Intelligent Video for Protecting Crowded Sports Venues

N. Krahnstoever P. Tu T. Yu K. Patwardhan

D. Hamilton B. Yu C. Greco G. Doretto

GE Global Research
1 Research Circle, Niskayuna, NY, 12309, USA
{krahnsto|tu|tyu|patwardh|hamilton|byu|grecoc|doretto}@crd.ge.com

Abstract

Intelligent video in urban settings can be challenging due the presence of crowds, clutter, poor camera placement and continuously changing light conditions. The surveillance of sports venues is particularly difficult, because thousands of people can enter or exit a venue in short periods of time. This paper presents a case study of successfully monitoring a sports venue using a multi-camera multi-target tracking system. The system performed site-wide tracking throughout a network of calibrated cameras and was able to accurately track thousands of people in real-time under challenging conditions. The extracted tracking information was used to detect a range of real-time events such as crowd formation, left luggage, and loitering. In addition all video, track and event information was indexed and stored to allow operators to perform playback and forensic search. This paper will present an overview of the deployed system and discuss the challenges that were encountered during the deployment.

1. Introduction

Video has emerged as one of the key enablers that can significantly enhance city-wide surveillance and security. Consequently, many townships and municipalities are at various stages of installing their own modern large-scale video surveillance systems. However, the ability to effectively make use of such infrastructure remains elusive. Thus many cities and government agencies are looking for computer vision-based approaches to improve the effectiveness of these installations. Automatic surveillance of whole cities and specific urban sites is challenging for reasons such as size, occlusion, clutter, and population density. Another problem is that urban environments are highly variable. For example, no single algorithm can be expected to satisfy the surveillance needs of both a subway platform and a parking lot. For this reason, this paper will focus on a specific type of urban environment, namely sporting venues. Sporting venues are a great concern to both police and homeland security because many such venues frequently host high-profile events attracting audiences as large as 100,000 patrons. Even small venues are at risk, because an attack, even at a small venue, could have a large psychological impact on society.

This case study describes the installation and operation of an experimental tracking and event detection system that was installed in collaboration with the US Department of Homeland Security at a medium sized ice-hockey arena.

The goal of this study was to monitor patron traffic in and out of the arena and to detect a set of events that could be indicative of suspicious activities in real-time. A multi-camera, multi-target person tracking system that has the ability to detect and track people through crowded environments was installed. The deployed system was based on the algorithms described in [6, 10, 12] and was augmented with functionality directed at robust person detection, event detection, and storage.

As this paper will describe in more detail, the deployed system was able to successfully track individuals under extreme levels of activities and varying environment conditions while providing a good level of event recognition to system operators.

This paper is organized as follows: In the next section we will provide a brief overview of the technical components of the system, in particular person detection, tracking and event recognition. In section 3 we will describe the system installation, in particular camera placement and calibration. Results will be presented and discussed in section 4. The paper is concluded in section 5.

2. System Overview

In the following sections we will briefly describe the main technical components used for this deployment. The

978-0-7695-3718-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AVSS.2009.87

IEEE computer society

installed system was based on a multi-target tracking algorithm that is designed to detect people in real-time from multiple cameras. The system follows a detect-and-track approach, where the tasks of detection and tracking are strictly separated. The reason for this approach is that it is computationally very efficient; except for the task of raw person detection, no further image processing is required. This makes multi-sensor fusion of the data relatively straight-forward. The system requires complete knowledge about the geometry of the camera network [13], i.e., full metric camera calibration with respect to a global world reference frame. Although our general system architecture supports non-stationary PTZ (Pan Tilt Zoom) cameras, this study relied solely on stationary cameras.

In the following sub-sections we will describe the main computer vision components, namely person detection, tracking, and event detection. We will also discus integration issues such as storage and retrieval.

2.1. Person Detection

Our approach to person detection [7] is based on the work described in [6]. To obtain a probabilistic foreground map that is needed for this approach, we conduct background learning based on non-parametric kernel density estimation akin to [3] in gray-scale space. This is followed by a series of post-processing steps for false-alarm suppression as well as shadow and light change handling. False alarm suppression is necessary to reduce the influence of non-stationary objects such as trees, flags and small camera motion. Shadow-suppression was necessary for this deployment due to the occurrence of severe shadows at night, sunset and sun-rise (see Figs 6-7). Rapid light changes, one of the most difficult issues for many computer vision systems, occurred frequently on windy, cloudy days.

Our approach to person detection is targeted at explaining a probabilistic foreground map via the placement of geometric shapes that are similar to the size and shape of people and small objects. The approach, described in more detail in [6], makes a range of simplifying assumptions that allow for the efficient evaluation of person likelihoods using integral images. Our system takes advantage of the fact that (i) only people and small objects are expected to be observed in the system field of view, and (ii) that the expected targets can be well approximated by rotationally symmetric upright ellipsoids and spheres. As we will discuss in the results section, these assumptions were occasionally violated by small children and people in wheelchairs. These individuals can be closer to the size and shape of left-behind objects instead of people and hence were neither tracked as people by our system nor detected as left-behind objects (because other conditions necessary to trigger left-behind object detections were violated).

2.2. Tracking

For computational reasons our system employs an approach that separates person detection from tracking. One person detector operates independently per camera and projects the locations and location uncertainties of detections into a world groundplane. The detections from each sensor are time ordered and sent to a centralized tracker that performs data association and tracking. Details of the employed approach have been reported in [6, 10, 12]. For the current deployment the tracking system was augmented with additional components that manage the occurrence of false alarms caused by subjects that move from a location after they have remained stationary for extremely long periods of time and may have hence merged into the background. Due to the nature of the operational environment (stadium entrance with ticket counter), this was not an unusual occurrence (see Figure 8).

The crowds observed at this site, had on occasion extremely high densities (see Figures 8 and 10). This made it necessary to adjust a number of the geometric parameters of the system with regards to how close subjects were allowed to move together.

2.3. Event Detection

For the purposes of demonstrating the ability to potentially identify suspicious individuals based on behavior, our system was equipped with a collection of "Event Generators". The following is a brief description of each type of event that could be recognized by the deployed system. Most of the event generators relied on analyzing the trajectory information of targets in the groundplane [8, 1].

Loitering Person: Our approach to loitering is based on trajectory analysis. If a detected person maintains a speed lower than a given threshold for greater than a specified amount of time, a loitering event is raised.

Fast Person: Two types of fast moving events were instantiated for this system. The first approach simply determines whether or not the speed of a given individual is at any time faster than a given threshold. The second approach determines whether or not the speed of a given individual is significantly greater than the speed of its neighboring individuals.

Counter Flow: A crossing event was raised if the direction of motion of a given individual is opposite to that of the majority of its neighboring individuals.

Group Formation and Dispersion: Our definition of a group formation is that several people come to a single spot and stay together for a significant period of time. Similarly, group dispersion is the event of people in a crowd departing from each other within a short period of time. The detection of these two types of events requires a method to detect and track groups of people and describe groups dy-

namics (merging and splitting) over time. We achieve this motion group concept by tracking and clustering [9, 2] individuals that move in a spatially and temporally consistent way, where consistency is measured by a pre-defined distance metric. In crowded scenes, people interact with each other frequently, leading to motion groups that continually change their membership. Our system captures these transitions by building and expanding group interaction models over time. These models are characterized by a directed graph structure, where nodes represent individual groups, and directed links indicate group interactions (merging and splitting). Given this model, a group dispersion event is detected by first identifying a node with a large number of individual persons, and conducting a graph search to count how many descendant nodes this group has within a given period of time. An event is raised when the number of descendants is larger than a specified threshold. A group formation event was detected in a similar way.

Click and Track: If an operator of the system determines that a given individual may warrant further analysis, the corresponding subject can be manually selected by the operator. This flags the individual as a target of interest and as such is given additional consideration by the tracking system.

Line and Zone Crossings: The system also supported standard video analytic features such as zone and line crossings, person counting and other region based events.

Left Behind Object: As previously described, our system is designed detect a set of foreground object types with specific geometric proportions that are representative of people and objects (i.e., left behind objects). All objects that are detected sufficiently close to a previously tracked person are labeled as object targets. If such object targets become stationary and the associated person leaves the object for more than a specified amount of time, a left-behind object event is raised. Details regarding this approach to left luggage detection can be found in [6].

2.4. Storage and Retrieval

In addition to performing target tracking and event detection our system was also responsible for display, video storage and retrieval. Operators could observe site activity in a single operational view (see Fig. 2) obtained by combining image content from all available cameras into a single birds-eye-view.

The system also performed video compression and storage on all views. Meta-data associated with tracking and event information was simultaneously indexed and stored in a SQL database for later retrieval. The additional task of video and database storage as well as display placed a significant burden on computational resources but the system still managed to achieve frame rates of around 15 fps due to the use of multi-core CPUs and a careful multi-threaded system design. Event retrieval is enabled via a simple but effective user interface that supported a range of queries such as line crossing, dwell time and zone intrusion events.

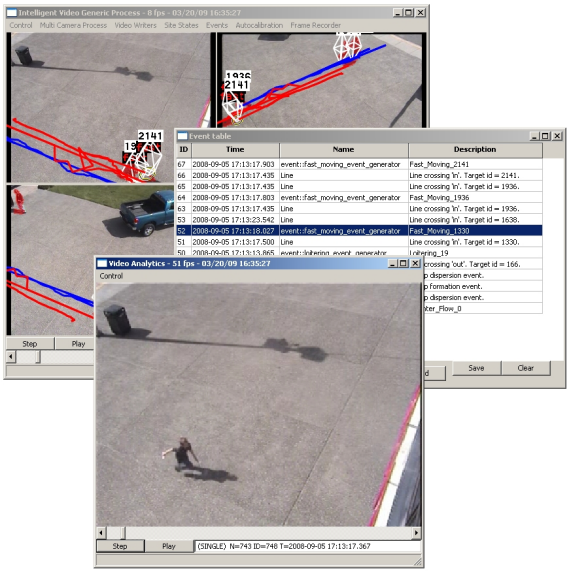

Figure 1. **Event Playback**: The system provided the ability to retrieve and plaback previously stored events.

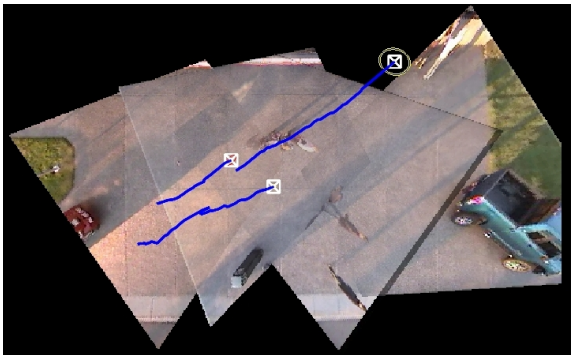

Figure 2. **One World View**: The operator could interact with the system in a virtual top-down view of the site, obtained by composing ground-plane image data from all available cameras into a single view.

3. Setup and Installation

Camera placement was constrained by the availability of lamp-posts and building structures. It was decided to place three cameras on the facade of the building and one camera on a lamp-post in front of the building (see Figure 3). The site was calibrated using methods similar to [5, 4] with an additional processing step to obtain highly accurate site-wide calibration via bundle adjustment.

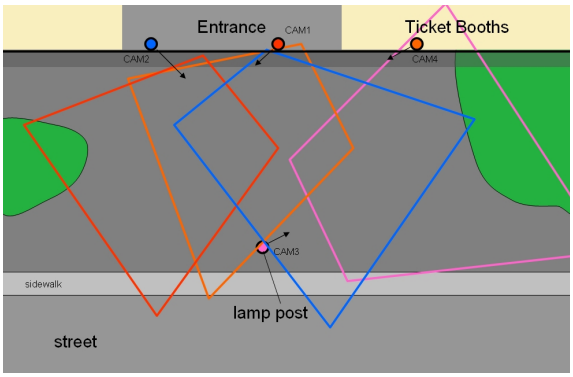

Figure 3. **Camera Placement**: Four cameras were used to secure the sports arena entrance. Three cameras were installed on the roof of the stadium, and one camera on a lamp post near the stadium.

One issue that arose during site-calibration was that the ground surface of the observed area had a slant of a few degrees. It was decided to calibrate with respect to a coordinate system where the z-axis was aligned perpendicular to the ground plane, as opposed to the direction of gravity, which is typically assumed by autocalibration systems that rely on people as calibration objects.

4. Results and Discussion

During each sporting or entertainment event, the test site had to be able to manage up to 4000 people per hour. In general our system was able to provide reasonable tracking results and was observed to degrade gracefully when exposed to a number of real world challenges. Figures 4 are examples of the performance of the installed tracking system. In the rest of this section we discuss how the system responded to a number of significant operational challenges.

Crowds: The system was able to successfully track individuals in small groups, such as shown in figure 5 (left), however as larger groups of people were encountered, the foreground patches tended to be under-segmented (see figure 5 (right). Since a detect and track approach was used, errors in detection can sometimes result in track switching. Crowd segmentation approaches such as [11] may be used to address this issue. In addition we believe that signature-based tracking approaches will be able to increase track fidelity in the future.

Shadows: When moderate shadows are encountered, the system was able to distinguish between shadows and true foreground objects by considering the covariance of the image intensity and the intensity predicted by the background model. An example of such shadows is shown in Figure 6 (left). When the shadows became too strong or there was insufficient background texture, it became difficult to distinguish between foreground objects and shadows. If the shape of the shadow is inconsistent with the geometric model of

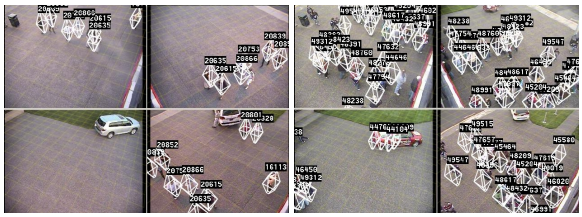

Figure 5. Left: Even larger groups were often handled gracefully by the system. Right: Example of under segmentation for large crowds

a person, false detections can still be avoided (see Figure 6, middle). However if a wrongly classified shadow was consistent with the geometric model used for person and object detection, then false detections can be observed (see Figure 6, right). This problem could be addressed by requiring that all available image views must be consistent with the observation of a person (as it the case in Figure 6, top right view). However, this approach could lead to increased missed detections.

Low Lighting: For many of the events hosted at this sporting venue, patrons would start to arrive in the early evening and continue to arrive for some time after sun-set. After sun-set, the only major source of lighting was the parking lot lights. This resulted in only partial illumination of the site. For regions that received sufficient lighting, tracking was maintained. However near the periphery of the site, where lighting was very limited, the ability to accurately detect individuals was diminished - see figure 7. To address this issue, thermal or IR cameras could be considered or additional light sources could be deployed.

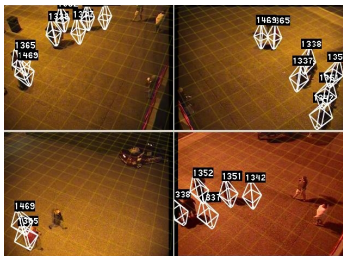

Figure 7. Example of system performance under low-lighting conditions.

Queues At certain points in time, large queues of people would form in front of the facility entrance. These queues produced a large number of detections. However these detections were often unstable and as such the ability to accurately track specific individuals in the queue became problematic. This problem was amplified by the very long dwell times of individuals within the queue, since this sometimes caused targets to merge into the background. Figure 8 illustrates this issue. We believe that the ability to track people

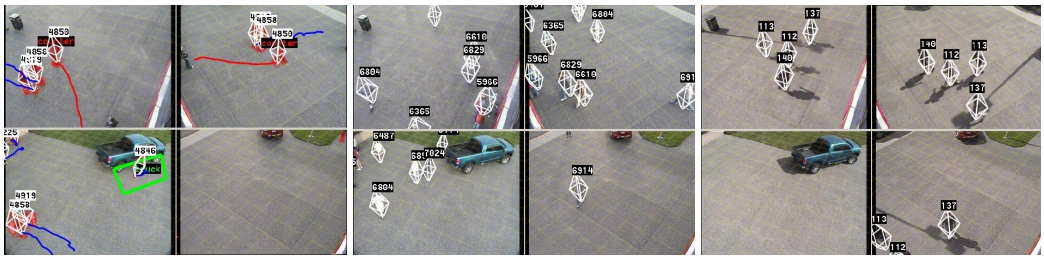

Figure 4. Example of typical tracking performance. One can see from the IDs of the tracks, that the system performs site-wide tracking across the camera views. The system was able to easily disambiguate small groups of individuals approaching the sports venue. In the left image a counter-flow event can be seen.

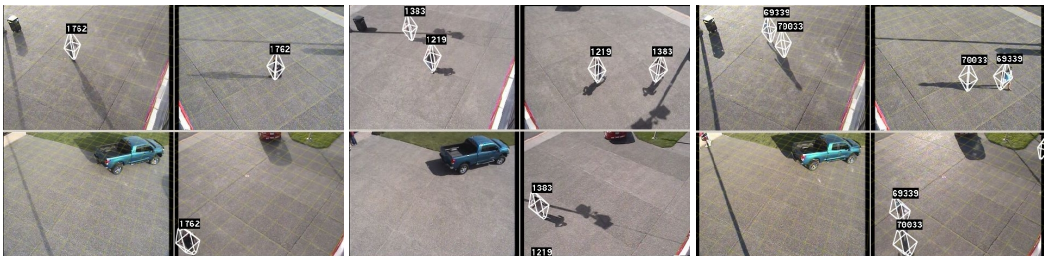

Figure 6. Left: Example of accurate discrimination between shadows and foreground objects. Middle: Example of inaccurate discrimination between shadows and foreground objects, however false detections are suppressed due to geometric person models. Right: Example of inaccurate discrimination between shadows and foreground objects. In this case false detections are encountered since the shadows are consistent with a person model.

in dense queues remains an open research problem.

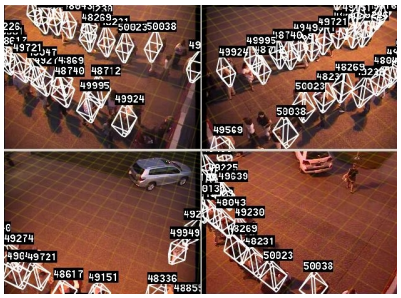

Figure 8. Example of tracking performance with dense queues

Children: In this application, a single geometric person model was used. For this reason, children and people in wheel chairs were sometimes not detected (see figure 9). This issue could be addressed by instantiating additional object models at the potential cost of additional false positive detections and class confusion.

Event Detection: Most of the trajectory based events such as loitering, fast moving person, path crossing and group formation, were accurately detected if track fidelity was maintained (see figures 10 and 10). However, due to the complexity of the site, the parameters of the left-behind object event detection system had to be set conservatively in order to avoid false left-behind object events.

Figure 9. Example of a missed detection associated with children.

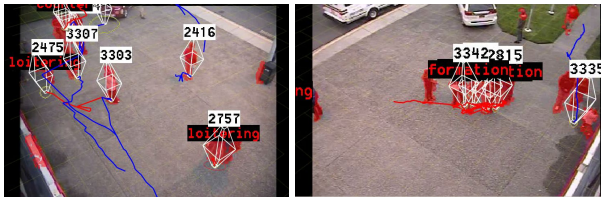

Figure 10. Example of detected events.

Camera Malfunction: During the course of the experiments, cables connecting one of the cameras was inadvertently severed, rendering it inoperable. Since the installers and developers were no-longer on site, the system was operated with only three of the four cameras supplying real

image data. The system was able to remain operable and continued to provide site-wide tracking services for the remaining regions that had camera coverage (see Figure 11). It should be noted that the tracking system was unaware of the fact that one of the cameras was no-longer operational.

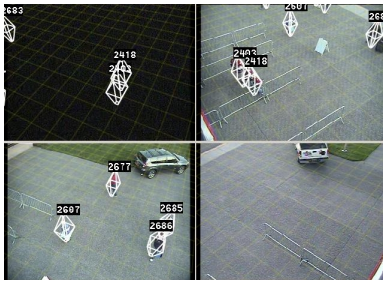

Figure 11. Example of tracking after a camera malfunction.

5. Conclusions

Automatic surveillance in crowded urban environments is a challenging problem. However, as we have shown in this work, modern computer vision systems can actually perform fairly well with respect to the detection and tracking of people, even under a variety of challenging conditions. However, more research is still required and issues of robustness need to be addressed. Such issues include varying environment conditions, shadows, night-time operation, severe crowds and the need for long-duration tracking. It should be mentioned that in practice, the drive for robustness often has to be balanced with time and budgetary constraints. Even though there are a large number of academic efforts currently focused on many of these issues, the commercial integration and development of such capabilities is limited by cost and schedule constraints.

In general, we believe that the systems developed here must be augmented with more advanced event and behavior recognition capabilities. Standard video analytics features such as person counting, intrusion events and left-behind objects have an easy to understand value proposition. However, when it comes to the problem of detecting suspicious activities that are indicative of a possible terrorist threat, much more research needs to be performed. It is extremely difficult to formulate the manner in which such events and behaviors manifest themselves, let alone how to develop approaches needed for their modeling and recognition. The burden is on the computer vision community to show that our systems will be able to provide real and tangible results.

6. Acknowledgement

We would like to thank the Science and Technology division of the Department of Homeland Protection for sponsoring this work under the Standoff Technology Integration Demonstration Program (STIDP). As site administrators, the team from Pacific North-West National Laboratories (PNNL) performed all site preparation, installation as well as staffing for the operation and evaluation of our system. We would also like to thank the team from LM TSS for assisting in pre-installation testing of the system.

References

[1] M. T. Chan, A. Hoogs, Z. Sun, J. Schmiederer, R. Bhotika, and G. Doretto. Event recognition with fragmented object tracks. In *In proc ICPR*, 2006.

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001.

[3] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. In *Proc. of IEEE*, volume 90, pages 1151–1163, July 2002.

[4] N. Krahnstoever and P. Mendona. Bayesian autocalibration for surveillance. In *Proc. of IEEE International Conference on Computer Vision (ICCV'05), Beijing, China*, October 2005.

[5] N. Krahnstoever and P. Mendona. Autocalibration from tracks of walking people. In *Proc. British Machine Vision Conference (BMVC), Edinburgh, UK, 4-7 September*, 2006.

[6] N. Krahnstoever, P. Tu, T. Sebastian, A. Perera, and R. Collins. Multi-view detection and tracking of travelers and luggage in mass transit environments. In *Proc. Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2006.

[7] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. CVPR, San Diego, CA*, 2005.

[8] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 2(50):203–226, 2002.

[9] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[10] P. Tu, G. Doretto, N. Krahnstoever, A. A. G. Perera, F. W. Wheeler, X. Liu, J. Rittscher, T. B. Sebastian, T. Yu, and K. G. Harding. An intelligent video framework for homeland protection. Proceedings of SPIE Defence and Security Symposium - Unattended Ground, Sea, and Air Sensor Technologies and Applications IX, April 9–13, 2007. (Invited paper).

[11] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu. Unified crowd segmentation. In *Proceedings of European Conference on Computer Vision*, 2008.

[12] P. Tu, F. Wheeler, N. Krahnstoever, T. Sebastian, J. Rittscher, X. Liu, A. Perera, and G. Doretto. Surveillance video analytics for large camera networks. In *SPIE Newsroom*, 2007.

[13] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208– 1221, September 2004.