RC24067 (W0610-011) October 3, 2006
Computer Science

IBM Research Report

On Sensor Sampling and Quality of Information: A Starting Point

Chatschik Bisdikian
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

On Sensor Sampling and Quality of Information: A starting point

Chatschik Bisdikian
IBM T. J. Watson Research Center
Hawthorne, NY, USA
bisdik@us.ibm.com

Abstract— **Numerous sensor-based applications depend upon the detection of certain events. The proper operation of these applications depend on the quality of the information (QoI) that they receive from their sensor-based event detectors. In this paper, we establish relationships that tie the QoI attributes of timeliness and confidence to the operational characteristics of a sensor system and the events they detect. By building upon the Neyman-Pearson hypothesis testing procedure, we study the dependence of these characteristics and attributes on each other and establishing their theoretical performance boundaries.**[1]

I. Introduction

With the introduction of autonomous, battery-operated, and especially wireless communication capable sensing platforms, sensor-based systems are becoming very powerful and flexible sources of data that support a wide collection of applications.

Typical sensor-based applications comprise lower-level *sensing* modules that take environmental measurements, and high-level *fusing* modules that transform these measurements into useful (in some sense) information that support various decision making processes [1]. A decision maker will make a decision (and cause an action), if the information derived from sensed data indicates that an event of interest has occurred, e.g., when acoustic or seismic measurements indicate that an explosion (might) have occurred. In deciding how to proceed, decision makers take decisions based on the confidence they place upon the *quality of information* (QoI) available to them.

Regarding sensor-collected data, the quality of the sensed data is captured via a collection of attributes that includes [2]:

- *timeliness*: which describes how timely the data are provided to be useful to applications;
- *accuracy*: which describes the level of detail (precision) in the sensed data;
- *reliability*: which describe how much *confidence* can be placed in the sensed data;
- *throughput*: which describes the rate at which data are provided to user-applications;
- *cost*: which described the cost collecting the sensed data.

As sensed data are processed (fused) and raw data turn into useful and actionable information, the above quality attributes of sensed data affect similar higher-level QoI attributes (timeliness, accuracy, and reliability), as well as higher-level concepts, like *completeness*, *relevancy*, and *usability* [3].

To determine whether an event of interest has occurred sensor measurements are correlated and support a hypothesis test as to whether the event has occurred or not. The outcome of this test represents the information derived through the fusion of the sensed data. The quality of this information can be captured by how fast this outcome can be made –timeliness of information–, and the probability of correct detection and false alarm (i.e, the probability to declare the event has happened when the opposite is indeed true) –accuracy of or confidence about the information.

The study of the above hypothesis testing is part of the time-honored detection theory [4]. With the increased interest in wireless sensors, the topic of hypothesis testing is experiencing a resurgence of sorts. For example, recent studies for wireless sensor systems have looked into centralized, distributed, and hybrid data fusion architectures and decision making schemes based on hypothesis testing, where measurements are fused either locally or remotely or in a combination of the two [5], [6], [7], as well as collaborative schemes [8] where groups of sensors first collaborate to "improve" the outcome of their local fusion prior to fussing centrally the outcomes of the locally fused data.

Prior studies have focused on system designs and algorithms that are energy-aware and/or improve the detection capabilities of the system. While these are certainly key design objectives for a wireless sensor system, to the best of our knowledge, prior studies in this area have not yet investigated the fundamental relationships and trade-offs that could exist between the events whose occurrence the sensor system is to detect and the operational characteristics of the sensing system, namely its sampling rate. As an example, on the one hand, if sampling were to occur in accordance to the Nyquist frequency, not only detection but even a pretty accurate facsimile of the original signal (the event signature) could be constructed. On the other hand, when dealing with time-limited signals, very sparse sampling will result in missing the event entirely. So, something interesting must be happening for sampling frequencies in the middle. This observation fact has motivated this first study on this topic, with an initial contribution

[1] Research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation heron.

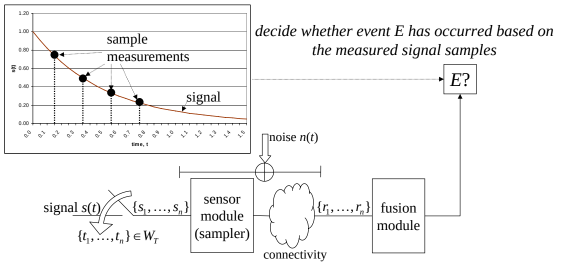

Fig. 1. A high-level system model.

the discovery of relationships between event detection and sampling. In view of the prior discussion, an additional interest and contribution of this study is mapping these relationships to the QoI attributes of timeliness and confidence.

The paper is organized as: In section II, we present the system model to be studied. In section III we present our analysis employing the Neyman-Pearson hypothesis testing technique. In section we present our numerical results for a special class of event signatures, and finally we conclude in section with some concluding remarks.

II. System model

Fig. 1 shows a high-level model of the system that we consider in this study. Specifically, we consider a system comprising a *sensor module* that observes "its" environment in search of an event of interest E, and a *fusion module* that process these observations. The outcome of the fusion process supports a decision maker that decides whether the event E has occurred or not.

Based on cost and other design considerations, the sensing and fusion modules may be physically collocated within the same device or be apart of each other. In either way, the connectivity cloud in the figure is assumed to represent an idealized communications path between the two modules and we will not consider it further in this study.

As shown in the figure, the sensor takes measurements, or samples, at discrete times. Let $t_i \in W_T$ represent the time the sensor takes the i-th measurement and r_i the *observable* outcome of this measurement; $W_T = [0, T]$ represents the *observation window* We assume that during the observation window W_T there are N measurements taken, i.e., we have an N-dimensional observation space represented by the observation vector variable $\mathbf{r}_N = \{r_1, \ldots, r_N\}$[2]. Finally, we assume that each measurement is corrupted by an additive noise process; we do not distinguish between different sources of noise or error in the measurements.

We assume that the event of interest generates a signal $s(t)$, and let s_i represent the value of $s(t)$ during a sampling instance ($s_i = s(t_i)$). While $s(t)$ can have any profile, of particular interest in the paper will be events that are transient in nature, e.g., an explosion whose, say, acoustic energy $s(t)$ is sampled by an acoustic sensor. Regarding the noise process, for the numerical results later on, we assume that the additive noise component to each measurement constitute an i.i.d. sequence of independent, zero-mean Gaussian random variables with variance σ^2.

III. The solution approach

With the set-up described in II, we formulate two hypotheses depending whether the event occurred or not as follows:

$$\begin{aligned} H_1 &: \ r_i = s_i + n_i, \qquad i = 1, \ldots, N, \\ H_0 &: \ r_i = \phantom{s_i +{}} n_i, \qquad i = 1, \ldots, N, \end{aligned} \tag{1}$$

where s_i and n_i are the signal (when present) and noise components of the i-th (observable) measurement. Let $f_{A|Hi}(\cdot)$ represent the probability density of some random variable A conditioned on the hypothesis H_i, $i = 0,\ 1$. From the classical hypothesis testing analysis, the decision test dependents on the *likelihood ratio* of the conditional probabilities [4]:

$$\Lambda(\mathbf{R}_N) = \frac{f_{\mathbf{r}_N|H_1}(\mathbf{R}_N)}{f_{\mathbf{r}_N|H_0}(\mathbf{R}_N)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{2}$$

where $\mathbf{R}_N = \{R_1, \ldots, R_N\}$ represents the collection of the N actual measurements observed; $\mathbf{R}_N$ should be contrasted with $\mathbf{r}_N$ which is a vector random variable for the observed data. The threshold η in (2) depends on decision test criteria.

The decision test criterion that we have elected to use is the *Neyman-Pearson test*, leaving other test options for future studies. The Neyman-Pearson test maximizes the probability of correct detection P_D, while constraining the probability of false alarm P_F. It uses only conditional probabilities (P_D and P_F are conditional probabilities by definition) and avoids using the usually unown and/or hard to derive a priori probabilities for the hypotheses, like the more popular Bayesian test criterion uses. Also, the use of the conditional probabilities actually is pretty pertinent to our QoI study. Specifically, assuming that action at some higher level takes place only when the decision maker reports that an event of interest has occurred, the QoI conveyed by this decision increases the sooner the report is made relative to the time the event occurs. It, of course, also increases the higher the probability P_D of correctly detecting the event becomes, while maintaining the probability of false alarm P_F "under control," i.e., bounded by a specified false alarm rate $P_F \leq \alpha$.

With the noise process assumed an i.i.d. Gaussian process, we can easily calculate the ratio in (2):

$$\begin{aligned} \Lambda(\mathbf{R}_N) &= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left((R_i - s_i)^2 - R_i^2 \right) \right\} \\ &= \exp\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{N} s_i \left(2R_i - s_i \right) \right\}. \end{aligned} \tag{3}$$

[2] We borrow much of our notation from [4].

Taking the logarithm on both sides of (3) and rearranging terms, the decision test in (2) becomes:

$$\sum_{i=1}^{N} s_i R_i \underset{H_0}{\overset{H_1}{\gtrless}} \eta^\star = \sigma^2 \ln \eta + \frac{1}{2}\sum_{i=1}^{N} s_i^2. \tag{4}$$

Note that the decision test depends only on the weighted sum of the measurements $l = \sum_{i=1}^{N} s_i R_i$; the scalar l represents the *sufficient statistic* for this test.

The Neyman-Pearson decision test maximizes the probability of correct detection P_D by setting the false alarm rate P_F at its maximum acceptable value α. According to (4), the probability of correct detection P_D is equal to:

$$P_D = \Pr(l \geq \eta^\star | H_1) = \int_{\eta^\star}^{\infty} f_{l|H_1}(w)\, dw\,, \tag{5}$$

Under hypothesis H_1, l is the sum of N independent Gaussian, random variables with the i-th random variable having mean s_i^2 and variance $\sigma^2 s_i^2$. Let $N(0,1)$ represent a normalized Gaussian random variable, and let $E_{S_N} \triangleq s_1^2 + \cdots + s_N^2$, i.e., when the signal is indeed present, E_{S_N} is representative of the energy of the signal measured. Then, the sufficient statistic l is distributed as a $N\big(E_{S_N}, \sigma\sqrt{E_{S_N}}\big)$ random variable, or equivalently, the random variable $y_1 = (l - E_{S_N})/\sigma\sqrt{E_{S_N}}$ is distributed as a $N(0,1)$ random variable. Therefore, the probability of correct detection P_D in (5) is given by:

$$\begin{aligned} P_D = \Pr(l \geq \eta^\star | H_1) &= \Pr\left(y_1 \geq \frac{\eta^\star - E_{S_N}}{\sigma\sqrt{E_{S_N}}} \right) \\ &= 1 - \Phi\left(\frac{\eta^\star - E_{S_N}}{\sigma\sqrt{E_{S_N}}} \right), \end{aligned} \tag{6}$$

where $\Phi(\cdot)$ represents the cumulative distribution of a $N(0,1)$ random variable.

Through a similar procedure, it can be found that the false alarm probability P_F is equal to:

$$\begin{aligned} P_F = \Pr(l \geq \eta^\star | H_0) &= \int_{\eta^\star}^{\infty} f_{l|H_0}(w)\, dw \\ &= 1 - \Phi\left(\frac{\eta^\star}{\sigma\sqrt{E_{S_N}}} \right). \end{aligned} \tag{7}$$

According to the Neyman-Pearson decision test, the false alarm rate is set at its upper bound value α, which implies from (7) that:

$$\eta^\star = \sigma\, \Phi^{-1}(1-\alpha)\sqrt{E_{S_N}}. \tag{8}$$

Substituting $\eta^\star$ in (6) with (8) yields:

$$P_D = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{E_{S_N}}}{\sigma} \right). \tag{9}$$

According to (9), P_D depends on a measure of the *signal-to-noise ratio* (SNR) $(\sqrt{E_{S_N}}/\sigma)$. Specifically, as it would have been expected, P_D increases with SNR. Since $N(\min\{s_i\})^2 \leq N\overline{S}_N^2 \leq E_{S_N} \leq N(\max\{s_i\})^2$, where $\overline{S}_N$ represents the average value of the N (uncorrupted) signal samples, (9) results in:

$$\begin{aligned} \alpha \leq 1 - \Phi &\left(\Phi^{-1}(1-\alpha) - \frac{\overline{S}_N\sqrt{N}}{\sigma} \right) \\ &\leq P_D \leq 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\max_{\{1\leq i\leq N\}}\{s_i\}\sqrt{N}}{\sigma} \right). \end{aligned} \tag{10}$$

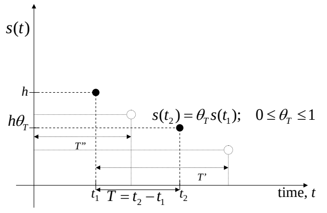

Fig. 2. A simple two sample set-up.

The bounds in (10) capture the amount of variability of the probability of detection P_D when compared to a signal that would had the same (constant) value during the observation interval. Notice that the upper and lower bounds coincide when $\max\{s_i\} = \overline{S}_N$, i.e., when the signal, actually its measured samples, have no variability.

Studying the limiting behavior of P_D, we observe that when $\overline{S}_N\sqrt{N} \to \infty$ with N, e.g., when the average value of the samples remains bounded away from 0 uniformly with respect to N (i.e., when $\overline{S}_N \geq \delta > 0$ for all N larger than some N_0), then P_D increases towards 1 with increasing N. Notice that the latter condition does not hold true when the signal is time-limited and the sampling rate is upper bounded by a positive number. Finally, when the measurement-dependent portion of (9) and (10) becomes infinitesimally small, e.g., when the noise variance σ increases, P_D reduces toward α.

IV. QoI and detection performance

In this section, we use the expressions for P_D and P_F from section III to study the impact of system and signal parameters on the QoI attributes of timeliness and accuracy.

For simplicity, but without lack of insigthtfulness, in this paper, we consider the simple set-up shown in Fig. 2. Specifically, we assume a time decaying, transient signal representing a transient event. We assume that two measurements are taken at times t_1 and t_2 that are $T > 0$ time units apart. If the event had indeed occurred, then in the absence of any measurement noise, the measurement values would have been h and $\theta_T h$, respectively; if $h = 1$, we may consider that the measurements are normalized with respect to the first measured value. The parameter θ_T represents the decay of the signal over the

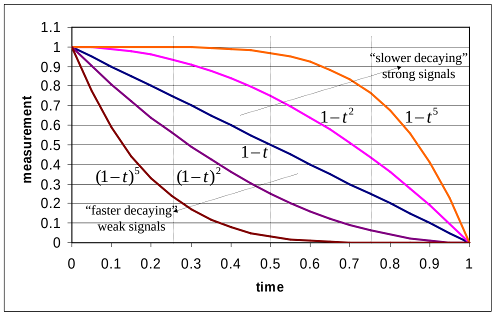

Fig. 3. A broad family of transient decaying signals.

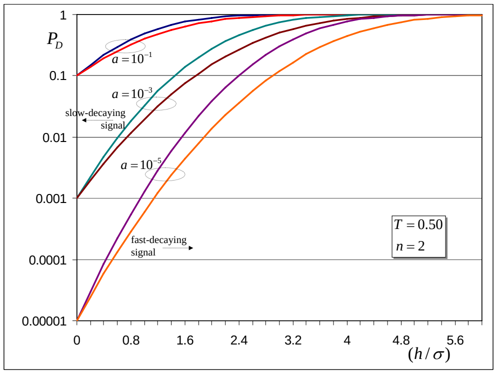

Fig. 4. Probability of detection P_D as a function of the signal-to-noise ratio (h/σ) for various false alarm rates α (log graph).

"observation" interval T. Since we assume that the signal, when present, is decaying with time, θ_T decreases with T with $\theta_0 = 1$ and $\theta_\infty = 0$.

Before proceeding further with the analysis, we feel the need to provide an additional interpretation of Fig. 2. While this initial study looks only on two measurements, in the grand scene of things, these two measurements may represent the first and second, the first and last, etc., measurements from a collection of multiple measurements. Thus as part of this simple(r) case study, we also search for any hints of relationship that may exist between the decaying-value measurements and the frequency of sampling (shown in the figure via the inter-sample intervals T, T', T'', etc.), and how these may reflect upon the QoI information attributes of interest.

For this set-up, the probability of detection is given by, see (9):

$$P_D = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{h}{\sigma}\sqrt{1+\theta_T^2}\right), \qquad (11)$$

which shows that P_D dependents on the "primary" SNR component (h/σ), i.e., the SNR component due to the first measurement only. The impact of the additional measurement is through the term $\sqrt{1+\theta_T^2}$. It follows that as long as the noise variability σ is normalized relative to first sample value h, we can assume that $h = 1$ without lack of generality.

Fig. 3 shows the test signals that we use in this analysis. The exact form of the signals is immaterial at this stage, but we have nevertheless selected them to cover a wide range of shapes of interest. We have a "reference" signal that decays linearly, $s(t) = 1 - t$, over its lifespan of duration 1, and a collection of signals that decay symmetrically around the reference signal: $s(t) = 1 - t^n$ for the signal above the reference signal and $s(t) = (1-t)^n$ for the signals below the reference one; note that in either case when $n = 1$, we obtain the reference signal. We also mark in the figure time instances, like $T = 0.25$, 0.5 and 1.0, that the second sample is taken. We will refer to the signals above the reference as the slow-decaying, strong signals, while those below the reference signal as the fast-decaying, weak ones; the parameter n represents the decay parameter.

For the class of signals in Fig. 3, the probability of detection in (11) becomes:

$$\begin{aligned} P_D(T) &= 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{2}}{\sigma}\sqrt{1+T^n\left(\frac{T^n}{2}-1\right)}\right), \\ &= 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{1}{\sigma}\sqrt{1+(1-T)^{2n}}\right), \end{aligned} \qquad (12)$$

for the fast- and slow-decaying signals, respectively. Of course, when $n = 1$ the two expressions in (12) coincide.

Figures 4–6 show the probability of detection P_D as various system and signal parameters vary. Fig. 4 shows how P_D behaves as a function of the primary SNR component (h/σ) for various values of false alarm rate α, when the second sample is taken at time $T = 0.5$. We consider a fast- and a slow-decaying signal both with a decaying parameter of $n = 2$. First of all, we notice that when SNR=0, $P_D = \alpha$ as discussed at the end of section III. Again as previously discussed, and expected, P_D approaches to 1 when SNR increases. It is interesting to notice that the higher the false alarm rate α the higher the probability of detection becomes. This again is to be expected, since when we always decide in favor of hypothesis H_1, then we always detect the event correctly $(P_D = 1)$ but at an unreasonably false alarm rate $(P_F = 1)$.

Fig. 5 shows how P_D behaves as a function of T, the time that the second sample is taken, when $\alpha = 10^{-3}$, $n = 2$, and for various SNR values. The figure reveals how much better the confidence on the detection result becomes the "stronger" the second measurement is relative to the first, which happens the closer the second measurement is taken to the first one. In other words, the figure captures how the timeliness of the detection decision impacts the confidence on the decision taken. We see that for the faster-decaying signal no substantial improvement is made in P_D when $T > 0.5$ (even when n is just 2). This fact becomes even more prominent the smaller the SNR becomes. The P_D for the slower-decaying signal behaves more favorably with T, with the difference in P_D between the slower- and faster-decaying becoming more prominent with increasing SNR, which is expected since the higher the SNR,

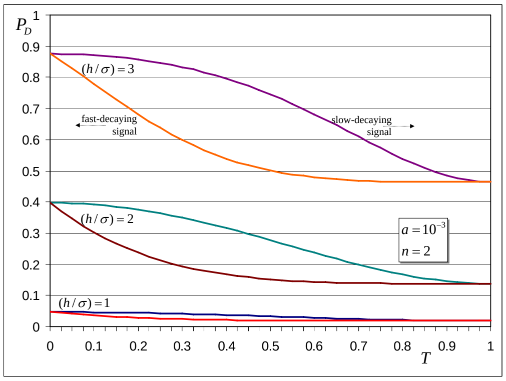

Fig. 5. Probability of detection P_D as a function of the sampling time T of the second sample for various values of the SNR (h/σ).

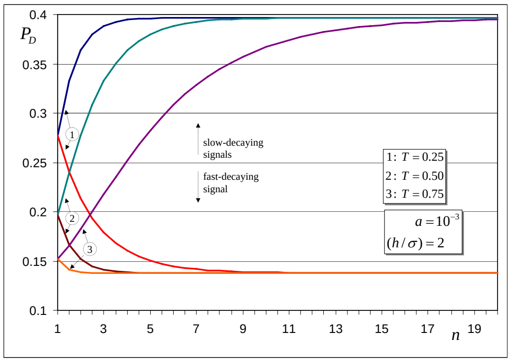

Fig. 6. Probability of detection P_D as a function of the signal decay parameter n for various sampling times T.

the more "elbow-room" exists for improvement. Note that for each value of the SNR, the curves start from the same point at $T = 0$ and monotonically decline diverging initially but eventually converge to the same point when $T = 1$. This behavior can easily be explained due to the shape of the curves considered, see Fig. 3.

Finally, Fig. 6 shows how P_D behaves as a function of n, the decaying parameter, when $\alpha = 10^{-3}$, SNR=2, and for various values of T. Recall that when $n = 1$, the fast- and slow-decaying signal coincide and hence the corresponding pairs of curves start at the same value of P_D. However, they quickly diverge from each other, as the second measurement of the slow-decaying signal becomes more and more equal to the first measurement, while for the fast-decaying signal the second measurements hardly "registers" with increasing n, something that becomes even more prominent the later the second measurement is taken, i.e., as T also increases. As expected, the left and right convergence points in the SNR=2 curve in Fig. 5 coincide with the upper and lower convergence levels for P_D in Fig. 6.

V. Concluding Remarks

In this paper, we have made a first attempt to discover and quantify relationships between: (a) the characteristics of events that a sensor system tries to detect; (b) the operational characteristics of the sensor system; and (c) the attributes the capture the quality of the information that the sensor system provides to its users. Starting with classical detection theory analysis, we have extended that work to derive simple, elegant, and insightful analytical formulaethat are applicable to a large number of signal profiles. To gain an even deeper understanding on these relationships, we have applied the formulae to a broad family of decaying signal profiles and studied a simple, but non-trivial subcase comprised of two measurements.

The performance trends revealed in our analysis are not unexpected. The results basically show that the stronger the signal is the higher the probability of detection becomes. There is nothing surprising with this observation and we dare to say that if something out of the ordinary were shown during our analysis, then either our analysis or our intuition would have been at fault. The benefit from this analysis though is not merely to confirm our intuition but to provide for he first time simple, fully quantifiable insights for how the signal characteristics, e.g., its decaying time parameter, and the system operational parameters, e.g., the sampling rate, impact the quality of the information produced by a sensor system. In a sense, our work captures the theoretical limits for the QoI of a sensor-based detector system parameterized on the events that the detector tries to detect.

In closing, this study reveals how our confidence on the event detector changes as new samples are incorporated in the decision making process. While related, this study is distinct from traditional sequential detection procedures for, under our analysis assumptions, the amount in detection improvement with each new sample datum is exactly quantified. This facilitates the development of stopping rules for sampling to be made other than reaching a predetermined event detection threshold, for example, a stopping rule may be reflective of the "accumulated" confidence and timeliness QoI attributes.

References

[1] H. Carvalho, W. Heinzelman, A. Murphy, and C. Coelho, "A general data fusion architecture," in *6th Int'l Conf. on Information Fusion (FUSION)*, Cairns, Queensland, Australia, July 8-11, 2003, pp. 1465–1473.

[2] E. Blasch and S. Plano, "DFIG level 5 (user refinement) issues supporting situational assessment reasoning," in *8th Int'l Conf. on Information Fusion (FUSION)*, Philadelphia, PA, USA, July 25-28, 2005, pp. xxxv– xliii.

[3] S. Ehikioya, "A characterization of information quality using fuzzy logic," in *18th Int'l Conf. of the NA Fuzzy Information Processing Society (NAFIPS)*, New York, NY, USA, June 10-12, 1999, pp. 635 – 639.

[4] H. L. van Trees, *Detection, Estimation, and Modulation Theory, part I*. John Wiley & Sons, 1968.

[5] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Energy-efficient detection in sensor networks," *IEEE JSAC*, vol. 23, no. 4, pp. 693–702, April 2005.

[6] L. Yu and A. Ephremides, *Mobile, Wireless, and Sensor Networks*. Wiley-Interscience, 2006, ch. Detection, Energy, and Robustness in Wireless Sensor Networks, pp. 145–172.

[7] L. Yu, L. Yuan, G. Qu, and A. Ephremides, "Energy-driven detection scheme with guaranteed accuracy," in *Proc. IPSN'06*, Nashville, TN, USA, April 19-21, 2006.

[8] N. Katenka, E. Levina, and G. Michailidis, "Local vote decision fusion for target detection in wireless sensor networks," in *Joint Research Conference on Statistics in Quality Industry and Technology*, Knoxville, TN, USA, June 7-9, 2006.