Physics-based Cooperative Sensor Fusion for Moving Object Detection

Sohail Nadimi and Bir Bhanu
Center for Research in Intelligent systems
University of California, Riverside
{sohail, bhanu}@cris.ucr.edu

Abstract

A robust moving object detection system for an outdoor scene must be able to handle adverse illumination conditions such as sudden illumination changes or lack of illumination in a scene. This is of particular importance for scenarios where active illumination cannot be relied upon. Utilizing infrared and video sensors, we propose a novel sensor fusion algorithm that automatically adapts to the environmental changes that effect sensor measurements. The adaptation is done through a new cooperative coevolutionary algorithm that fuses the scene contextual and statistical information through a physics-based method. Our sensor fusion algorithm maintains high detection rates under a variety of conditions and sensor failure. The results are shown for a full 24 hour diurnal cycle.

1. Introduction

Over the past several decades, many approaches have been developed for moving object detection for indoor and outdoor scenes. Moving object detection methods fall into two categories: (a) feature-based methods [10], and (b) featureless methods (e.g., image subtraction, optical flow, statistical modeling) [1, 3-5,9, 12]. Each of these methods offers advantages that are exploited for different applications. For example, image subtraction is simple and may suffice for indoor type illuminations, temporal differencing can be adopted for slow moving objects, optical flow is useful for a moving camera platform and statistical modeling can capture the background motion.

Some of the shortcomings of the current approaches for moving detection are: 1) None of these approaches address the problem of low light or no light conditions, 2) No contextual information is used to update the Guassian parameters, 3) Generally, a large number of observations are required before a background model can be learned effectively, 4) All the previous algorithms have been applied to a single sensing modality (usually visible or near-infrared) and no results have been shown for extreme conditions, for example, no illumination, sunset, or sunrise condition. In order to overcome illumination conditions such as low or no light conditions, other sensing modalities such as cameras operating in near or longwave IR have been utilized [2]. However, these sensing modalities could still fail due to similar conditions in their respective bandwidth. For example, in a longwave (thermal) IR a subject's temperature could reach that of the background, thus having limited contrast which may cause detection failure.

Multisensor fusion attempts to resolve this problem by incorporating benefits of different sensing modalities. The advantages of multisensor fusion are improved detection, increased accuracy, reduced ambiguity, robust operation, and extended coverage. Sensor fusion can be performed at different levels including signal or pixel level, feature level and decision level.

Our algorithm provides a novel sensor fusion algorithm that fuses longwave (thermal) and visible sensors in a unified manner. By utilizing the IR signal, we can overcome some of the limitations of the visible cameras and by combining the visible and IR signal we improve the detection under variety of conditions.

The salient features of our approach presented in this paper are given below: a) *consistent data representation:* all sensing modalities are represented by a matrix of mixture of Gaussians in a consistent manner. b) *physical models*: sound physical models are used for each sensing modality (e.g., visible and IR) to provide prediction for each signal. c) *evolutionary-based fusion*: a cooperative coevolutionary algorithm is developed to systematically fuse and integrate information from both statistical and physical models into a unified structure for detection. d) *context-based adaptation:* environmental conditions such as ambient air temperature, wind velocity, surface emissivities, etc., are directly incorporated into the detection algorithm and influence the fusion strategies.

2. Related Work and Motivation

Sensor fusion approaches generally fall into one of the following categories: statistical-based, AI-based, algorithmic-based and physics-based. The AI and algorithmic-based paradigms are less suited for dynamic conditions whereas the statistics and physics-based paradigms are the method of choice for integrating sensor information that can change over time.

We provide a new sensor fusion technique that combines the statistical and physics-based fusion paradigms through an evolutionary process. We overcome the disadvantage of each of these paradigms by including suitable sensor

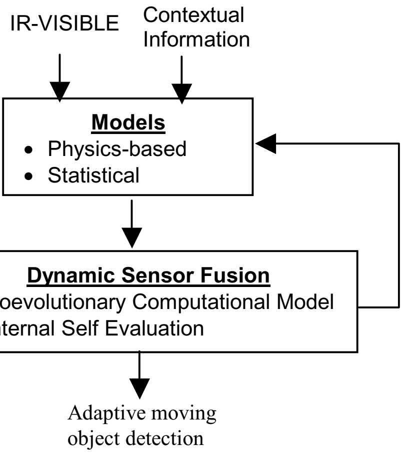

IR-VISIBLE
Contextual Information
Models
• Physics-based
• Statistical
Dynamic Sensor Fusion
• Coevolutionary Computational Model
• Internal Self Evaluation
Adaptive moving object detection

Figure 1. Sensor Fusion Architecture.

models that have enormous generalizing power. This generalizing power is then used to complement the limited available sensor data that is required by the statistical methods. The fusion is performed at the pixel level where the information loss is minimal.

3. Technical Approach

The sensor fusion architecture for moving object detection is depicted in Figure 1. Observations from the sensors along with the external conditions, which carry the contextual information, are used to build statistical (mixture of Gaussian) background model. The contextual information is also used to update values of internal physical models. Physical models include reflectance models for predicting image intensity values and thermal models for predicting background surface temperature values. Unlike the previous work that updates the background models solely based on the current observations, we incorporate our physical models into the adaptive loop. The physical models are integrated with the statistical models through a cooperative coveolutionary process [7]. The cooperative coevolutionary process estimates the best representation for the background per pixel. This is done through genetic based evolutionary process that searches for the optimal representation based on the current, and recent past observations and detection results in addition to the predictions given by the physical model.

Our representation of a matrix of mixture of Gaussians (described in section 3.1) includes Gaussian parameters for the infrared and visible sensors (including RGB channels). A population of this representation is maintained as a pool of individuals for the evolutionary process. Once the evolutionary process is stopped the best individual represents the background model of that pixel. In this manner, the contextual information plays an active role in contributing to the most ideal sensor for a particular condition. The detection algorithm in Figure 1 requires a model of the background. This model is estimated by a mixture of Gaussians per pixel. Table 1 shows this process. The details are explained in the following subsections.

3.1 Representation

The probability of a pixel classified as a background drawn from a probability distribution can be estimated by a mixture of density functions. Assuming the parametric form of the mixture is Gaussian, probability of observing background is:

$$\mathrm{P(x)} = \sum_{i=1}^{g} \mathrm{W_i}\, \eta(\mathrm{x}, \mu_i, \Sigma_i)$$

Where x is the pixel value, W is the prior, g is the number of Gaussians, and η is the Gaussian form characterized by the mean μ and covariance Σ. Each pixel is then defined by its first order statistics in the 4D vector (R,G,B,T) as follow: $\hat{\mathrm{I}} = <\mathbf{w}_{\mathbf{IR}_g}, \mu_{\mathbf{IR}_g}, \sigma_{\mathbf{IR}_g}, \ldots, \mathbf{w}_{\mathbf{VC}_g}, \mu_{\mathbf{VC}_g}, \sigma_{\mathbf{VC}_g}>$ where **IR** = infrared, **VC** = Video Channel ∈ {R,G,B}, and **g** = number of Gaussians. We assume that R, G, B, T are independent. Î represents a solution instance; we maintain a matrix (rows of Î) to represent the solution space.

3.2 Physical Models

The algorithm in Table 1 uses the physics-based predictions in its evaluation phase. Models of bi-directional reflectance distribution functions (BRDF) and thermal equilibrium based on conservation of energy are used to predict surface color and temperature in the visible and longwave IR. The models are briefly described here.

3.2.1 Physical Models of Reflectance

Several reflectance models including the lambertian, Phong, dichromatic [8] and Ward [11] models have been developed to describe the reflectance due to normal, forescatter and backscatter distributions. We utilize the dichromatic model:

$L(\lambda, \hat{e}) = L_i(\lambda, \hat{e}) + L_b(\lambda,\hat{e}) = m_i(\hat{e})\, c_i(\lambda) + m_b(\hat{e})\, c_b(\lambda)$;

where L is the total reflected intensity, L_i and L_b are reflected intensities due to surface and subsurface respectively, m_i and m_b are geometric terms, c_i and c_b are relative spectral power distribution (SPD) of the surface and subsurface respectively, and $\hat{e}$ is a vector representing incident and reflected angles with respect to surface normal. The dichromatic model is useful in describing the reflection from inhomogeneous opaque dielectric materials (e.g., plastics). It is also useful in describing material colors since the SPD of the reflected light due to subsurface is decoupled from the geometric terms. To calculate the invariant body color, the image is segmented into regions with uniform reflectivity. For each region, pixel values in the RGB space are formed into a matrix M of size n × 3 where n is the number of rows (pixels) and 3 represents R,G, and B values. Singular value decomposition is then applied to M and the singular vector corresponding to the largest singular value is selected as the body color (c_b), which is the predicted surface color [6].

3.2.2 Thermal Physical Model

For predicting surface temperatures in the longwave IR, the following conservation of energy model is used. $E_{in} = E_{out}$; $E_{out} = E_{rad} + E_{cv} + E_{cd}$; Where E_{in} is the input energy, E_{out} is the output energy described by three phenomenon E_{rad} (energy radiated), E_{cv} (energy convected), and E_{cd} (energy conducted). Models for each energy flux is described in details in [6]. Briefly the following models are used to describe each of the above fluxes:

$\mathbf{E_{in}} = E_{direct} + E_{skylight} + E_{atm}$

$\mathbf{E_{direct}} =(1089.5/ma)\ e^{(-0.2819\ m_a)}$

$\mathbf{E_{atm}} = E(BB,Ta)\ \{1\text{-}[0.261\ e^{-7.77 * 10\text{-}4\ (273\text{-}Ta)2}]\}$ where E_{direct} = direct irradiation due to sun, $E_{skylight}$ = irradiation due to sky ≈ (40-70 W/m^2), E_{atm} = irradiation due to upper atmosphere, m_a = The number of air masses ($m_a \approx$ secant(Z)), T_a = Air temperature, $E(BB,T_a)$ = radiation of a blackbody at T_a temp, and Z = sun's Zenith angle.

E_{rad} is estimated based on Stephen-Boltzman law: $E_b = \sigma\ T^4$, where $\sigma = 5.669 \times 10^{-8}$ watts/m^2 Kelvin^4 and the subscript *b* is for blackbody which is capable of 100% absorption (or emission) of energy.

The convected heat flux is given by: $E_{cv} = h_{cv}\ (Ts - T\infty)$ where h_{cv} is the convective heat transfer function which is a complex phenomena, Ts and T∞ are surface and fluid temperatures respectively. For laminar flow, h_{cv} can be roughly estimated by the following empirical model:

$\mathbf{h} = 1.7\ \ |Ts - Ta|^{1/3} + (6\ Va^{0.8})\ /\ L^{0.2}$ where Va = wind speed; L = characteristic Lateral dimension of surface, Ts and Ta are surface and air temperature respectively.

The conducted heat flux is described by:

$E_{cd} = A\ (T2 - T1)\ /\ (L\ /\ k)$, where A is the area, T2-T1 is the differential temperature and L/k is called the thermal resistance or R-value and is tabulated for many materials. The above equilibrium model is solved for Ts which is the predicted temperature.

3.3 Background Model Estimation

The cooperative coevolutionary (CC) algorithm in Table 1 is used to select an optimal representation for the background based on the recent past observation detection results, and physics-based predictions.

The CC algorithm utilized here is a recent evolutionary paradigm that has been applied to optimization problems [7]. The success of CC depends on 4 criteria: 1) problem decomposition, 2) interdependability, 3) credit assignment, and 4) population diversity. Our sensor fusion algorithm satisfies all four criteria's since a) our problem is naturally decomposed (IR and video), b) our representation (matrix of mixture of Gaussians) provides interdependencies between subcomponents, c) the objective (or fitness) function minimizes the physics-based prediction in both IR and video, and d) population diversity is maintained by roulette wheel selection method.

An important part of the evolutionary algorithm is the evaluation function, referred to as the fitness function. We provide a suitable fitness function that integrates the statistics collected by the system and the physical models that are directed by the contextual information (environmental conditions).

Table 1. Algorithm for learning background model.

```
Evolutionary Adaptive Background Modeling
T = Training set which includes prediction, observation,
and previous detection results per pixel;
Note: An organism represents a solution.
---------------------- CC ALGORITHM ------------------
For each pixel
  Create and initialize 4 subpopulations (see Section 3.1).

  Loop
    Build 4 organisms (e.g., solution space)
    Evaluate organisms using the training set T
    Store the best organism
    For each subpopulation
      Evolve each subpopulation (Selection, Mutation,
                                 Crossover)
    EndFor
  Until stop Condition
    Return the best organism
EndFor
```

3.3.1 Fitness Function

For each channel, a population of individuals (see 3.1) is initially created randomly. These individuals are maintained both for the IR (temperature) and the video sensor (RGB). Let an individual I in a population be represented as: $I = <w_1, \mu_1, \sigma_1, . . ., w_m,\mu_m, \sigma_m>$

Let: T represent Temperature or RGB values

T_{ob_j} = Observed values, j = 1 .. n ; n = window in the past

T_{p_j} = Predicted values by Physics

P(T) = The probability distribution.

We keep a moving window of recent past n frames. This recent window is used as groundtruth, G, for training examples. Unlike most other work that only uses the last or current observation (frame) to update the mixture of Gaussians, we elect to keep a window of frames. Let

$$\underset{\{j=1..n\}}{G_j} = \begin{cases} 1 & \text{Background} \\ 0 & \text{Foreground} \end{cases}$$

and individual's statistical estimation *F(I)*:

$$F(I) = \frac{1}{n}\sum_{j=1}^{n}\left[G_j P(T_{ob_j}) + (1 - G_j)(1 - P(T_{ob_j}))\right]$$

The above function is only based on the past and current statistics. To tie the knot with the physics, we introduce the following function, named *credibility function*. This function provides a credibility measure as to how close our observations come to the predicted value.

Our physics prediction should be more credible if the observed value is detected as background, when the physics-based prediction also agrees with the observed values for the background. Moreover, if the physics predicts a very different value for the background and our system has actually detected the pixel value as the foreground, then the physics may still be credible. On the other hand, if we have classified a pixel as background where our physics is predicting otherwise, we must be able to assign a low credibility to our physics prediction. Similarly if the physics prediction is very close to that of the observed value but the system has detected the pixel as foreground, then the physics prediction may not be reliable and a LOW credibility must be assigned. To realize the above, the following credibility function, C, is provided:

$$C = e^{-\alpha\left[\frac{1}{n}\sum_{j=1}^{n} G_j \frac{\left|T_{ob\,j} - T_{P\,j}\right|}{T_{ob\,j} + T_{P\,j}} + (1-G_j)(1-\frac{\left|T_{ob\,j} - T_{P\,j}\right|}{T_{ob\,j} + T_{p\,j}})\right]}$$

where vectors G, T_{ob} and T_p are defined as before and α decides the rate of credibility function. As the observed values T_{ob} agree closer with the predicted values T_p for a particular decision G, then the value of the credibility approaches 1. For example, it is easy to verify that in the extreme case where all the previous n frames were background, and that the predicted values matched the observed values, the sensor will get a credibility of 1.

Given $F(I)$ and the credibility function C for individuals for both IR and video, then, a fitness function for an organism (solutions) made of both IR and video species can be realized as follows:

$$F_{\mathrm{organism}}\ (<I_{\mathrm{video}},\ I_{\mathrm{IR}}>) = C_{\mathrm{video}}\ F(I_{video}) + C_{\mathrm{IR}}\ F(I_{IR})$$

Above equation is used for evaluating the organisms formed by the video and IR signals. The parameter α adjusts the importance of the role the credibility function plays in the fitness function. α can be adjusted depending on how fast the credibility function is desired to be influenced by the agreement between the physics prediction and actual observations. Furthermore, roulette wheel selection method, a single point crossover operator with a crossover rate of 0.8, and mutation rate of 0.01 for a population size of 60 per organism are used.

4. Experiments

The data was gathered at a typical urban location with the latitude 33:50:06 and longitude 117:54:49, from 15:30:00 on January 21, 2003 till 14:24:00 January 22, 2003. Initially, from 15:30:00 till 17:07:04, data was collected at the rate of 1 frame every 2 seconds, then the temporal resolution was changed to 1 frame per 10 second for the rest of the data collection period. Two cameras, a FLIR system thermal camera operating at 7-13 μm and an Intel web-cam operating in the visible range were utilized for data acquisition. The thermal camera was fully radiometric, which means that the pixel values obtained by the camera were thermal. The thermal camera included self-calibration that at specified intervals adjusted to internal thermal noise. The radiation-to-temperature conversion was done automatically by the camera for the default values of emissivity = 0.92, air and ambient temperatures = 280 Kelvin, distance to target = 100 m, and humidity = 50%.

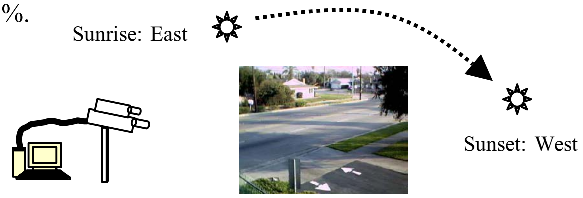


Figure 2. Position of the cameras with respect to the scene and the direction of the sun's path.

The video camera was attached to top of the thermal camera on a tripod (see Figure 2). Both cameras were located 20 feet above the ground looking downward at the scene at an angle of approximately 25°. In addition to the thermal and the video cameras, a complete weather station was utilized to obtain weather data every minute. The weather station included an anemometer, humidity sensor, wind direction, two temperature sensors, and a barometer sensor. All sensors and the cameras were controlled by a PC. Data collection between the cameras and the weather data was all synchronized through software control.

For spatial registration affine transformation was applied and to avoid temporal registration, both cameras were triggered simultaneously and in parallel. For predicting correct reflectance and thermal predictions, a split and merge algorithm initially segmented the image where a user initially labeled the segments into 5 regions, asphalt, concrete, grass, bush, and unknown. Only statistical properties were utilized for the unknown surface types.

4.1 Physical Model Estimation and Predictions

For surface color estimation, the dichromatic model was utilized. The results for the four different pre-segmented surfaced is given in terms of unit vectors in the RGB space. The values were obtained for various times and are given in Table 2 at an hourly illumination condition. The concrete and asphalt had similar vectors due to their neutral color attributes. On the other hand, the chloroform in the vegetation such as grass and bush causes the vectors to be shifted toward green.

The *average,* and *standard deviation* for the reflectance vectors for the four surfaces were: Asphalt = (0.8°, 0.5°), Concrete = (1.1°, 0.6°), Grass = (1.7°, 1°), and Bush =(3.9°, 2.9°), where the first number represents the average and the second represents standard deviation. Since vegetation include chlorophorms, the higher variation in reflectance of grass and bush are contributed to their surface specularity, which is not modeled by our algorithm.

For surface temperature prediction, the thermal models of Section 3.2.2 were used. These predictions were used by the fitness function in section 3.3.1. Figure 3 shows the result of predictions superimposed on actual measurements by the thermal camera. As shown, the models were able to track temperature fluctuations for 4 different surface types closely. The average difference between the prediction and measurement for all surfaces were about 2°c with standard deviation of 1.87°c.

Table 2. Surface body color estimation (c_b).

Time	Asphalt			Concrete		
	R	G	B	R	G	B
8:30	.5727	.5726	.5867	**.5813**	**.582**	**.5687**
9:30	.5714	.5716	.5889	**.5791**	**.5797**	**.5732**
10:30	.5773	.5714	.5862	**.5824**	**.5824**	**.567**
11:30	.5669	.5676	.597	**.5737**	**.5745**	**.5838**
12:30	.5695	.5695	.5927	**.5686**	**.5749**	**.5884**
13:30	.5682	.568	.5954	**.5767**	**.5753**	**.5801**
14:30	.5741	.572	.5859	**.5681**	**.5752**	**.5886**
15:30	.5635	.552	.6025	**.557**	**.5723**	**.6019**
16:30	.5623	.5684	.6006	**.5572**	**.5802**	**.594**
17:30	.5544	.5668	.6095	**.5566**	**.5813**	**.5935**

Time	Grass			Bush		
	R	G	B	R	G	B
8:30	.6336	.726	.2672	**.5718**	**.6239**	**.5327**
9:30	.6343	.7189	.2844	**.5893**	**.624**	**.5132**
10:30	.6369	.7128	.2938	**.5662**	**.6368**	**.5234**
11:30	.632	.7193	.2883	**.5476**	**.625**	**.5563**
12:30	.6256	.7376	.2543	**.543**	**.637**	**.5471**
13:30	.6249	.7364	.2591	**.5749**	**.6404**	**.5093**
14:30	.621	.7391	.2611	**.5968**	**.6338**	**.4921**
15:30	.606	.7505	.2636	**.5639**	**.6421**	**.5193**
16:30	.604	.7572	.2486	**.6567**	**.6369**	**.4039**
17:30	.6231	.738	.259	**.6321**	**.6357**	**.4431**

4.2 Detection Results

Moving object detection is performed after an initial background model is built. Once new thermal and video frames are available, they are registered. The registered image then contains red, green, blue, and temperature values at each pixel location. The cooperative coevolutionary algorithm is used to build the background model. Each pixel is updated independently. The background model is periodically updated to track the environmental changes. The following parameters were used in the cooperative coevolutionary algorithm to update the background models: number of species = 4; population size = 60; crossover = single point; crossover rate = 0.8; mutation rate = 0.01; maximum number of generations = 60; training data = 20 frames; number of Gaussians per sensor = 3; α = 0.5.

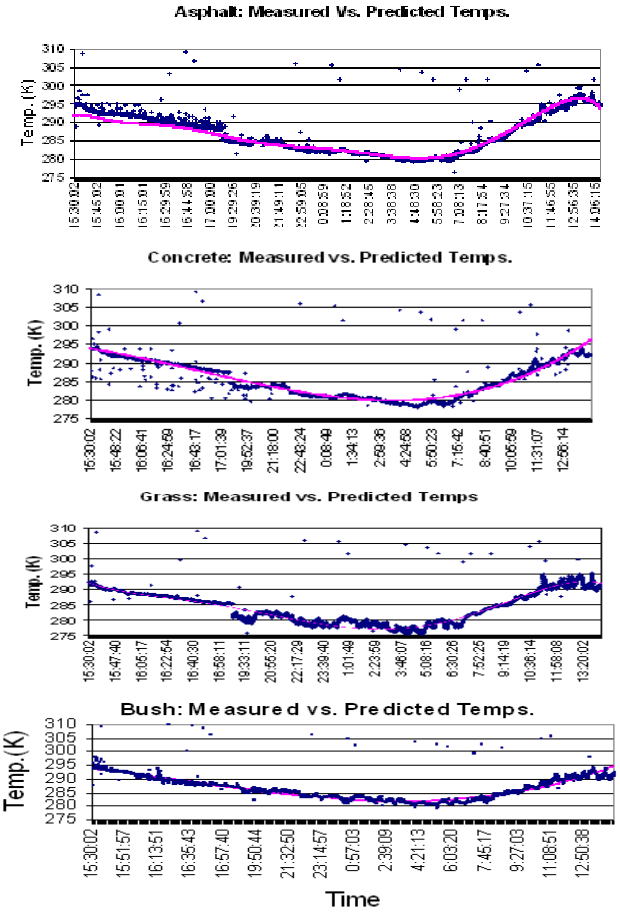
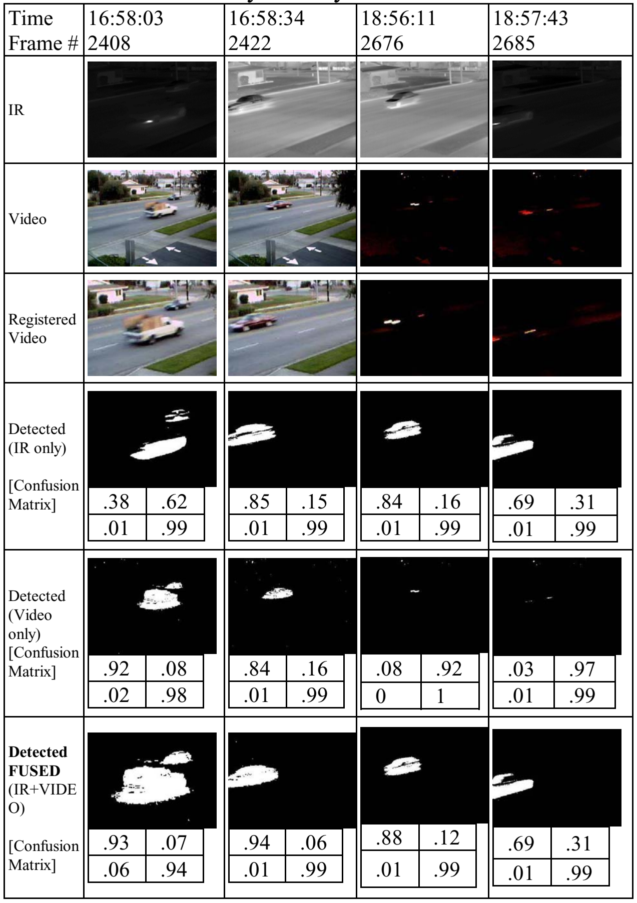


Figure 3. Measurement (blue) vs. predicted (red) surface temperature values.

Once the background model is available, for each incoming frame, each pixel is compared to its corresponding model and if its value is within 3 standard deviation of any of its models, it is classified as background. This information is kept in a binary image where a detected moving pixel is a binary 1 (white) and a background pixel is 0 (black). These binary frames provide training data for the next background model update. In the following examples, in addition to the thermal IR and video frames, detection for each sensor and the fused detection for the registered image are provided. Moreover, the following confusion matrix is given for the results:

% moving Obj correctly detected	% moving Obj missed
% Background missed	% Background correctly detected

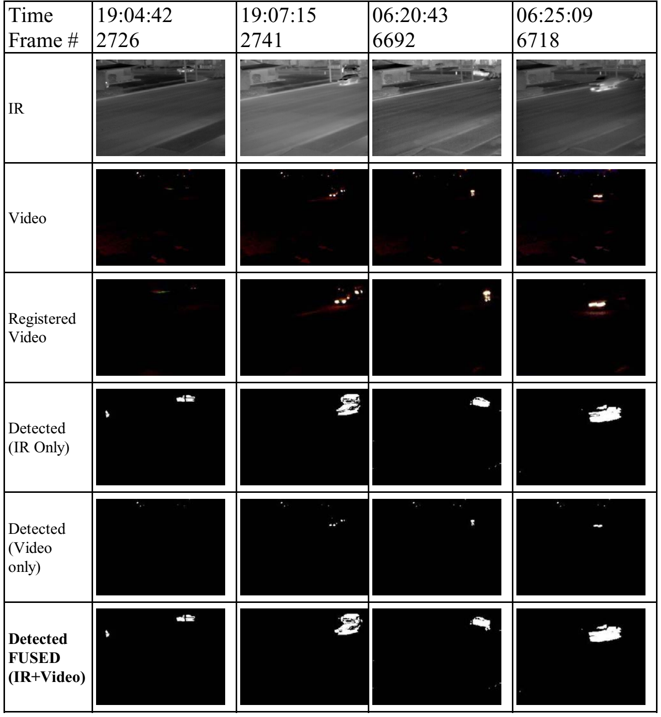
- **Example 1:** Figure 4 shows example frames detected in the afternoon and early evening hours. During this period, illumination and heat exchanges are rapid. Depending on the heat stored and reradiated by an object and the background the object may be observed having very similar temperatures as the background (IR frames 2408 and 2685) or very different (IR frames 2422 and 2676). In frame 2408, video signal was much stronger, providing sharp contrast for the moving objects. Despite the lower performance of the IR, the objects were recovered by the video. Similarly, in frame 2422, the detection result of the IR was further enhanced by the registered video as is shown in the fused detected frame. Frames 2676 and 2685 are during early evening hours. The video camera had a 25 lux minimum illumination requirement; therefore, although the scene was not totally dark, the video signal during the night time was very weak. This was compensated by the strong IR signal; however, the maximum detection was obtained by IR only.



Time Frame #	16:58:03 2408	16:58:34 2422	18:56:11 2676	18:57:43 2685
IR				
Video				
Registered Video				
Detected (IR only) [Confusion Matrix]	.38 .62 .01 .99	.85 .15 .01 .99	.84 .16 .01 .99	.69 .31 .01 .99
Detected (Video only) [Confusion Matrix]	.92 .08 .02 .98	.84 .16 .01 .99	.08 .92 0 1	.03 .97 .01 .99
Detected FUSED (IR+VIDEO) [Confusion Matrix]	.93 .07 .06 .94	.94 .06 .01 .99	.88 .12 .01 .99	.69 .31 .01 .99

Figure 4 Example 1: Mixed-good and bad IR and video at various times in the afternoon and evening.

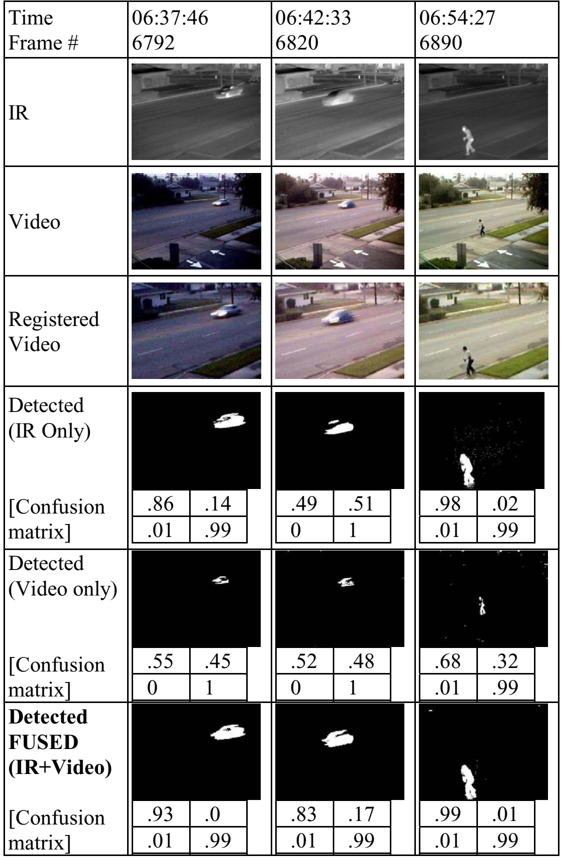
- **Example 2:** Figure 5 is an example where the detection algorithm relied heavily on one sensor, IR. Due to lack of illumination and video sensor's low sensitivity, objects could not be detected by video only. A good example is frame 2726 where a car and a person were in the scene. These were not observed in the video; however, they were present in the IR image and were clearly detected in both IR and the fused frame. Frames 2741, 6692 and 6718, indicate that the detection was not influenced by the video. The lights from the vehicles were very visible and were detected as part of the moving object, but the surface reflection of the lights clearly did not contribute to misdetection. This is due to the fact that the physics-based prediction assigns low credibility to the video signal; hence, low reflections are not detected. In effect this plays a role in deciding how important a sensor's observations are. If a video pixel gets a low credibility, then its values are less meaningful; therefore, in order to observe a change, the signal must be very strong (e.g., front head lights of the cars). Since the front head lamps of most vehicles are halogen and radiate heat, they are also observed as part of the vehicle in the IR image, thus, they are also being detected as part of the vehicle.



Time Frame #	19:04:42 2726	19:07:15 2741	06:20:43 6692	06:25:09 6718
IR				
Video				
Registered Video				
Detected (IR Only)				
Detected (Video only)				
Detected FUSED (IR+Video)				

Figure 5 Example 2. Good to excellent IR signal, bad video signal at night. (*Note: Due to lack of video contrast no groundtruth could be obtained.*)

- **Example 3:** Figure 6 is an example of dramatic illumination changes during the early sunrise and early morning hours. During these periods, the environment changes radically change due to the energy of the sun. The sensors must adapt to these rapid changes. Figure 8 shows the thermal changes on different surfaces that are tracked by the physics-based models. As shown, the slope of the temperature values change radically during this period. However, the physics-based models are able to follow these changes and provide high credibility values that affect the background models build by the algorithm. As the illumination reaching the video camera is increased, the detection due to video gets better. This is shown in frames 6792 and 6820 where the video camera began participating in the detection. This is indicated by the increase in the detection in the fused image versus the IR or video only images.

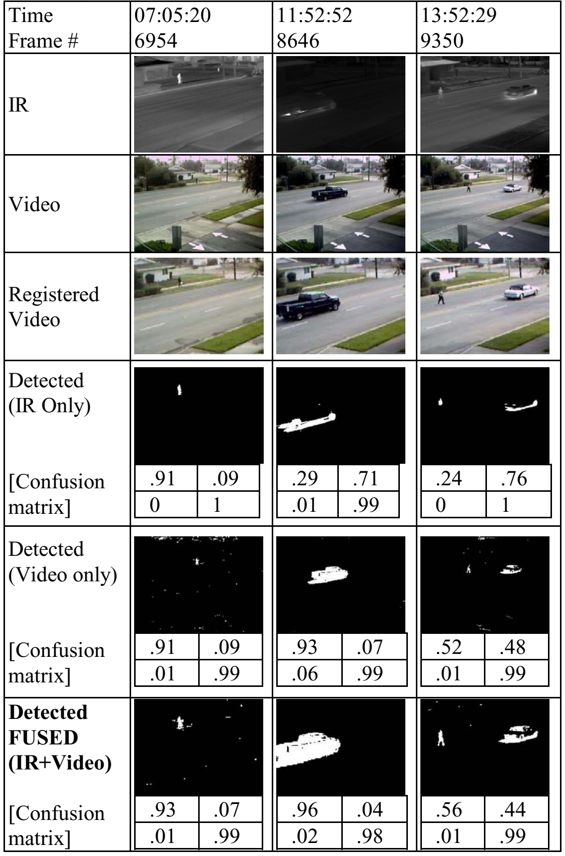


Time Frame #	06:37:46 6792	06:42:33 6820	06:54:27 6890
IR			
Video			
Registered Video			
Detected (IR Only) [Confusion matrix]	.86 .14 .01 .99	.49 .51 0 1	.98 .02 .01 .99
Detected (Video only) [Confusion matrix]	.55 .45 0 1	.52 .48 0 1	.68 .32 .01 .99
Detected FUSED (IR+Video) [Confusion matrix]	.93 .0 .01 .99	.83 .17 .01 .99	.99 .01 .01 .99

Figure 6. Example 3: Fusion while illumination changes at sunrise.

- **Example 4:** Figure 7 is an example of early morning, noon and early afternoon hours. As the sun comes up, the surfaces are heated up by the incoming energy from the sun, the increase in the surface temperatures approaches closer to the temperatures of some moving object surfaces. Depending on the moving object surface temperatures and emissivities, the contrast in the IR can be radically different from frame to frame. This is obvious between frames 6954 and 8486 for example. Frame 6954 represents an image in the morning with a person in the scene. Surface temperatures are still lower than that of the human body; moreover, human body's emissivity is high (0.98) compared to the background surfaces. The human is clearly visible in the IR image. Although not very visible in the video image of frame 6954, the human is also in that image; this is clearer in the registered image. Both sensors provided excellent contrast in this case and the person was clearly detected.

Frames 8486 and 9350 show moving objects later in the day when surfaces have reached higher temperatures. In this case, it is possible to have a moving object that may have closer temperature to the background surface as is indicated by both of these frames. On the other hand, video provided excellent signal and contrast. Many pixels were missing from the detected IR only, but the final fused detection recovered most of these missed pixels on moving objects.



Time Frame #	07:05:20 6954	11:52:52 8646	13:52:29 9350
IR			
Video			
Registered Video			
Detected (IR Only) [Confusion matrix]	.91 .09 0 1	.29 .71 .01 .99	.24 .76 0 1
Detected (Video only) [Confusion matrix]	.91 .09 .01 .99	.93 .07 .06 .99	.52 .48 .01 .99
Detected FUSED (IR+Video) [Confusion matrix]	.93 .07 .01 .99	.96 .04 .02 .98	.56 .44 .01 .99

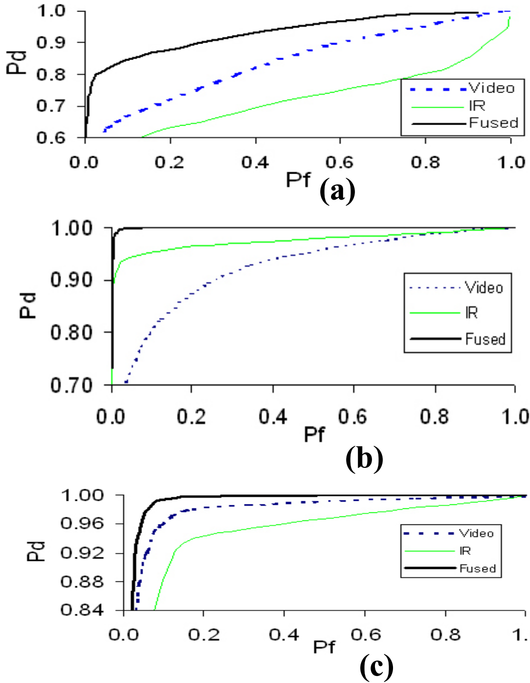
Figure 7. Example 4: Mixed IR and good video signal.

4.3 Performance Analysis

To compare the performance of the detection algorithm for sensor fusion, we utilize the Receiver Operating Characteristic (ROC) curves and define the probability of detection as percentage of moving object correctly detected and probability of false alarm as percent of background classified as moving object.

We selected frames representing afternoon, early morning and high noon for this analysis. The nighttime was not selected since no video signal was available at night and the detection algorithm relied only on the IR sensor; this was explained in example 2 above. The first ROC curve, Figure 8(a), represents an afternoon time. An example of this is frame 2408 in Figure 7. As is indicated by example 1 frame 2408 and this ROC curve, the video signal operated at a higher rate than the IR signal. The fusion method operated at a higher level than both the video and the IR.

The ROC curve of Figure 8(b) is an example of early morning hours. This figure is a contrast to that of Figure 8(a) in the afternoon. In this case the detection rates for both the IR and the fused image were high where the video sensor operated only nominally. This is again due to the fact that during early morning hours, a large gradient may exist between natural surface temperatures and those of animated objects with internal sources of energy such as vehicles and humans. This is due to the fact that a great deal of energy has been dissipated to the environment throughout the night. In addition, the video signal, as indicated in Figure 6, example 3, is rapidly changing due to the illumination changes when sun is coming up.



(a)

(b)

(c)

Figure 8. ROC curves for various periods of the day. (a) afternoon-evening, (b) early morning, (c) morning-noon.

The third ROC curve, Figure 8(c), is an example of how fusion can enhance the detection when both sensors may be operating at lower rates, yet the fused version will be able to detect at higher rate. This is an example of when cooperation between sensors can play a complementary role. The fused detection in this case operates at higher level than each single sensor alone. Partly this is due to the fact that different sensors may detect different parts of an object. So, one ought to expect sensor fusion to do much better in detecting more pixels on the object than one sensor alone. This is also observed from frames 8646 and 9340 of example 4 in Figure 7 when for example, the detected IR and video frames have detected different part of the same object.

These ROC curves also indicate that as the time of day changes, the dynamic sensor fusion introduced here can automatically adapt to environmental changes. This adaptation is also in the form of adapting to the best sensor at the time. The cooperation among sensors can also take on a complementary role when different sensors are able to detect different part of an object that may not be visible to one another. This adaptation is done in a cooperative manner where sensors have already participated in the model-building phase.

5. Conclusions

In this paper a novel physics-based sensor fusion technique for moving object detection was introduced. The sensor fusion architecture integrated the statistical and phenomenology of the sensors in the visible and longwave IR through an evolutionary computational model. Our representation, matrix of mixture of Gaussians, along with the cooperative coevolutoionary search algorithm integrated the contextual information through the physics-based and statistical models. We showed that our fusion model adapted to various illumination conditions and is suitable for detection under variety of environmental conditions.

Acknowledgement – This work was supported by grants F49620-02-1-0315, F49620-01-1-0343, and DAAD19-01-0357. The contents and information do not necessarily reflect the position or policy of U.S. Government.

References:

[1] M. Cristani, M. Bicego, and V. Murino, "Integrated region and pixel-based approach to background modeling," *IEEE Workshop on Motion and Video Computing*, pp. 3-8, Dec. 2002.

[2] J. Han and B. Bhanu, " Detecting moving humans using color and infrared video," *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 228-233, July 2003.

[3] I. Haritaoglu, D. Harwood, and L. Davis, "W4: real time surveillance of people and their activities," PAMI 22(8), 809-830, 2000.

[4] T. Horprasert, D. Harwood, and L.S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," *Proc. FRAME-RATE Workshop held in conjunction with Intl. Conf. on Computer Vision*, pp. 1-19, 1999.

[5] A.J. Lipton, H. Fujiyoshi, and R.S. Patil, "Moving target classification and tracking from real-time video," *IEEE Workshop on Applications of Computer Vision*, pp. 8-14, 1998.

[6] S. Nadimi and B. Bhanu, "Physics-based models of color and IR video for sensor fusion," *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 161-166, July 2003.

[7] M.A. Potter and K.A. DeJong, "A cooperative coevolutionary approach to function optimization," *Proc. of the 3rd Conference on Parallel Problem Solving from Nature*, pp. 249-257, 1994.

[8] S. A. Shafer, "Using color to separate reflection components," *Color Research and Application*, 10(4), pp. 210-218, 1985.

[9] C. Stauffer and W.E.L. Grimson "Learning patterns of activity using real-time tracking," *PAMI* 22(8), pp 747-757, 2000.

[10] M.M.D. Viva, and C. Morrone, "Motion analysis by feature tracking," *Vision Research*, Vol. 38, pp. 3633-3653, 1998.

[11] G. Ward, "The RADIANCE lighting simulation and rendering system," *Computer Graphics (SIGGRAPH' 94)*, pp. 459-472, July 1994.

[12] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: real-time tracking of the human body," PAMI 19(7), pp. 780-785, 1997.