Information fusion for CB defense applications: Challenge and opportunity

Citation: Braun, J.J. et al. "Information fusion for CB defense applications: Challenge and opportunity." Technologies for Homeland Security, 2009. HST '09. IEEE Conference on. 2009. 493-502. © 2009 Institute of Electrical and Electronics Engineers.

As Published: http://dx.doi.org/10.1109/THS.2009.5168078

Publisher: Institute of Electrical and Electronics Engineers

Persistent URL: http://hdl.handle.net/1721.1/55374

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Information Fusion for CB Defense Applications: Challenge and Opportunity [1]

Jerome J. Braun, Yan Glina, Austin Hess, Timothy J. Dasey, Edward C. Wack

Lincoln Laboratory, Massachusetts Institute of Technology
244 Wood Street, Lexington, Massachusetts 02420-9185, USA
Phone: (781) 981-5316, Email: jbraun@LL.mit.edu

Abstract —*Information fusion aims at a synergistic exploitation of multiple information sources, to enable decision-making or to enhance its performance. The sources include sensors of various modalities and characteristics, as well as sources which are not sensors. We discuss the challenges of information fusion, technical directions and approaches that can cope with those challenges, and the opportunities those approaches enable. Several research and development efforts conducted at MIT Lincoln Laboratory over the recent years in the area of information fusion for chemical-biological defense serve as the backdrop for that discussion.*

1. INTRODUCTION

Demanding decision-making tasks in chemical-biological defense (CBD) can benefit from, and some of them require, reliance on multiple information sources of diverse modalities. Information fusion aims at a synergistic exploitation of multiple information sources, to enable decision-making or to enhance its performance. The sources include sensors of various modalities and characteristics, as well as sources which are not sensors.

Detection and identification are fundamental among the CBD decision-making tasks. Accordingly, a significant portion of our information-fusion efforts has been in those areas, in particular for biodefense. However, the role of information fusion extends to other CBD decision-making tasks. They include threat mapping, tracking, and propagation prediction. Threat mapping aims to provide information about the agent cloud shape and structure, and the tracking about its motion. The goal of propagation prediction is forecasting the cloud dispersion, its future position, shape and motion, based on past data such as sensor observations. Other decision-making tasks include consequence prediction and course-of-action (CoA) guidance or response guidance, i.e., the determination of the effects of attack, and the discovery of response options and tradeoffs.[13]

CBD information fusion presents significant challenges due to multisource data imperfections such as uncertainty, inconsistency, conflicts, incompleteness, ambiguity, *a priori* knowledge and model reliability limitations. Different algorithmic techniques, ranging from conventional to advanced machine-intelligence based approaches, vary in their ability to cope with those challenges and consequently exhibit varying levels of performance robustness and adaptability. For example, certain conventional approaches can be brittle due to their dependence on assumptions or models whose validity is problematic in a given application. Other approaches may not be able to accommodate information that is highly imperfect or uncertain. Yet others may not be able to cope with multiple information sources that are sufficiently dissimilar in some aspects. Some approaches can perform well when carefully tailored for a specific application but may not scale well or may not be feasibly transferrable to other applications.

The trade-offs involving the above factors are complex. Moreover, the capabilities and limitations of various algorithmic approaches are complex, and often require a high level of expertise, experience, and understanding of difficult concepts underlying a given algorithm. In addition, the performance of many algorithms cannot be predetermined on theoretical grounds but instead must be

[1] This work was sponsored by the Department of the Air Force under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Part of this material is based upon work that was supported by the National Science Foundation under Grant No. 0329901. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

978-1-4244-4179-2/09/$25.00 ©2009 IEEE

determined by properly designed and conducted computational experiments.

In this paper we discuss some of the information fusion challenges, we point out the limitations and advantages of some approaches, and we argue for those we consider particularly promising. Our discussion starts with examining the essence of information fusion in section 2. In section 3 we take a closer look at the challenges and fundamental issues in information fusion. In the recent years, MIT Lincoln Laboratory conducted several research and development efforts focused on CBD information fusion, and in section 4 we briefly outline some of the approaches we have developed. The three different approaches outlined in that section all involve to a varying degree the methods of machine-intelligence, i.e., machine learning or reasoning. Our choice to outline those was motivated by our belief that the machine-intelligence based information fusion paradigm holds particular promise. The discussion of the path to realizing the potential of information fusion for decision support and decision-making systems is expanded in section 5.

The treatment of topics in this paper reflects our intent to provide a broad overview for the science and technology professionals of the homeland-defense community and to make the paper accessible to a broad range of readers in that community. Consequently, we focus on concepts while refraining from the detail and mathematical aspects underlying the topics discussed. Those aspects can be found in our publications and other works listed in the references section.

2. INFORMATION FUSION: NATURE AND SCOPE

Informally, information fusion can be viewed as the process of synergistic combining of data from multiple sources. This definition, while correct, brings about the following fundamental questions. What is the goal of such synergy? Under what circumstances does it make sense to combine information from different sources? Are there instances in which such combining does not make sense? This section helps to answer some of these questions.

Information Fusion and Decision-Making. At first glance it might appear that there is a wide range of possible goals of information fusion. However, in actuality, any use of data sources serves making decisions of various types. In conceptually-simplest cases these decisions are merely solutions to specific constrained problems or performing relatively simple actions such as repetitive operations of a manufacturing robotic device on a factory floor. Other applications, for instance those of an aircraft autopilot, combat-aircraft fire control system, or a flight system in a UAV, require a much more complex decision-making.

In those instances where the role of automation is to assist a human – the role often referred to as decision support – decision-making involves generating the answers or guidance that is provided to the human who then makes decisions at the next, higher level. Increasing automatic decision-making robustness allows decisions that previously required a human decision-maker to be made automatically. Those decisions then become decision support to the next level in a potentially long hierarchy of decision-making. For some applications autonomous systems can be built with the current technologies. In other cases it is possible to at least envision such total autonomy with the existent or near-future technologies. Finally, for those tasks that require human-like intelligence, it is possible to envision a cognitive machine-intelligence system based on conceivable future technologies (or to pronounce that such tasks will always be left to humans – dependent on one's degree of skepticism).

Therefore, information sources ultimately serve as inputs in a chain of progressively larger-scope decision-making tasks, including decisions that lead to actions. Thus, the common underlying goal of information fusion is decision-making.

Combining information sources outside the decision-making context is not a well-defined problem. Within the decision-making context the utility of combining can be defined as that of improving the decision-making quality which can be measured by relevant measures of performance (MoP) such as Receiver Operating Characteristics (ROC) curves.

Could the fusion process be used independently of decision-making just to improve the fidelity of one type of data by combining it with another? While such improvement could be quantified by, say, some data-precision MoP, this would constitute an exceedingly narrow scope (and even then it is making decisions about the values of data). More importantly, however, decision-making has an obvious intrinsic value as a crucial enabler of recognition or proper actions. In contrast, improving the quality of data in itself does not serve any specific purpose. If – for any given task, application, or activity, whether performed by human or by machine – proper decisions and actions can be determined with limited-quality data, improving the quality of those data serves no meaningful purpose and is unnecessary.

Thus, defining information fusion as a concept outside the decision-making context is not meaningful. Rather, information fusion is a form of decision-making, based on information from multiple sources which provide often-imperfect data. Essential decision-making aspects such as automatic learning, inference and reasoning are its inextricable components.

Information fusion models. Currently prevailing information-fusion models are in agreement with the above viewpoint. Currently the most influential of those models – largely due to its acceptance in the DoD circles – is the so-called JDL model. The JDL model effort was initially started in the 1980s by the Joint Directors of Laboratories (JDL) organization, hence its name. It has, since then, undergone significant revisions. At the present time, the

JDL fusion model consists of five main entities, referred to as *levels*. The sketch of the five levels provided below follows [15] and the details can be found there.

The constituent levels of the JDL model delineate the types and the end-goals of tasks that comprise them. All levels assume the use of information from multiple sources. Level 0 is devoted to the initial processing of the multisource data, and may include input standardization and conditioning, preliminary filtering, and the like. Level 1, called object refinement or assessment, focuses on individual "objects" or entities of interest. In the CBRN domain, an "object" might be a biological or chemical agent cloud. The main outcomes of Level 1 are estimates of entities' presence, identity, location, and other attributes.

Level 2, called situation refinement or assessment, considers objects in context of other objects. Namely, relationships among entities and their relationships to the environment are established, and the interpretation of the evolving situation that comprises multiple objects is developed. In particular, Level 2 functionality includes not only spatial and temporal aspects, but also environmental ones such as weather. It may involve multiple perspectives, such as those of friendly and adversary forces, and possibly the neutral perspective. Level 3, called threat refinement or assessment, extends the goals by projecting into the future to assess risks, threats and impacts. This includes estimating force capabilities, predicting adversary's intent, identifying opportunities and predicting implications – as in Level 2, possibly including multiple perspectives.

The goals of Level 4, called process refinement, are essentially internal to the information fusion system itself. Its goals include monitoring the system to improve performance, modeling sensors, optimizing utilization of the information sources (sensor tasking for example), and optimizing algorithms. Measures of performance are among the goals of this level. Level 5, called cognitive refinement, is concerned with interaction between the system and its users. Its goals include assessing and optimizing the quality of human/system interface facilities to improve the human-machine interaction (HMI). Its focus is the management of that interaction to enhance the user's cognitive performance.

JDL model interpretation. We emphasize that the JDL model is *not* a computational architecture. Neither is it a general design of a fusion system, nor is it a system or information flow. Rather, the model attempts only to partition information fusion tasks, so as to facilitate understanding and communications among the various communities, from theoreticians, through developers and evaluators, to managers, sponsors, and users of the fusion techniques.[15] This is envisioned, in turn, to result in more successful and cost-effective design, development, and operation of information fusion systems.

3. FUNDAMENTAL ISSUES AND CHALLENGES

To discuss the fundamental challenges of information fusion in certain demanding applications we return now to the CBD domain. The challenges facing CBD information-fusion and decision-making systems are significant. They include high clutter of the ambient background (biological or chemical) and unknown and unexpected changes in that clutter. Sensing for agent presence in such conditions can often result in low signal-to-noise ratios. The information provided by the potentially relevant multiple sources can be uncertain, disparate, conflicting or incomplete. In some cases the information can be inherently imprecise or vague, such as in case of descriptive data or human language statements (e.g., describing a threat level). Given the present state of knowledge, some phenomena (certain biological phenomena for instance) are poorly understood, and the reliability of relevant models is limited. In the remainder of this section we discuss some of these challenges.

Uncertainty. In biological and chemical sensing, and in many other applications, the data available from the sensors and other information sources exhibit various levels of uncertainty. Numerical imprecision of sensor data is one example of information uncertainty. Detection, identification, and other decision-making tasks listed in section 1 all must cope with uncertainty, to be able to make good decisions based on inferior-quality data. Therefore, a suitable representation of uncertainty, and algorithmic provisions for recognition and reasoning under uncertainty, are among the key issues in information fusion.

At the first glance, it may appear that all that is needed to represent and deal with uncertainty is conventional probability. In fact, approaches based on Bayes' theorem are amongst the most widely applied mechanisms. Such approaches are not always sufficient. One of the limiting factors is the problem of data scarcity, discussed in the next subsection.

There are, however, much more fundamental problems that in some instances limit the applicability of probabilistic methods as the mechanism for dealing with uncertainty. For sufficiently imprecise or vague information elements it may be difficult, or even impossible, to provide a valid probabilistic description of their uncertainty.

CBD data scarcity. Availability of ample amounts of data pertaining to the entities of interest in the context of a given decision-making task is paramount for many approaches to information fusion. Some approaches place particularly high demands with regards to the amount of data they require.

For example, traditional probabilistic approaches require sufficient amounts of data to construct probability distributions, both the priors and the conditionals. In the CBD realm, some of these are not determinable in a frequentist manner at all. The *a priori* probability of attack belongs in that category because of the fortunate rare-event character of CB attacks. Measurements involving simulants of agents and measurements of backgrounds, i.e., ambient aerosol levels, can in principle provide data needed for conditional (attack or no-attack given sensor indications) probability distributions. However, a proper estimation of probability distributions requires sufficient statistics, i.e., availability of sufficient amounts of data for the distribution estimation process.

In CBD realm, acquiring data is challenging. Simulant release tests involve observing the simulant cloud with relevant sensors as it progresses during the release. These tests are relatively difficult technically and logistically, and therefore are relatively expensive. That constitutes a practical constraint on the amounts of release data that can be obtained.

Technical and logistic difficulties and cost issues also exist in the case of background characterizations. Background conditions are characteristic of a specific setting or locale. Therefore, separate background data collection experiments are needed for different settings. Moreover, backgrounds are non-stationary and may exhibit short-term as well as long-term variability. Therefore, a comprehensive characterization of background for a given setting requires a continuous sensing for extended time duration. Deploying sensors of appropriate modalities for durations that would allow collecting sufficient quantities of background data can be costly and logistically difficult in practice.

The foregoing implies that information fusion approaches and algorithms in the CBD realm must often face a challenge of the scarcity of data available for their development. That limits the suitability of certain approaches in the CBD realm. As mentioned above, a traditional approach involving probability distribution estimation as a precursor for a direct application of Bayes' theorem is an example of directions that would be particularly challenged by such conditions. The absence of sufficient data necessary for a valid estimation of distributions and for the verification of the assumptions underlying that approach may result in brittleness of performance or render the approach invalid. Certain more advanced approaches are better equipped to cope with the data scarcity challenge.

Models. Reliable theoretical models of the phenomenology pertinent to a given information-fusion task have a twofold benefit. First, such models could constitute a solution to the decision-making process itself. For instance data association and estimation methods such as Kalman filters excel when the process models are well known. Second, good-quality models could be used to provide ample quantities of data needed for evaluating the probability distributions discussed in the previous subsection.

However, modeling of the CBD-realm phenomena is particularly challenging. For example, modeling of biological agent or pathogen properties, effects, gene expression responses, and the like, is limited by the current state of biological knowledge. The physics of agent transport and dispersion of agent plumes involves complex nonlinear phenomena, making the modeling task much more challenging than in many other application domains where more straightforward models, e.g., based on classical mechanics may be sufficient.

Furthermore, for CBD information fusion tasks the transport and dispersion model should be able to provide data on specific instances of the plumes, including their individual histories as they progress and meander. In principle, such information can be obtained from computational fluid dynamics (CFD) models. However, these models are extremely computationally-intensive, thus requiring formidable computing resources and time. Finally, the validation of transport and dispersion models is also an issue.

Disparity. Difficult decision-making tasks typically require exploiting information sources of multiple modalities. However, sensing modality differences do not necessarily imply disparity.

In [2] we investigated the essence of the disparity problem. We argued that disparity pertains to the information the source delivers in the context of a specific application, rather than to the modality or the physical characteristics of the sources. We proposed three types of disparity we referred to as the data characteristics disparity, hypothesis space disparity, and information disparity.[2] The latter two can be further decomposed into multiple categories. The specifics of these types and categories are beyond the scope of this paper and can be found in [2].

Therefore, from the information fusion viewpoint, dissimilar information sources should not always be considered disparate. The disparity problem exists when the data they provide in context of a given task meet some of the conditions for disparity types we specified in [2]. Not all information-fusion approaches can cope adequately with those conditions. Furthermore, certain approaches can cope with one type or category of disparity, but not with another. In [2] we have discussed the potential of one of the approaches we have developed (the FLASH approach we outline briefly in section 5) to deal with the various facets of disparity.

Unstructured problems and vague information. Even detection and identification information fusion tasks may require moving beyond the well-structured formulations such as probabilistic approaches. For instance, to establish the *a priori* probability of an attack may require unstructured concepts, such as those related to social networks of actors that may underlie the likelihood of a given attack. Furthermore, the exploitable information may require a less precise than probabilistic representation.

In progressively more demanding CBD information-fusion tasks at JDL levels 2 and 3, these unstructured aspects become increasingly more dominant. For example, CoA guidance or intent prediction tasks require considering a complex and potentially large number of details related to things such as an expected behavior of actors or crowds. Here, the uncertainty aspects become even more acute. Information fusion requires methods suitable for recognition and reasoning with such highly imprecise and vague data.

4. Information Fusion in CBD: Selected Approaches

Approaches to information fusion range from conventional decision-theory techniques to advanced state-of-the-art machine-intelligence methods. In the recent years, MIT Lincoln Laboratory conducted a number of research and development efforts focused on CBD information fusion. These efforts have been described in our publications, some of which are listed in the references section of this paper.

In this section we provide a brief overview of some of those efforts. The intent of that overview is to serve as a backdrop for the discussion of information fusion challenges and of some of the technical directions that could enable realizing the significant opportunity information fusion can offer to decision support and decision-making systems.

The more conventional approaches can constitute valid solutions especially for relatively simple problems. The higher-end, machine intelligence approaches can offer more robustness and versatility, and hold significant promise for the decision-support and decision-making systems of the future. Consequently, in this section we focus our attention on machine-intelligence approaches, briefly outlining three such approaches we developed.

Subway aerosol anomaly detection. Anomalies in the aerosol content in an environment can be an indication of a biological attack. Such anomalies include unexpected or inexplicable changes in particle concentrations. Sensors for measuring air particulate concentration are commercially available and some of them are relatively inexpensive. Many of such sensors, commonly called particle counters, can provide concentration data as a function of particle size. Particle counters belong to a category of so-called *point sensors*, because their measurements reflect the value of the sensed phenomena only at the location of the sensor.

Basic particle counters, especially the more economical commercial models, are non-specific in the sense that they do not offer information regarding the nature of the particles. Thus they can neither determine bioagent particle identity, nor whether the particles are biological. Given their limitations, one potential use of particle counters for bioattack detection is as a triggering mechanism for activation of more robust and specific means to confirm the bioattack prior to initiating response measures.

However, if an aerosol anomaly could be determined with sufficiently high fidelity, such determination could be used to initiate preliminary “low regret” measures such as ventilation system activation in indoor environments. Moreover, theoretically such determination could even be used for taking more drastic measures if the detection and false-alarm performance reach appropriate levels. This is problematic by means of a single particle counter, but might be achievable by the fusion of data from multiple such sensors and other information sources.

In one of our early biodefense-related projects we developed a machine-intelligence based approach for aerosol anomaly detection in a subway environment.[3][4] The sensing environment consisted of particle counter sensors placed at three different locations within a subway station, sonic anemometers for air velocity (wind) measurements, as well as rudimentary low-cost radars and simple optical beam type sensors for train traffic monitoring.

The key challenges represented in this effort were as follows. The background particulate concentrations in a subway setting are high and non-stationary. This amounts to low SNR, or more precisely low signal-to-clutter. Some of the variability is due to the effects of airflow induced by the train motion. The computational approach had to accommodate sensors of multiple modalities. Finally, aerosol phenomena in general, and for the dynamic settings such as a subway in particular, are too complex to be modeled adequately.

The anomaly-detection machine intelligence approach involved a set of two feedforward neural networks (FNN). One of the networks was responsible for the detection in presence of train passage, the other for quiescent times. The network selection process was automatic. Supporting the above machine-learning component, were a multi-sensor feature extraction component and a feature-space dimensionality reduction component based on principal component analysis (PCA).

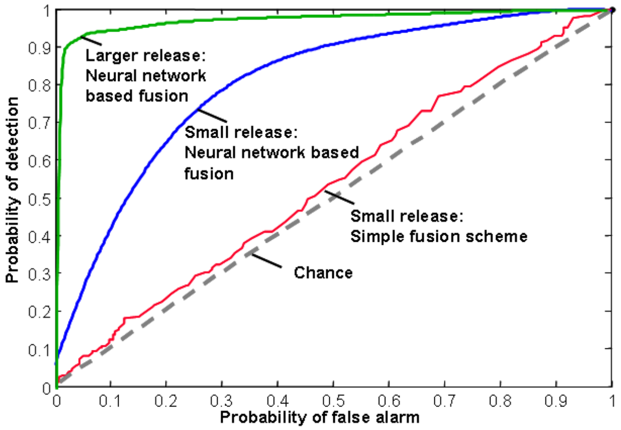

Figure 1 – FNN fusion for anomaly detection [4][13]

Computational experiments we have carried out with the developed approach and using the data collected in the subway showed the advantage of the approach we proposed.

This is illustrated in Figure 1, which summarizes the ROC curves we presented in [4].

Pathogen identification. Another of our prior efforts aimed at the identification rather than detection task.[5][6][7] The overall approach involved gene expression, microarray-based sensing, and machine-intelligence based information fusion. In combination, that approach had a number of significant potential advantages. These included applicability to various types of pathogens (bacteria, viruses, toxins, etc.), operation without the knowledge of pathogen's genome sequence, and the potential for identification of uncataloged, i.e., emerging pathogens. The biological genomic aspects of this effort are beyond the scope of this paper and can be found in [6]. The brief sketch that follows touches only on the aspects pertinent to the information fusion topic.

The key challenges in this effort were as follows. The information sources in this case were the microscopic DNA probes that comprised a cDNA microarray. Each microarray contained approximately 10,000 of such probes. Each probe conveys information about the level of expression of a particular gene, potentially indicative of the pathogen identity. Thus, each probe acts as a separate sensor. The fusion process, therefore, involves fusing a massive number of these sensors, while any single microarray, each containing thousands of the DNA probes, constitutes a single case or exemplar. Therefore, the data for each exemplar form a vector whose cardinality is that of the number of microarray probes, leading to an extremely high dimensionality of the data space (on the order of 10,000 in our case).

It should be pointed out that the the microarray probes cannot be viewed simply as pixels in an image. This is because pixels in a non-random image are presumably related via a higher level structure corresponding to the entities the image represents. In contrast, for the microarray probes such underlying structure cannot be assumed in general. The relationships between the probes are often unknown, because of the unknown nature of the gene expression aspects for the pathogens to be identified. The same reason implies the unavailability of predictive models to aid the identification process. Additionally, due to various biological and technical reasons, microarray probe data exhibit significant levels of noise and uncertainty.

The only realistic approaches for pattern recognition in the microarray domain appear to be those of machine learning. However, supervised learning methods require an ample number of exemplars available for the classifier training process. This however leads to the following difficult trade-off. Either most of the probe data for each sample are neglected as irrelevant, or the feature space dimensionality is much higher than the number of training exemplars (because collecting a very large number of microarray data is impractical for cost and effort reasons). In absence of relevant biological models, the former involves the risk of discarding valuable information, while the latter poses a serious challenge for classifier training.

The fusion and recognition architecture we developed in the context of this effort involved algorithmic mechanisms for partitioning of the high-dimensional feature space into smaller, though still high-dimensional subspaces. The feature vectors from those subspaces constituted inputs to multiple Support Vector Machine (SVM) based recognizers, one per subspace.[8][19][18] Our choice of SVMs was motivated in part by the advantages they can offer in terms of dealing with data scarcity and multi-dimensional spaces. The methods we introduced for characterizing the efficacy of the feature subspaces were used to further fuse the results of the multiple subspace SVM recognizers into the final identification outcome using techniques based on the Dempster-Shafer theory of evidence [17][20][12].

Performance studies for the above approach included identification of four pathogens including bacterial, viral, and toxin. The results of the computational experiments with the approach we developed [5][6] indicated that, within the constraints of the data we had in that effort, high levels of identification accuracy can be obtained using that fusion approach.

FLASH. Each of the many possible machine learning and reasoning paradigms has its specific strengths and limitations. The differences include such aspects as the ability to support feature-level vs. decision-level fusion, robustness to data uncertainty, disparity, suitability for processing time-series such as the sensor signals, and more. A key idea underlying the FLASH (Fusion Learning Adaptive Super-Hybrid) information fusion approach and architecture we developed in another one of our recent past efforts [1][2][8][9][10] was to combine multiple method types in a cohesive integrated hybrid structure. This creates powerful synergies between the various constituent methods, and enables exploitation of their respective strengths and compensation for their respective weaknesses.

Since some of the FLASH constituent parts are or can be hybrids themselves, FLASH constitutes a hybrid of hybrids. The methods that comprise FLASH are not simply a collection of multiple techniques. They were selected judiciously, based on the specifics of their respective functionality and roles. Their integration and collaboration within FLASH were among the significant aspects of that research.

Cognitive processing orientation is another noteworthy aspect of FLASH. This stems from the fact that its most essential constituents are machine-learning and reasoning components, and from the fact that some of its aspects are rooted in or inspired by certain advances in neuroscience and human cognition studies.[1] The major constituents of FLASH range from low-level processing (signal processing, feature extraction and selection, to name just a few) to multiple progressively higher recognition and reasoning levels. This structure is not, however, strictly hierarchical, in principle forming a multiply-interconnected set of entities. In particular, the recognition components – which may perform detection or identification of events, objects, assertions, or patterns of interest – operate at multiple levels of detail, from more specific or local to more general or global. A more detailed discussion of the cognitive nature of FLASH can be found in [1].

FLASH has multi-purpose applicability. However, its first embodiment, implemented in the effort outlined here was in the context of a bioattack detection task. A multisensor testbed was constructed to serve as a source of input data.[14] The testbed included aerosol particle sensors, biological point sensors, and airflow sensors. Some additional contextual information inputs such as the threat-level estimates were simulated.

Briefly, FLASH-1 consisted of two principal recognition stages at progressively higher levels, referred to as the instance level recognition and the time-series recognition. The former aimed at recognizing events within a limited temporal scope, and involved Support Vector Machines (SVMs). The latter aimed at considering a wider temporal scope; that level involved Hidden Markov Models (HMMs). The instance level recognition stage was preceded by lower-level processes such as feature extraction and selection, and several other important auxiliary components. One of those auxiliary components, referred to as background clustering was tasked with the mining of multisensor data to categorize possible types of backgrounds (ambient air content types). Adaptive Resonance Theory (ART) type neural network methods were used for that purpose. Prior to its delivery to the time-series recognition stage, the outcomes of the instance-level recognition were fused using techniques based on the Dempster-Shafer theory formalism. The outcomes of the time-series recognition stage were further fused with additional context information using fuzzy reasoning methods.

The results of FLASH-1 performance experiments were presented in [1]. Those results, shown in Figure 2, indicated a steady performance increase as major stages of FLASH-1 were progressively added. This constituted the proof of concept for FLASH and its structure.[1][2]

The key challenges in the FLASH effort included multiple information sources of various types, data uncertainty and disparity. We have demonstrated in that effort the promise of addressing all of these challenges with an advanced cognitive-processing oriented machine-intelligence approach. The adaptability potential, for instance to changes in the local conditions or system deployment settings, is another potential benefit of FLASH. Finally, although FLASH-1 was developed in the context of bioattack detection, we believe that the FLASH concepts and architectural principles have a broad multi-purpose potential for variety of CB and other homeland defense decision-support tasks.

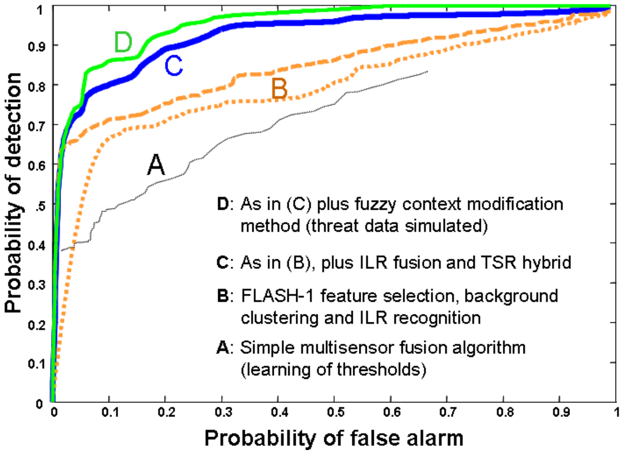

Figure 2 – FLASH-1 computational experiments [1]

5. Realizing the Opportunity

Overcoming the challenges. The efforts we briefly outlined in the preceding section 4 suggest strongly that the challenges involved in the exploitation of imperfect multisource data towards robust decision-making can be overcome with appropriate information-fusion algorithmic methods.

When multi-purpose suitability and scalability are not a concern, the algorithmic approach choice and design can be driven by the nature and the requirements of the specific application. For some applications, basic "cookbook"[1] probabilistic approaches might be sufficient, but in some cases a more careful examination might reveal that the application specifics stress or violate the assumptions needed for the operation of those techniques. Their potential brittleness of performance is therefore a concern.

For cases where the use of basic probabilistic methods is problematic, heuristic approaches can be amongst potentially viable alternatives, and they should not be viewed as universally less rigorous than their conventional probabilistic counterparts. Heuristic approaches generally attempt to embody some aspects of human expert knowledge. Such knowledge is often rich and complex. A heuristic approach may subsume the information that could be represented in terms of probability distributions, and it can be less brittle especially if those distributions were constructed in a non-robust way, e.g., from insufficient data.

Decision-making tasks include recognition of entities and events, and reasoning with the recognition results and other information. Bayesian approaches, e.g., certain types of Bayesian belief networks, and a variety of probabilistic methods can be exploited in many cases. For some problems involving high levels of uncertainty, certain non-Bayesian theories are promising. Some non-Bayesian

[1] Term used by [16] in a related topic discussion.

reasoning formalisms can represent uncertainty in a more complex way than probabilistic constructs. For example, the Dempster-Shafer theory formalism involves the concepts of belief and plausibility functions, both of which characterize the uncertainty of a given information element. Fuzzy sets theory and its many variants are among other examples of promising non-Bayesian paradigms. Such approaches can also be useful in cases that require human expert knowledge since human experts find it often easier to provide their knowledge in terms less precise than probabilities.

Certain advanced machine-learning paradigms are capable of dealing particularly robustly with the problems of data scarcity and *a priori* knowledge (modeling) limitations. Some of these were mentioned in section 4 in context of our information-fusion efforts.

The next-generation decision-making and decision support systems – in CB defense, other homeland defense applications, and other application domains – will need to be highly adaptive and multi-purpose. For systems such as bio-detection, the former includes such obvious aspects as changes in conditions during the system operation, or flexibility in deployment locations. The multi-purpose functionality importance goes beyond economics in terms of meeting multiple needs with the same development investment. This is because next-generation systems will need to interoperate and collaborate, and the feasibility of achieving that with disjoint "single application" system solutions is not clear.

Machine-intelligence based approaches at sufficiently high level of sophistication can meet the challenges we discussed in this paper. We believe that hybrid approaches such as FLASH are amongst those that hold particular promise.

Developing machine intelligence systems. Advanced machine learning and reasoning algorithms are in most cases technically complex. The theoretical concepts and formalisms that underlie them are typically complex as well. Consequently, development of those algorithms requires a substantial level of expertise. This is true even when working with one of the many commercially or publicly available algorithmic toolkits. Proper setting of numerous algorithmic options and parameters requires a significant understanding of their underlying theoretical aspects. An inappropriate reliance on the toolkit's default settings can lead to disappointing performance or lack of robustness. The design and preparation of training datasets suitable for a given problem is also non-trivial.

One of the major themes in machine learning is the notion of *generalization*, defined as the ability of the algorithm to operate reliably with data not seen in the training dataset. The factors that can impact generalization robustness include the choice of approach, options and parameters. The developer must have a thorough understanding of the aspects involved in machine-learning algorithm testing, the requirements a specific technique may impose on the training datasets and their content, and of the testing and validation procedures needed to ensure generalization robustness and superior performance. When these desiderata are fulfilled, machine learning systems do not present insurmountable validation concerns.

Thus, the development of machine intelligence systems requires significant expertise, some of which is acquired by experience with the relevant techniques. However, it should be emphasized that, while machine-intelligence approaches place higher expertise demands on developers as the price for their power and promise, no such additional demands are placed on the end-users. On the contrary, properly designed machine-intelligence systems can be easier to use than conventional information systems.

Performance measures. Measures of effectiveness (MoE) of information fusion systems are fundamentally application-dependent. Furthermore, they depend on many operational aspects that may change from moment to moment. For example, selection of the optimal CoA may depend on specific goals – the best CoA to minimize casualties may differ from the best CoA to maximize achieving certain objectives of a given mission. Similarly, an alarm issuance following, say, a bioagent detection event may depend on the specific circumstances in which the system is used.

However, the measures of performance (MoP) of information fusion systems can be determined in a more objective ways and their determination must be part of the development process. This is done by establishing and evaluating MoPs appropriate for a given technique or system. There are many potential MoP choices available to the developer. Receiver Operating Characteristics (ROC) curves and Confusion Matrices are very useful. The former is most suitable for detection tasks but in some cases can also be used for identification tasks. Other useful MoPs include: cumulative match characteristics, cross-entropy, accuracy, sensitivity, precision, F-score, and more.

Performance studies leading to MoP evaluation require the "ground truth", that is the data for which the true values, such as the object or event identity, are known. Sometimes the ground truth data are difficult or infeasible to acquire, such as the agent release data in biodefense. In those cases appropriate simulations can be used as a reliable source of ground truth, and we have demonstrated this in our past work for the detection and identification tasks, e.g., [1][2][8][9][10]. However, as we argued in [11], the simulation-based approach can also be used in more unstructured tasks such as in JDL Level-3 impact assessment systems. It should be pointed out that in some cases the ground truth data may have to be subjective. For example, in some cases of the CoA guidance systems, the optimal CoA can only be available as a human expert opinion.

Development datasets. A necessary supporting element of a successful information-fusion system development effort is a body of appropriate multisource data for development and testing of the fusion algorithms. Generating such datasets often requires significant resources and expenditures that must be absorbed by a particular information fusion effort. As we mentioned in section 4, in one of our efforts we were compelled to include a multisensor testbed construction to gain the needed sensor background data. To obtain release data we developed appropriate release simulation capabilities.[3][4][1]

If datasets pertaining to selected cardinal homeland-security application domains were available, these supporting activities could be avoided. Collecting multisource data for information fusion efforts and the construction of appropriate datasets is non-trivial. It can be resource-intensive and costly. Furthermore, the datasets must be designed in the way appropriate for information fusion studies – not every collection of sensor data meets those suitability conditions. As mentioned earlier, information fusion work also places specific demands on the CB agent release simulation models. Specifically, individual instances of releases, as opposed to average representations, are required. Since the use of CFD models for generating the amounts of release simulations could be prohibitively expensive computationally, we developed a simpler and computationally efficient model for simulation of individual releases and we used that model in two of the efforts outlined in section 4.[3]][4][1] Such approaches are among the potential alternatives to full CFD simulation models for use in information fusion efforts.

Initiatives focused on collections of multisource data for information fusion efforts, appropriate simulation models, and on generation of standardized multisource information fusion datasets for relevant homeland-security sub-domains are clearly desirable. In addition to the obvious benefits for the fusion algorithm developers, such common and standard datasets will facilitate a more objective validation and performance comparisons of different approaches and systems.

Trade-offs and way forward. The foregoing discussion suggests a number of significant trade-offs that must be considered in information-fusion development efforts. More elementary approaches can offer good solutions mostly for relatively short-term goals and well-constrained problems and requirements. Within those confines, they require more moderate developer expertise and can be implemented with moderate effort. However, their utility will generally be limited, as will their scalability beyond those confines. Even for the short-term solutions, however, the basic "cookbook" approaches should be viewed with caution. As we discussed earlier, the assumptions underlying such solutions should always be thoroughly examined, but unfortunately such deliberations are too often compromised to meet the short-term cost and schedule goals.

Machine intelligence based information fusion appears to offer the best prospect in the long term and should be considered as the primary direction for the next-generation systems. The advanced hybrid approaches are particularly promising. They can alleviate the limitations of particular single-paradigm machine learning and reasoning methods.

As exemplified by our efforts outlined in section 4, machine intelligence based information fusion can outperform its conventional counterparts in such tasks as bioattack detection. As the machine intelligence approaches continue to evolve, they may start approximating human-level decision-making performance in a growing number of tasks and applications.

6. CONCLUSION

CBD information fusion poses significant challenges, ranging from prerequisites such as the ground-truth data generation to the foundational issues of information fusion algorithmic methods that could cope with CBD information characteristics, uncertainty and disparity. In this paper we discussed some of these challenges. We touched on limitations of some standard techniques. Drawing on some of our past information-fusion efforts, we discussed how the challenges of information fusion can be overcome by certain algorithmic directions such as the machine-intelligence based information fusion.

With appropriate algorithmic approaches and appropriately resolved tradeoffs, information fusion can offer to the CBD decision-making and decision support applications the potential of reaching performance that would be difficult, if not impossible, to attain otherwise. Thus, information fusion represents a significant opportunity for the CB defense and homeland security realm.

REFERENCES

[1] J. J. Braun and Y. Glina, "Hybrid methods for multisource information fusion and decision support," *Proc. of SPIE*, vol. 6242, 2006.

[2] J. J. Braun, Y. Glina, and L. Brattain, "Fusion of disparate information sources in a hybrid decision-support architecture," *Proc. of SPIE*, vol. 6571, 2007.

[3] J. J. Braun, Y. Glina, J. K. Su, "Urban Biodefense Oriented Aerosol Anomaly Detection using Sensor Data Fusion," *Proc. 1st Joint Conf. Battle Mgt. Nuclear, Chem., Biolog. and Radiological Defense*, November, 2002.

[4] J. J. Braun, Y. Glina, J. K. Su, T. J. Dasey, "Computational Intelligence in Biological Sensing," *Proc. of SPIE*, vol. 5416, 2004.

[5] J. J. Braun and S. P. Jeswani, "Information Fusion of Large Number of Sources with Support Vector Machine Techniques," *Proc. of SPIE*, vol. 5099, 2003.

[6] J. J. Braun, Y. Glina, N. Judson, K. D. Transue, "Biological agent detection and identification using pattern recognition," *Proc. of SPIE*, vol. 5795, 2005.

[7] J. J. Braun, Y. Glina, N. Judson, R. Herzig-Marx, "Multiclassifier information fusion methods for microarray pattern recognition," *Proc. of SPIE*, vol. 5434, 2004.

[8] J. J. Braun, “Sensor Data Fusion with Support Vector Machine Techniques,” *Proc. of SPIE*, vol. 4731, 2002.

[9] J. J. Braun, Y. Glina, D. Stein, E. Fox, “Multisensor Information Fusion for Biological Sensor Networks and CBRN Detection,” *Proc. Conf. Science and Technology for Chem-Bio Information Systems*, October, 2004.

[10] J. J. Braun, Y. Glina, D. W. Stein, P. N. Skomoroch, E. B. Fox, “Information fusion and uncertainty management for biological multisensor systems,” *Proc. of SPIE*, vol. 5813, 2005.

[11] J. J. Braun, “Remarks on Performance Evaluation for Impact Assessment Systems,” Panelist position presentation, SPIE Defense, Security and Sensing Symposium 2008 (unpublished).

[12] J. J. Braun, “Dempster-Shafer Theory and Bayesian Reasoning in Multisensor Data Fusion,” *Proc. SPIE*, vol. 4051, 2000.

[13] T. J. Dasey and J. J. Braun, “Information Fusion and Response Guidance”, Lincoln Lab. J., vol. 17, no.1, 2007.

[14] Y. Glina, J. J. Braun, P. N. Skomoroch, K. D. Transue, “Multisensor data analysis and aerosol background characterization,” *Proc. of SPIE*, vol. 6218, 2006.

[15] D. L. Hall, S.A.H. McMullen, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, 2004.

[16] R. Mahler, “Random Set Theory for Target Tracking and Identification,” in *Handbook of Multisensor Data Fusion,* ed. Hall D.L. and J. Llinas, CRC Press LLC, 2001.

[17] G. Shafer, *Mathematical Theory of Evidence,* Publisher, Princeton University Press, Princeton, N.J., 1976.

[18] V. N. Vapnik, *Statistical Learning Theory,* John Wiley & Sons, Inc., New York, NY, 1998.

[19] V. N. Vapnik, *The Nature of Statistical Learning Theory,* Springer-Verlag New York, Inc., New York, NY, 2000.

[20] R. R. Yager, J. Kacprzyk, M. Fedrizzi, *Advances in the Dempster-Shafer Theory of Evidence,* J. Wiley & Sons, Inc., New York, N.Y., 1994.