Industrial Surveillance Day in conjunction with
8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2011

AVSS 2011 demo session: A Systems Level Approach to Perimeter Protection

Peter Tu, Ting Yu, Dashan Gao
General Electric

Ram Nevatia, Sung Chun Lee
University of Southern California

Hale Kim, Phill Kyu Rhee
Inha University

Joong-Hwan Baek
Korean Aeronautics University

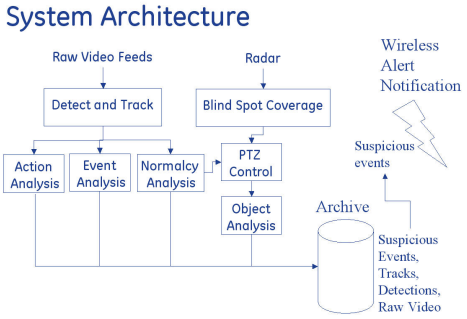

Figure 1. System Diagram

1. Introduction

The primary technical goal of this work is the development of a comprehensive approach to perimeter protection for critical infrastructure and industrial sites. The approach taken is to combine both video and non-video sensors so as to produce a real-time system capable of tracking objects of interest and of detecting potential events that may warrant the attention of security officials. A summary of the system architecture is shown in Figure 1. System modules include:

- The ability to detect and track people using both visual and radar based sensors.
- The detection of articulated motions as well as complex and abnormal events.
- The application of object recognition to detected left behind objects.

.

2. Multi View Tracking

A multi-camera tracking system was deployed for the purposes of maintaining space-time trajectories of observed individuals. During system installation, an initial one-time geometric calibration of all cameras is performed. The main person detection algorithm continuously estimates a background model. The foreground/background modeling is performed using non-parametric kernel density estimation in gray-scale space. The foreground map is refined using a series of post-processing steps so as to reduce false-alarms resulting from non-stationary objects, shadows and lighting changes. Once the foreground map has been computed, the person detection approach relies on explaining foreground patches using geometric regions that have size and shape similar to people. In this approach, people are modeled as rotationally symmetric upright ellipsoids. In order to address the issue of false detections nominated by the foreground analysis module, the system leverages a set of machine learning person classifiers that do not rely on motion cues, but instead use shape and appearance features to distinguish between people and non-people. All person detections from the background modeling approach are verified using these machine-learning classifiers. For computational efficiency the system separates person detection from tracking. After person detection has been performed, the location and location uncertainty of each detection are projected onto the ground plane. A centralized tracker processes the time-ordered detections and is responsible for assigning detections to tracks. To perform long-duration tracking through occlusions and other periods of track loss, the system makes use of signature-based track linking.

3. Blind Spot Coverage

For various reasons, it is often not possible to achieve site coverage with video cameras. However, blind spots can be monitored using inexpensive Radar systems such as the RCR50. In general such devices can only report the presence of a moving object. However the range of these sensors can be changed dynamically. Thus by sweeping the sensor range and recording the first instance at which a detection is observed, one can determine the distance of a moving person from the radar sensor. Given multiple radar sensors, one can then compute the position of the target by intersecting the distance arcs associated with each detec-

978-1-4577-0845-9/11/$26.00

tion. Thus radar-based tracking methods can be used to augment tracks produced by the vision-based systems. In addition, radar detections can be used to drive PTZ cameras resulting in the capture of high resolution imagery that can be presented to security staff. With the current implementation, it takes 2 seconds to perform a complete range sweep.

4. Articulated Action Analysis

Given the trajectory of a tracked individual, the purpose of this module is to determine whether or not a given articulated action such as digging or climbing is currently being observed. In our approach, a sparse point representation of a human action is provided using a set of spatiotemporal interest points detected from a spatiotemporal volume that is extracted from the tracked individual. we choose to extract both motion and shape features at each detected interest point. In particular, for motion features, we depend on 3-D spatial-temporal Gabor filters, which include motion-sensitive Gabor filters of different spatial and temporal frequencies and orientations. At each interest point, filter responses at different spatial and temporal frequencies are computed. In addition, their variances are computed within rectangular regions defined by a shape-context like descriptor. A space time cube is attached to each tracked individual. These cubes are continuously classified as being either an instance of the action of interest or not. Boosting is used to construct a strong classifier that is responsible for this discrimination task. The set of weak classifiers used to construct the strong classifiers are parameterized by defining a sub-cube in the person centric space time cube. All interest points found in the sub-cube are used to construct an average description vector and the classification decision is made using Fisher's linear discriminant. During the deployment of this system, four actions of interest were defined: climbing, digging, throwing and the placing of an object.

5. Complex Event Analysis

Based solely on track information generated by the main system, various events of interest can be inferred. For this application, events of interest include: illegal entry, line formation, person collision and left object detection. One of the major requirements for this application is the capacity to process a continuous stream of track data without causing system stoppage. To support this requirement, we divided our system into two modules that perform both real time and semi-real time tasks. The real time task consisted of a trajectory based event detection method that can process in real time. Having nominated a potential event, specialized person detectors sensative to a variety of poses our used to scour the associated video at which point a decision regarding whether or not the event of interest has occured can be made.

6. Normalcy Analysis

Given track information generated by the tracking mechanisms, the normalcy modules must determine whether or not such tracks constitute normal or abnormal activity. Each trajectory is analyzed and the results are stored into a hierarchical historical ontology. Based on observed historical frequencies, all elements of the ontology can be viewed as either normal or abnormal. New observed trajectories are mapped to the ontology. Based on this mapping a normal or abnormal classification can then be made. In addition, new trajectories are also used to update the structure of the ontology. Clustering methods are used to achieve this update process. In this way new activities that are frequently observed will form new nodes in the ontology that may eventually be designated as normal activity. Intrinsic to this process is the ability to compare trajectories. This is achieved using normalized measures of location, speed and direction. During testing, it was found that the system was able to distinguish between normal and abnormal variants of forward motion, u-turns and wandering.

7. Object Recognition

Given the hypothesis that a left behind object event has occurred, PTZ cameras are tasked with capturing high resolution imagery of the left behind object. The next task, denoted by 'object recognition', is to determine the presence of any instance of a given set of specific object classes. The major steps in our approach are: (i) Image signature generation: We explore various descriptors such as Scale Invariant Feature Transforms (SIFT), Speeded Up Robust Features (SURF) and Color Histograms. Individually SIFT, a texture based descriptor, outperforms other descriptors. However, since color is also an important cue for specific objects, the SIFT descriptor is augmented with a color histogram extracted from a patch centered at the SIFT interest point. The Bag of Words (BoW) model is then used to generate a global signature for the image based on the set of locally extracted descriptors. The BoW model uses k-means clustering to generate a visual codebook. (ii) Object model generation: for each object class, we create a model. This is achieved by summing all of the image descriptors for each training image in each class and normalizing them so as to generate a representative object model. (iii) Classification: Given an image descriptor, the posterior probability of each object class is calculated using the Naive Bayes assumptions. The Maximum A Posterior (MAP) among all classes determines the image/object class. During the learning stage we generate object models and a codebook of visual words. When online we use the codebook to generate image signatures and objects models for classification.