Confidence Fusion

Mel Siegel, *Fellow, IEEE*, Huadong Wu

Abstract –The field of sensor fusion is well developed, but approaches to describing the confidence one ought to have in the fused result as a function of the confidence one has in the individual measurements is not. This speculative paper is intended to stimulate discussion in the robotics and instrumentation communities regarding how to approach "confidence fusion" in a way that can be practically applied in robotics applications and is intuitively satisfying with respect our feelings regarding how confidence in measurements ought to propagate. Preliminary ideas are developed in the context of a scenario wherein two one-bit-output sensors in which we have different confidence are used together. The results are compared with numerical experiments simulating a real-life experiment that we have described in the field of sensor fusion for context-aware computing.

***Index Terms* — confidence fusion, context-aware computing, Dempster-Shafer theory of evidence, sensor fusion**

I. Introduction

In several recent manuscripts [1][2][3][4] we introduced the Dempster-Shafer (D-S) "theory of evidence" to the field of sensor fusion for context-aware computing, and we provided several pragmatic extensions to the basic approach that have proven useful in this application area. As an outgrowth of this research, we have recently begun to investigate an area for which we suggest the term "confidence fusion"[1]. We hope that by presenting our earliest thoughts on this topic in the speculation-tolerant ROSE-2004 workshop medium we will stimulate discussion and new ideas in the robotics and instrumentation and measurement communities for whom sensor fusion and related technologies are so crucial.

To illustrate what we mean by confidence fusion consider this simple scenario: we have two sensors, each with a one-bit-binary ("dichotomous") output for reporting the presence of weapons of mass destruction (WMD). One sensor, *S1*, has been demonstrated to be correct 95% of the time, the other, *S2*, has been demonstrated to be correct 85% of the time. If we are allowed to make one and only one measurement using *S1* and one and only one measurement using *S2*, it is obvious that we will always take the word of *S1*, the better of the two sensors. However our confidence with be increased if the two sensors agree, and it will be diminished if they disagree. Quantitatively, if they agree we will believe *S1* with something greater than 95% confidence, whereas if they disagree we will believe it with something less than 95% confidence. Exactly how to do the arithmetic of quantifying the respectively increased and decreased confidences requires some thought, and perhaps some assumptions, but the qualitative principle is beyond dispute.

Now what if we have three sensors, *S1* correct 95% of the time, *S2* correct 85% of the time, and *S3* also correct 85% of the time; each makes one and only one measurement of the same target, and in these measurements *S2* and *S3* agree with each other but they disagree with *S1*? Two very bad sensors clearly cannot outvote one very good sensor, but exactly where is the dividing line beyond which we will feel better allowing two slightly worse sensors to outvote one slightly better sensor? And so on for N sensors.

In this paper we report our early efforts toward refining these questions and beginning tentatively to answer them. Refining is necessary for several reasons, not the least of them being that "correctness" is too simple a description of sensing system reliability when, for example, the human consequences of a false negative are very much more serious than the human consequences of a false positive. We are thus led to consider the basic question of confidence fusion again in the context of "Type-I" and "Type-II" errors, i.e., false positives and false negatives.

In developing these ideas we in a sense come full circle, recognizing a complementarity between the D-S theory and the "receiver operating characteristic" (ROC)[2]-based set of correctness measures that we first noted [5] at SCIMA-2003. a speculation-tolerant workshop similar to ROSE-2004.

The paper is organized as follows: Section II A FIRST PASS AT DEFINING CONFIDENCE further explores the problem quantitatively defining confidence in scenarios where "good" and "not-so-good" dichotomous sensors replicate or complement each others measurements; Section III SENSOR FUSION BACKGROUND reviews from a confidence fusion perspective the sensor fusion for context aware computing application mentioned in the introduction; Section IV SENSOR FUSION EFFECTIVENESS EXPERIMENTS describes the design of numerical experiments intended to compare the ideas regarding confidence fusion that were developed analytically for simple cases in Section II with the complex real-world sensor fusion experiments described in Section III. Section V CONCLUSIONS AND FUTURE WORK summarizes the what has been accomplished to date toward understanding confidence fusion, and outlines work to be done in the future.

[1] The term "confidence fusion" actually already appears occasionally in the literature, but so infrequently and in such different settings with such different meanings that we feel it is not inappropriate to adopt it to describe the path we are pursuing herein.

[2] ROC curves are generated parametrically in the two dimensional coordinate space of false positives and false negatives as the threshold or "cut-point" that dichotomizes the output of a test is varied through its range.

II. A First Pass at Defining Confidence

In this section we further explore the problem defining confidence in scenarios where "good" and "not-so-good" dichotomous sensors replicate or complement each other's measurements.

A. *Two Simple WMD Sensors*

Expanding the scenario presented in the introduction, say we send a mobile robotic probe to a desolate remote region with the aim of detecting clandestine WMD. The probe is equipped with two simple[3] WMD sensors each of which gives a one-bit response when pointed at a target under investigation: *not_WMD (0)* or *_WMD (1)*. Neither sensor is perfect: *S1* has been correct fraction *f1* of the time when evaluated on known WMD and innocuous targets, and *S2* has been correct fraction *f2* of the time. Both sensors examine each target once. There are four possibilities: *(0,0)* indicating both sensors report *not_WMD*, *(0,1)* indicating *S1* reports *not_WMD* and *S2* reports *_WMD*, *(1,0)* meaning *S1* reports *_WMD* and *S2* reports *not_WMD*, and *(1,1)* meaning both sensors report *_WMD*.

The following proposition – call it *Proposition-1* – seems intuitively obvious and satisfying: If the sensors agree then we should conclude that the target under investigation is correspondingly *not_WMD* or *_WMD* with confidence greater than our confidence in the better of the two sensors alone, whereas if the sensors disagree then we should conclude that the target under investigation is *not_WMD* or *_WMD* in agreement with which response is reported by the better of the two sensors, but our confidence should be lower than our confidence in the better sensor alone.

There are eight possible situations: the target of interest may or may not be WMD, and the two binary sensors together can report four possible conditions, two in agreement and two in disagreement. Let us now evaluate the corresponding probabilities.

If the target of interest is actually *_WMD* then the probability *p_WMD(0, 0)* of both sensors agreeing that it is *not_WMD*, the probability *p_WMD(0, 1)* of *S1* saying it is *not_WMD* and *S2* saying that it is *_WMD*, etc., are enumerated exhaustively by:

```
p_WMD(0,0)=(1-f1)(1-f2)   [(wrong,wrong)->wrong]
p_WMD(0,1)=(1-f1)f2       [(wrong,right)->wrong]
p_WMD(1,0)=f1(1-f2)       [(right,wrong)->right]
p_WMD(1,1)=f1 f2          [(right,right)->right]
```

Similarly, if the target of interest is actually *not_WMD* then the probabilities *p_not_WMD(0, 0), p_not_WMD(0,1),* etc., are enumerated exhaustively by:

```
p_not_WMD(0,0)=f1 f2      [(right,right)->right]
p_not_WMD(0,1)=f1 (1-f2)  [(right,wrong)->right]
p_not_WMD(1,0)=(1-f1)f2     [(wrong,right)->wrong]
p_not_WMD(1,1)=(1-f1)(1-f2) [(wrong,wrong)->wrong]
```

Now we can calculate the fraction of right and wrong answers given by the fusion of *S1* and *S2*. Let *P_WMD* and *P_not_WMD* designate the actual probabilities that the target of interest is *_WMD* and *not_WMD* respectively, of course with the constraint *P_WMD* + *P_not_WMD = 1*:

```
P_right = P_WMD(p_WMD(1,0)+ p_WMD(1,1))+
  P_not_WMD(p_not_WMD(0,0)+ p_not_WMD(0,1)
P_wrong = P_WMP(p_WMD(0,0)+ p_WMD(0,1))+
  P_not_WMD(p_not_WMD(1,0)+ p_not_WMD(1,1)
```

Expanding gives *P_right = f1 and P_wrong = (1-f1)*. This is a very surprising result if we believe *Proposition-1*. It says that two sensors observing the same dichotomous variable are no better than the better sensor alone, provided only that we adopt the rule that if the sensors disagree we take the word of the better one. On further thought it is, however, not surprising at all: the rule that if sensors disagree we take the word of the better one is equivalent to saying we use only the better sensor, *S1*, in which case by definition *P_right = f1* and *P_wrong = (1-f1)*, exactly the result we just calculated.

Nevertheless, our intuitions continue to demand that we should have more confidence than *f1 = MAX(f1, f2)* in the outcome of the measurements in which *S1* and *S2* agree and less confidence than *f1* in the outcome of the experiments in which they disagree. The fault, then, must lie in our tacit assumption that "confidence" is quantified solely by *P_right* and *P_wrong*. Both logic and algebra show that *P_right* and *P_wrong* depend only on *f1*, i.e., explicitly on the characteristics of *S1* but not on *f2*, i.e., not on the characteristics of *S2*. Yet our intuitions continue to demand that our confidence in the result reported by *S1* must reasonably by colored by what we know *S2* says, i.e., on *f2*. We next visit the issue of quantifying "confidence".

B. *What do we mean by confidence?*

We use the word "confidence" a lot, but we have not yet defined it quantitatively, even though we routinely assign numerical values to our confidences in the judgments we make as well as to the data our sensors generate. To illustrate with an example close to home, the standard review form for workshops like ROSE-2004 sponsored by IEEE Instrumentation & Measurement Society ask the reviewer to rate each submission on several criteria of quality and relevance, to make several recommendations regarding acceptance and mode of presentation, and, finally for the reviewer to rate his/her own "confidence in my review" on a scale of from "out of my field" to "excellent". In short, we have excellent qualitative understandings of what confidence means, and we do not hesitate to treat it as a quantitatively.

It is thus surprising that, whereas statisticians define many quantitative measures that incorporate the word "confidence", e.g., "95% confidence interval", there is no standard quantitative definition of "confidence". So in this section we will consider several definitions of our own.

To be both convenient and intuitive a definition of confidence must have a several obvious properties: (1) it should range between *0* and *1*; (2) complete certainty

[3] We use "simple" to mean that, at this stage of the discussion, we consider only "right" and "wrong", not distinguishing between false positive wrongs and false negative wrongs, i.e., "Type I" and "Type II" errors.

should correspond to confidence *1*; (3) complete uncertainty should correspond to confidence *0*.

Our expectations about the linearity of this scale are less obvious, and probably less consistent from one application scenario to another. Humility demands that we refrain from claiming certainty, so it should be difficult or impossible to score unit confidence. In some scenarios we would have no qualms about acknowledging that we have no useful knowledge, i.e., that our confidence is zero; we would not hesitate to acknowledge that we have zero confidence in a purported sensor that was shown by careful experiments to be no better than a random bit generator. On the other hand, when two equally good binary-output sensors disagree is our confidence in the outcome of the experiment as a whole zero, or would it be more useful to say that our confidence is 50% in each of the outcomes proposed by the two sensors together? These are questions that we cannot answer, but about which we want to stimulate discussion.

In some very simple cases it seems we can propose answers that most practitioners of the measurement science and art will find satisfying. For example, if we have only one dichotomous sensor that has been shown to be correct fraction *f* of the time, then a sensible measure of confidence is the difference between the fraction that are classified correctly minus the fraction that are classified incorrectly: $f – (1 – f) = 2 f – 1$; this confidence measure's value is *0* when $f = 0.5$, i.e., when the sensor is just guessing randomly, its value is *1* when the sensor is always right, and it is an intuitively satisfying *0.5* when the sensor is right *3* times out of *4*, this being *50%* better than random guessing, which would be right *2* times out of *4* if the dichotomous outcomes were equally probable, e.g., if half the targets of interest were actually WMD.

Note, however, that we still need to be careful with this sort of definition when the dichotomous outcomes, e.g., *not_WMD* and *_WMD*, are far from equally probable. In an environment where WMD is absent this measure would give us very high confidence in a completely bogus sensor, e.g., one that didn't really sense anything, but that just reported *not_WMD* for every target of interest.

As has been illustrated above, the scenario only needs to get a little more complicated, e.g., two dichotomous sensors looking at the same target of interest, before we being to question the validity of our intuition.

Statisticians have some standard tests that are potentially applicable to scenarios of the sort we are considering, e.g., McNemar's Test [6], which is typically used to evaluate the influence of question order on response by looking for inequality of matrix elements analogous to the cross-terms $p\_WMD(0,1) = (1–f1)f2$ and $p\_WMD(1,0) = f1(1–f2)$ in the preceding discussion. The potential value of McNemar's and other standard statistical tests in the context of our interest in confidence fusion is still and open question.

C. Correlation and complementarity

Another important issue that we need to consider is whether *S1* and *S2* both respond to the same underlying characteristic of the target of interest. If the answer is no then the statistics of their agreement and disagreement is determined entirely by chance. But if the answer is yes then *S1* and *S2* will agree more often than chance predicts. Given an expectation of better-than-chance agreement based on knowledge that the two sensors are looking at the same property of the target of interest bolsters our intuition that we should have more confidence in the outcome when the two sensors agree and less when they disagree.

In more detail, when we say, via *Proposition-1*, that *f1* and *f2* represent the fractions of the respective experiments in which *S1* and *S2* have been correct we do not intend these fractions to represent probabilities that originate in the randomness of the underlying measured continuum variable. The origin of these correctness-measuring fractions lies rather in the setting of a cut-point or threshold in a comparator circuit or algorithm that is hidden inside the sensor. Thresholding casts some underlying raw measured continuum variable into one or the other of the binary outcomes *_WMD* or *not_WMD*. When we presume that if the sensors agree we should have more confidence in the agreed-upon report than in the report of the better sensor alone, and when we presume that if the sensors disagree we should have less confidence in the combined report than we have in the better sensor alone, we are in essence presuming that the sensors should agree more often than randomly. The origin of this implicitly presumed agreement is a deeper implicitly presumed correlation between the sensors. The origin of the presumed correlation is, more deeply still, the presumption that in this scenario both sensors respond to the same underlying measured continuum property of the targets of interest.

This point is further clarified by examining a case in which we would *not* be naturally inclined to make an implicit presumption of correlation. Consider, for example, the case where *S1* reports *_WMD* or *not_WMD* based on the temperature differential between the examined object and its surroundings – radioactive objects are presumed to be warmer – and *S2* reports *_WMD* or *not_WMD* based on the direct detection of radioactivity. Since we can question each underlying assumption independently – warmth could be generated by a heap of rotting leaves and radioactivity could be generated by a heap of rocks bearing naturally radioactive elements – we are inclined to back off from the presumption that if the sensors agree we will be more confident of the outcome. When sensor responses are manifestly uncorrelated, as in this example, we are likely to adopt the perspective that the better sensor is better not because it is inherently a better measuring instrument, but rather because it is based on a better underlying classification principle, so only the better sensor really matters.

To end this section we note that, on the other hand, *S1* and *S2* may be complementary in the sense that they look at different characteristics of the target of interest either of which would be sufficient to classify it as WMD. This is in fact especially well illustrated by the particular case of WMD, since WMD is actually the union of three separate and distinct classes, chemical (C), biological (B) and nuclear (N) weapons of mass destruction. The sensor technologies appropriate for detecting the three WMD classes are as different as the underlying threats. Thus a perfect sensor for C-WMD would likely be quite useless for B-WMD and N-WMD, and so on for the other classes. To resolve conflicts in this sort of case we would probably invoke not a "better sensor decides" rule but instead an "any sensor sounds the alarm" rule, i.e., an OR scenario vs. a hierarchical scenario.

III. Sensor fusion background

In this section we review from a confidence fusion perspective the sensor fusion for context aware computing application mentioned in the introduction.

A. Basic Sensor Fusion Architecture Type

Sensor fusion architectures vary greatly depending on the details of the particular application. From an information processing perspective, sensor fusion approaches can be grouped roughly into three categories [7]: (a) raw data fusion, (b) feature fusion, and (c) decision or identity declaration fusion. The block diagrams of Figure 1 illustrate and contrast the processing steps involved in each category.

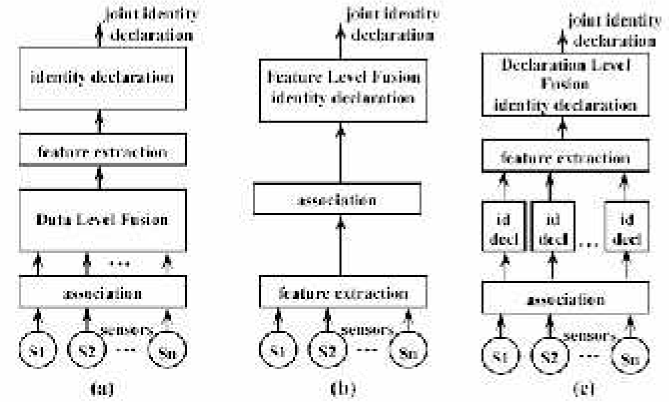


Figure 1. Typical sensor fusion architectures: (a) direct data fusion; (b) feature fusion; and (c) decision fusion

If the underlying sensors observe the same physical entity then usually the raw sensor data can be combined directly. But in many applications the sensors draw indirect high-level inferences based on very different sorts of low-level observations. In these cases information can be fused only at the feature or decision levels. For example, an intrusion detector that combines sonar and short-range radar sensors might very productively fuse both echograms in the low-level range-azimuth space, but one that combines a pyroelectric sensor for moving warm bodies and a microphone for detecting footfalls can be fused only at a much higher symbolic level.

Even when low-level data from multiple sensors are – in the physical sense – commensurate, it is often most productive to do some or all of the fusion a high "semantic" level. For example, in image processing for object recognition, sensor-level data fusion is typically applied at the image pixel level, but it is often effective then to employ feature level fusion operating on extracted features – luminosity and color patches, boundaries and edges, etc. – with decision-level fusion algorithms then tagging recognized objects based on their shapes or other appearance-features.

In decision level fusion, each sensor makes a preliminary determination of a targeted object's identity – or some other attribute of interest – and the high-level fusion algorithm combines the individual outcomes to generate a fused determination of higher accuracy or confidence or both.

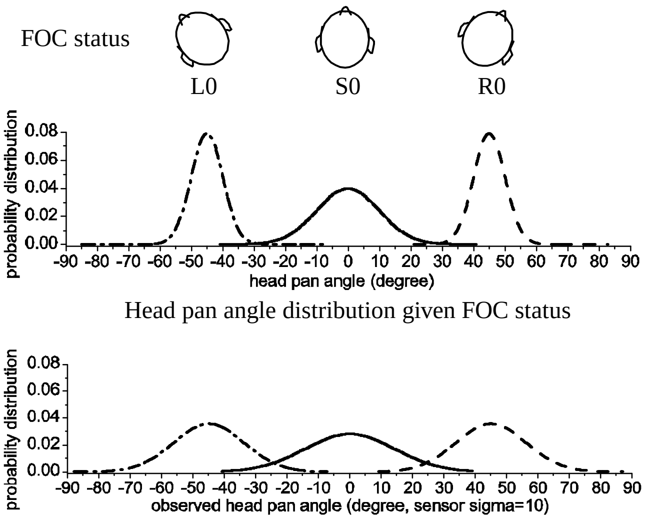
From sensor- to feature- to decision-level sensor fusion the communication bandwidth and computation power requirements become less demanding as the value of each bit retained and transmitted is corresponding reduced. The price, of course, is that this compression is inevitably lossy: some of the originally-available information is irretrievably lost when similar but not identical data are coalesced into categories. Nevertheless, decision level fusion may be the only alternative when the system is composed of multiple, independent, geographically distributed components.

B. A Focus-Of-Attention Case Study

In the references already cited, we have previously reported on the design of a focus-of-attention study for which we have a large historical data base whose temporal playback simulates a live data stream. We used this setup to compare and contrast several architectural and algorithmic approaches to sensor fusion on a perfectly reproducible data stream. We briefly review this background.

Four people are meeting around a small table. We want to decide automatically the current focus-of-attention (FOC) of each participant. We restrict the possibilities to *L0, S0,* and *R0,* an arbitrary one of which we designate by index *i*. The specific labels *L0, S0, R0* mean that *individual-0*'s focus-of-attention is on the person to his/her left, straight across the table, or his/her right respectively. The sensors – audio, video, and ground-truth reported by a human observer – each report all three possibilities with assigned probabilities p_{L0}, p_{S0}, and p_{R0} respectively.

individual-0's head pan angle θ is modeled by normal probability distributions: $N[\theta_{L0},\sigma_{L0}]$, $N[\theta_{S0},\sigma_{S0}]$ and $N[\theta_{R0},\sigma_{R0}]$ corresponding to the three focus-of-attention possibilities, shown in Figure 2, in which $\theta_{L0}=-45°$, $\theta_{S0}=0°$, and $\theta_{S0}=45°$.



Head pan angle distribution given FOC status

Observed head pan angle distribution given FOC status

Figure 2. Meeting-participant's focus-of-attention (FOC) estimation based on head-pan angle measurement

Suppose we use two sensors, *S1* and *S2* to measure the meeting-participant's head pan angle to estimate the FOC. A sensor's measurement error is comprised of two parts: the deterministic part $dft_S(t)$ reflects the sensor's static and dynamic (drift) calibration error, and the random part reflects the uncontrollable measurement noise n_S, which can typically be modeled by a zero-mean normal distribution. Assuming that sensors *S1* and *S2* have this typical measurement error probability distribution, described as $\varepsilon_{S1} \sim N[dft_{S1}(t), \sigma_{S1}]$ and $\varepsilon_{S2} \sim N[dft_{S2}(t), \sigma_{S2}]$ respectively, at time t. If the actual FOC situation is *i* and the actual head pan angle is $\theta(t)$, then the head pan angles reported by the sensors are:

$$\begin{array}{ll} \theta_{S1}(t) = \theta(t) + dft_{S1}(t) + n_{S1} \quad : & N[\theta_i + dft_{S1}, \sqrt{\sigma_i^2 + \sigma_{S1}^2}] \\ \theta_{S2}(t) = \theta(t) + dft_{S2}(t) + n_{S2} \quad : & N[\theta_i + dft_{S2}, \sqrt{\sigma_i^2 + \sigma_{S2}^2}] \end{array} . \quad (1)$$

As illustrated in the lower frame of Figure 2, because of the noise on the sensor measurements, the distribution of observed head pan angles is broader and consequently of lower peak amplitude than the noise-free distributions. Noise consequently increases the FOC identification error rate.

C. Low-Level and High-Level Data Fusion

In this section we assume for simplicity that sensors *S1* and *S2* are always recently calibrated and drift-free, so the deterministic error can be ignored. Low-level fusion combines multiple head pan angle measurements of the meeting-participant to estimate his/her FOC index, and high-level fusion combines the sensors' FOC index estimation probability values directly.

1) Low-level sensor fusion

Let θ_{S1} and θ_{S2} be *S1* and *S2*'s measured head pan angles respectively. Combine them via a weighted average:

$$\theta_{S12} = \frac{\sigma_i^2 + \sigma_{S2}^2}{2\sigma_i^2 + \sigma_{S1}^2 + \sigma_{S2}^2}\theta_{S1} + \frac{\sigma_i^2 + \sigma_{S1}^2}{2\sigma_i^2 + \sigma_{S1}^2 + \sigma_{S2}^2}\theta_{S2} . \quad (2)$$

Since the combined head pan angle measurements are distributed normally it follows that:

$$\theta_{S12} \sim N[\theta_i, \frac{(\sigma_i + \sigma_{S1})(\sigma_i + \sigma_{S2})}{\sqrt{2\sigma_i^2 + \sigma_{S1}^2 + \sigma_{S2}^2}}] \quad (3)$$

Equation (3) shows that the standard deviation of the combined "virtual sensor" head pan angle measurement is smaller than the standard deviation of either real sensor's head pan angle. Thus we see that statistical-weight based low-level sensor fusion gives a tighter estimate of FOC than does either sensor alone.

The weighted linear combination of sensor head pan angle measurements provides the most effective sensor fusion for each FOC index. In practice we use a linear sum approximation to calculate the weights in Equation (2):

$$\sigma_i = pL0 \cdot \sigma_{L0} + pS0 \cdot \sigma_{S0} + pR0 \cdot \sigma_{R0} \quad (4)$$

2) High-level sensor fusion

In our high-level sensor fusion scenario sensors *S1* and *S2* first make independent FOC probability estimates based on their own head pan measurements:

$$\begin{array}{l} P(i\,|\,\theta_{S1}) = \dfrac{P(i)P(\theta_{S1}|i)}{P(L0)P(\theta_{S1}|L0)+P(S0)P(\theta_{S1}|S0)+P(R0)P(\theta_{S1}|R0)} \\ P(i\,|\,\theta_{S2}) = \dfrac{P(i)P(\theta_{S2}|i)}{P(L0)P(\theta_{S2}|L0)+P(S0)P(\theta_{S2}|S0)+P(R0)P(\theta_{S2}|R0)} \end{array} , \quad (5)$$

Then, we use the D-S "theory of evidence" calculus to combine the FOC probabilities. Using $P_{S1}(i)$ for $P(i\,|\,\theta_{S1})$ and $P_{S2}(i)$ for $P(i\,|\,\theta_{S2})$, the combined FOC probability is:

$$P_{S12}(i) = \frac{P_{S1}(i)P_{S2}(i)}{1 - \sum_{j \neq k} P_{S1}(j)P_{S2}(k)} . \quad (6)$$

The D-S method can be viewed as a generalization of the classic Bayesian method that includes the latter as a special case. The described FOC estimation is an instance of this special case, so the result is the same as would be obtained by applying the Bayesian method. The nature of the special case is that the hypotheses are mutually exclusive [1]. The main reason we use the D-S vs. the equivalent Bayesian formulation is that the D-S formulation includes a calculus for confidence management that is absent in the Bayesian formulation. The mechanism for confidence management in the D-S formulation is via the introduction of new concepts of "belief" – the aggregate of all information in support of a hypothesis or proposition measured up from 0 – and "plausibility" – the aggregate of all information against the hypothesis or proposition measured down from 1. The gap between belief and plausibility is a regarded as a confidence interval within which the true probability is sure to be found. The Bayesian formulation drops out as the special case when the gap between belief and plausibility goes to zero.

IV. Sensor Fusion Effectiveness Experiments

In this section we use numerical experiments to show that sensor fusion does not always decrease measurement error or increase confidence in the result. These numerical experiments verify in the complex scenario of a real-world experiment the generality of the conclusion that was drawn in Section II by exhaustive enumeration of all cases and algebraic summing of corresponding probabilities for the simple case analyzed there.

A. Simulated sensor data generation

We use a Monte Carlo construction to compare in a pristine environment the effectiveness of several of sensor fusion approaches developed and compared using the real-world recorded video data described in [1] and [2]. We generate the simulated data stream as follows. An FOC index is generated using the given prior probabilities p_{L0}, p_{S0}, and p_{R0} for a random time interval between 5 to 15 seconds. Once per second the algorithm generates one "true" head pan angle θ from the normal probability distributions $N[\theta_{L0}, \sigma_{L0}]$, $N[\theta_{S0}, \sigma_{S0}]$ and $N[\theta_{R0}, \sigma_{R0}]$ correspondingly to the three allowed FOC means $\theta_{L0} = -45°$, $\theta_{S0} = 0°$, and $\theta_{S0} = 45°$. For each head pan angle θ, ten observations are generated by sensors *S1* and *S2* as per Equation (1).

B. Sensor fusion effectiveness and confidence

Table I summarizes the above described numerical experiments via a spreadsheet that examines the calculated outcome and calculated confidence in that outcome for sensor fusion algorithms based on D-S, weighted D-S, and dynamically weighted D-S formulations. Experiments were done for a fixed total uncertainty budget distributed 5-to-1, 2-to-1, and 1-to-1 between the two sensors. For this analysis we define confidence as the largest probability value minus the next largest probability value:

$$c = p_{\max} - p'_{\max} \qquad (7)$$

Table 1 shows that the classical D-S algorithm does not always produced a result that is better than the best sensor alone, but it substantially increases confidence in the result. The weighted D-S algorithm, in which the classical D-S combination algorithm is modified by introducing weights related to individual sensor reliabilities, consistently produces results that are better than the best sensor alone, and it consistently improves confidence in the results. The dynamically weighted D-S algorithm, in which the weights are made adaptive to differential drift of the sensors, is in both respects only marginally better than the statically weighted D-S algorithm, but if increased drift were injected into the simulation its performance would be relatively stable in comparison with the static algorithms.

V. Conclusions and Future Work

We have identified confidence fusion as an issue in sensor fusion that has to date received little attention. We have defined the issue and qualitatively discussed our expectations for how confidence fusion ought to behave in experiments involving two one-bit binary-output sensors in which we have different confidence. We have suggested some quantitative definitions of confidence in this context, but point out that each is somehow unsatisfying to our intuitions in some regime of the parameters. We have conducted numerical experiments that simulate previously described experiments in sensor fusion for context aware-computing, and shown that the numerical results for this complex scenario are consistent with the analytical results we obtained in the two one-bit-binary-output sensor scenario. Future work is needed to obtain consensus of the applications-oriented community regarding practical-to-implement and intuitively satisfying definitions of confidence in sensor fusion scenarios.

References

[1] Huadong Wu, "Sensor Data Fusion for Context-Aware Computing Using Dempster-Shafer Theory," Ph.D. dissertation, Carnegie Mellon University, December 2003.

[2] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang, "Sensor Fusion Using Dempster-Shafer Theory," in IEEE International Measurement Technology Conference (IMTC) 2002. Anchorage AK USA: IEEE, 2002, pp. 7-12.

[3] H. Wu, M. Siegel, and S. Ablay, "Sensor Fusion for Context Understanding," in IEEE International Measurement Technology Conference (IMTC) 2002. Anchorage AK USA: IEEE, 2002, pp. 13-18.

[4] H. Wu, M. Siegel, and S. Ablay, "Sensor Fusion Using Dempster-Shafer Theory II: Static Weighting and Kalman Filter-like Dynamic Weighting," in International Conference on Measurement Technology (IMTC'2003). Vail, Colorado: IEEE Instrumentation and Measurement Society, 2003.

[5] M. Siegel and H. Wu, "Objective Evaluation of Subjective Decisions," in SCIMA-2003. Provo, UT: IEEE, 2003.

[6] See http://www.fon.hum.uva.nl/Service/Statistics/McNemars_test.html, or any standard statistics textbook.

[7] David L. Hall and James Llinas (editors), "Handbook of Multisensor Data Fusion," CRC Press, June 2001, ISBN 0849323797.

Table 1: Comparison of outcomes of numerical experiments on sensor fusion and confidence using two sensors and sensor fusion algorithms based on Dempster-Shafer, weighted Dempster-Shafer, and dynamically weighted Dempster-Shafer approaches.

sigma(S1)	sigma(S2)	prior probability	S1 correct	S1 confidence	S2 correct	S2 confidence	fused correct	fused confidence	DS correct	DS confidence	wDS correct	wDS confidence	dwDS correct	dwDS confidence
5	25	not-biased (0.33, 0.34, 0.33)	99.10%	0.9821	73.84%	0.5024	98.69%	0.8784	99.03%	0.9844	99.13%	0.9729	99.13%	0.9739
5	25	biased (0.3, 0.6, 0.1)	99.00%	0.9801	76.13%	0.5359	98.36%	0.8204	98.79%	0.9791	99.02%	0.9686	99.09%	0.9714
10	20	not-biased (0.33, 0.34, 0.33)	94.98%	0.8995	80.27%	0.6148	95.95%	0.7008	96.08%	0.9379	96.04%	0.8767	95.95%	0.8778
10	20	biased (0.3, 0.6, 0.1)	94.81%	0.8963	81.40%	0.6325	93.25%	0.6619	95.61%	0.9325	95.82%	0.8733	95.91%	0.8778
15	15	not-biased (0.33, 0.34, 0.33)	87.71%	0.7559	87.72%	0.756	94.07%	0.6288	94.15%	0.9078	94.06%	0.8036	93.50%	0.8065
15	15	biased (0.3, 0.6, 0.1)	87.96%	0.76	87.96%	0.76	90.49%	0.6091	93.64%	0.9056	93.63%	0.8095	93.25%	0.8145