Video Surveillance at Night

Mark R. Stevens, Joshua B. Pollak, Scott Ralph, and Magnus S. Snorrason
Charles River Analytics, Cambridge, MA

ABSTRACT

The interpretation of video imagery is the quintessential goal of computer vision. The ability to group moving pixels into regions and then associate those regions with semantic labels has long been studied by the vision community. In urban nighttime scenarios, the difficulty of this task is simultaneously alleviated and compounded. At night there is typically less movement in the scene, which makes the detection of relevant motion easier. However, the poor quality of the imagery makes it more difficult to interpret actions from these motions. In this paper, we present a system capable of detecting moving objects in outdoor nighttime video. We focus on visible-and-near-infrared (VNIR) cameras, since they offer low cost and very high resolution compared to alternatives such as thermal infrared. We present empirical results demonstrating system performance on a parking lot surveillance scenario. We also compare our results to a thermal infrared sensor viewing the same scene.

1. INTRODUCTION

In this paper, we focus on the problem of automated parking lot surveillance at night using visible-and-near-infrared cameras (VNIR). These sensors are low cost (100s of US dollars) when compared with thermal infrared sensors (1000s of dollars) and have a much higher resolution. Thermal imaging may seem an obvious choice for nighttime surveillance since it enables detection of people and warm cars without any ambient light. However, most parking lots are never completely without ambient light, with typical illumination levels between 5 and 50 lux even on moon-less nights. In fact, parking lots next to buildings with strict security tend to be well lit as a deterrent to illegal activity. Intuitively, cameras optimized for low-light are therefore just as appropriate as thermal cameras for this scenario. Long-wave infrared (LWIR) 160x120 cameras are the least expensive thermal solution, but they still cost anywhere from $7,000 to $15,000. Finally, the ability of thermal cameras to detect cars generally depends on the engine being warm, hence a cold car starting up and driving away can be difficult to see.

We call our research system VANESSA, for Video Analysis for Nighttime Surveillance and Situational Awareness. VANESSA is capable of: 1) enhancing a video stream, 2) determining moving objects, and 3) rejecting false motion due to vehicle headlights. This paper focuses on identifying the key difficulties associated with our chosen domain and camera technology. In addition, we present algorithms to correctly detect and segment moving objects while maintaining a low false alarm rate. This is a non-trivial task given the difficult nature of the imagery.

The remainder of this paper is laid out as follows. In Section 2, we present a brief overview of previous work in the field, Section 3 discusses the challenging problems associated with video surveillance in VNIR, and Section 4 discusses our data collection. In Section 5, we present our motion detection algorithm along with mechanisms for suppressing false motion due to headlights. Section 6 presents empirical results on several video sequences along with a simple performance comparison to LWIR data.

2. PREVIOUS WORK

The computer vision literature is rich with valuable contributions to the field of video surveillance and tracking. The previous work discussed in this section is not an attempt at a complete literature review, but rather focuses on the key bodies of work most helpful when designing and implementing the VANESSA system. As such, the review does not focus on specific papers, but rather long-running research programs spanning multiple publications.

The DARPA funded Video Surveillance And Monitoring (VSAM) work at MIT has lead to the development of a sophisticated system for tracking and identifying pedestrians from a collection of sensors spanning a large area. Their approach begins with background modeling over time [1]. The technique they have formalized uses a mixture of Gaussians to represent each pixel in the image. (We have chosen to use this technique for our background modeling work, see Section 4). Once moving objects are segmented, they are labeled as pedestrian or clutter using a trained classifier [2]. The classifier is constructed in an unsupervised fashion and learns to delineate frequently detected parts of moving people (such as torso). Pedestrians are tracked over time using a simple Kalman filter. The tracks formed are then used

to register all sensors viewing the object into a common reference frame [3]. The described system uses color imagery and has not been extended to nighttime use.

The DARPA funded VSAM work at CMU focused on reliably tracking moving objects from a moving platform. In this context, simple background modeling is not appropriate, so the video stream was first stabilized before frame differencing could be applied to detect moving objects [4]. To track objects from frame to frame, infinite impulse response filters (IIRF) were used to build a dynamic appearance model, or image chip, of the 2D object [5]. This technique allows a short time window of memory when building the template, and greatly increases the robustness when tracking multiple vehicles and individuals. In addition, they explored the use of skeletonization for classifying pedestrian motion [6]. This system has not been extended to nighttime use.

The University of Maryland has constructed a series of surveillance systems to track people in video imagery. They use color and grayscale imagery as well as IR video for nighttime operation. Their system begins with a non-parametric background-modeling algorithm [7] to detect moving objects. Once segmented, row and column projections are examined as well as points of high curvature on the object boundary [8]. These features are used to localize pedestrian heads and are tracked over time. Their appearance models are based on histograms, IIRF models, and motion history templates [9]. Color information is also exploited to track objects in groups as well as segment shadows from moving figures.

NIGHTTIME VIDEO SURVEILLANCE

Before developing the VANESSA system, we collected and analyzed VNIR video from our parking lot. We used this data to identify several issues that make VNIR based video surveillance difficult. We then developed our system to deal with as many of these issues as possible. The issues we determined to be relevant are:

- *Pixel noise*: low light levels require high gain to produce an image with acceptable brightness, but the high gain also amplifies the sensor's noise. This causes high variance in pixel values between frames. This can be problematic to any background modeling approach since pixels of the same surface change significantly from frame to frame.
- *Blooming*: vehicles at night use headlights, turn signals, and reverse-gear lights that cause blooming. These effects distort the segmented shapes of objects, confusing our tracker. On the other hand, a heuristic-based image-processing algorithm could take advantage of such features in the image to detect vehicles and even significant motion events (turns, backing up, *etc*.).
- *Glare/Specular reflections*: car headlights cast a wide beam that reflects off any surface the light irradiates. To complicate matters, parking lots contain surfaces designed to reflect light, such as retro-reflective parking markers and signs. Unless compensated for, these effects cause numerous false detections in the background model.
- *Rapid lighting changes*: since vehicles move quickly across the field of view, shadows appear/disappear quickly, headlights and taillights appear/disappear quickly and reflections shift along reflective surfaces quickly.
- *Shadows*: since the scene is artificially lit by many point-light sources, people walking through the scene cause sharp shadows with changing orientation (with respect to the person) over time. This is problematic for motion detection since the shadows often have similar luminance as the person.
- *Low local contrast*: pixel values tend to form two distinct clusters corresponding to well-lit surfaces and poorly lit surfaces. This causes most pixels to appear either very light or very dark. When a bright object moves across a dark background, segmentation is easy. However, more often a dark object appears against dark background, making segmentation very difficult. This contrasts with daytime operation where many different gray levels are present in the image making segmentation more robust.

In addition, our datasets contain many of the same difficulties that plague automated video surveillance algorithms operating on data collected in the daytime. Two of these issues are:

- *Camera motion*: our initial rooftop data collections used a tripod-mounted camera. Wind shook the camera during several collections, causing many false alarms in a background-modeling algorithm designed to detect motion.
- *Occlusions*: both cars and pedestrians tend to move in-between parked vehicles causing large levels of occlusion. This is problematic for tracking since the apparent object size in the image plane changes rapidly at the start and end of each occlusion event.

- One of our key concerns has been the feasibility and robustness of VNIR-based video surveillance at night. VNIR needs justification in particular since IR-based analysis might be thought of as the most obvious sensor choice for this domain. Based on the VNIR and LWIR data we have collected, we have formulated the following list of factors affecting the choice of sensor for this domain:
- *Resolution*: VNIR cameras use CCD or CMOS sensors of various sizes, but even low-end cameras generally have VGA (640x480) or better resolution. The Sony camera we used in our work has a sensor resolution of 720x480, and our initial experience shows that the data has sufficient resolution to identify crude facial features at 45 meters. Low-end LWIR sensors typically have 160x120 sensing elements and up-sample or interpolate to produce data at 320x240. Both cameras operate at 30fps.
- *Cost*: VNIR sensors are typically on the order of a few hundred dollars compared to LWIR priced at many thousand dollars. While the cost of LWIR sensors has been steadily decreasing over the years, a comparable decrease in VNIR sensor prices has also occurred and LWIR is unlikely to catch up in the foreseeable future.
- *Light levels*: LWIR sensors can operate in complete darkness. VNIR sensors require ambient lighting that produces light levels on the order of 1–10 lux. Therefore, VNIR is appropriate for partially lit parking lots, but not environments without any ambient lighting.
- *Day/Night Transition:* we have collected several VNIR sequences at dusk and right after sunset. There is a short time window where the local contrast between people and cars substantially decreases, making tracking of people harder as they move in front of or behind cars. This time window is much shorter than the transition period in LWIR where pavement and asphalt takes several hours to cool down from solar loading, during which vehicles have very similar temperature as the pavement.
- *Segmentation:* pedestrian segmentation in LWIR is trivial. In many cases, a simple threshold can be used to cleanly segment people from cool backgrounds. This does not hold for cars with cold engines, they appear the same temperature as the cold pavement (after solar loading has dissipated). Apart from that, LWIR is not plagued by the segmentation problems discussed above: headlight blooming, glare, and shadows are not sensed.

3. DATA COLLECTION

We have collected 80 video clips from three different sensors covering 55 events on 6 separate days spanning 10 hours. Our goal was to produce a data set that could be used both to determine the difficult operating conditions for VNIR-based video surveillance, as well as to provide data for algorithm development and evaluation. provides a map of the building and the area of the parking lot that can be seen from the camera location currently being used. Our current camera placement is approximately 12.8 meters above the ground (measured using a laser rangefinder). Measurements to other features in the scene were also made to give a better feel for scale, typical distances to moving pedestrians and cars were between 30 and 60 meters.

The first two sensors used in the collection were very similar: Sony digital video (DV) camcorders, models DCR-TRV33 and DCR-TRV17. The CCD sensor in each camcorder has peak sensitivity in the NIR range (around 0.8μm), which requires an internal IR blocking filter for normal daylight usage. The NightShot™ feature removes that filter from the optical path, thereby significantly increasing the total number of photons reaching the CCD. This extends the camera's operational capability into lower lighting conditions, at the expense of color. Sony camcorders also offer a second mode called Super NightShot™, which produces significantly brighter images than the NightShot™ mode, presumably because of longer exposure times (Sony does not disclose such technical details). While the better sensitivity of the Super NightShot™ mode is clearly preferable, tracking can become problematic if the resulting shutter speed is so slow that individual rapidly moving objects have significant motion blur.

The third sensor was a LWIR imager rented from Bodkin Design (www.bodkindesign.com). It is a 160x120 sensor that produces a 320x240 interpolated image stream. The LWIR data was all collected on the same night. The environmental conditions were optimal for LWIR sensing: 1) very cold air temperature (10 degrees Fahrenheit) and 2) data was collected several hours after sunset. This data is used to estimate the performance difference of our tracking algorithms between the two modalities (LWIR and VNIR).

Once the video clips were recorded, the data was converted to AVI files for processing. The VNIR images were cropped to 640x480 and resampled to 320x240 to ensure that the VNIR and LWIR images were at comparable resolutions.

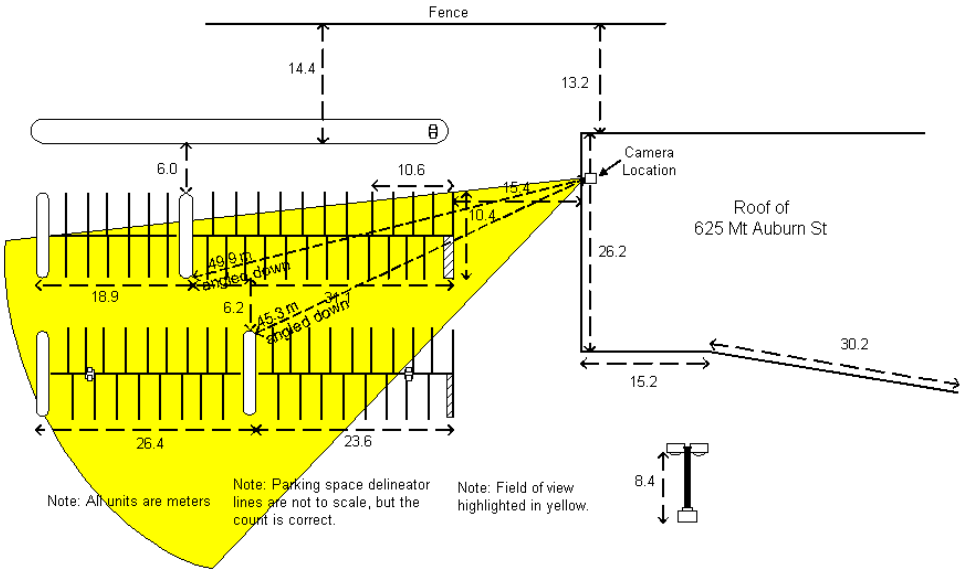

Figure 1. Data collection layout

4. MOVING OBJECT DETECTION

Figure 2 shows the VANESSA system architecture. This system was built to perform automated parking lot surveillance at night, addressing many of the issues presented in Section III. The system takes as input sequential frames from a video sequence. During our preprocessing stage (discussed in Section 5.1) a contrast enhancement operator is first applied to deal with the low local contrast. Also in this stage, both spatial and temporal smoothing is used to reduce the effects of random variance at the pixel level (sensor noise and camera motion).

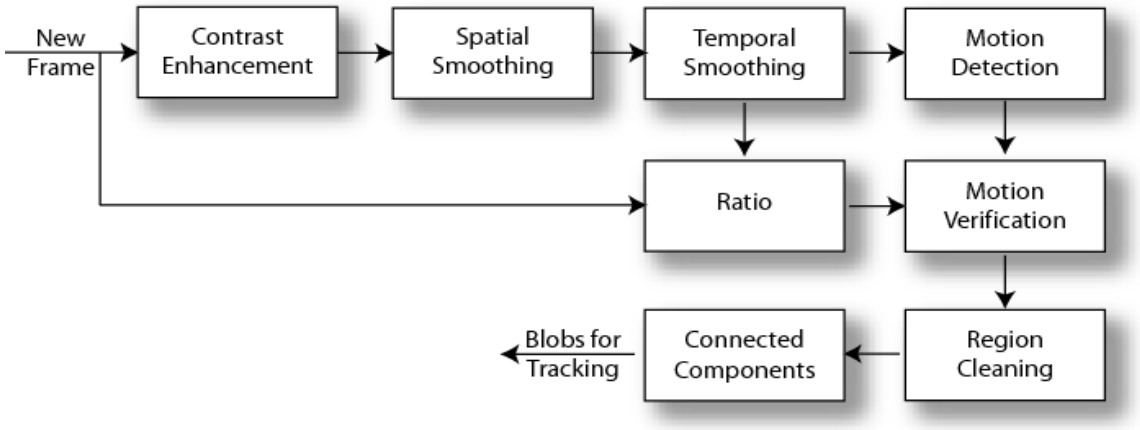

Figure 2. VANESSA System Architecture

A motion detection algorithm is then used to identify objects of interest in the scene. Motion detection is a common approach to video surveillance [4, 5, 11] where some form of background representation, or model, of the scene is computed. The easiest method to implement is background subtraction where the current frame is subtracted from the mean of the last N frames [5]. More complicated non-parametric [8] and parametric models [1] have been explored to build an adaptive representation that is more robust to objects stopping and starting (such as pedestrians) and reoccurring events (such as fluttering foliage).

We have implemented the Gaussian Mixture Model (GMM) technique presented in [1]. This technique is outlined in Section 5.2. For nighttime surveillance, we found this technique performed poorly when applied directly to the (non-preprocessed) video data. We attribute this to the fact that the on-line learning algorithm assumes that the various foreground and background distributions have very little overlap. Since the raw data has very low contrast, learning to separate foreground from background was impossible. However, our preprocessing stage increases the dynamic range as well as reduces the high pixel variance (which also interferes with the GMM technique). Even with these modifications, the technique still produced a large number of false alarms during rapid illumination changes (typically caused by glare from vehicle headlights). To correct for this, we then verify the detected motion by checking for a corresponding change in texture in an image-patch centered on each pixel labeled as moving [10]. If texture change is insignificant then we assume the "motion" was really just a sudden change in local illumination, such as headlight glare. Region culling and connected components are then used to compute blobs for tracking. The remainder of this section provides details for each step of this pipeline.

IMAGE PREPROCESSING

The first step in the process is to compute a histogram for each frame. The cumulative distribution of the pixel data is then computed and a parameter is introduced to clip 2% of the data at the beginning and end of this distribution. A gamma function is then fitted to the clipped cumulative density, followed by the computation of a lookup table that both increases the brightness in the image and produces a higher dynamic range (increased contrast). The parameters of the gamma function are tracked over time so as not to produce flicker.

Once the data is both contrast and brightness enhanced, spatial smoothing is applied. We are currently using a 5x5 Gaussian kernel with a standard deviation of 1.4 and exploit the separable nature of the filter to reduce convolution time. The result is then temporally smoothed using a global infinite impulse response filter. The filter's output is a weighted average of the current frame and a buffer containing an exponential history of all past frames. The single parameter of this filter determines how rapidly the memory of past frames is extinguished from the buffer. Best results were obtained with a value of this parameter = 0.5. The result is then passed to the background-modeling algorithm, which labels pixels as foreground/background.

Each of these preprocessing steps was designed to deal with different environmental conditions specific to nighttime surveillance (see Section 3). The contrast/brightness enhancement compensates for the low local contrast. The spatial and temporal smoothing compensate for high levels of pixel noise. Finally, temporal smoothing reduces the effects of camera jitter. shows an example of an original image (left) and the preprocessed result (right).

GAUSSIAN MIXTURE MODELS

Algorithms designed to detect events in dynamic environments must adapt to ensure an acceptably low false alarm rate without missing important events. For example, a detection system trained to operate during the day requires a different model of acceptable background content than a system operating at night. Hard-coding a static model for a specific set of operating conditions will ensure drastic failures when a fielded system is presented with even slightly different conditions.

Figure 3. Data preprocessing result

We used the GMM technique to learn a representation of an event-free environment. When new data arises that deviates from the model, we assume that the deviation is the result of an event of interest. Furthermore, this model adapts over time to deal with an ever-changing environment. Since multiple models are used, the system also learns the different states of re-occurring events (such as the same pixel sometimes showing a bright leaf and sometimes a dark leaf in wind-ruffled foliage). Simple differencing of current measurements with a running average of previous measurements is not able to cope with such repetitive events.

The method we used was presented in [11]. In their mixture model, each pixel is represented using a collection (Stauffer suggests 3 to 5) of weighted Gaussian distributions. These distributions are parameterized by the mean μ and the covariance ∑. In the following notation, we will use the subscript *i* to denote a specific mixture of a random variable and *t* to denote the current snapshot in time. The shape of the composite distribution is then:

$$P(x_t) = \sum_i w_{i,t} \cdot G(x_t, \mu_{i,t}, \sum_{i,t}) \quad (1)$$

where $G(x_t, \mu_{i,t}, \sum_{i,t})$ is the Gaussian model, *x* is the measurement, *P(x)* is the probability the pixel is part of the model, and w_i is used to weight the distributions in the mixture. When a new measurement *x* is made, it is compared to the current set of mixtures. If a *match* is found, the distribution parameters associated with that match are updated.

To verify that our implementation of the GMM technique was correct, we generated several synthetic datasets of five Gaussian distributions with different means and standard deviations. A biased sampling mechanism was then used to ensure that the learning technique could estimate acceptable weightings for the different distributions. shows a simulation in which, all five synthetic distributions were set so that there was very little overlap. The red lines represent the synthetic data, and the blue represent the learned model. Note that the model has learned the correct mean and variance of each distribution, and has a good approximation of the weighting.

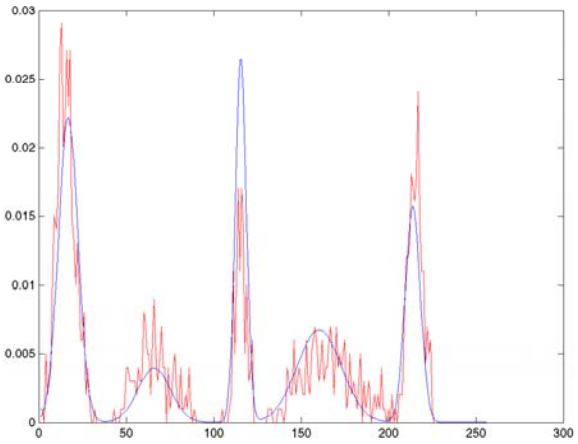

Figure 4. Evaluating the GMM learning technique

After each pixel is labeled moving/stationary, the image is convolving with a 16x16 square uniform kernel and threshold to produce a cleaner version of the motion mask. A connected-components algorithm is then applied to individuate each moving region.

These regions are then tracked over time. shows an example of the motion detection algorithm being applied to two different scenes containing a car and a pedestrian. The left column shows the binary detection labeling produced by the GMM technique, after the 16x16 blurring. The middle column shows the connected component output of the data. The right column shows the input images with tracked objects indicated with a box overlay. Note that the false motion produced from the headlight glare on the pavement easily fools the algorithm.

In the simple examples shown in , the problems associated with nighttime surveillance are painfully evident. The headlights cause a drastic change in the pixel values across a large portion of the ground and object shadows cause inaccurate segmentations. In addition, the headlights cause flashes of highly specular reflection across the parked cars. The

literature is filled with approaches to deal with illumination changes that cause a motion detection algorithm to produce false alarms. The technique we have chosen is based on the notion that false alarms should not produce a corresponding change in texture in the image. Intuitively, an object moving over an image should cause both a change in pixel intensity and a change in the local texture of the image [10].

Our approach first computes a new image representing the ratio of the background model to the new frame. Each pixel marked as moving by the background subtraction algorithm is examined in this ratio image to determine if a change in texture also exists. A change in texture is defined as a large variance in a 5x5 window in the ratio image. shows two images with the same detected object shown in both. In the left image, the pixels labeled as moving by the motion segmentation are shown in red. Note the large number of false alarm pixels. The right image shows only those pixels passing the motion verification step. This verification step has essentially eliminated the false motion due to headlight glare in this example. The red and blue boxes represent the output of a temporal tracker performing data association across time. shows another example of both vehicles and pedestrians being detected.

Figure 5. Motion Verification

6. SYSTEM EVALUATION

To evaluate the system, we manually image-truthed 10 LWIR and 10 VNIR sequences of the same scene. We then compared the true bounding boxes of each moving object to the regions produced by the motion detection algorithm. For the LWIR evaluation, we did not apply the false motion suppression algorithm, as headlights do not show up in thermal IR. We then combined the per frame score into a score across the entire lifetime of each tracked object.

Fig. 6 shows the percentage missing tracks (as a percentage of actual tracks) and the count of extra tracks in each sequence. Each point in the figure represents a single track, and both LWIR and VNIR modalities are shown. In this plot, we can clearly see that motion detection algorithm in LWIR produces fewer false alarms than the motion detection operating on VNIR. The main reason for this is the specular glare from headlights cast onto vehicles that caused false motions not removed by our motion verification. These extra regions can be rejected as false motion by a higher-level tracker as they have very distinct motion characteristics. Also note, that the motion detection missed comparable number of objects in both LWIR and VNIR. This implies that while VNIR has a slightly higher false alarm rate, the probability of detection for both sensors is comparable.

7. CONCLUSIONS

In this paper, we have presented an automated video surveillance system developed to detect moving pedestrians and vehicles at night. We have explored the use of both LWIR and VNIR imagery and presented several techniques to reduce the false alarm rate associated with VNIR artifacts (such as headlight glare on pavement). Our preliminary results demonstrate the significant potential of lower cost VNIR sensors for parking lot surveillance.

8. REFERENCES

[1] Stauffer, C. and E. Grimson, Learning Patterns of Activity Using Real-Time Tracking. PAMI, 2000. 22(8): p. 747-757.

[2] Stauffer, C. and E. Grimson. Similarity templates for detection and recognition. in CVPR. 2001.

[3] Stauffer, C. and K. Tieu. Automated multi-camera planar tracking correspondence modeling. in CVPR. 2003.

[4] Collins, et al., A System for Video Surveillance and Monitoring: VSAM Final Report. 2000, Robotics Institute, Carnegie Mellon University.

[5] Lipton, Fujiyoshi, and Patil. Moving Target Classification and Tracking from Real-time Video. in IEEE Workshop on Applications of Computer Vision (WACV). 1998.

[6] Fujiyoshi, H. and A. Lipton. Real-time human motion analysis by image skeletonization. in WACV. 1998: IEEE.

[7] Elgammal, A., D. Harwood, and L. Davis. Non-Parametric Model for Background Subtraction. in European Conference on Computer Vision (ECCV). 2000. Dublin, Ireland.

[8] Haritaoglu, I., D. Harwood, and L. Davis, W4: Real-Time Surveillance of People and Their Activities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. 22(8).

[9] Bobick, A. and J. Davis, The Representation and Recognition of Action Using Temporal Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. 23(3): p. 257-267.

[10]Bevilacqua, A., Effective Shadow Detection in Traffic Monitoring Applications. WSCG, 2003. 11(1): p. 57--64.

[11]Stauffer, C. and W. Grimson, Adaptive Background Mixture Models for Real-Time Tracking. IEEE Computer Vision and Pattern Recognition, 1999.

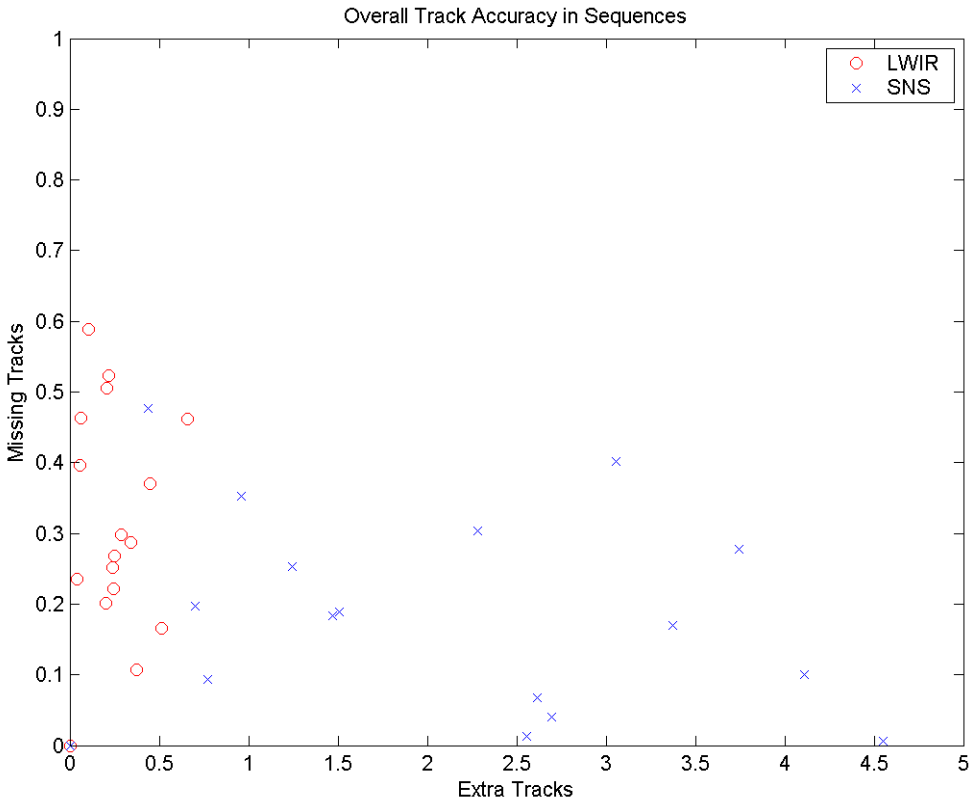

Fig. 6. Comparing LWIR and VNIR Tracking performance

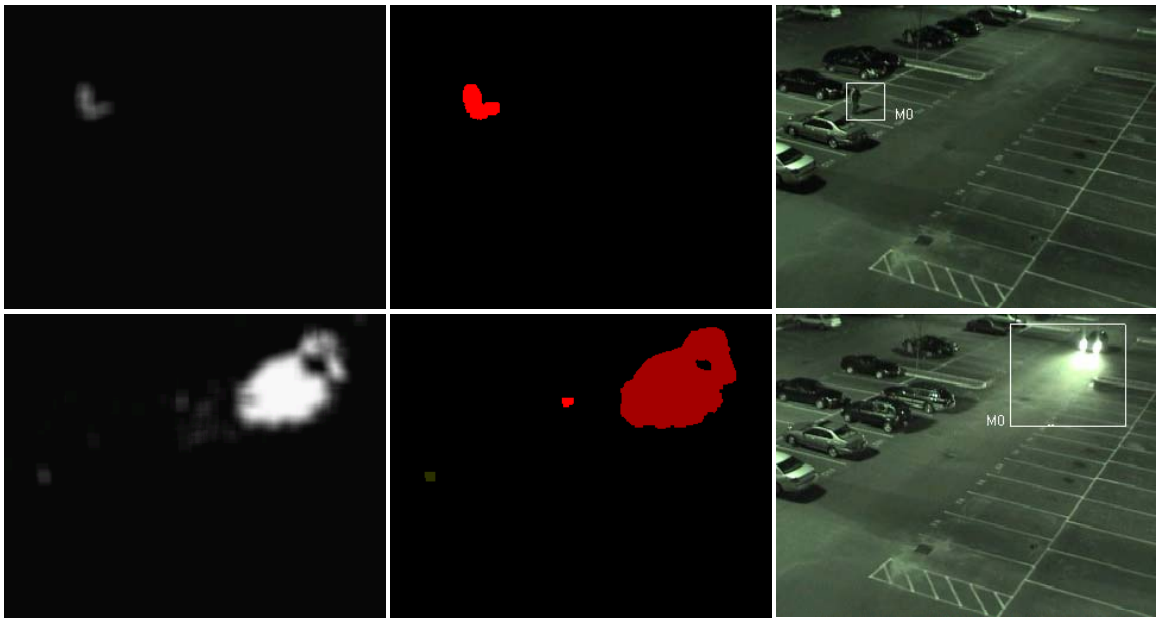

Fig. 7. Examples of moving target detection

Figure 8: Example of tracking vehicles and pedestrian