Adaptive Fusion for Diurnal Moving Object Detection

Sohail Nadimi and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, CA, 92521, USA
{sohail, bhanu}@cris.ucr.edu

Abstract

Fusion of different sensor types (e.g. video, thermal infrared) and sensor selection strategy at signal or pixel level is a non-trivial task that requires a well-defined structure. In this paper, we provide a novel fusion architecture that is flexible and can be adapted to different types of sensors. The new fusion architecture provides an elegant approach to integrating different sensing phenomenology, sensor readings, and contextual information. A cooperative coevolutionary method is introduced for optimally selecting fusion strategies. We provide results in the context of a moving object detection system for a full 24 hours diurnal cycle in an outdoor environment. The results indicate that our architecture is robust to adverse illumination conditions and the evolutionary paradigm can provide an adaptable and flexible method for combining signals of different modality.

1. Introduction

Multisensor fusion attempts to improve object detection/recognition 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.

It is generally desirable to fuse sensors at the signal or pixel level where the information loss is minimal; however, sensor fusion at this level may be difficult for sensors of different type. A unified representation must be obtained whereas all sensors can be put into this representation. Furthermore, operations on the data representation must be defined. These operations must encapsulate the process for which the sensor fusion is taking place. Once a representation and its operators are defined a set of rules are generally developed to adapt the fusion strategies to changes in the signal caused by the environment in which the system is operating. An example is detecting moving objects under a variety of environmental conditions in outdoors. Suitable representation and automatic adaptation methods are applied to detect and track changes [8] in the scene for robust detection; however these methods do not take into account any contextual information and rapid environmental dynamics which can greatly affect the sensor selection and detection strategy. For more robust detection under adverse illumination condition for example no light at night, other sensing modalities that can operate under those conditions must be introduced.

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 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. Our sensor fusion and strategy selection algorithm is performed at the pixel level where the information loss is minimal.

The salient features of our approach described 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) *Evolutionary-based strategy selection*: A cooperative coevolutionary algorithm is developed to systematically fuse and integrate information from both statistical and physical models into a unified structure for sensor selection and detection. c) *physical models*: Sound physical models are utilized for each sensing modality (e.g., visible and IR) to provide prediction for each signal and include the contextual information into the evolutionary process. d) *Contextual-based adaptation:* Environmental conditions such as ambient/air temperatures, wind and fluid velocities, surface emissivities, etc., directly influence the fusion strategies. e) *Experimental results:* Results are obtained for a full 24 hours diurnal cycle for a moving object detection system fusing long-wave infrared (IR) and video.

2. Previous Work

There have been several approaches to moving object detection, including feature-based methods [1], and featureless methods such as statistical background subtraction [2, 3, 8]. Inherent problems of feature-based methods are due to noise, articulation or occlusion of background

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

and foreground objects, which make the correspondence problem intractable. In the featureless methods the background is modeled based on statistical properties of observed signal at pixel level. Outdoor scenes provide more challenging scenarios such as sudden changes due to cloud or swaying motion of the background such as leaves. In [8] the idea (of representing a pixel based on a Gaussian distribution) is extended to a mixture of Gaussians density functions that can represent any arbitrary distribution functions. Assuming independence, this idea is exploited among sensors and fusing the results through multistrategy models such as cooperative, competitive and Dempster-Shafer [4].

To maintain and track the dynamics of a scene, each of these approaches provide a statistical-based adaptive procedure solely based on current or past observations. In some cases, for example in [9], the adaptive procedure is augmented with an object or region-based procedure to update the background models. The mixture model can adapt to slow illumination changes over time. Some of the shortcomings of the current approaches are: 1) All the previous approaches use a fixed recursive filter to adapt to scene changes which requires fixed parameters, learning rates and thresholds, 2) None of these approaches address the problem of low light or no light conditions, 3) No contextual information is used to update the Guassian parameters, 4) Generally, a large number of observations are required before a background model can be constructed and maintained effectively, and 5) 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.

Our algorithm explained below overcomes these shortcomings by providing a novel sensor fusion algorithm that fuses longwave (thermal) and visible sensors in a unified manner. Adaptation is done automatically, and is guided by contextual information, which also provides constraints for strategy selection. 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 a variety of conditions.

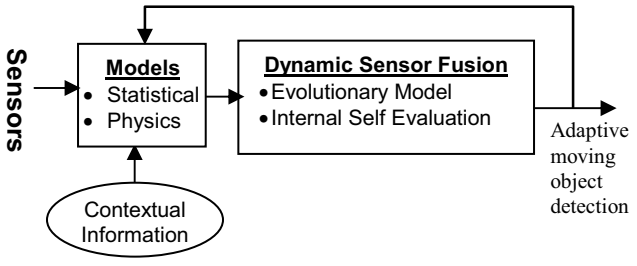

Figure 1. Sensor Fusion Architecture.

3. Technical Approach

The detection algorithm in Figure 1 requires a model of the background. This model is estimated by a mixture of Gaussians per pixel.

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

Where x is the pixel value, W is the prior, g is the number of Gaussians, and η is the Gaussian with mean μ and covariance Σ. Each pixel is then defined by its first order statistics for each sensor S∈ {R, G, B, T} as an "*individual*" where R, G, B are color and T is temperature:

$$I_S = \langle W_{S_1}, \mu_{S_1}, \Sigma_{S_1}, \ldots, W_{S_g}, \mu_{S_g}, \Sigma_{S_g} \rangle$$

For optimal representation, a population of individuals of the form I_S called a *sub-population* is maintained for each sensor. To form the complete representation of a pixel, a 4D vector is used by concatenating each I_S from each sensor: $\hat{I} = \langle I_R, I_G, I_B, I_T \rangle$. We assume independence between channels. $\hat{\mathbf{I}}$ represents a solution instance or *organism*; we maintain a matrix: $M = [\hat{I}_1, \ldots, \hat{I}_n]^t$ to represent the solution space where *t* is the transpose operation.

Background Model Estimation: A model of background must be built and maintained before detection is performed. Unlike the previous work that infers this model solely based on available statistics, we introduce physics-based prediction that complements observations for each respective channel. This provides an enormous generalizing power not available with any existing purely statistical-based approaches.

To estimate and select the optimal representation for each pixel, we introduce an evolutionary method, the cooperative coevolution (CC) algorithm [6], into the adaptation loop. In CC, a solution (i.e., organism) is formed by selecting the best "individual" from a sub-population and concatenating it with individuals of other sub-populations. Each sub-population of individuals is maintained separately and does not represent a solution by itself. The solution is selected from a population of organisms (i.e., matrix ***M***) based on an evaluation function, referred to as the fitness function. The CC algorithm is given as follows:

Initialize sub-populations
Loop
Build organisms (e.g., solution space)
Evaluate and Store the best organism
For each subpopulation
Evolve each subpopulation
EndFor
Until stop Condition
Return the best organism (e.g., solution)

- **Fitness Function:** The CC algorithm requires an evaluation mechanism for selecting the best solution. We provide a suitable fitness function for this evaluation. The

fitness function integrates the statistics collected by the system and the physical models that are directed by the contextual information (environmental conditions). Let

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

be an individual's statistical estimation, where P(.) is the probability distribution function, S_{ob_j} is the jth observation in the past for sensor S, and G_j is defined as follows:

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

and n represents a window in the past. To tie the knot with the physics, we introduce the following function, named credibility function:

$$C_S = e^{-\alpha\left[\frac{1}{n}\sum_{j=1}^{n} G_j \frac{\left|S_{ob_j} - S_{P_j}\right|}{S_{ob_j} + S_{P_j}} + (1-G_j)\left(1 - \frac{\left|S_{ob_j} - S_{P_j}\right|}{S_{ob_j} + S_{P_j}}\right)\right]}$$

where SPj is physics-based prediction for sensor S (described below) and α is the adaptation rate. The credibility function adjusts the influence and role of the physics-based predictions into the adaptive loop. For example, if the prediction is close to observation, when it is expected, high credibility is assigned to that sensor. It is easy to verify for example in the extreme case when the predictions and observations match the value of CS will be 1 (or 100%). We can define the following fitness function:

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

The higher the value of F, a better representation is captured by the individual sensor. Note that all the equations are functions of time.

- **Physics-based Prediction:** The fitness function above requires a predicted value (S_{P_j}) for each sensing modality S. We have integrated a number of these models for both the visible and thermal IR, that can predict reflectance and thermal radiance for urban type materials [5]. We adapt the dichromatic reflection model [7] as follows:

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

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. C_b can be robustly calculated using singular value decomposition for a surface. Similarly, we use a thermal model based on the conservation of energy $E_{in} = E_{out}$, where E_{in} is the input energy flux due mainly to sun's irradiation and $E_{out} = E_{rad} + E_{cv} + E_{cd}$ where output loss is due to radiation, convection and conduction to the environment respectively. Various empirical models are obtained for each energy flux for thermal prediction.

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. Two cameras, a thermal camera operating at 7-13 μm and a web-cam operating in the visible range were utilized for data acquisition. The thermal camera was fully radiometric and the radiation-to-temperature conversion was done automatically by the camera for the default values of emissivity = 0.92, ambient temperatures = 280° K, distance to target = 100m, and humidity = 50%.

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 a background 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. The following parameters were used in the CC algorithm to update the background models:

Number of species = 4; Population size = 60; Crossover = Single point; Crossover rate = 0.8; Recombination rate = 0.7; Mutation rate = 0.01; Maximum number of generations = 60; Training data = 20 frames; Number of Gaussians per sensor = 3; $\alpha = 0.5$.

Detection Results: 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. These binary frames provide training data for the next background model update. Figure 2 shows several frames at different times of the day. It shows results for single sensors and the result of the fusion. To quantify the results, the following confusion matrix is given for each frame:

% moving obj correctly detected	% moving obj missed
% background missed	% background correctly detected

The results indicate that the fusion tracks the illumination and thermal changes in the scene. These changes are particularly adverse during sunset (e.g., frames 2422, 2676) and sunrise (e.g., frames 6820, 6890, 6954).

Another advantage of our method is shown when one sensor does not perform well under a particular instance. For example in frame 2676 lack of illumination in the scene caused the failure of the video channels (notice only the vehicles own lights are visible); on the other hand, frame 8646 represents the noon time when surface temperature can reach as high as the vehicle body's temperature where the lack of thermal contrast has caused the IR to fail in detecting the vehicle whereas the fused detection performed better than each sensor alone. It worth noting that the detection process is independent of the type of objects; for example, frames 6890, 6954 and 9350 include humans

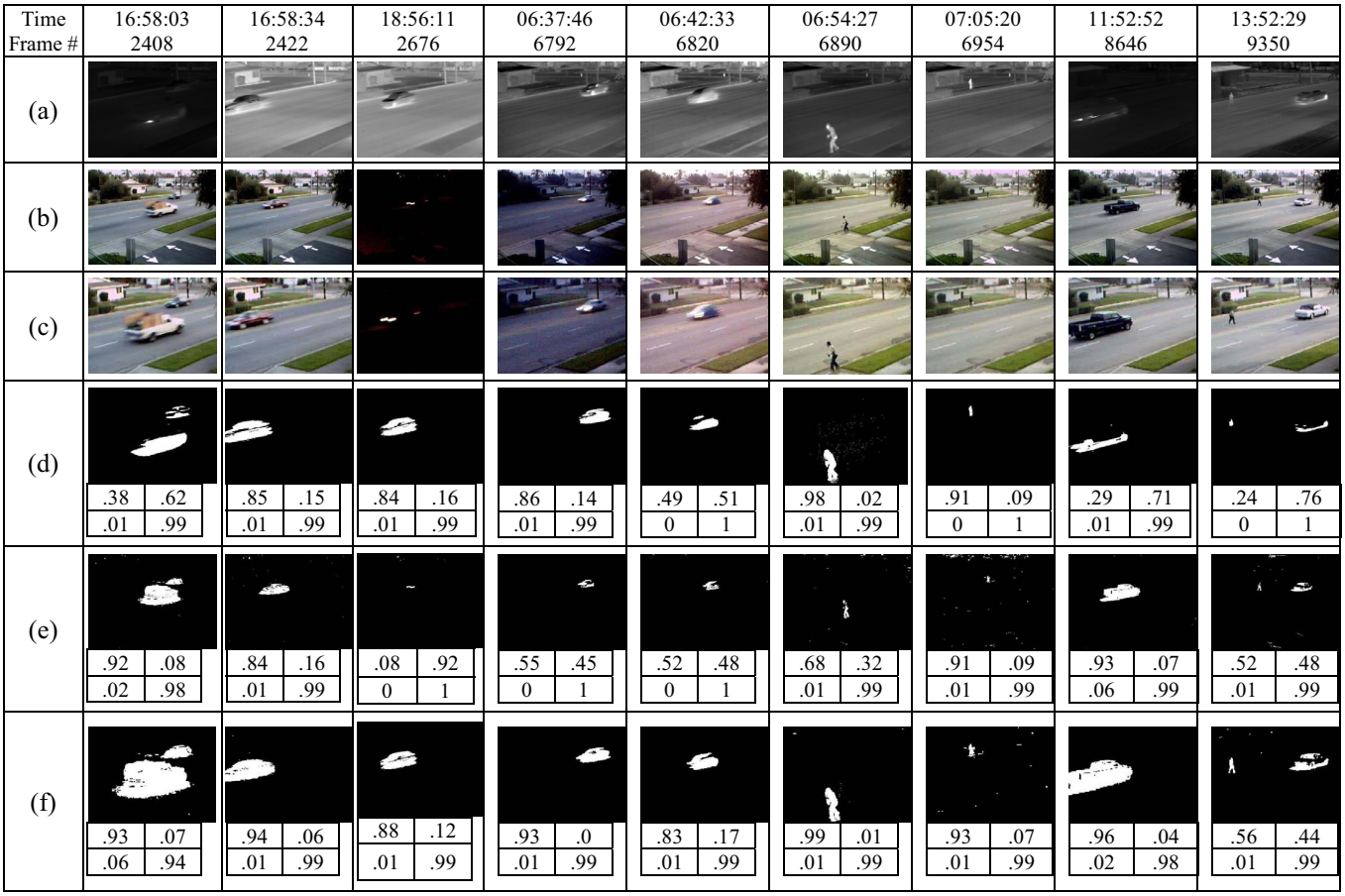

Figure 2. Selected frames for moving object detection during a diurnal cycle. (a) IR frames, (b) video frames, (c) registered video, (d) detected IR, (e) detected video, (f) detected in fused IR+ video.

at different distances as well.

5. Conclusions

We introduced a novel fusion technique that incorporated both a physics-based and statistical method for fusion. We introduced a cooperative coevolutionary algorithm in the adaptive phase and provided suitable evaluation function for optimally searching for the best mixture representation. The results show that our method robustly detects various objects while adapting to environmental changes. Our method can be further extended to model other extreme conditions such as snow, rain, etc.

ACKNOWLEDGEMENTS

This work was supported in part by grants F49620-02-1-0315 and DAAD19-01-0357; the contents and information do not necessarily reflect the position or policy of the US Government.

References:

[1] R. Mehorta "Establishing motion-based feature point correspondence," *Pattern Recognition*, Vol. 31, No. 1, pp. 23-30, 1998.

[2] I. Haritaoglu, D. Harwood, and L. Davis, "W4: real time surveillance of people and their activities," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22 (8), pp. 809-830, 2000.

[3] 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.

[4] S. Nadimi and B. Bhanu, "Multistrategy fusion using mixture model for moving object detection," *Proc. Intl. Conf. on Multisensor Fusion & Integration for Intelligent Systems*, pp. 317-322, August 2001.

[5] S. Nadimi and B. Bhanu, "Physics-based models of color and IR video for sensor fusion," *Proc. Intl. Conf. on Multisensor Fusion & Integration for Intelligentl Systems*, pp. 161-166, July 2003.

[6] 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.

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

[8] C. Stauffer and W.E.L. Grimson "Learning patterns of activity using real-time tracking," *IEEE Transaction on Pattern Analysis and Machine Intelligence Vol.* 22, No. 8, pp 747-757, Aug. 2000.

[9] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 780-785, July 1997.