A Multi-Disciplinary University Research Initiative in Hard and Soft Information Fusion: Overview, Research Strategies and Initial Results

Dr. James Llinas and Dr. Rakesh Nagi
Department of Industrial and Systems Engineering
University at Buffalo
Buffalo, New York, U.S.A.
llinas@buffalo.edu, nagi@buffalo.edu
Dr. David Hall
School of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania, U.S.A.
Dr. John Lavery
Army Research Office
Research Triangle Park, North Carolina, U.S.A.

Abstract – The University at Buffalo (UB) Center for Multisource Information Fusion (CMIF) along with a team including the Pennsylvania State University (PSU), Iona College (Iona), and Tennessee State University (TSU) is conducting research to develop a generalized framework, mathematical techniques, and test and evaluation methods to address the ingestion and harmonized fusion of Hard and Soft information in a distributed Level 1 and Level 2 data fusion environment. The primary Research Thrusts addressed are framed around the major functional components of the JDL Fusion Process; these include:

1. ***Source Characterization of Soft Data input streams including; human observation—direct, indirect, open source inputs, linguistic framing, and text processing.***
2. ***Common Referencing and Alignment of Hard and Soft Data, especially strategies and methods for meta-data generation for Hard-Soft data normalization.***
3. ***Generalized Data Association Strategies and Algorithms for Hard and Soft Data.***
4. ***Robust Estimation Methods that exploit associated Hard and Soft Data.***
5. ***Dynamic Network-based Effects on Hard-Soft Data Fusion Architectures and Methods.***
6. ***Test and Evaluation Methodology Development to include Human-in-the-Loop.***
7. ***Extensibility, Adaptability, and Robustness Assessment.***
8. ***Fusion Process Framework.***
9. ***Technology Concept of Employment.***

This program is a large, 5-year effort and considered distinctive in being a major academic thrust into the complexities of the hard and soft fusion problem. This paper summarizes the research strategy, the early technology decisions made, and the very early results of both design approaches and prototyping.

Keywords: Hard and soft fusion, information fusion, fusion process design, graphic theoretic methods, computational linguistics.

1 Introduction and General Research Philosophy

Experiences in Iraq and Afghanistan and other places in the world in dealing with insurgency/counter-insurgency problems have required the (ongoing) formulation of new paradigms of intelligence analysis and dynamic decision-making. Depending on the phases of counter-insurgency ("COIN") operations [1], the nature of decision-making ranges from conventional military-like to socio-political. Because of this wide spectrum of action, the nature of information support required has an equally wide range. Since automated Information Fusion (IF) processes provide some of the support to such decision-making, requirements for IF process design must address these varying requirements, resulting in considerable challenges in IF process design. Further, these experiences have also shown that some of the key observational and intelligence data in COIN operations comes from dismounted soldiers reporting on their patrol activities. (From the US Army Field Manual on COIN [1]: "Intelligence for current operations comes from a variety of sources, but operations reports are particularly important. This is because current enemy activities are more often reported by patrols, units conducting raids, or observation posts than they are by dedicated intelligence collectors.") These data are naturally communicated in language in the form of various military and intelligence reports and messages. Such data finds its way into IF processes as unstructured digitized text, and this input modality creates new challenges to IF process designs, contrasted with more traditional IF applications involving the use of highly-calibrated, numerically precise observational data. However, the input side also includes data from the usual repertoire of Hard-sensor data from various radio frequency (RF) sensors, video and other imaging systems, as well as SIGINT and satellite imagery. The current state of knowledge necessary to explore and prototype techniques for effective processing of Hard and Soft Data in a Fusion context is very limited [2]. Literature searches conducted by our Team have verified the limitations of the literature in this area.

One initial and important question in developing a research approach is to ask: "At what point in the processing flow of Hard and Soft Data streams do you fuse the data?"; i.e., what is the architectural framework? Although Framework structure is itself a research concentration area for this program, it can often be argued that the data should be joined at the closest point to the source observing mechanism, i.e. where the data are of a "raw" nature. This approach is often advanced on the basis of an information-theoretic argument, claiming that any operation on data loses valuable information. In this case however, we feel that, initially, there are two overarching, mitigating factors: (1) the fact that we know reasonably well how to process Hard data and fuse it (see for example [3], [4], [5]), and (2) that we hardly know at all how to process and fuse Soft data. This means that if we choose to combine the data modalities early in the processing stream, we immediately encounter high risk. Thus, our evolving approach will *initially* separate fusion processes for each data modality and architect the process as an Estimation-Fusion process analogous to a Track Fusion approach versus a Measurement Fusion approach. This approach allows immediate investigation of Hard-Soft fusion at the report level (while including parallel research of issues related to source characterization, data association, etc.) without immediately requiring general cross-domain mathematical models that span the hard (physical observation domain) and the human observation domain (see also Section 4.5). The approach can also take advantage of existing, early prototypes for each of Hard and Soft data that exist across the Team. Cross-domain models are challenging enough in hard sensor fusion, and indeed still do not exist across very heterogeneous sensor sources. We view our strategy as a Technology-Risk-Mitigation approach for the initial research path for the program. Given that progress can be made on the needed science and on Framework definition, a Measurement Fusion-based or other architecture can be explored at a later point in the program; additional ideas for processing will of course come from the Framework study task.

One critical path in our research approach is that toward achieving a capability to fuse multiple, disparate Soft Data sources. One core Soft data source is the content of human report messages; we see a second Soft source as coming from Open Sources such as web-based information [6]. One primary candidate and broadly-relevant Open Source for military/security-related problems would be Broadcast News/Newswire-type feeds; of course access to the web enables all types of additional Soft data options. There are at least two challenges toward realizing an automated capability: (1) designing an automated Web Harvesting function that provides the means to fetch Newswire feed (or possibly other) data, and (2) designing a Relevance Filter that determines what information that is streaming in the feeds is relevant to that in the human-based observational data. News data will provide both "pseudo-observational" data (i.e. entity-specific) as well as contextual information for both consistency-checking and amplification of the fused estimates resulting from the human observational data. Note that the contextual information may also aid in forming an improved Soft data estimation algorithm, coupled with the human observational data (in a kind of "a priori" mode, i.e., incorporating the contextual information in the algorithm design), as well as *subsequent* to the formation of a fused estimate from that data, for consistency checking (in an "a posteriori" mode, examining the consistency of state estimates with yet additional contextual information).

In this MURI program, we are leveraging another of CMIF's ongoing research programs in Contextual Exploitation with the Office of Naval Research. That program has already defined the architecture for a priori/a posteriori contextual exploitation, as well as a "Relevance Filter" and the reasoning logic for checking a fused hypothesis for consistency with related contextual information. These processes have already been prototyped using extensions of semantic network based reasoning techniques. In this approach, we now have a *multiple-modality Soft Data Fusion process*; the details are shown later in Figure 2.

Similar to the Soft data case, our initial Hard Data Fusion process is being evolved from our Tennessee State University (TSU) Teammate's existing laboratory prototype (funded by the ARO Center of Excellence in Battlefield Sensor Fusion), that has Visual, Acoustic, and RF sensors linked into a asymmetric-like test scenario problem environment (Figure 1). TSU researchers have developed the ability to process visual, acoustic and RF sensors to perform target location, identification and characterization and to observe targets such as moving vehicles and humans. Algorithms have been developed to perform signal characterization (e.g., image processing, extraction of information from video images, signal processing, etc.) and feature extraction which leads to fusion processing at both the feature level and at the report or decision level.

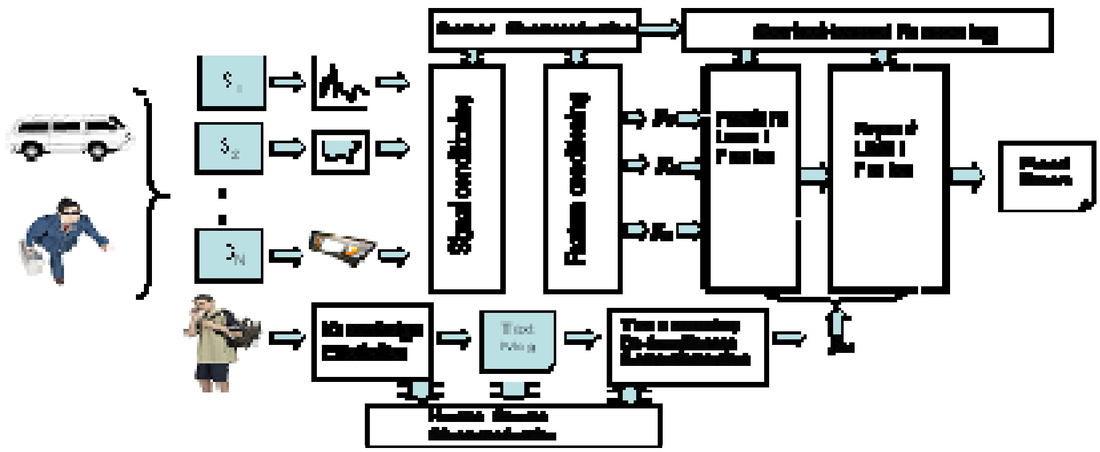

Figure 1Nominal Hard Data Fusion

The overarching Hard + Soft data fusion notional architecture that combines these capabilities is shown in Figure 2. Here it can be seen that this is a data-modality-specific approach as argued for above, with the final state-estimate fusion being done at the network level. It is envisioned that data-specific fusion nodes would be reflective of echelon-specific fusion capabilities that communicate their estimates to a network enterprise bus for system-level fusion.

2. Innovation

This project is addressing a very complex problem that will require employment and extension of some known methods, creation of new methods, and hybrid use of multiple methods. Our approach involves two threads of innovation for this program: (1) extensions of already-implemented and prototyped innovations in Soft Data Fusion using graphical-science-based methods, Fuzzy Divergence and new semantic similarity techniques for soft-data association, extended semantic network-based reasoning, automated methods for semantic labeling in imagery, and Feature Energy-logic Assessment Modeling (FEAM) technique by TSU for multi-modality distributed Hard/Soft sensor agents, and (2) a multidisciplinary, leveraged research approach involving social science, linguistics (e.g. word sense disambiguation, latent semantic analysis), mathematics (e.g. clouds, generalized p-boxes, finite-set statistics), estimation (transferable belief model, Dezert-Smarandache Theory, Bayes with Imprecise Probabilities, Conceptual Spaces), as well as extensions of CMIF current research on methods for Test and Evaluation ([7], [8]) involving for example large-factor-space experimental design techniques ([9]), and research on new metrics such as trustworthiness of fusion processes. The high level functional architecture for this approach is shown in Figure 2.

3. Critical Research Choices

There are also some critical research investment decisions that need to be made when addressing the challenges described for this effort:

I. How to approach the need for a computational-linguistics capability that processes the textual Soft data to an adequate level of semantic clarity without reinventing the extensive existing technologies of the natural language processing (NLP) and computational linguistic communities; this will require a "make-buy" decision that exploits existing technology while making new contributions to this area of research.

II. An assessment of how to approach the quantification of human observational capability that allows such observational data to be qualified in terms of accuracy and reliability; this has to do with the notions of trusted agents and an ability to characterize such data in a sufficiently general way without having to engage in an extensive series of controlled experiments to obtain the needed quality measures.

III. Developing a strategy for an adequately-rich test data set that is truthed and that represents enough problem-domain variability that it allows testing developed technologies for both accuracy and robustness. The ARO is providing some data for this program but we are also developing an adjunct data set that needs to be defined in a way that satisfies the above accuracy/robustness criteria.

IV. Developing an approach to research that yields near-usable technologies that fit into viable Army concepts of employment. While this is a basic research program, we are working with the US Army to understand how these technologies may fit into either existing operational frameworks or into forecasted, new operational frameworks. In spite of this being an academic research program, we want to avoid developing technical capability that would need to be significantly reworked for practical use.

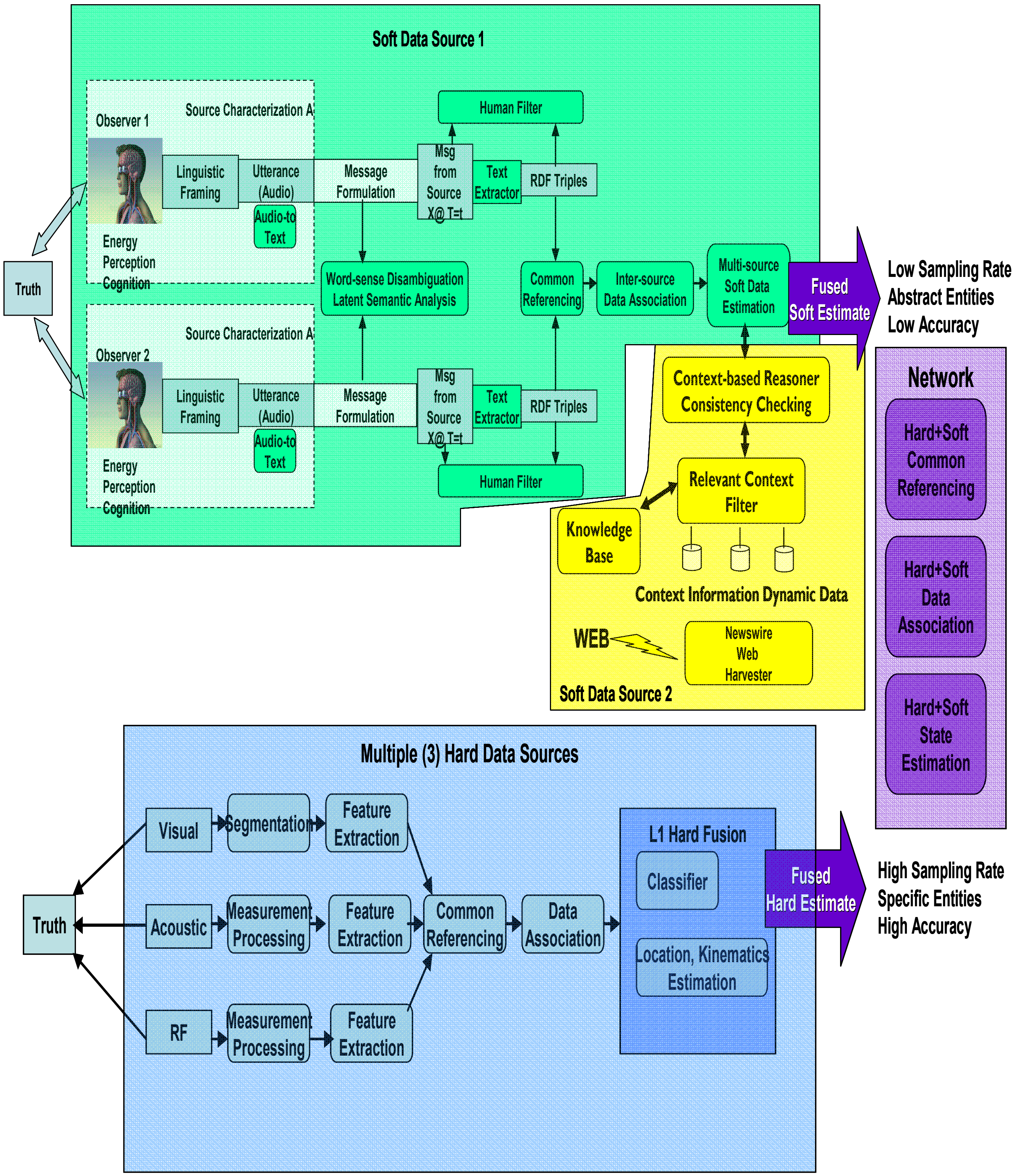

Figure 2. Initial Overarching Notional Soft and Hard Data Fusion Processing Concept

4. Representative Research in Progress

While this multi-year research project has just begun, we have already obtained some initial results. A brief description of these results is provided below.

4.1 Common Referencing

In a data and information environment as described here that involves a broad range of disparate input and reference-data types, the initial IF function of Common Referencing (aka Alignment) will be very important for normalizing data into common frames of reference. One important aspect will be the alignment of different representational forms of uncertainty. Here, Professor Yager of our IONA College teammate has already developed an initial position paper on a measure based approach to the fusion of possibilistic and probabilistic uncertainty forms. Developing a framework for smooth acceptance of both probabilistic uncertainties typical for representing uncertainties in electronic/hard sensor systems, and possibilistic uncertainties typical for representing uncertainties in linguistic/soft data, is considered a fundamental requirement of this MURI. This work takes a view based upon the use of a type of monotonic set measure, often referred to as a fuzzy measure [10, 11]. In this development, it is shown that the probability measure is a special fuzzy measure. This paper also develops the disjunctive and conjunctive forms of the aggregation operators, as well as the forms for means and Yager's Ordered Weighted Average form [12] that may be involved in IF combining operations. The simpler aggregation operations can be shown to take advantage of various t-norms in forming the fused result. The more complex aggregation types are shown to be manageable with entropy-based combining operations.

4.2 Processing Framework

As illustrated in Figure 2, an initial processing framework has been developed. The framework was extended to include three key processing streams; (1) information from traditional physical sensors (e.g., images, signals, vector and scalar quantities), (2) information from unstructured text-based human reports, and (3) combined hard and soft data from web sources. For each of these threads, initial processing algorithms have been selected and are currently being implemented. While it is not the intent of the MURI project to duplicate or replicate the extensive legacy of hard sensor fusion algorithms, we seek to develop sufficient techniques to allow demonstrations and experimentation with combined hard data, soft data and information from the web. Additional details on the architecture are provided in chapter 12 of [6], and specific algorithms are described in [13], [14] and [15].

4.3 Test and Evaluation Approach

Test and evaluation of hard and soft fusion systems is challenging because we must develop both representative test data (involving both physical sensors and human observers) and test environments to evaluate the performance of the hardware, software and humans-in-the-loop. Our teammates at the Pennsylvania State University have developed an experimental facility (our extreme events laboratory [16]), a test and evaluation approach, and are collecting evolving test data sets for evaluation of human-centered information fusion systems for situation awareness. The data sets include both synthetic data as well as data obtained using human subjects in campus wide experiment. The details of this approach are described in [17].

The extreme events laboratory is a faculty that allows use of the entire Penn State University Park campus as a laboratory in which students can be enlisted as participatory observers for staged activities and events. The textual reports (via Twitter, e-mail and solicited reports) are input to a central data visualization and analysis facility (see Figure 3). In addition, hard sensor data observations can also be input to the central facility for fusion.

Figure 3: Central visualization/processing center for the Penn State Extreme Events Laboratory

Three basic types of data sets are being prepared for test and evaluation of new fusion algorithms. These include; (1) the HASTEN-1 data set, originally developed by the U. S. Army Research Office via a contract to BAE [18], (2) an enhanced set of data developed around the U. S. Army's Soft Target Exploitation and Fusion (STEF) program [19], and (3) data collected via our EEL on the Pennsylvania State University campus involving human subjects [16]. As they evolve the latter two data sets will be provided for researchers throughout the information fusion community.

4.4 Graph Theoretic Methods in Hard-Soft Fusion

For higher levels of fusion, Directed Attributed Relationship Graphs (DARGs) facilitate the representation of situational information, and graph matching methods are helpful in situational understanding. Our past work has aimed at developing inexact graph matching algorithms for situational understanding [20]. A situation of interest or complex hypothesis of interest to the analyst is encoded in the form of a "template graph", which is matched against a "data graph" that represents the accumulation of sensor and evidence data. Since this problem is NP-hard, our effort has been on developing fast, limited state space heuristic search algorithms. We have demonstrated that large data graphs can be searched for situations of interest with reasonable accuracy and speed.

In the current MURI effort and joint work with ARL, CMIF is enhancing these graph matching methods for linguistically framed soft information that comes with semantic vagueness and uncertainty associated with human observations. In [13] we are discussing uncertainty representation and transformation methods, and graph matching under these uncertain representations of situations of interest as well as corpus of evidence (relevant data graph). Probabilistic and possibilistic information is transformed into a common fuzzy representation and fuzzy matching is performed in place of crisp matching. In addition to uncertainty issues, in [14] we are focusing on temporal issues in graph matching to develop a temporal understanding of evolving situations. Often times in forensic analysis the analyst is not only interested in the situational match, but how the situation evolved through time. Leveraging on general temporal theory and temporal algebra, we are hoping to develop an Intelligence Running Estimate (IRE), which is of great interest to Army and other defense analysts at the current time.

4.5 Conceptual Spaces in Hard-Soft Fusion

Typically the soft world is characterized by symbols and words, while hard sensing is a numerical and feature space. How does one fuse these disparate modalities? It turns out that the Artificial Intelligence community has also faced similar challenges when working with Symbolic (rule-based) or Associationist (feature-based) cognitive models. To draw upon their advantages and combine them in a unifying framework, Gardenfors [21] has proposed Conceptual Spaces. Conceptual Spaces are geometric representations of how humans understand concepts. Concepts are broken down into dimensions, domains, properties, etc. and brought together into geometric figures. Our proposed fusion framework, in addition to estimate level fusion of the soft and hard processing, is looking at novel approaches of conceptual spaces in soft-hard information fusion at the observation level. More specifically, we are extending the conceptual spaces modeling framework to incorporate fuzzy geometric representation of linguistic terms that patrol reports and other soft sources might provide. Corresponding to the modeling enhancement we are working at new processing frameworks that will deal with uncertainty in hard and soft information. This is a fertile ground for new research in high level information fusion.

5. Summary

The Multidisciplinary University Research Initiative is addressing a wide range of issues for hard and soft fusion, from information collection and calibration of human observations, to basic architectures and processing flows for hard and soft fusion, and new algorithms for text-based information extraction, representation (including uncertainty), and fusion methods based on advanced graph matching techniques. Data being generated for test and evaluation of these new methods includes synthetic data and human-in-the-loop data collected using human subjects. It is anticipated that the program will establish advances in the area of hard and soft fusion and provide legacy architectures, test and evaluation approaches, algorithms and test data for the information fusion community to address this increasingly important area.

Acknowledgement

We gratefully acknowledge that this research activity is supported by a Multidisciplinary University Research Initiative (MURI) grant (Number W911NF-09-1-0392) for "Unified Research on Network-based Hard/Soft Information Fusion", issued by the US Army Research Office (ARO) under the program management of Dr. John Lavery

References

[1] U.S. Department of the Army. *Counterinsurgency: Field Manual 3-24*. Washington, DC: Author. Retrieved March 10, 2010, from http://www.usgcoin.org/library/doctrine/COIN-FM3-24.pdf , 2006, p1-15

[2] D. Hall, M. McNeese, J. Llinas, J., and T. Mullen, "A framework for hard/soft fusion," *Proceedings of the 11th International Conference on Information Fusion,* Cologne, Germany, June 30 – July 03, 2008

[3] E. Waltz and J. Llinas, *Multisensor Data Fusion*, Artech House,

[4] D. L. Hall and S. A. McMullen, S. A. *Mathematical Techniques in Multisensor Data Fusion*, 2nd edition, Artech House, Inc. Norwood, MA, 2004

[5] M. Liggins, D. L. Hall, D. L., and J. Llinas, *Handbook of Multisensor Data Fusion*, 2nd edition, CRC Press, Boca Raton, FL, 2008

[6] D. Hall and J. Jordan, *Human Centered Information Fusion*, Artech House, 2010

[7] Llinas, J. and Sambhoos, K., "Designing a Performance Evaluation Methodology for Data Fusion-capable Tactical Aircraft V", *AIAA Air Force T&E Days Conf.*, Destin, FL, Feb 2007

[8] Sambhoos, K., Llinas, J., Dastadar, S., and Bowman, C., "Performance Evaluation Methods for Network Centric Data-Fusion Capable Tactical Platforms", *Natl. Symp on Sensor and Data Fusion*, Washington, DC, June 2006

[9] Zentner, J.M., "A Design Space Exploration Process for Large Scale, Multi-objective Computer Simulations", Dissertation, School of Aerospace Engineering, Georgia Institute of Technology, August 2006

[10] Sugeno, M., "Fuzzy measures and fuzzy integrals: a survey," in *Fuzzy Automata and Decision Process*, Gupta, M.M., Saridis, G.N. & Gaines, B.R. (eds.), Amsterdam: North-Holland, Pub, 89-102, 1977.

[11] Klir, G. J., *Uncertainty and Information*, John Wiley & Sons: New York, 2006.

[12] Yager, R. R., "On ordered weighted averaging aggregation operators in multi-criteria decision making," *IEEE Transactions on Systems, Man and Cybernetics* 18, 183-190, 1988.

[13] G. Gross, R. Nagi and K. Samboos, "Soft information, dirty graphs and uncertainty representation/processing for situation understanding", in *Proceedings of the 13th International Conference on Information Fusion, Edinburgh*, UK, July, 2010

[14] D. McMaster, R. Nagi and K. Samboos, "Temporal issues in soft data processing and graph matching," in *Proceedings of the 13th International Conference on Information Fusion, Edinburgh*, UK, July, 2010

[15] M. Hannigan, J. Llinas, K. Sambhoos and M. Sudit, "Data association in soft data streams," in *Proceedings of the 13th International Conference on Information Fusion, Edinburgh*, UK, July, 2010

[16] Hall, D., Hellar, B., and McNeese, M. D., (2009) The Extreme Events Laboratory: A cyber infrastructure for performing experiments to quantify the effectiveness of human-centered information fusion," *Proceedings of the 2009 International Conference on Information Fusion (Fusion 2009),* Seattle, Washington, July, 2009

[17] D. Hall, J. Graham, L. More and J. Rimland, "Test and evaluation of soft/hard information fusion systems: an experimental environment, methodology and initial data sets,"in *Proceedings of the 13th International Conference on Information Fusion,* Edinburgh, UK, July, 2010

[18] M. A. Pravia, P. K. Prasanth, P. O. Aranbel, C. Sidner and Chee-Yee Chong, "Generation of a fundamental data set for hard/soft information fusion,", *Proceedings of the 11th International Conference on Information Fusion*, Cologne, Germany, June 30 – July 03, 2008

[19] Sambhoos, K., Llinas, J., and Little, E., "Graphical Methods for Real-Time Fusion and Estimation with Soft Message Data", *Proceedings of the 11th International Conference on Information Fusion*, Cologne, Germany, June 30 – July 03, 2008

[20] K. Sambhoos, R. Nagi, M. Sudit, M. and A. Stotz, "Enhancements to High Level Data Fusion using Graph Matching and State Space Search," accepted to *Information Fusion*, October 2009.

[21] P. Gardenfors, (2000). Conceptual Spaces: The Geometry of Thought. Cambridge, MA, MIT Press.