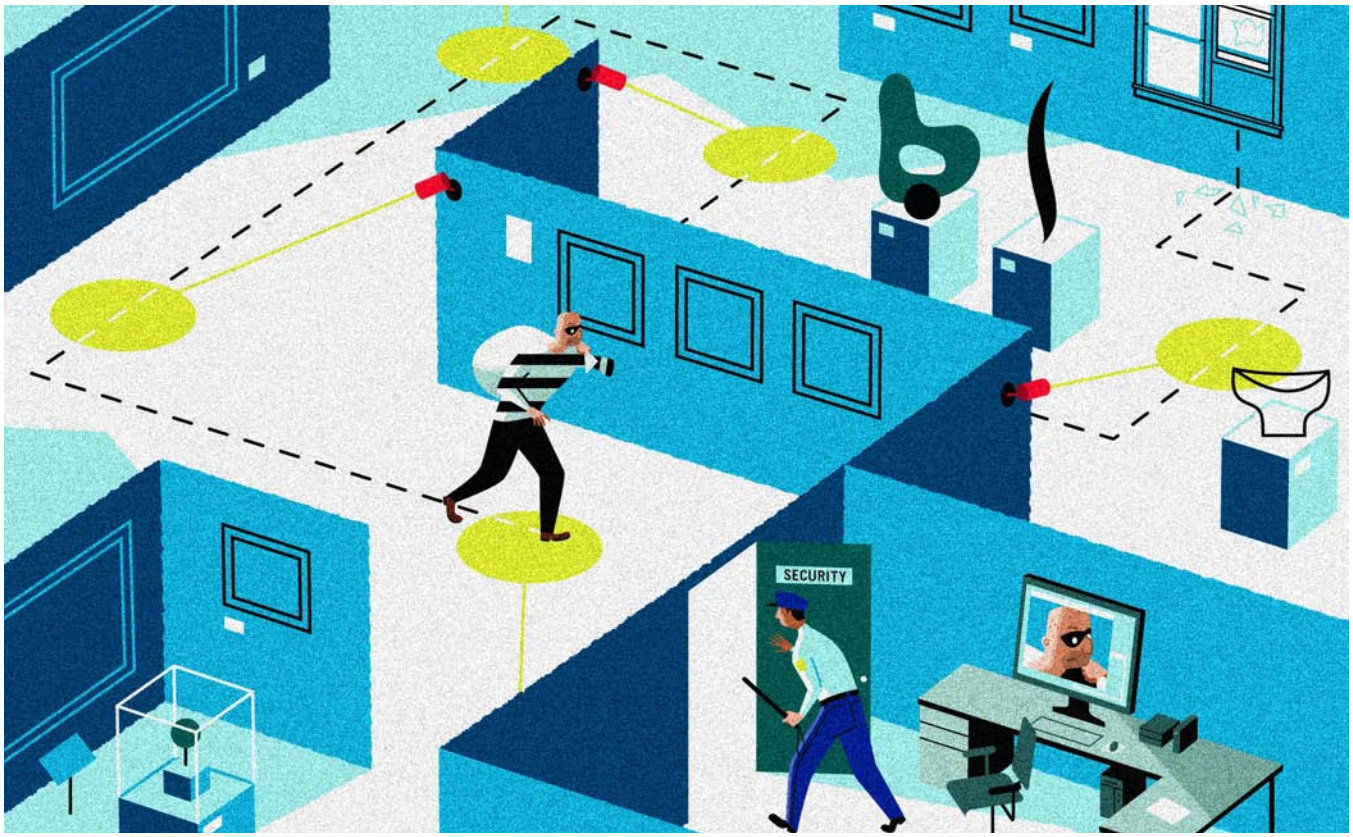

ILLUSTRATION BY RON CHAN

Looking for Targets

BY ARSLAN BASHARAT, ASAAD HAKEEM, AND MUBARAK SHAH OF THE UNIVERSITY OF CENTRAL FLORIDA; AND ABHIJIT MAHALANOBIS OF LOCKHEED MARTIN CORP.

Automatic target detection and recognition for video sensors requires a different method for each class of topology.

Video sensor networks play a vital role in unattended wide-area surveillance. Most computer vision research in this area deals with stationary electro-optical sensor networks, which have topologies with overlapping and non-overlapping fields of view (FOV) between the sensors. Recently there has been an increased interest in networks with sensors on mobile platforms such as mobile robots, all-terrain vehicles, and unmanned aerial vehicles (UAVs). Because of the data and time involved, automatic target detection and recognition (ATD/R) is becoming increasingly important for both stationary and mobile sensor networks.

The goal of an effective video surveillance system is to detect targets in the scene and find their correspondence across frames in a video. Issues inherent in video surveillance include rapidly changing lighting conditions (i.e., as a result of cloud cover), the presence of shadows, target occlusion, and the detection of target entry and exit.

Targets can either be moving or stationary in an observed scene. One technique for detecting stationary targets is the maximum-average-correlation-height approach. In this technique, the video frame is processed by a bank of linear correlation filters that are optimized to respond to the

presence of a specific target object by producing a peak at the corresponding location. Each filter is synthesized using representative training images of a respective target. This allows the filter to exhibit distortion tolerance over a limited range of target orientations.

Moving objects are typically detected using background subtraction. The techniques discussed in this article use multiple levels of processing during background subtraction. The first level is pixel-level processing, which uses color- and gradient-based distributions separately to find pixels that belong to the foreground (target) or the background. The second level is region-level processing, which integrates the gradient and color information. A connected-component algorithm groups all foreground pixels into regions. Once the target silhouettes are obtained, target tracking depends on the network topologies.

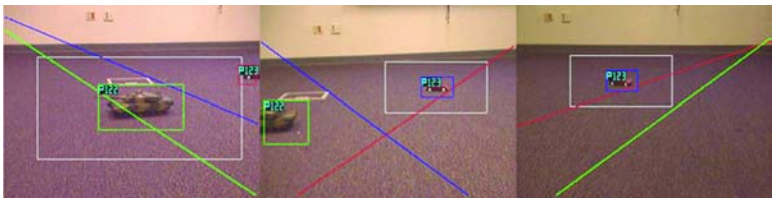

Figure 1 The automatic FOV line estimation for three overlapping cameras (red, green, and blue) shows consistent labeling for two models of vehicles. Output after target detection, recognition, and global correspondence generates unique target IDs across the sensor network.

Stationary Sensor Topologies

We can class network topologies as stationary sensors with overlapping FOVs, stationary sensors with non-overlapping FOVs, or mobile sensors with overlapping FOVs. The test networks in the work described here were based on commercial off-the-shelf components.

Let's start with the case of the stationary sensor network with overlapping FOVs, which is the simplest among the three network topologies. Tracking targets across multiple cameras involves establishing correspondence between the detected targets in each camera. We achieve this by estimating spatial correspondence between cameras using the camera FOV lines. During training, a single person walks around in the network of overlapping sensors and the algorithm recovers the FOV lines automatically. Whenever a person enters or exits a camera's FOV and he or she is still visible in another camera, the system marks that point as a candidate point (see figure 1). Two such points are sufficient to mark a camera FOV line, although more points give a better estimate of the FOV lines using a Hough transform.

After training, the algorithm solves target correspondence by giving each target consistent labeling during entry and exit, using the spatial correspondence (FOV lines) between cameras. We performed experiments for both human and vehicle targets. The results in figure A show consistent labeling across three cameras for two models of vehicles in the overlapping FOV of the sensors.[1]

Wide-area surveillance problems do not always offer the luxury of overlapping FOVs, so we have developed a method for ATD/R in a network of non-overlapping stationary sensors.[2] This method exploits the redundancy in paths that people and vehicles tend to follow (i.e., roadways and walkways/trails). The algorithm learns the network topology using target appearance and spatio-temporal probability models (see figure 2).

During training, we model the target appearance in each camera using color histograms and calculate the change in target appearance between cameras using a distance measure. This distance measure is estimated using the modified Bhattacharya coefficient, a measure of similarity between two data samples, and is approximated as a Gaussian distribution.

We estimate intercamera space-time probabilities using a statistical method for density estimation. The feature vector used for learning the space-time probability density functions from camera C_i to C_j is a 7-D vector consisting of a 2-D exit location from C_i, a 2-D entry location from C_j, a 2-D exit velocity, and the time taken between exit and entry.

We achieve target correspondence across cameras using a maximum *a posteriori* (MAP) estimate of the observation sequence to maximize target appearance and spatio-temporal probabilities. Our experiments were conducted with two different network topologies, one consisting of two sensors and the other consisting of three sensors in an outdoor scene, as shown in figure B. In the second experiment, we trained the system using a 10-minute video sequence with multiple people in the environment. Testing was performed on a 15-minute video with 45 intercamera transitions, all of which

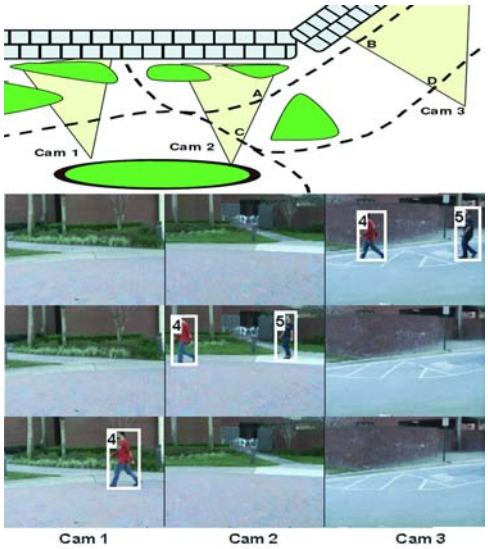

Figure 2 The software learns the sensor network topology using the target appearance and spatio-temporal probability models (top). As a result, it can label objects consistently across cameras using the MAP estimate of all of the targets (bottom); each row of images represents the frames from three sensors at a specific time.

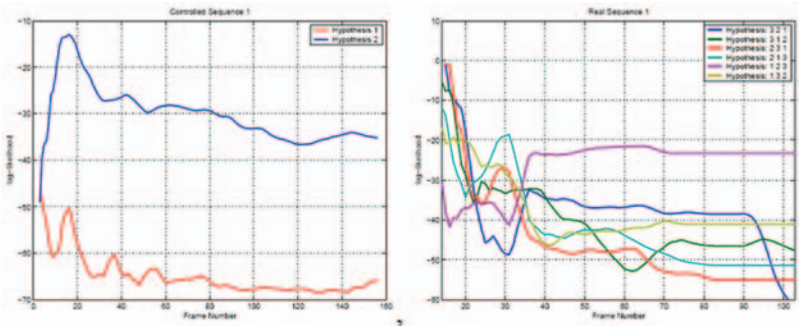

Figure 3 An experiment with only two targets in two sensors has two possible hypotheses (left), making the correct solution obvious on plots of log likelihood versus time. It takes longer to detect the unique hypotheses in a scenario with two sensors, three targets, and six hypotheses (right) because of the geometrical similarities at the start of the 3-D trajectories.

were detected correctly. As can be seen in figure 2, the system is able to label targets across cameras consistently.[2]

Going Mobile

The above-mentioned frameworks perform well for static network topologies but are not adaptive to network topology changes. Because of variations in camera motion and position over time, conventional methods that exploit appearance or positional similarities cannot be used. Rather, our mobile sensor method benefits from the geometrical similarities among target trajectories captured across different cameras. These geometric similarities are used to solve global target correspondence.

Given a video for a single sensor, we first generate target trajectories by projecting (warping) all of the target locations to a reference frame (i.e., the first frame) using interframe homography. We match each trajectory in a given sensor to all the trajectories in the rest of the sensors using a maximum likelihood estimation of the trajectory similarity measure. This measure is based on a cost function that uses algebraic and geometric distances between trajectories. Because trajectories of the same target in two different sensors are generated from a common 3-D trajectory on the ground plane, we can compute a homography between these two trajectories with a direct linear transform algorithm.

The likelihood of correspondence between two target trajectories is based on the re-projection distance (based on intertrajectory homography) between them. In the case of two moving sensors, we can find the global correspondence using maximum matching of a complete bi-partite graph, in which all of the nodes in a bi-partition represent the trajectories from a particular sensor. The edge weights between these bi-partitions constitute the correspondence likelihood estimates.

A more complex scenario involves multiple sensors detecting several targets simultaneously. In order to obtain a globally optimal trajectory correspondence, the solution is equivalent to finding maximum matching of the split G* of the directed acyclic weighted graph D.[3]

We performed experiments on controlled indoor sequences and the outdoor UAV videos, assessing the effectiveness of the likelihood maximization estimate for the global correspondence hypotheses (see figure 3). The method was able to detect the same three targets on the ground in sample frames from two different detectors (see figure 4).

Mosaic images provide better visualization of the tracking scenario. Although the videos in question had short temporal overlap (less than a minute), the estimated target correspondence was correct. The final results are also shown as the blended image of two mosaics using a quadratic color transfer function. The trajectories of three cars are shown in red, blue, and green.

As unattended sensors play an ever-greater role in security, the need for ATD/R increases. We have developed and tested three different approaches for ATD/R in video sensor networks with different network topologies of stationary and mobile sensor nodes. Future challenges include ATD/R scenarios with mobile sensor nodes that have non-planar FOVs. In this case, planar trajectories generated by target motion would not be a viable option. This is significant not only to low-flying airborne sensors but also to sensors aboard mobile robots.

Another challenge is ATD/R with mobile sensor nodes having non-overlapping FOVs. One approach is to benefit from supplementary calibration data, such as gyro and GPS, for finding the solution. **oe**

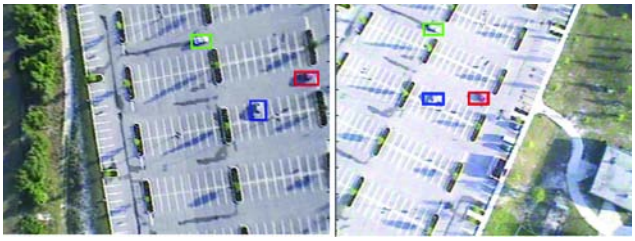

Figure 4 In sample-time-synchronized frames from UAV 1 (left) and UAV 2 (right), the method was able to recognize the same targets across different sensors (red, green, and blue bounding boxes).

Arslan Basharat and Asaad Hakeem are PhD students at Computer Vision Lab and Mubarak Shah is the Agere Chair Professor and Director of the Computer Vision Lab at the University of Central Florida, Orlando, FL. Abhijit Mahalanobis is a fellow at the Lockheed Martin Corp., Orlando, FL. For questions, contact Basharat at arslan@cs.ucf.edu.

References

1. A. Mahalanobis et al., Proc. SPIE #5440, p. 1 (2004).

2. O. Javed et al., Proc. 9th IEEE International Conference on Computer Vision vol II, p. 952 (2003).

3. Y. Sheikh and M. Shah, "Object Tracking Across Multiple Independently Moving Cameras," to be published in Proc. 11th IEEE International Conference on Computer Vision, China (2005).