SAND2013-1873C

One-Class Multiple-Look Fusion: A theoretical comparison of different approaches with examples from infrared video

Anonymous CVPR submission

Paper ID ****< replace **** here and in header with paperID>

Abstract

Multiple-look fusion is quickly becoming more important in statistical pattern recognition. With increased computing power and memory one can make many measurements on an object of interest using, for example, video imagery or radar. By obtaining more views of an object, a system can make decisions with lower missed detection and false alarm errors. There are many approaches for combining information from multiple looks and we mathematically compare and contrast the sequential probability ratio test, Bayesian fusion, and Dempster-Shafer theory of evidence. Using a consistent probabilistic framework we show how to transform results from one approach to the other and show results for an application in infrared video classification.

1. Introduction

There have long been multiple competing approaches for accomplishing multiple look sensor fusion. By multiple look fusion we assume we can make multiple measurements on an object, as it passes through the field of view of the sensor. For example, multiple frames in an infrared video or the extraction of multiple high-resolution-range profiles from ground-moving-target-indicator radar.

Here, we will theoretically compare and contrast three data fusion approaches: sequential probability ratio test (SPRT), Bayesian, and Dempster Shafer (DS). To accomplish this we will use a common probabilistic framework that is useful in real-world pattern recognition problems in unconstrained environments. Our goal is not to say one is better than the other, but to find commonalities between the various approaches and use one approach to find insights into others. This can eventually lead to a unified approach in sensor fusion.

2. One Class Classifiers

A structure one might choose for a classification problem is based on a *one-class* classifier [1]. For one class classifiers, we are interested in one specific target θ_1 represented by the alternative hypothesis H_1, and the null hypothesis H_0 represents the non-target $\overline{\theta_1}$ class. While this might seem like an over simplification, one could argue that for a multi-class problem with other targets of interest $\theta_2,...,\theta_m$ one would design a one-class classifier for each of them. For the θ_1 one-class classifier we can further divide the nontargets into two groups: the other targets of interest $\theta_2,...,\theta_m$ and the unknown class θ_0. This allows us to further distinguish between two types of false alarm errors: *between-class* and *out-of-class*. Between-class errors occur when alarming on another target $\theta_2,...,\theta_m$ by calling it the target θ_1. Out-of-class errors occur when alarming on an unknown signature θ_0 by calling it the target θ_1. In making any decision, we want to control two types of errors: *missed detection errors* and *false alarm errors*. Missed detection errors result from missing a target signature by calling it a nontarget, and false alarm errors result from alarming on a nontarget signature by calling it a target.

For example, suppose we are interested in classifying moving objects in infrared video as humans, vehicles, or unknown. Here, θ_1 would represent the human class and θ_2 the vehicle class. The unknown class θ_0 would represent all moving objects not in θ_1 or θ_2. This would be wind-blown objects (tumble weed, boxes, trash cans, etc.) or animals. These unknown moving objects, a possible source of the out-of-class errors, are a significant problem in real-world pattern recognition problems in unconstrained environments. A Bayesian classifier approach designed for the human vs. vehicle problem, while minimizing the between-class errors, would require models of all the possible objects that could be imaged by the sensor to control the out-of-class errors. Otherwise it would classify an animal as human or vehicle. Modeling "the whole world" of possible objects is untenable for most realistic systems deployed in unconstrained environments. Instead, we use a goodness of fit (GOF)

classifier to control the out-of-class errors, and power analysis [8] to model the unknown class.

Whereas Bayesian classifiers minimize the between-class error, they do nothing to control the out-of-class errors. Figure 1 illustrates this potential problem. The figure shows a two-dimensional feature space, with samples from two targets: Target *A* represented by stars and Target B represented by filled circles. Assuming normal distributions and equal covariance matrices for the targets, the Bayes decision boundary has a linear form (Figure 1a). Whereas the Bayes classifier minimizes the between-class errors of the *A* and the *B* targets, it does not control the out-of-class errors caused by unknown objects represented by "*x*" symbols. Depending on which side of the boundary the nontarget falls, the classifier will assign the unknown to one of the known classes and make an out-of-class error. Figure 1b shows a GOF classifier that tries to surround the target class. Here, the unknown objects, that have widely differing features from the target class ("*x*" symbols), will be classified correctly. In general, the GOF classifier has improved out-of-class errors, but the between-class errors will increase, since it is not necessarily an optimal Bayes classifier.

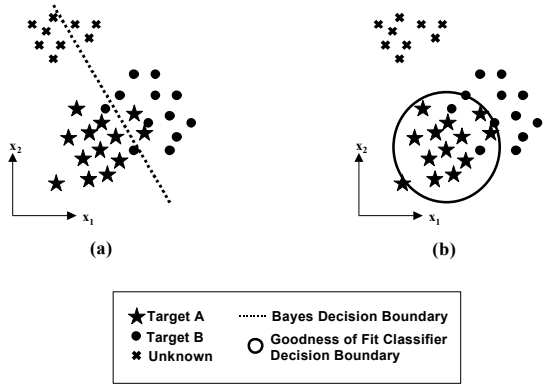

Figure 1: Comparison of Bayes and goodness of fit (GOF) classifiers. (a) Bayes classifier. (b) GOF classifier.

3. Probabilistic Framework

We will use the same probabilistic framework for comparing and contrasting the fusion algorithms. We start by assuming a stream of observations represented as random variables $x_1, x_2, \ldots$ These samples result from the best match scores in the GOF metric. We also assume knowledge of the functions $p(x_i \mid T)$ and $p(x_i \mid \overline{T})$ which represent the probability density function (PDF) of an observation x_i, given the target T and the nontarget $\overline{T}$, respectively. The PDF's can be discrete or continuous and be determined through theoretical means or empirical modeling.

The PDF $p(x_i \mid T)$ is usually straight forward to determine, since one has knowledge of the target of interest. This information can come from data collections or modeling and simulation using CAD descriptions and a physics-based sensor signature prediction software such as Irma [9]. The PDF $p(x_i \mid \overline{T})$ is usually more problematic. One approach for modeling the nontarget class uses *statistical power analysis* [8] to model the *worst case nontarget*. The approach has some similarities to that taken by [1] for modeling composite hypotheses by determining the *least favorable choice*. Power analysis assumes the tested effect is linear and the measured effect size (small, medium or large) is known. Typically, power analysis allows the statistician to determine if enough samples were collected to give the test a high power. While the exact form depends on $p(x_i \mid T)$, we show an example from [5] where $p(x_i \mid T)$ is $N(0,1)$. This is convenient in problems where the central limit theorem can be applied. In [5], it was shown that the worst case nontarget PDF is $N(\mu_N, 1)$. Here the location parameter μ_N represents the smallest acceptable effective difference between the target and nontarget.

4. Sequential Probability Ratio Test

One approach for combining decisions as they become available is to use the Wald sequential hypothesis test or SPRT [11]. After n observations the likelihood ratio is:

$$\Lambda(n) = \prod_{i=1}^{n} \lambda_i, \quad \lambda_i = \frac{f(x_i \mid T)}{f(x_i \mid \overline{T})} \tag{1}$$

Often it is more numerically convenient to work with the log-likelihood:

$$Z(n) = \log(\Lambda(n)) = \sum_{i=1}^{n} z_i, \quad z_i = \log\left(\frac{p(x_i \mid T)}{p(x_i \mid \overline{T})}\right) \tag{2}$$

The SPRT uses two decision boundaries (a, b) to make a decision:

$$\begin{array}{ll} \text{Reject } H_0 & \text{If } Z(n) \geq a \\ \text{Accept } H_0 & \text{If } Z(n) \leq b \\ \text{Get more data} & \text{If } b \leq Z(n) \leq a \end{array} . \tag{3}$$

One desirable property of the SPRT is that the decision boundaries can be determined from the chosen error rates. Thus, these decision boundaries can be obtained using the desired false alarm rate, α, and the missed detection rate, β:

$$a = \log\frac{1-\beta}{\alpha} \text{ and } b = \log\frac{\beta}{1-\alpha}. \tag{4}$$

It has been shown that the SPRT, on average, uses the smallest number of observations to make a decision [11]. It is interesting to note that the a-priori probabilities $p(T)$ and $p(\overline{T})$ do not appear in the log-likelihood ratio (2).

5. Bayesian Fusion

A sequential update formula can be derived from Bayes formula:

$$p(T \mid x_1) = \frac{p(T)p(x_1 \mid T)}{p(T)p(x_1 \mid T) + (1 - p(T))p(x_1 \mid \overline{T})} \tag{5}$$

The quantity $p(T \mid x_1)$ is the likelihood of target after one observation x_1. By substituting $p(T \mid x_1)$ for the prior $p(T)$ and a new observation x_2 for x_1 in (5) we get the likelihood of a target after two observations as:

$$\frac{p(T)p(x_1 \mid T)p(x_2 \mid T)}{p(T)p(x_1 \mid T)p(x_2 \mid T) + (1 - p(T))p(x_1 \mid \overline{T})p(x_2 \mid \overline{T})} \tag{6}$$

Using $B(n)$ to represent the Bayesian likelihood after n observations or $p(T \mid x_1 \ldots x_n)$ we get

$$Y(n) = \frac{\lambda_0 \Lambda(n)}{\lambda_0 \Lambda(n) + 1} \tag{7}$$

where $\Lambda(n)$ is the Wald likelihood ratio in (1) and λ_0 represents the a-priori likelihood ratio:

$$\lambda_0 = \frac{p(T)}{(1 - p(T))}. \tag{8}$$

The Bayesian likelihood is always between 0 and 1. By (1) and the definition of a PDF $0 \le \Lambda(n) < \infty$, so as $\Lambda(n) \to \infty$ then $Y(n) \to 1$ and if $\Lambda(n) = 0$ then $Y(n) = 0$. Also if the PDF's are discrete then we can compute a Bayesian probability of a target. We can use the SPRT stopping conditions to determine thresholds on the Bayesian likelihood. If

$$a = \log(A) \text{ and } b = \log(B) \tag{9}$$

then A and B are the upper and lower stopping conditions for $\Lambda(n)$, respectively. Thus the Bayesian stopping rule becomes:

$$\begin{array}{ll} \text{Reject } H_0 & \text{If } Y(n) \ge C \\ \text{Accept } H_0 & \text{If } Y(n) \le D \\ \text{Get more data} & \text{If } C \le Y(n) \le D \end{array} \tag{10}$$

where

$$C = \frac{\lambda_0}{1 + \lambda_0} A = \frac{\lambda_0}{1 + \lambda_0} \frac{(1 - \beta)}{\alpha} \tag{11}$$

and

$$D = \frac{\lambda_0}{1 + \lambda_0} B = \frac{\lambda_0}{1 + \lambda_0} \frac{\beta}{1 - \alpha}. \tag{12}$$

Note the $\lambda_0 / (1 + \lambda_0)$ basically tweaks the threshold according to the a-priori information. For $\lambda_0 > 1$ the thresholds will go lower to make target calls slightly more probable and if $\lambda_0 < 1$ the thresholds will go higher to make nontarget calls slightly more probable.

6. Dempster-Shafer Theory of Evidence

The Dempster Shafer (DS) theory [10] is a mathematical theory of evidence that allows combining evidence from different sources to arrive at a degree of belief. It models uncertainty by not requiring one to assign all of one's belief to a proposition.

The main assumption we make is that evidence is *consonant*. This allows us to use the probabilistic framework that we established in section 3. Shafer defines consonant evidence as evidence that points in a single direction and only varies in the precision of focus [10]. This fits well with the GOF metric. The GOF describes the difference between stored knowledge, for example a template of the target, and the measured data. Thus it points only in the direction and focus of the hypothesis represented by the stored knowledge.

For the one-class problem the frame of discernment is $\Theta = \{T, \overline{T}\}$. From [10] the support function for the target T is

$$\left.\begin{array}{l} m_x(\overline{T}) = 0 \\ m_x(T) = 1 - \dfrac{p(x \mid \overline{T})}{p(x \mid T)} \\ m_x(\Theta) = \dfrac{p(x \mid \overline{T})}{p(x \mid T)} \end{array}\right\} \text{if } \frac{p(x \mid T)}{p(x \mid \overline{T})} > 1 \tag{13}$$

and the support for the nontarget $\overline{T}$ is

$$\left.\begin{array}{l} m_x(\overline{T}) = 1 - \dfrac{p(x \mid T)}{p(x \mid \overline{T})} \\ m_x(T) = 0 \\ m_x(\Theta) = \dfrac{p(x \mid T)}{p(x \mid \overline{T})} \end{array}\right\} \text{if } \frac{p(x \mid \overline{T})}{p(x \mid T)} > 1. \tag{14}$$

where $m(\cdot)$ represents the DS basic probability assignment (BPA) function and $m_x(\Theta)$ represents the amount of uncertainty in the observation x.

For what we want to show it is simpler to work with DS's weight of evidence. If $m(A)$ represents the BPA for the proposition $A \subseteq \Theta$ then weight of evidence $w(A)$ is:

$$w(A) = -\log(1 - m(A)). \tag{15}$$

In terms of weight of evidence equation (13) becomes:

$$\left.\begin{aligned} w_x(\bar{T}) &= 0 \\ w_x(T) &= \log\left(\frac{f(x\,|\,T)}{f(x\,|\,\bar{T})}\right) \end{aligned}\right\} \text{if } \frac{f(x\,|\,T)}{f(x\,|\,\bar{T})} > 1 \qquad (16)$$

and equation (14) becomes

$$\left.\begin{aligned} w_x(\bar{T}) &= \log\left(\frac{f(x\,|\,\bar{T})}{f(x\,|\,T)}\right) \\ w_x(T) &= 0 \end{aligned}\right\} \text{if } \frac{f(x\,|\,\bar{T})}{f(x\,|\,T)} > 1 \qquad (17)$$

Using equation (2)

$$\left.\begin{aligned} w_x(\bar{T}) &= 0 \\ w_x(T) &= z \end{aligned}\right\} \text{if } \frac{f(x\,|\,T)}{f(x\,|\,\bar{T})} > 1 \qquad (18)$$

and

$$\left.\begin{aligned} w_x(\bar{T}) &= -z \\ w_x(T) &= 0 \end{aligned}\right\} \text{if } \frac{f(x\,|\,\bar{T})}{f(x\,|\,T)} > 1 \qquad (19)$$

where z represents the log-likelihood ratio in the SPRT. As long as we separate the evidence and combine only evidence supporting the same proposition, then we have the homogenous weight of evidence combination rule where the weights of evidence combine additively. For the total support of the target class T, let w^+ represent the total amount of positive weight of evidence. Similarly define w^- for the total support for the non-target class $\bar{T}$. Here

$$w^+ = \sum_{i=1}^{n} w_{x_i}(T) = \sum_{i=1}^{n} z_i, \text{ for } \frac{f(x\,|\,T)}{f(x\,|\,\bar{T})} > 1 \text{ or } z_i > 0 \qquad (20)$$

and

$$w^- = \sum_{i=1}^{n} w_{x_i}(\bar{T}) = \sum_{i=1}^{n} z_i, \text{ for } \frac{f(x\,|\,\bar{T})}{f(x\,|\,T)} > 1 \text{ or } z_i < 0 \qquad (21)$$

Combining conflicting weights (w^+ and w^-) of evidence cannot be done by simple addition. From [10] the weight of evidence for the contradictory propositions T and $\bar{T}$ becomes:

$$w(T) = \log\left(\frac{e^{w+} + e^{w-} - 1}{e^{w-}}\right) \qquad (22)$$

and

$$w(\bar{T}) = \log\left(\frac{e^{w+} + e^{w-} - 1}{e^{w+}}\right). \qquad (23)$$

Using (15) in reverse we can get the corresponding BPA for propositions T and $\bar{T}$

$$m(T) = \frac{e^{w+} - 1}{e^{w+} + e^{w-} - 1} \qquad (24)$$

and

$$m(\bar{T}) = \frac{e^{w-} - 1}{e^{w+} + e^{w-} - 1} \qquad (25)$$

The DS uncertainty is the BPA assigned to Θ or

$$m(\Theta) = \frac{1}{e^{w+} + e^{w-} - 1} \qquad (26)$$

Since $m(T)+m(\bar{T})+m(\Theta)=1$ for a one-class classifier. Note the uncertainty of the system's belief is driven down to 0 as evidence is collected or as w^+ and/or w^- increase. Thus high but equal w^+ and w^- would give low uncertainty, but an uninformed decision.

The most obvious decision rule is

$$\begin{aligned} &\text{Decide } T \quad \text{If } w(T) - w(\bar{T}) \geq 0 \\ &\text{Decide } \bar{T} \quad \text{otherwise} \end{aligned} \qquad (27)$$

For the one-class problem, this turns out to be equivalent to the three unambiguous decision rules proposed by Kim in [4]. After some algebraic manipulation one can show that (27) is equivalent to

$$\begin{aligned} &\text{Decide } T \quad \text{If } w^+ - w^- \geq 0 \\ &\text{Decide } \bar{T} \quad \text{otherwise} \end{aligned} \qquad (28)$$

This is equivalent to a forced SPRT decision if one is unwilling or unable to wait for any more observations.

7. Results

We show results on a one-class problem for video motion classification (VMC). The target class T represents upright-human dismounts and the nontarget class $\bar{T}$ is any other mover detected by the video motion detection (VMD) algorithm. The imager is an uncooled DRS E3500 infrared camera with 320×240 resolution and 8-bit precision. VMD was accomplished with background subtraction [1] and tracking with an alpha-beta tracker. The features for VMC were histograms of oriented gradients (HOG) [2] and the GOF metric was based on a multinomial pattern matching (MPM) [6].

Figure 2 shows a mosaic of chips (subimages containing the detection) collected of a runner at about 135 meters. The chips are of size 29×24 which is very small, on the order of 100 pixels on target. Performance on only one observation per frame is mediocre, especially when the runner is obscured by a brighter object as seen in the first row of chips in the mosaic.

Figure 3 shows the result of multilook fusion using the SPRT, Bayesian Fusion, and DS on the outputs of the MPM GOF classifier. Since the MPM is designed to produce $N(0,1)$ scores for HOG features from a target, $p(x_i\,|\,T)$ is set to $N(0,1)$. As discussed in Section 3 we use $N(\mu_N,1)$ for $p(x_i\,|\,\bar{T})$ where μ_N is empirically set to 5. The desired error rates α and β are set to 1×10^{-3}.

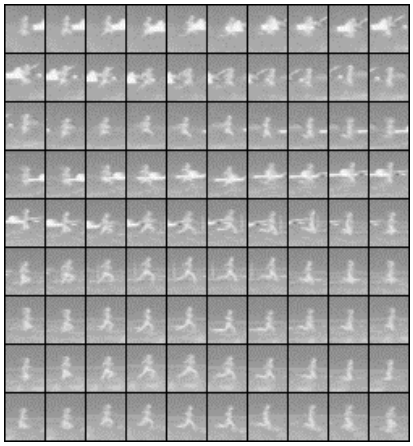

Figure 2: Infrared detections of a runner tracked in an uncooled infrared video imager.

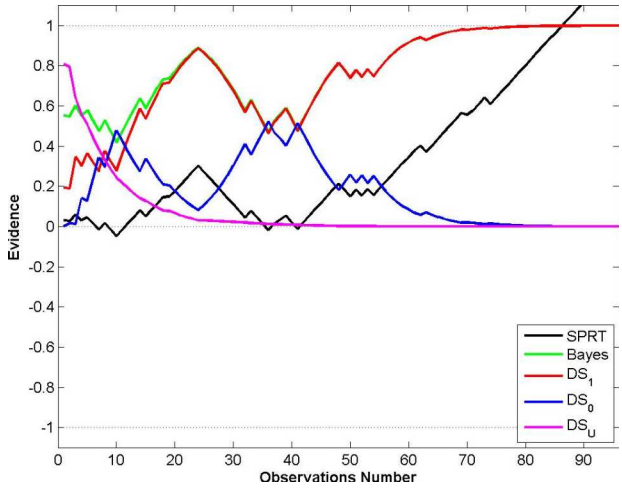

Figure 3: Results of different multilook fusion algorithms of the runner GOF scores from runner chip sequence.

The black line in Figure 3 shows the SPRT cumulative log likelihood $Z(n)$ (2) normalized by the upper SPRT threshold a (4). This puts the SPRT on the same scale as the Bayesian likelihood and DS BPA and gives a target / nontarget declaration when $Z(n)/a$ passes $1/-1$ (since $a=|b|$) . The green line shows the Bayesian likelihood $B(n)$ (7) with λ_0 set to 1 (equal priors). The DS BPA is shown by three curves. The red curve is the BPA for the T class $m(T)$, the blue curve is $m(\bar{T})$, and the magenta curve represents the uncertainty of the belief $m(\Theta)$.

Note that as the uncertainty $m(\Theta)$ goes to zero the Bayesian likelihood $B(n)$ approaches $m(T)$. Also when the current SPRT likelihood points to a nontarget or $Z(n)<0$ then $m(T)<m(\bar{T})$ and then for $Z(n)>0$ $m(T)>m(\bar{T})$. This supports the result in (28) that SPRT and DS make the same forced decision. There is also a similar relation between the Bayes likelihood $B(n)$ and $Z(n)$. When $Z(n)<0$, $B(n)<0.5$, and $Z(n)>0$, $B(n)>0.5$. This becomes evident from (7) when $\lambda_0=1$ and the knowledge that $Z(n)=0$ corresponds to $\Lambda(n)=1$.

It is interesting to note that between observation 30 and 40 the DS uncertainty $m(\Theta)$ is close to 0, but so is the SPRT log-likelihood. Any decision at this point would be an uninformed decision at low uncertainty. When $Z(n)/a$ goes above the threshold of 1, then we can make a high confidence decision that corresponds to low error rates of α and β set to 1×10^{-3}. Also note the Bayesian likelihood of $B(n)$ and D.S. BPA $m(T)$ of a target are very close to 1.

Figure 4 shows a mosaic of a chicken tracked by the system. The chip size is 51×34. Even though the chip size is different than that of the runner the use of the HOG features with the same number of horizontal and vertical blocks gives a system that is scale invariant.

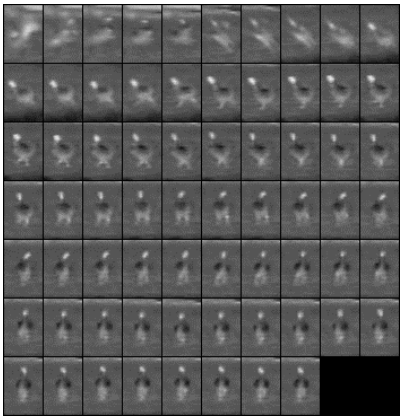

Figure 4: Infrared detections of a chicken tracked in an uncooled infrared video imager.

Figure 5 shows the multilook fusion results for the tracked chicken against the upright-human dismounts classification system. When the SPRT log likelihood $Z(n)/a$ falls below the threshold of -1, we can make a nontarget decision with low error rates. Also $B(n)$ and $m(\bar{T})$ go to zero indicating a nontarget.

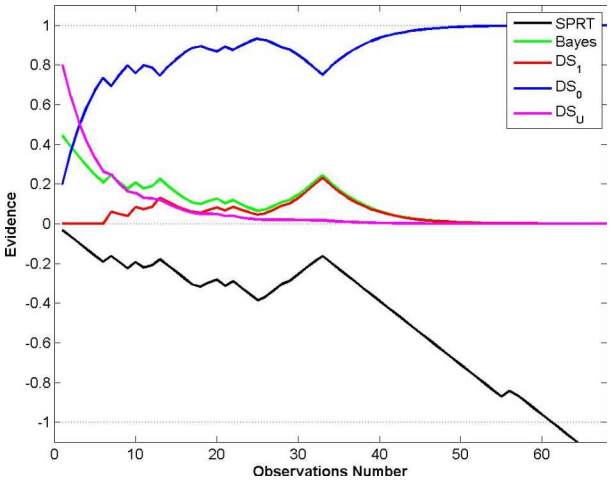

Figure 5: Results of different multilook fusion algorithms of the chicken GOF scores.

8. Conclusion

One of the main conclusions is that the sequential probability ratio test (SPRT), Bayesian fusion, and Dempster Shafer (DS) theory of evidence all make the same the forced decision for the one-class problem when the consonant evidence is represented by probability density functions and there are equal priors for the target and nontarget classes. This is rather surprising, since Dempster Shafer incorporates the uncertainty of a belief in its belief combination rule.

While a one-class classifier approach may seem severely limiting one could solve the multi-class problem by designing a one-class classifier for each class.

Each fusion approach brings a different element to the problem of multiple look fusion. The SPRT approach allows one to select decision thresholds that control the two main errors used to measure a system's performance: the probability of missed detection and the probability of false alarm. Bayesian likelihoods become probabilities if the PDF's are discrete. Dempster Shafer incorporates belief uncertainty into evidence combination rule.

By theoretically analyzing each under a common probabilistic framework we are able to mathematically transform one into another and use or improve the best features of each.

References

[1] E. Boult, X. Gao, R. Micheals, and M. Eckmann, "Omni-directional visual surveillance," *Image and Vision Computing*, vol. 22, pp. 515, 2004.

[2] N. Dalal, and B. Triggs, "Histogram of oriented gradients for human detection," *Computer Vision and Pattern Recognition Conference,* June 2005.

[3] J. D. Gibson and J. L. Melsa, *Introduction to Nonparametric Detection with Applications*. New York: IEEE Press, pp. 25, 1996.

[4] H. Kim, and P. H. Swain, "Evidential reasoning approach to multisource-data classification in remote sensing," *IEEE Transactions on Systems, Man, and Cybernetics*, **24**, No. 8, 1257-1265, 1995.

[5] M. W. Koch, G. B. Haschke, and K. T. Malone, "Classifying acoustic signatures using the sequential probability ratio test," *Sequential Analysis Journal*, vol. 23, 4, pp. 557-583, 2004.

[6] M. L. Koudelka, J. A. Richards, and M. W. Koch, "Multinomial pattern matching for high range resolution radar profiles," Algorithms for Synthetic Aperture Radar Imagery XIV. Edited by Zelnio, Edmund G.; Garber, Frederick D.. Proceedings of the SPIE, Volume 6568, pp. 65680V (2007).

[7] M. Moya and D. Hush, "Network constraints and multi-objective optimization for one-class classification," *Neural Networks*, vol. 9, 3, pp. 463-474, 1996.

[8] K. Murphy, and B. Myors, Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests. New Jersey: Lawrence Erlbaum Associates, 1998.

[9] J. Savage, et. al., "Irma 5.2 multi-sensor signature prediction model," Proc. SPIE 6564, Modeling and Simulation for Military Operations II, April 2007.

[10] Shafer, G., *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[11] A. Wald, *Sequential Analysis*, John Wiley & Sons Inc, New York, 1947.