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

Abstract

Effective perimeter protection mechanisms for industrial sites and critical infrastructure must contend with a large variety of potential threats as well as with the fact that normal site activity can be both complex and diverse. This paper documents the development of a system level approach capable of functioning under such challenging conditions. A multi-view tracking system is used to provide real-time site wide trajectories of all observed individuals. A Radar-based system is also used for tracking if and when camera coverage of various regions is not available. Track information is then analyzed with respect to articulated motion analysis, complex event analysis and normalcy analysis. In addition, object recognition is used to classify left behind objects using high resolution PTZ imagery. A real-time integrated version of this comprehensive approach to perimeter protection was deployed using a single standard off-the-shelf desktop computer.

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 1. The ”Detect and Track” module initially consumes imagery collected by a network of video cameras. This module is responsible for the tracking of all visible people in a real time fashion. Tracks and the associated regions of interest are then passed to the action analysis, event analysis and normalcy analysis modules. The action analysis module is responsible for detecting suspicious articulated motions such as placing, throwing, and climbing. Based on the real-time track data, the event analysis module must initially detect the possibility of complex events such as a person leaving behind a piece of luggage, the collision of individuals or a person intruding on to the site. It may then request a short-term video volume associated with these events from the main system archives and attempt to verify these observations by employing a set of computationally intensive algorithms in a semi-real-time fashion. The normalcy analysis module must first determine a probabilistic model for activity routinely observed at the site. It must then identify events that cannot be explained by these models, which are then flagged as abnormal events. Certain types of events such as left-luggage events will trigger a set of pan-tilt-zoom (PTZ) commands that cause the PTZ cameras to capture high resolution imagery of the objects of interest. In the case of a left-luggage event, object recognition algorithms are applied so as to determine whether or not the targeted object is an instance of a set of specific object instances such as a known bicycle or ladder. Since comprehensive camera coverage may not be feasible, radar based motion detectors are used to augment the system’s person detection capabilities. Such radar-based detections are also used to invoke the capture of high resolution PTZ camera imagery. All captured imagery is compressed and stored on to disk. All track information is continuously inserted into an SQL database. All events are inserted into an XML document. In addition, events of interest result in the wireless alert notification of security personnel. A graphical user interface (GUI) allows the user to continuously observe captured imagery augmented with track overlay information. The user can view a GUI log table of detected events and playback the video clips associated with these events via GUI interactions.

The remainder of this paper provides details regarding the following system modules: Multi-View Tracking, Blind Spot Coverage, Articulated Motion Analysis, Complex Event Analysis, Normalcy Analysis and Object Recognition. The paper concludes with a system level perspectives associated with this ambitious deployment.

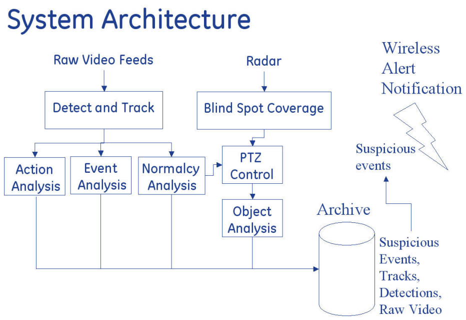


Figure 1. System Diagram

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 [3]. 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 light changes. Once the foreground map has been computed, the person detection approach relies on explaining foreground patches using geometric shapes that have size and shape similar to people [9]. 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 [13]. 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 [14]. To perform long-duration tracking through occlusions and other periods of track loss, the system makes use of signature-based track linking. To perform track linking, all tracks that terminate are kept in abeyance for a pre-specified time and are compared with new tracks to determine whether or not they should be linked based on spatial, temporal and appearance based considerations. For the purposes of signature matching, each camera produces a signature for each tracked individual. These signatures are composed of a representation that encodes the spatial distribution of color. The system employs a multi-cue signature maintenance framework, which allows for multiple visual features to be used to define the appearance signature of each target. All detections received from multiple cameras are assigned to tracks by considering both their location and the similarity of their appearance signatures to those maintained by the trackers. Under this mode of operation, camera-to-camera handoff between overlapping views is automatically achieved via our ground plane tracking methods. Figure 2 shows an example of the tracking performance achieved by this system. As a measure of the tracking robustness with respect to changing environmental conditions, a number of data sequences were collected and the NIST statistics [8] were calculated using manual groundtruthing as the reference (See Table 1).

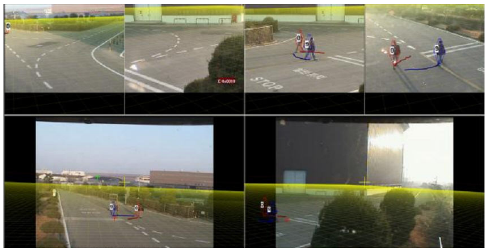


Figure 2. Tracking

	Cloudy Morning	Cloudy Noon	Bright Noon	Rainy Afternoon
SFDA	0.69	0.64	0.57	0.65
ATA	0.22	0.34	0.23	0.23
MODA	0.79	0.68	0.54	0.52
MODP	0.68	0.61	0.59	0.51
MOTA	0.78	0.67	0.46	0.42
MOTP	0.67	0.61	0.59	0.50

Table 1. NIST Tracking Metrics.

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 (see figure 3). 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 detection. 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

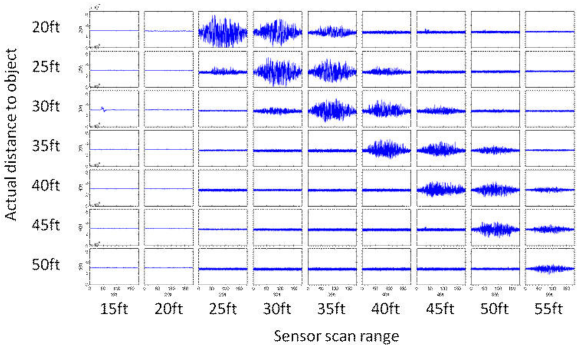


Figure 3. Radar scan responces.

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. It has been shown in various human psychophysical studies that human subjects can easily recognize human actions represented by sets of lights attached to the joints of a human body [1]. Thus a sparse representation of the image sequence using locally informative features may be sufficient for recognition of the underlying action.

In our approach, the 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. Various types of spatiotemporal interesting points have been proposed [10, 4]. In this work, we have adopted a method based on spatial 2-D Gaussian smoothing and temporal Gabor filtering [2], which is briefly reviewed in the following.

This detector assumes a stationary camera and applies separable spatiotemporal filters to the motion sequence to obtain the response function as follows:

$$F(x,y,t) = (I * g * f_{even})^2 + (I * g * f_{od})^2, \quad (1)$$

where $g(x,y;\sigma)$ is a spatial 2-D Gaussian kernel which smoothes each frame along only the (x,y) spatial dimensions, $f_{even}(t;\tau,\omega)$ and $f_{odd}(t;\tau,\omega)$ are real and imaginary components of a 1-D Gabor filter applied in the time dimension, t. From these filter responses, local maxima are found and used as interest points. While the detector has been designed to produce maximum response to repetitive motions, it also responds vigorously to many forms of complex motion.

Following neuroscience evidence that both motion and form pathways in the human brain contribute to the recognition of biological motion [5], 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.

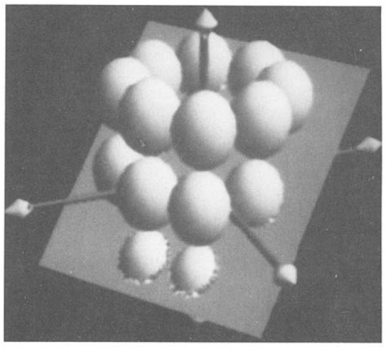


Figure 4. An illustration of the power spectra of 12 spatiotemporal Gabor filters at four orientations and 3 temporal scales. The figure is reproduced from [6]

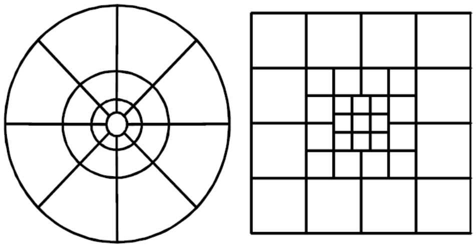
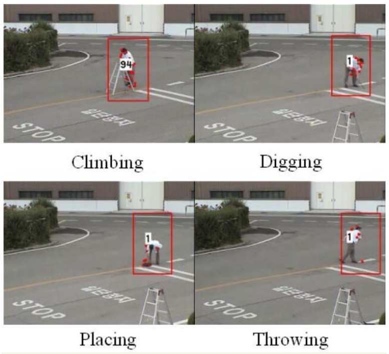


Figure 5. An illustration of the shape context descriptor pattern. (a) Each image region is partitioned into log-polar sections. (b) For efficient computation, the circular pattern in (a) is approximated by rectangular partitions.

In Figure 4 we illustrate the power spectra of 12 motion-sensitive Gabor filters (4 orientations and 3 temporal scales) in the spatiotemporal-frequency domain (reproduced from [6]). Each of these Gabor filters is tuned to a specific motion direction and speed of a certain image pattern. For example, one of the illustrated filters is most sensitive to the rightward motion of vertically oriented patterns, while another is most sensitive to leftward motion. In practice, a motion-sensitive Gabor filter can be implemented by a linear combination of four separable spatio-temporal filters so as to achieve efficient computation [6].

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 [11] (Figure 5).

To capture human pose changes, we compute histograms of local orientations of image intensity using the shape-



Climbing Digging

Placing Throwing

Figure 6. The four example actions.

	climbing	digging	placing	throwing
climbing	0.9	1.0	1.0	1.0
digging	1.0	0.89	1.0	1.0
placing	1.0	0.94	0.89	1.0
throwing	1.0	0.81	0.97	0.9

Table 2. The accuracy matrix calcualated using the action sequences (columns) vs. the action models (rows).

context like descriptor pattern in Figure 5. Centered at each interest point, the local image region is first partitioned into log-polar sectors, and then the histogram for orientation in each sector is computed. Since location information is important for distinguishing between different body parts (e.g. arms and legs), the image coordinates of the interest points are also selected as features. To achieve efficient computation, as suggested in [11], the circular partition can be approximated by rectangular partitions (Figure 5 (b)) so that integral image techniques can be applied.

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 [12] 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 (see Figure 6). For testing purposes, actors performed at least 10 iterations of each task. The associated accuracy matrix for this experiment is found in Table 2.

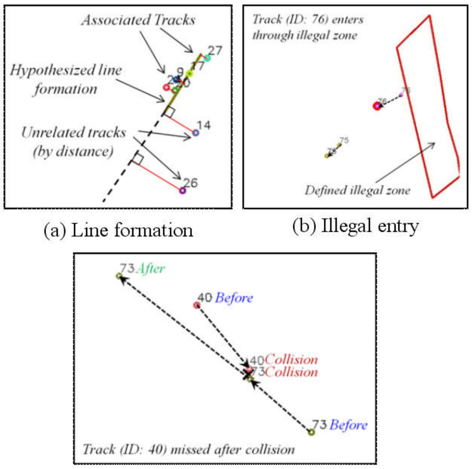
5. Complex Event Analysis

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. Some examples of our trajectory based event detection are shown in Figure 7. Unlike the events depicted in the figure, Left Luggage Event Module (LLEM) requires more complicated communication between system modules. To begin with, the main system triggers an initial left luggage event detection based on blob detection and the LLEM must then verify that the left object is a non-human object. For non-human object verification, the LLEM requests and receives VOD (Video On Demand) data from the main server. The LLEM then applies a 'human detector' to the received video to determine whether or not the left object is human. If the initial event is confirmed, the LLEM continues to monitor the owner's track (the closest track to the left object). When the owner leaves the left luggage, the LLEM verifies that the left object is still remains. The LLEM then requests and receives additional VOD data. A final check is performed by performing template matching between the initial and final left-luggage imagery.

Since ghosts or missed tracks may trigger the initial event detection. The system needs to verify the initial detections by once again performing intensive human detection methods. Confirmation of these events requires a intense analysis of the imagery associated with these events. At the heart of such an analysis is the ability to detect people in a variety of poses (standing, crouching and lying down). To this end methods similar to those proposed in [7] were used as a basis for this form of analysis. Following this approach a cluster boosted tree object detector, which is capable of automatically partitioning the positive sample space whenever the complexity of the current classifier becomes too high was constructed.

We collected a dataset of humans in varying poses from multiple cameras (multiple viewpoints). The data was divided into two sets, the training set and the test set. For training the detector, we collected annotations automatically generated via background subtraction. Adaptive background subtraction was run on the training videos and the bounding box of the foreground patches were extracted to obtain the training samples. The samples were normalized to the same orientation. A number of the training samples obtained in this fashion are shown in Figure 8. For poses such as lying down, a 90 percent detection rate was achieved with an associated 18% error rate. Figure 9 shows a number of events that were detected and subsequently validated


Associated Tracks
Hypothesized line formation
Unrelated tracks (by distance)
27
7
9
14
26
Track (ID: 76) enters through illegal zone
76
75
Defined illegal zone
73 After
40 Before
40Collision
73Collision
73 Before
Track (ID: 40) missed after collision

(a) Line formation

(b) Illegal entry

(c) People collision

Figure 7. Examples of methods used to perform real time trajectory based event detection.

Figure 8. Examples of training data used to construct a generalized person detector

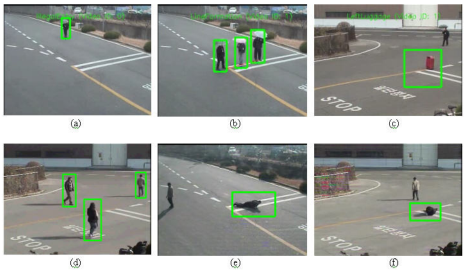


Figure 9. Detected Events

using these detectors.

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 (Figure 10). 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.

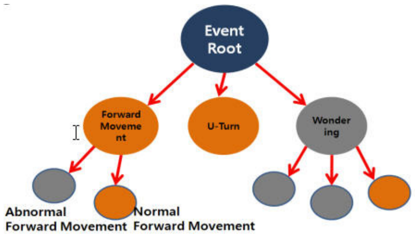


Figure 10. Ontology

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 if 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 Transform (SIFT), Speeded Up Robust Feature (SURF) and Color Histogram. 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 models 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.

Our object recognition module consists of object detection and classification. We apply spatial grouping on the interest points in the image plane. Each candidate group is then classified as a specific object or as background us-

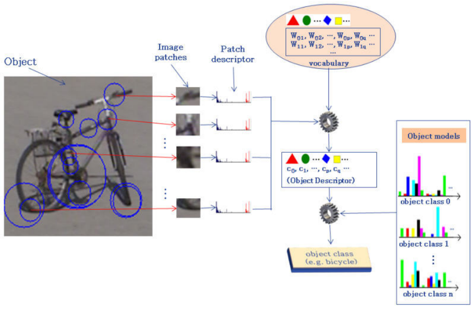


Figure 11. An illustration of object classification stage.

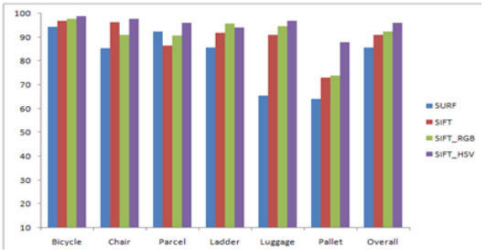


Figure 12. Object classification rates obtained by various descriptors.

ing the classification approach described above. The number of interest points in the group are counted for the obtained class. Majority voting on the number of interest points among the classes results in the object class designation and a minimum rectangle holding the groups belonging to the major class determines the object location. Figure 11 illustrates the object classification stage. Figure Figure 12 reports on the observed recognition rates for a variety of objects and local feature descriptors.

8. Discussion

A complex industrial site was selected for the development and testing purposes of the system illustrated in Figure 1. Four fixed cameras as well as two PTZ cameras were deployed (see Figure 2). In addition, three radar detectors were used for blind spot coverage. Using a single dual quad core PC, our system is able to operate the following modules simultaneously,

- perform site wide tracking across all camera views.
- perform articulated motion analysis and recognition on each tracked individual for actions including: climbing, digging, throwing and placing.
- perform complex event analysis for: abnormal person detection, line formation, collisions and left luggage detection.
- perform normalcy analysis on all observed trajectories.
- Automatic targeting of PTZ cameras onto left behind luggage followed by automatic classification of these objects.
- storage of all video to disk.
- storage of all track data to a SQL database.
- the logging of all events into a XML document.
- the electronic transmission of alerts to security personnel.

To our knowledge this exercise represents one of the most comprehensive approaches ever taken with respect to automatic perimeter protection.

References

[1] R. Blake and M. Shiffrar. Perception of human motion. *Annual Review of Psychology*, 58(1):47–73, 2007. 259

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 0:65–72, 2005. 259

[3] A. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction. In *6th European Conference on Computer Vision, Dublin, Ireland*, 2000. 258

[4] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7):1–18, 6 2008. 259

[5] M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci*, 4(3):179–192, March 2003. 259

[6] D. Heeger. Optical flow from spatiotemporal filters. *International Journal of Computer Vision*, 1(4):279–302, 1988. 259

[7] C. Huang and R. Nevaita. High performance object detection by collaborative learning of joint ranking of granule features. In *CVPR2010*, 2010. 260

[8] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 31:319–336, 2009. 258

[9] 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 *In Proc. Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2006. 258

[10] I. Laptev. On space-time interest points. *IJCV*, 64(2/3):107–123, 2005. 259

[11] H. Ning, W. Xu, Y. Gong, and T. Huang. Discriminative learning of visual words for 3D human pose estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 259, 260

[12] R. E. Schapire. The boosting approach to machine learning: An overview. *In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, editors, Nonlinear Estimation and Classification. Springer*, 2003. 260

[13] P. H. Tu, N. Krahnstoever, and J. Rittscher. View adaptive detection and distributed site wide tracking. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2007. 258

[14] T. Yu, Y. Wu, N. Krahnstoever, and P. Tu. Distributed data association and filtering for multiple target tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska*, June 2008. 258