Multistrategy Fusion Using Mixture Model For Moving Object Detection

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

Abstract

In a video surveillance domain, mixture models are used in conjunction with a variety of features and filters to detect and track moving objects. However, these systems do not provide clear performance results at the pixel detection level. In this paper, we apply the mixture model to provide several fusion strategies based on the competitive and cooperative principles of integration which we call OR, and, AND strategies. In addition, we apply the Dempster-Shafer method to mixture models for object detection. Using two video databases, we show the performance of each fusion strategy using receiver operating characteristic (ROC) curves.

Index terms: Fusion, Mixture Model, Dempster-Shafer.

1. INTRODUCTION

With the advent of newer, much improved and inexpensive imaging technologies, video has found its way into mainstream of computation and everyday life. Naturally, emerging technologies and advancements in signal/image processing and computer vision are providing applications that were not feasible a decade ago. A prevailing application is to use cameras to detect, recognize and track moving objects. For example, cameras are installed on a highway to inform any anomalies such as traffic jams. A typical parking lot can be surveyed, a train station can be monitored for vandalism, a high school perimeter can be checked for intruders and so on [1]. A complete automated surveillance system needs to address three major problems: Detection, Recognition and Event Handling. Regazzoni, et.al [2], provide a research overview in advanced video based surveillance systems. Motion detection is the key research area as the first step is applied to subsequent processing stages of recognition and event handling.

The state of the camera and the world can be divided into 4 categories: 1) Stationary Camera, Stationary Object (SCSO), 2) Stationary Camera, Moving Objects (SCMO), 3) Moving Camera, Stationary Object (MCSO) , 4) Moving Camera Moving Objects (MCMO). In most of the scenarios we mentioned, the SCMO is the most applicable.

One can simply apply, to a single frame, a feature-based segmentation algorithm and then find correspondence between frames to detect moving objects [3,4]. These techniques are not generally robust to illumination changes and noise. Since, video, at 30 frames a second provides a huge amount of data and statistics, one may be able to observe a great number of images in a short period of time and make some statistical estimation on how the scene is represented/modeled. More recently, mixture of Gaussian models [5] have been utilized to detect moving objects [6-8]. It is in this context that we would like to address the detection problem and how sensor fusion can enhance the results. The focus of this paper is the stage where moving objects must first be distinguished from their surrounding background.

The contribution of the paper is the development of sensor fusion techniques in the context of mixture model. We furnish detection results at the pixel level and show that our fusion strategies not only outperform a single sensor system, but also they compete with each other. Under different scenarios, our AND strategy outperforms the Dempster-Shafer (D-S) strategy. In addition, we provide the logic behind each strategy and in which strategy may be appropriate for various applications.

2. OBJECT DETECTION AND RELATED RESEARCH

Figure 1 shows a typical loop of a system using statistical modeling. The first step is to model each pixel. This is usually done by collecting statistics on some recent time window and estimating the distribution in some

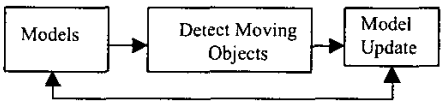

Figure 1. System for detecting moving objects

parametric form. Pixel distribution is estimated via mixture model and each component of the mixture can be estimated using clustering algorithms such as K-means or techniques such as Expected-Maximization. In the second step, as incoming pixels arrive, they must be checked against the background model. Usually a statistical test is used to classify the pixel as either a foreground or a background pixel. For example, a pixel value that is not within three standard deviation distance of any of the mixture models is considered foreground, otherwise, it is considered background. Depending on the outcome of this step, classifying pixels as foreground or background, the next step, model update, decides how to modify the parameters of the models to account for the new infor-

mation. The adaptive procedures account for the dynamics of the background, such as slow changes of sun's illumination or rapid appearance of cloud covers. This stage enables the system to adapt to certain changes in the scene and continue to function under environmental changes.

The result of the detection is normally a binary image (or a mask of the moving pixels) that is noisy and requires some kind of filtering. Most systems enhance the output by incorporating features and tracking algorithms. Even though good macro (Object) results have been reported at close ranges, it is not known whether the detection result is as good at the micro (pixel) level. The problem is that even though the whole system may work well, one may not know how much contribution each part of the system provides. For example, if we do not utilize any ad hoc filtering, it is possible to get erratic recognition and tracking results due to the noise.

In our research we investigate the performance of such algorithms at the pixel level and provide fusion algorithms operating at different points of receiver operating characteristics (ROC) curve.

2.1 Statistical Based Approaches

In a perfect static world, an image of a static background would look the same at all times. One can simply take a snapshot image of the background and then subtract it from the incoming images to distinguish new objects entering or moving in the scene. This simple technique has been used to detect objects [9]. However, the world is neither perfectly static nor noise free. Statistical models have been proposed to model the noise and dynamics of the scene.

One of the early works utilizing the statistical modeling of the background is *Pfinder* [6]. In *Pfinder*, people are detected and tracked in an indoor scene with controlled illumination environment. Each pixel in an image of a video stream is viewed as an independent statistical process; therefore, assuming Gaussian noise, a background pixel is modeled as

$$p(o) = \frac{exp\left[\frac{1}{2}(o-\mu)^T K^{-1}(o-\mu)\right]}{(2\pi)^{\frac{m}{2}}|K|^{\frac{1}{2}}}$$

where μ and K are the mean and covariance of the distribution of a pixel in YUV plane. *Pfinder* accommodates for the camera and ambient noise in an indoor environment; however, it cannot account for large dynamics (in the background) encountered in outdoor scenes where a slew of physical phenomenon such as wind, temperature, cloud covers, sun angle , rain, snow etc. can affect the sensor readings, not to mention the noise due to the sensor itself. A closely related project to *Pfinder*, called *VSAM* [7], tries to address these problems for an outdoor scenario where background could be highly dynamic. Like *Pfinder* it views each pixel as an independent statistical process; however, each pixel is modeled based on the fit of recent observations [1..t-1] to N Gaussian mixture models. Formally, if x_t is the intensity of a pixel at time t, then,

$$p(x_t) = \sum_{i=1}^{N} \pi_i G_i(x_t, \varphi)$$

where, $p(x_t)$ is the probability of observing value x at time t, π_i is the weight , or prior, of the G_i-th distribution and φ characterizes the distribution. New pixels are checked against the pdf and if they do not fit any of the model's, they are classified as foreground. Furthermore, incoming pixels can contribute to the models and each pixel's model can change and adapt to the dynamics of the background. It is worth noting that the results are usually given in the form of the number of correct people or moving objects detected. It is not clear what the performance is at the pixel level. We provide several sensor fusion techniques that could be applied to this statistical modeling. We observe that each fusion technique operates at a different point on the ROC curve. Moreover, Dempster-Shafer technique is utilized with the mixture model to provide a statistically sound fusion technique. .

3. TECHNICAL APPROACH

Our approach is similar to the system shown in Figure 1. We employ several fusion strategies to improve the results.

3.1 Motivation For Fusion

An operator of a surveillance system monitoring pedestrians in close view of a camera may tolerate many undetected pixels and some incorrectly detected pixels since the size of the objects compensate for it. On the other hand, an observation post cannot afford loosing any object pixels, since the objects it is monitoring are far away and they have only a few pixels on them. It may tolerate only a few incorrectly detected pixels, it is desired to have low false alarms while not loosing any of the object pixels.

Imagine now that we have several sensors looking at the same object and their signals are co-registered. In the pedestrian monitoring system, the operator may be satisfied if only one of the sensors strongly suggests that the pixel is background since there are plenty of opportunities to find object pixels and it is not desired to process too many noisy pixels. On the other hand, an observation post, will not like to miss an object that is about to harm it. Therefore, it is desired to have higher confidence that the object (pixel) is background, so one may require all or majority of the sensors to strongly agree that the observed pixel is background before dismissing it. This provides opportunity to incorporate fusion techniques into the

detection phase of a mixture model. Of course, we are making the following underlying asumptions, a) we have multiple sensors, b) each sensor signal is independent, and c) all sensor signals are registered at the pixel level.

3.2 *Fusion Approach*

Sensor fusion [10] has been exploited in many domains such as image enhancement , target detection, medical diagnostics etc. The methods proposed for fusion are as diverse as the applications. Some methods simply access and manipulate signals while others apply sophisticated statistical theories at the decision level. Brooks and Iyengar [11] describe three main sensor configurations, complementary, competitive, and cooperative. In our experiments we use sensors that have different spectral sensitivities to Red, Green and Blue components of the color camera signal. Each sensor is looking at the same phenomenon in the scene and is providing independent measurements; hence, sensor configuration can be thought of as either cooperative or competitive. The result of the measurements are in the form of the mixture of Gaussian probability density functions (pdfs); therefore, each sensor provides a set of models which represents some aspect of the world (dynamics of the background of the physical world).

Formally, the following procedure is performed :

Recent history of each pixel $\{X_1,...X_t\}$ is modeled by a mixture of K Gaussians :

$$P(X_t) = \sum_{i=1}^{K} w_{i,t} \times \eta\left(X_t, \mu_{i,t}, \Sigma_{i,t}\right)$$

$$\eta(X_t, \mu, \Sigma) = \frac{1}{2\pi^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu_t)^T \Sigma^{-1}(X_t-\mu_t)}$$

Mixture Model for a Pixel
For each Pixel
- Collect initial data (**F** Frames)
- Cluster data to **K** clusters (K-means)
- For each Cluster Fit a Gaussian (μ, Σ)

end

3.2.1 *Fusion Strategies*

After the initial model building, we are at the second phase, the detection phase (see Figure 1). Unlike other approaches, the sensitivity of our system can be determined based on the user's need. We provide the following fusion strategies and explain the logic behind them.

G_c = { Gaussian Parameters } for channel **c** where **c** $\in$ {R,G,B},
P_R = R Value of the incoming pixel P,
P_G = G Value of the incoming pixel P,
P_B = B Value of the incoming pixel P,
S_3 : Means within *3 Standard* Deviation (of R,G, or B),
t : Some threshold.

Strategy 1 –(OR) Competitive
Let $Prob_c(P_c)$: probability of pixel P for channel **c** where **c** $\in$ R,G,B
if $\exists$ **c** , such that $Prob_c(P_c) \geq t$ then **P** $\in$ Background
Strategy 2 –(AND) Cooperative
if $\forall$ c , $Prob_c(P_c) \geq t$ then **P** $\in$ Background
where $Prob_c(P_c)$ can be estimated from the current mixture model.
Strategy 3 :
if **(P_R S_3 G_R) OR (P_G S_3 G_G) OR (P_B S_3 G_B) Then P** $\in$ Background
Strategy 4 :
if **(P_R S_3 G_R) AND (P_G S_3 G_G) AND (P_B S_3 G_B) Then P** $\in$ Background
Strategy 5 :
if : **{ (P_R S_3 G_R) AND [(P_G S_3 G_G) OR (P_B S_3 G_B)] OR (P_G S_3 G_G) AND [(P_B S_3 G_B) OR (P_R S_3 G_R)] OR (P_B S_3 G_B) AND [(P_R S_3 G_R) OR (P_G S_3 G_G)] } Then P** $\in$ Background

The following observation can be made from the above strategies. *First*, strategy 4 is a special case of strategy 5. *Second*, both strategy 3 and strategy 4 are really special cases of strategy 1 and strategy 2. One may ask, then why we have introduced strategies 3,4, and 5. In real time applications such as video surveillance, one can save a great deal of computation time by making simple standard deviation test. However, Startegies 1 and 2 provide (t) which is a sensitive parameter that can be adjusted by the user to operate the detection module at different points on the ROC curve. Generally the class of strategies such as strategy 2 are more desirable where penalty for missing object is high. This is in contrast with classes falling into strategy 1 when we expect lower detection.

3.2.2 *Dempster-Shafer (D-S) Strategy*

So far, we have assumed two classes, foreground and background. Our decision, proposition, has been simple, a pixel that is not foreground is considered background or vice versa. In another words :
prob(backround) + prob(foreground) = 1.
Since we are dealing with pdfs, there are some uncertainty issues involved. In other words, the decision that a pixel is foreground or background is uncertain. We can only classify the data according to our confidence. Dempster-Shafer [13] provide a natural approach to sensor fusion. In the context of mixture model, the threshold, **t**, can be associated with confidence interval. For example, if our threshold is 3 standard deviation, then the confidence interval will be approximately 99.8%. Confidence interval values can be readily calculated given the threshold. In statistics, the confidence interval is the complement of the critical region; we associate the criti-

cal regions to the foreground and confidence intervals to uncertainty. Within the uncertainty of our model we can then define the probability of the background class. In other words, the closer the observed value of a pixel is to the mean, the higher its background probability and the lower the uncertainty probability, see Figure 2.

With this notion, each pdf contributes not only to the amount of belief that the pixel is background but also to uncertainties of the belief (or disbelief). Dempster-Shafer rule of combination can then be applied to each pixel's probabilities and decision will be based on the highest belief. Formally, we define the following: let θ be a frame of discernment, which is the set of mutually exclusive atomic propositions, exactly one of which corresponds to the truth, Foreground or Background. Then the probability mass function **m** in the power set of θ is defined as follows : Let A be a proposition,

1) $0 <= m(A) <= 1$ for every A in θ

2) $m(\phi) = 0$

3) $\Sigma\, m(A) = 1 \quad : \forall A \subseteq \theta$

A Belief function B is then defined according to the following rules :1) $Bel(\phi) = 0$; 2) $Bel(\theta) = 1$

3) $A \subseteq B \subset \theta \rightarrow Bel(A) \leq Bel(B)$

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

The following, Dempster's rule of combination, for three sensor is then defined :

$$m_{123}(D) = \frac{\sum_{A \subseteq \theta, B \subseteq \theta, C \subseteq \theta, A \cap B \cap C = D} m_1(A)\, m_2(B)\, m_3(C)}{\sum_{A \subseteq \theta, B \subseteq \theta, C \subseteq \theta, A \cap B \cap C \neq \phi} m_1(A)\, m_2(B)\, m_3(C)}$$

Where $\{A,B,C,D\} \subseteq \theta$ are atomic propositions. The atomic propositions in our case are Foreground ***F***, Background ***B*** and Uncertainty ***U***. We obtain each sensor's probability contribution to each of the propositions, as follows :

$$p(F) = 2 \times \sum_{i=1}^{n} \pi_i \int_{-\infty}^{t_1} N_i(\varphi)$$

$$p(B) = \begin{cases} 2 \times \sum_{i=1}^{n} \pi_i \int_{t_1}^{x} N_i(\varphi) & \text{if } t_1 \leq x \leq t_2 \\ 0 \quad \text{Otherwise} \end{cases}$$

$$p(U) = 2 \times \sum_{i=1}^{n} \pi_i \int_{x}^{0} N_i(\varphi)$$

where n is the number of mixture models, $N_i(\varphi)$ is the ith, Gaussian density function in the mixture model with parameters φ, π_i is the weight or prior of the ith pdf in the mixture and is measured by the cluster size contributing to that pdf, t is a threshold and x is the observed value of the pixel. As shown in Figure 2, for a given threshold, the larger the probability of the background, the smaller the uncertainty is and vice versa. Given a value of x for a pixel we can compute the probabilities (regions in Figure 2) and apply the Dempster-Shafer combination rule to classify the pixel as background or foreground.

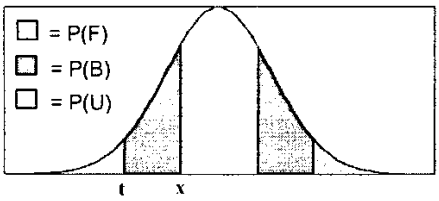

Figure 2. Associated probabilities for background P(B), foregrond P(F), and uncertainty P(U), x = pixel value, t = threshold, for a single Gaussian density in the mixture. (Note: for simplicity only a single Gaussian is drawn. The probability regions for all Gaussians in the mixture model are defined similarly).

An example:

For sake of simplicity and clarity we use two sensors with n=1. Let's say two sensors, make the following probability observations about a pixel.

Sensor 2

U=0.1			
B=0.4			
F=0.5			
	F= 0.1	B = 0.2	U=0.7

Sensor 1

F, B and U are probabilities of Foreground, Background and Uncertainty, respectively. They correspond to the lighter shade, darker shade and white region of Figure 2 respectively. The shaded area in the table above represents contradiction; since contradictions do not represent any reality of the world they are not considered in the belief functions. Applying the dempster's combination rule: $m_{12}(B) = 0.44$ and $m_{12}(F) = 0.47$ and the pixel is classified as forground.

4. EXPERIMENTAL RESULTS

4.1 Data

We used two outdoor video sequences taken at different time, with different object distance and environmental conditions. We opt for short video shots where environmental conditions do not radically change so, there is no need for adaptation for the number of Gaussians used in the background model. The first sequence of 900 frames, (Figure 3) was taken in an afternoon sunny summer day

with subject in close proximity to the camera. The second sequence of 1200 frames (Figure 4) was taken on an overcast day and the moving object was approximately 300 feet away. We obtained the ground truth (Figure 5) by randomly selecting a frame from the MPEG stream, then carefully drawing a contour over the moving object(s) circumscribing all the pixels that had changed including moving shadows. A Sony DCR-VX1000 digital camera with dichroic prism and three CCD sensors, each with different spectral sensitivity to the red, green and blue region of the spectrum, was utilized.

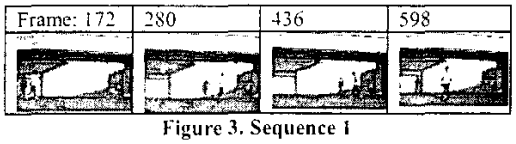

Figure 3. Sequence 1

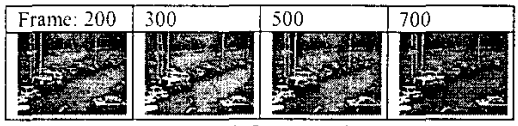

Figure 4. Sequence 2

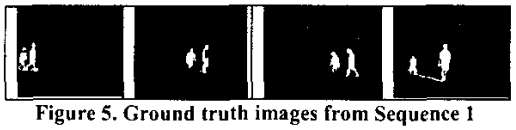

Figure 5. Ground truth images from Sequence 1

4.2 Performance Measure

We show the receiver operating characteristic curves for detection based on mixture model and our fusion strategies. Let

N = Number of moving object pixels in Ground Truth image,

α= Number of moving object pixels that the algorithm did not detect (i.e., missed),

B = number of Background pixels (Ground Truth),

β = number of background pixels that were detected as foreground.

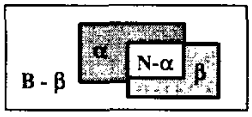

Figure 6. Illustration of background and foreground resgions.

Therefore, confusion matrix can be given as :

	Ground Truth Foreground	Ground Truth Background
Detected Foreground	N - α	β
Detected Background	α	B - β

We can define the Probability of detection and Probability of false alarms as :

$$\mathbf{Pd = (N - \alpha) / N} \qquad \mathbf{Pf = \beta / B.}$$

We used the above equations to obtain the ROC curves for detection performance of a single sensor, and using OR, AND and D-S strategies.

4.3 Experiments

We have developed a detection system based on the mixture model described previously. In sequence 1 (Figure 3), the first 120 frames were used for initial clustering and building the mixture model. The number of Gaussians, were user selectable; we experimented with 2, 5 and 10. A binary mask of the moving object and its shadow was then obtained for four randomly selected frames (Figure 5). Different Fusion algorithms were then run on the whole MPEG video sequence. For comparison we also implemented and tested the Dempster-Shafer (D-S) strategy. For each strategy and a single sensor we obtained ROC curves for four selected frames for sequence 1 (Figure 3).

Figure 7, indicates that both the OR and the AND fusion strategies outperformed the single sensor. We observed that the AND strategy had detected the most number of object pixels with comparable false alarms.

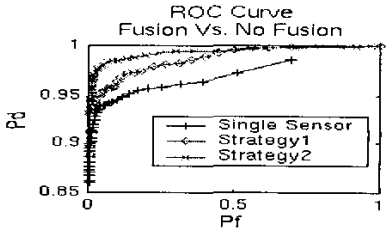

Figure 7. Performance of single sensor vs. multiple sensors with OR (Strategy1) and AND (Stategy2)

As indicated by the ROC curves (Figures 11 and 12), the D-S strategy performed between the OR and the AND strategies. Moreover, the AND strategy detected more object pixels while still holding low false alarms compared to the other strategies. This is evident in the right most frame of Figure 8 and Figure 9, corresponding to frame 598 of Figure 3. The results were far more dramatic when object was relatively small. The AND strategy detected more object pixels than the other two with some small increase in the false alarm rate (Figure 10). These strategies provide the opportunity for the user to operate the system at different sensitivity points on the ROC curve. In general, the OR and the D-S strategy provided less noise but detected less object pixels vs. the AND strategy where most object pixels were detected but at a higher false alarm rate. This indicates that the AND strategy may be more suitable to scenes where objects are

small and the OR strategy may be more suitable to scenes with large objects. D-S strategy, on the other hand, provides a compromise between the two.

Figure 8. Result of the OR strategy for sequence 1

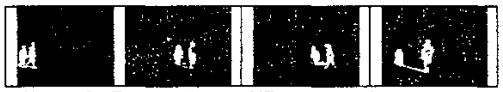

Figure 9. Result of the AND strategy for sequence 1

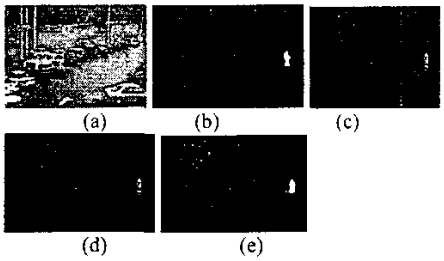

Figure 10. a) A frame from sequence 2, b) ground truth c) OR strategy, d) D-S strategy, e) AND strategy results.

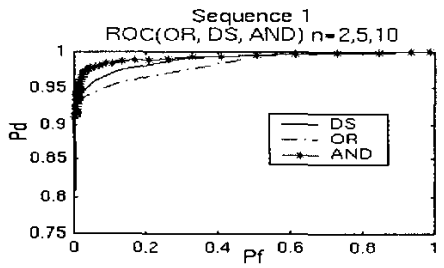

Figure 11. ROC curves for OR, AND and D-S strategies, averaged over n = 2,5,10 mixture models and frames.

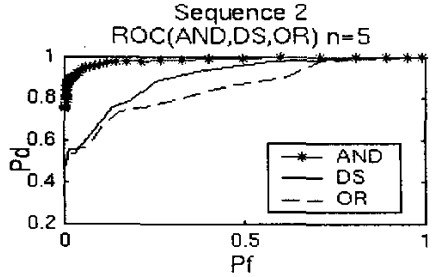

Figure 12. ROC Curve for sequence 2 for various strategies.

5. CONCLUSIONS

In this paper, we provided three basic fusion strategies in the context of mixture model for detection in outdoor scenes. Our results show that fusion is not only an effective approach to mixture model based detection system but also choosing different strategies affects the outcome of the final detection. We have shown that choosing a competitive, OR, vs. cooperative, AND and D-S, strategies, do make a difference. High detection at very low false alarm is achieved for the AND strategy. Our results also indicate that the AND strategy is a viable fusion strategy for relatively small moving objects at intermediate distances. A simple size filter can also be utilized to eliminate isolated pixels. This further reduces false alarms.

In our future work, we plan to investigate the effect of optimally choosing the number of models, and provide more sensors such as infrared (IR). We are also currently investigating the problem of moving shadows and how to eliminate them.

References

[1] A. McLeod, "The impact and effectiveness of low-cost automated video surveillance systems," *Proc. of IEEE Intl. Carnahan Conf. on Security Technology*, pp 204-11, 1996.

[2] C. Regazzoni, F. Gianni, and G. Vernazza, Advanced Video Based Surveillance Systems, Kluwer Academic, 1999.

[3] MMD. Viva, and C. Morrone, "Motion Analysis by feature tracking," *Vision Research*, Vol. 38, pp 3633-53, 1998.

[4] R. Chellappa, and Q. Zheng, "Automatic Feature Point Extraction and Tracking in Image Sequences for Arbitrary Camera Motion," *Intl. Journal of Computer Vision*, Vol. 15, pp 31-76, 1995.

[5] G.J. McLachlan and K.E. Basford, Mixture Models Inference and Applications to Clustering, Marcel Dekker, New York, 1988.

[6] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *Pattern Analysis and Machine Intelligence*, Vol. 19, (No. 7), pp 80-85, July 1997.

[7] W.E.L. Grimson, C. Stauffer, R. Romano, L. Lee, P. Viola, and O. Faugeras, "Forest of Sensors: Using adaptive tracking to classify and monitor activities in a site," *Proc. IEEE conference on Computer Vision and Pattern Recognition*, pp 22-29, 1998.

[8] M. Seki, H. Fujiwara, and K. Sumi, "A robust background subtraction method for changing background," *Proc. 5th IEEE Workshop on Applications of Computer Vision*, p.207-13, Dec. 2000.

[9] J.A. Freer, B.J. Beggs, H.L. Fernandez-Canque, and A. Goryashko, "Automatic Recongition of Suspicious Activity for Camera Based Securiy Systems," *IEE European Convention on Security and Detection*, pp 54-58, May 1995.

[10] R.C. Luo, and M.G. Kay,"A review of high level multisensor fusion: approaches and applications," *16th Intl. Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp 25-31, 1999.

[11] R.R. Brooks, and S. Iyengar, Multi-Sensor Fusion: Fundamentals and Applications with Software, Prentice Hall, 1997.

[12] B.V. Dasarathy, Decision Fusion, IEEE Computer Society Press, Los Alamitos, CA, 1994.

[13] A.P. Dempster, "A Generalization of Bayesian Inference", *J. Royal Statistical Society*, Ser. B. Vol. 30, pp 205-247, 1968.