

**Data Acquisition
and Exploitation**



Evolutionary Control of an Autonomous Field	47
Mark W. Owen (SSC San Diego) Dale M. Klamer and Barbara Dean (Orincon Corporation)	
Use of One-Point Coverage Representations, Product Space Conditional Event Algebra, and Second-Order Probability Theory for Constructing and Using Probability-Compatible Inference Rules in Data-Fusion Problems	58
I. R. Goodman (SSC San Diego)	
On Knowledge Amplification by Structured Expert Randomization (KASER)	70
Stuart H. Rubin (SSC San Diego)	
Establishing a Data-Mining Environment for Wartime Event Prediction with an Object-Oriented Command and Control Database	92
Marion G. Ceruti (SSC San Diego) S. Joe McCarthy (Space and Naval Warfare Systems Command)	
Thermal Pixel Array Characterization for Thermal Imager Test Set Applications	101
Ike Bendall, Ted Michno, Don Williams, Matthew Holck, and Richard Bates (SSC San Diego) José Manuel López-Alonso (Laboratorio de Termovision, Madrid, Spain) Robert J. Giannaris (Applied Technology Associates) Gordon Perkins and H. Ronald Marlin (The Titan Corporation)	
Hyperspectral Imaging for Intelligence, Surveillance, and Reconnaissance	108
David Stein, Jon Schoonmaker, and Eric Coolbaugh (SSC San Diego)	
Surface Plasmon Tunable Filter for Multiband Hyperspectral Imaging	117
Stephen D. Russell, Randy L. Shimabukuro, Ayax D. Ramirez, and Michael G. Lovern (SSC San Diego) Yu Wang (Jet Propulsion Laboratory)	
Knowledge Base Formation Using Integrated Complex Information	122
Douglas S. Lange (SSC San Diego)	
A Real-Time Infrared Scene Simulator in CMOS/SOI MEMS	129
Jeremy D. Popp, Bruce Offord, and Richard Bates (SSC San Diego) H. Ronald Marlin and Chris Hutchens (Titan Systems Corporation) Derek Huang (Advanced Analog VLSI Design Center)	

Evolutionary Control of an Autonomous Field

Mark W. Owen

SSC San Diego

Dale M. Klamer and Barbara Dean

Orincon Corporation

INTRODUCTION

The Office of Naval Research (ONR) established the Deployable Autonomous Distributed System (DADS) program (Figure 1) to demonstrate the feasibility of increased performance for an advanced tactical/surveillance system that operates as a field of underwater distributed sensor nodes. The goal of DADS is to demonstrate the feasibility of a cooperative field-level detection and data fusion system that increases performance at a reduced cost. Given limited power, the objectives are to use distributed detection and data fusion to increase the lifetime of the field (reduced power consumption), decrease the false alarm rate of the field over that of the individual nodes, increase the field-level detection, increase the probability of correct classification, and increase the accuracy of target position estimates [1, 2, and 3].

A DADS field consists of individual sensor nodes operating autonomously. Each sensor node uses a set of acoustic and electromagnetic sensors to provide coverage of a small area of interest. Each DADS sensor node uses a matched-field tracking algorithm to provide target detections consisting of position, velocity, and classification information. Once a detection is constructed at a sensor node, the data are transferred to a DADS master node where field-level data fusion is performed.

Detection Theory

In the DADS program, a need exists to identify what constitutes target detections from the field of autonomous sensor nodes. The DADS program also requires an optimization algorithm to route communication messages efficiently, using as little power as possible. A field-level control/detection scheme is sought to detect targets of interest at a given field-level probability and to route messages optimally by using a minimal amount of power. Control of an autonomous set of sensor nodes is needed to meet a desired probability of detection for the field and to extend the life of the field.

To construct a field-level detection, we now define what is required to call out a field-level detection. Each sensor node contains an acoustic sensor suite and an electromagnetic sensor suite. To report a detection, both the acoustic and magnetic sensors must detect a target at a sensor node. Once one node has detected the target, a second node nearby is cued and another sensor node must detect the target. Once this second sensor node detects and reports the target, a field-level detection is called and reported

ABSTRACT

An autonomous field of sensor nodes must acquire and track targets of interest traversing the field. Small detection ranges limit the detectability of the field. As detections occur in the field, detections are transmitted acoustically to a master node. Both detection processing and acoustic communication drain a node's power source. To maximize field life, an approach must be developed to control processes carried out in the field. This paper presents an adaptive threshold control scheme that minimizes power consumption while still maintaining the field-level probability of detection. The power consumption of the field of sensor nodes is driven by the false alarm rate and target detection rate at the individual sensor nodes in this problem formulation. The control law to be developed is based on a stochastic optimization technique known as evolutionary programming. Results show that by dynamically adjusting sensor thresholds and routing structures, the controlled field will have twice the life of the fixed field.

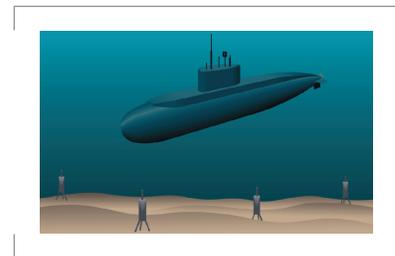


FIGURE 1. Field of DADS autonomous sensor nodes.

out by the master node for field-level fusion. Each sensor node has a threshold for the sensor suite given by an operating point on a receiver operating characteristic (ROC) curve as shown in Figure 2. The operating points on the figure are labeled R1 and R2 and represent different signal-to-noise ratio (SNR) levels for the sensor suite. Choosing different operating points on the ROC curve yields different probabilities of detection and probabilities of false alarm. A constant field-level probability of detection is desired for operation of the field of sensor nodes. By adjusting threshold levels at the sensor suite, that is, moving up and down operating points on the ROC curve at each sensor node, a constant field-level probability can be achieved.

Besides controlling the thresholds at the individual sensor suites at each node, another problem is to minimize the power consumption of the individual sensor nodes while meeting the field-level probability constraint. This issue addresses the routing of communication messages through the distributed field of sensor nodes. As messages are passed from sensor node to sensor node and finally arrive at the master node, the battery level is drained by the amount of communication power spent transmitting and relaying detections acoustically.

A field-level controller will adjust the detection threshold levels at each sensor node to meet the desired field-level probability of detection and to perform optimal routing of messages through the field. A typical example of a point on a ROC curve is shown in Figure 3.

A brief overview of detection theory is provided below [4]. In Figure 3, two possible hypotheses, labeled H_0 and H_1 , are shown. H_0 is the false alarm hypothesis and H_1 is the detection hypothesis. The threshold T is used to determine whether or not the SNR is high enough to call out a detection. The SNR in the figure is labeled γ . Under the two Gaussian curves, a probability of detection and a probability of false alarm can be determined. Integrating the H_0 probability density function (pdf) from T to ∞ , the false alarm probability is calculated. Integrating the H_1 pdf from T to ∞ , the probability of detection is calculated. Figure 4 shows several SNRs from a chosen ROC curve operating point. The objective of the field-level controller is to adapt the sensor node thresholds to acquire a target of interest and detect it successfully through the field. In the figure, the graph labeled nominal is shown to demonstrate a chosen operating point for the sensor node. The next two graphs show a decrease in SNR and an increase in SNR, respectively. As SNR levels vary, a target may become easier or more difficult to detect although the probability of false alarm remains constant across all three graphs. Only the probability of detection decreases or increases due to the SNR of the target. Our task is to adjust thresholds dynamically to make sure the target is acquired and tracked as it passes through the field. To do this, we will lower thresholds for subsequent cued detections to increase the detection range at a sensor node, but at the same time we increase the number of false alarms from a sensor node. When adjusting these thresholds at each sensor node, we must maintain a constant field-level probability of detection. A simple example of this threshold adjustment is to use a bathtub analogy. If one side of the bathtub water is pushed down, water on the other side of the tub will rise. This example shows what we will do when adapting thresholds: we will lower a certain set of sensor node thresholds while raising another set.

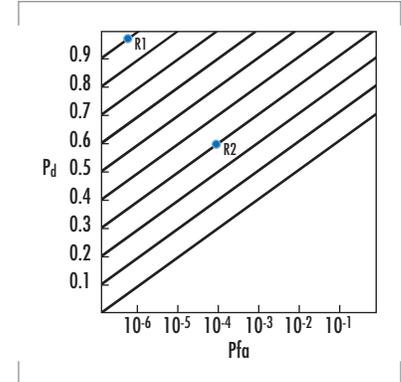


FIGURE 2. Typical ROC curve.

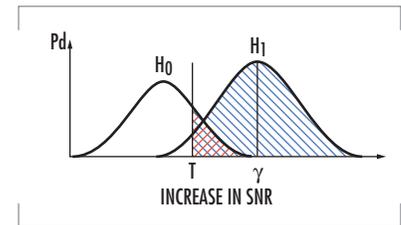


FIGURE 3. A single point from a ROC curve.

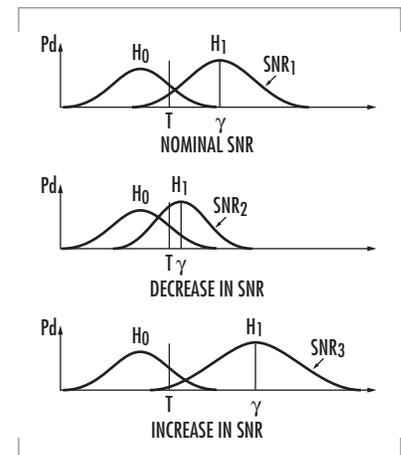


FIGURE 4. Possible detection curves.

Threshold Adaptation

Figure 5 shows a cookie cutter example of a field of sensor nodes. Each sensor node has a defined detection range given in red (small circles) for a high threshold (low false alarm rate, high SNR) and another detection range shown in blue (large circles) for a low threshold (high false alarm rate, low SNR). This figure demonstrates the adaptive process that must occur for the DADS field of sensor nodes to detect and continue to detect a target as it passes through the field.

If the field were static, the small red circles would dictate the area of coverage in which the field could pick up detectable targets. In the figure, a hypothetical target has been drawn by a black line with an arrow at the tip. If the threshold were held at this higher level, only one possible detection might occur as this target traversed the field of sensor nodes. By lowering the thresholds (larger blue circles), which is done by cueing the field, a broader coverage of the field is achieved. The figure shows that up to four possible detections on a target of interest can occur by lowering the sensor node thresholds. This improved detectability concept will improve the overall field-level data fusion by providing more contact information than previously capable with a static set of sensor node thresholds. By lowering the threshold though, a larger number of false alarms can occur and cause power to be drained from the sensor nodes. False alarms also make the data fusion problem at the master node more susceptible to miscorrelation. Therefore, dropping all of the sensor node thresholds is not acceptable because it will limit the system operation. As explained previously, we will lower thresholds and raise thresholds at individual sensor nodes to maintain the desired field-level probability of detection while maximizing the life of the field.

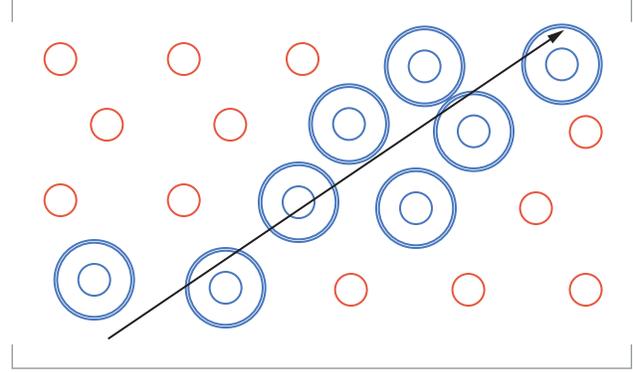


FIGURE 5. Sensor node threshold adjustments via field-level control.

2-of-2 Field Detector

To adjust thresholds, we propose to use a baseline model of a 2-of-2 detector. The detector will use communication costs, probabilities of detection and false alarm, node spacing of the field, and signal processing parameters used at the sensor node sensor suite. This formulation shows that false alarms as well as target detections drain the power at each sensor node. We will now present our baseline model equation for field-level control as derived in [5]. This formulation will allow the complete field to be controlled by the master node in the DADS system. The baseline model equation is as follows. The estimated power $\hat{P}^{(n)}$ consumed over a period of time T at each node n , $n = 1, \dots, N$, is given by

$$\begin{aligned} \hat{P}^{(n)}(T) = & \sum_{k=1}^{\rho_s T} C_{on} + [1 - (1 - F_1^{(n)} F_2^{(n)})^{N_p}] C_k^{(n)} \\ & + \sum_{n' \in R_k^{(n)}} [1 - (1 - F_1^{(n')} F_2^{(n')})^{N_p}] C_k^{(n)} \\ & + \sum_{n' \in B_k^{(n)}} [1 - (1 - F_1^{(n')} F_2^{(n')})^{N_p}] [1 - (1 - F_1^{(n)} F_2^{(n)})^{\rho_s \delta N_p P [1 + s D^2] / (\pi (r_d^{(2)})^2)}] C_k^{(n)} \\ & + \sum_{n' \in R_k^{(n)}} \sum_{n'' \in B_k^{(n)}} [1 - (1 - F_1^{(n'')} F_2^{(n'')})^{N_p}] [1 - (1 - F_1^{(n')} F_2^{(n')})^{\rho_s \delta N_p P [1 + s D^2] / (\pi (r_d^{(2)})^2)}] C_k^{(n)} \end{aligned} \quad (1)$$

$$\begin{aligned}
& + \rho_T r_d^{(n)} [1 - (1 - P_1^{(n)} P_2^{(n)}) (1 - F_1^{(n)} F_2^{(n)})^{N_p - 1}] C_k^{(n)} / D \\
& + \rho_T \sum_{n' \in R_k^{(n)}} r_d^{(n')} [1 - (1 - P_1^{(n')} P_2^{(n')}) (1 - F_1^{(n')} F_2^{(n')})^{N_p - 1}] C_k^{(n)} / D \\
& + \rho_T \sum_{n' \in B_k^{(n)}} r_d^{(n')} [1 - (1 - P_1^{(n')} P_2^{(n')}) (1 - F_1^{(n')} F_2^{(n')})^{N_p - 1}] \\
& [1 - (1 - P_1^{(n)} P_2^{(n)}) (1 - F_1^{(n)} F_2^{(n)}) [\rho_s \delta N_p P^{1+sD^2} / (\pi(r_d^{(2)})^2)]^{-1}] C_k^{(n)} / D \\
& + \rho_T \sum_{n' \in R_k^{(n)}} \sum_{n'' \in B_k^{(n')}} [1 - (1 - P_1^{(n'')} P_2^{(n'')}) (1 - F_1^{(n'')} F_2^{(n'')})^{N_p - 1}] \\
& [1 - (1 - P_1^{(n')} P_2^{(n')}) (1 - F_1^{(n')} F_2^{(n')}) [\rho_s \delta N_p P^{1+sD^2} / (\pi(r_d^{(2)})^2)]^{-1}] C_k^{(n)} / D
\end{aligned} \tag{1 contd}$$

where ρ_s is the basic sample rate and T is the time period of the estimated life of the node. The first term represents the power consumed C_{on} from the processor in the node. If the sensor node is on, a certain amount of processing power is drained from the battery. The second term represents the case that an initial false alarm is generated at node n , where $F_1^{(n)}$, $F_2^{(n)}$ are the probabilities of false alarm that are controlled by thresholds $T_1^{(n)}$ and $T_2^{(n)}$, and $C_k^{(n)}$ is the communication power used to transmit from node n to the next upstream node specified by the current communication route $R_k^{(n)}$ at time k . N_p is the size of the parameter space over which the detectors must test, e.g., if the detector must look over a discrete set of speed (say N_s) and closest point of approach (CPA), say N_{CPA} , thus giving $N_p = N_s N_{CPA}$. This is the second detection required for declaring a field-level detection from the field. The third term represents the case of a "downstream" node n' that generates a false alarm and node n is simply a passthrough; the communication route for node n at time k is specified by $R_k^{(n)}$. The fourth term represents the case that a false alarm is generated at node n as the result of being cued by another node n' in a set of neighboring nodes $B_k^{(n)}$. Specifically, P is the covariance of the track estimate at the time of the detection at the first node; $[1+sD^2]$ is the expansion factor for the track covariance until the second detection at the next node detection; $\pi(r_d^{(2)})^2$ is the area of the detection space for the second sensor node; and D is the length of the sensor field. The fifth term represents the case of a downstream node n' that generates a false alarm as a result of being cued, and node n is simply a passthrough. The last four terms deal with the cases of a target present; ρ_T is the target rate. The sixth term represents a target detection at node n , where $P_1^{(n)}$, $P_2^{(n)}$ are the probabilities of detection, again controlled by the thresholds $T_1^{(n)}$ and $T_2^{(n)}$. This is a true target detection and not a false alarm. The seventh term represents the case of a downstream node n' detection where node n is simply a passthrough for the initial condition. The eighth term represents the case that a target detection is generated at node n as the result of being cued by another node n' . The final term represents a downstream node n' that generates a target detection as the result of being cued, and node n is simply a passthrough.

Given the current power $P^{(n)}$ available at each node, the estimated remaining power is

$$\varepsilon^{(n)}(T) = P^{(n)} - \hat{P}^{(n)}(T).$$

The objective function for maximizing the life of the field is

$$\text{maximize } T,$$

subject to the constraints that each of the estimates of the remaining power is positive

$$\varepsilon^{(n)}(T) \geq 0, n = 1, \dots, N$$

and the field-level probability of detection is specified by

$$\text{PD} = N(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N) \pi(r_d^2) [1 - (1 - P_1^{(1)} P_2^{(1)}) (1 - F_1^{(1)} F_2^{(1)})^{N_p - 1}] \\ \times [1 - (1 - P_1^{(2)} P_2^{(2)}) (1 - F_1^{(2)} F_2^{(2)}) [\rho \delta_{N_p}^{P(1+sD^2)} / (\pi r_d^2)] - 1] / A(D)$$

where $N(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)$ is the number of nodes with nonzero power remaining and $\pi(r_d^2)/A(D)$ is the area covered by an individual node. The objective is to maximize field life T subject to meeting the field-level constraint by adjusting probability of detection / probability of false alarm threshold levels and varying communication routes (through $R_k^{(n)}$). By choosing appropriate thresholds at each sensor suite, the field-level probability of detection constraint can be met and the field life extended. An algorithm that will choose thresholds to meet the probability of detection constraint and extend the field life is discussed in the next section.

Evolutionary Programming

Evolutionary programming (EP) is a stochastic optimization technique applied in this paper to optimize routing of the sensor node message traffic at minimal power cost and to meet a field-level probability constraint. EP falls under the domain of Evolutionary Computation that contains other algorithmic techniques such as genetic algorithms (GAs), genetic programming, as well as others [6]. One of the main differences between EP and GAs is that EP performs a mutation operation while GAs perform a mutation operation and a crossover operation. Genetic algorithms also operate from the bottom up when finding a solution. EP is a top down approach to finding optimal solutions. An evolutionary algorithm is shown in Figure 6. In simple terms, an evolutionary algorithm starts out with a population of possible solutions to a problem. A population consists of parent solutions and their corresponding offspring solutions. This stochastic optimization technique allows the whole parameter space to be searched and evaluated for a best-fitting solution. In the figure, the initial solutions are called parents. Each parent solution can be a good first guess at the correct answer or a randomly chosen solution that may be very poor. Each parent has the ability to create a set of offspring solutions by mutation or by crossover if a genetic approach was used. Each parent solution is mutated by changing its state to form an offspring solution. This mutation can be Gaussian or some other linear or nonlinear deviation. Once the population of parents has been mutated and the offspring solutions are created, the population consisting of parents and offspring solutions is then scored, as shown in the figure. Scoring or evaluation of the population for our purpose is done to make sure the sensor nodes meet a defined field-level probability constraint with their defined threshold settings. A selection process is then performed whereby the next generation of parents are selected to evolve better and better solutions. This selection process chooses the solutions that passed the constraint in the scoring process by selecting the solutions that yield the largest amount of field life.

The standard EP approach consists of several steps (initialization, mutation, scoring, and selection) [6]. Initialization is performed by assigning thresholds to each sensor in the sensor suite (magnetic, acoustic) and using these thresholds, the sonar equation, and an error function to evaluate

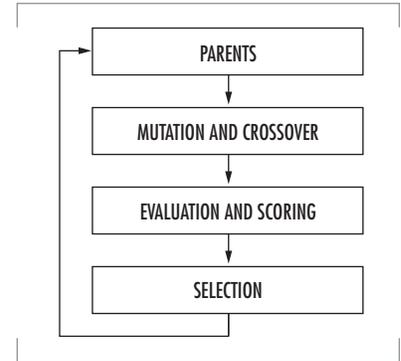


FIGURE 6. Evolutionary algorithm.

the probability of detection and probability of false alarm of the sensor node. This is done for each sensor node in the field given by

$$Pd(n) = 1/2*(1.0 - \text{erf}(T(n) - SL(n) + NL(n))) \quad (2)$$

and

$$Pfa(n) = 1/2*(1.0 - \text{erf}(T(n) + NL(n))) \quad (3)$$

where Eq. (2) initializes the probability of detection Pd for sensor node n given its threshold T , the target source level SL , and the noise level at the sensor NL . Eq. (3) initializes the probability of false alarm Pfa for sensor node n given its threshold T , and the noise level at the sensor NL . This is performed for each sensor node until all thresholds and probabilities of detection and false alarm have been initialized. This fully initialized field of sensor nodes is deemed as a parent solution in the EP language and is a possible solution for the field-life problem. Possible solutions are defined as parents and are given as

$$P(k) = S(Pd(n), Pfa(n), T(n), R(n)) \quad (4)$$

where $P(k)$ are the k number of parents in the population solutions. Each solution S is made up of a field of sensor nodes with independent thresholds T , which dictate a Pd and Pfa for the sensor node, and a routing table R for communication with other nodes in the field. Once the population of parent solutions has been initialized, the EP algorithm is able to perform the next three steps (mutation, scoring, and selection) iteratively to converge to the best possible solution given time constraints and memory requirements of the system. The first step is the mutation process whereby parent solutions generate offspring solutions. Offspring solutions have the possibility of generating a better solution than their parents. This is the evolutionary step in the EP process. One of the mutation steps is to change the threshold at each sensor at a sensor node to yield a better solution. This is defined by

$$O[T(m,n)] = P[T(k,n)] + N(0,1) \quad (5)$$

where $O[T(m,n)]$ is the mutated threshold at offspring m for sensor node n , $P[T(k,n)]$ is the threshold at parent k for sensor node n , and $N(0,1)$ is a Gaussian random variable with zero mean and unit variance. Eq. (5) changes each parent's threshold to generate an offspring's threshold. Another mutation step is to change the routing table for communications at each node. This is defined by

$$O[R(m,n)] = P[R(k,n)] \pm Urv * c \quad (6)$$

where $O[R(m,n)]$ is the mutated communication routes at offspring m for sensor node n , $P[R(k,n)]$ is the communication routes at parent k for sensor node n , Urv is a Uniform random variable, and c is the number of possible nodes for sensor node n to communicate with. The number of communication routes can increase or decrease according to Eq. (6). Eq. (6) changes each parent's communication route to generate an offspring's communication route. Each parent can perform these mutation steps and generate as many offspring as desired. Once this is done, the new population of parents and offspring are scored and evaluated against the system constraints. For example, if the desired field-level probability of detection is 0.8, each solution is evaluated using

$$\begin{aligned} PD = & N\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N \pi(r_d^2 [1 - (1 - P_1^{(1)} P_2^{(1)}) (1 - F_1^{(1)} F_2^{(1)})^{N_p - 1}] \\ & \times [1 - (1 - P_1^{(2)} P_2^{(2)}) (1 - F_1^{(2)} F_2^{(2)})]^{\lfloor \rho \delta_{N_p} P(1+sD^2) / (\pi r_d^2) \rfloor - 1} / A(D) \end{aligned} \quad (7)$$

which is the probability of detection for a field of sensor nodes defined above. (See 2-of-2 Field Detector.) We will use a simulated annealing approach to meet this constraint. For example, if 0.8 is desired, we may allow solutions to lie between (0.7, 0.9) in the beginning and slowly converge toward 0.8 while we iterate. All solutions that pass this field-level probability constraint are then passed to the selection process. Selection is done by picking the best k solutions that meet the constraint and minimize the power consumption defined from the baseline model from Eq. (1). These best k solutions then become the parents for the next iteration. The process continues until the best solution is found. This evolutionary process extends the field life by optimizing the thresholds of the field and planning the optimal routes for message passing.

RESULTS

Now we present some results of our EP solution to the adaptive threshold control problem. These results are for a complete field of sensor nodes. Each node has a set of thresholds solved for by the EP algorithm as well as the optimal routes for communication to extend field life.

Simulation Overview

As stated previously, the claim of this paper is that it can be shown that field life can be doubled by using a field-level controller to dynamically adjust thresholds and routing structures, as compared to a fixed field that uses static thresholds and routing structures.

The EP software written for this paper generates solutions that are representative of a field under the control of a field-level controller. To make the comparison to a fixed field, a fixed-field implementation had to be generated.

The Fixed Field

The fixed field required a nominal routing structure and a set of sensor thresholds, which would meet the field-level probability of detection. To generate the nominal routing structures, a field initialization scheme was emulated. The emulation of this field initialization scheme consists of the following steps:

1. The Master Node broadcasts a Wakeup Message.
2. Any node that can hear responds with a Wakeup Response message. In this case, any node within the cookie cutter range can hear.
3. Nodes that responded to the Master Node will be direct communication routes. This means that these nodes will relay their packets directly to the master node.
4. Nodes that heard the Master Node will broadcast to their neighbors.
5. Any node that can hear within the cookie cutter range will respond.
6. If the node that responds does not have a destination node yet, the node that broadcast will become the destination node.
7. This sequence is repeated until every node in the field has been assigned exactly one destination node.

The above sequence generated a nominal routing structure for a fixed field as shown in Figure 7. In conjunction with the routing structures, sensor thresholds that met the field-level probability of detection were

required. To obtain these thresholds, the EP model was run, and the thresholds from the optimal solution were used.

The Controlled Field

In the simulations, two types of results are generated for the controlled field. The first type is referred to as a "single optimized" solution. This solution is generated using the EP software. Once the EP algorithm finds an optimal combination of thresholds and routing structures, it uses that solution for the life of the field. Figure 8 shows the optimal routes found for the single optimized solution.

The second type of a controlled field solution is referred to as a "vector-optimized" solution. As with the single optimized solution, the EP algorithm finds a solution set, which maximizes field life. However, in this solution, the routes and thresholds can be adjusted every 24 hours, thus resulting in a vector of solutions. Because the control algorithm is run each day and the routes are potentially changed, it is not possible to show each daily graphical solution in this paper.

Field Laydown

Simulations were run for two field laydowns. In each laydown, the field consists of 30 sensor nodes and 1 master node arranged in a (56 by 28) unit grid. The difference between the two laydowns is the placement of the master node. In the first field laydown, the master node is a square box on the edge of the field as shown in Figures 7 and 8. In the second laydown, the master node is in the center of the field of sensor nodes.

Detector Types

The objective function defined previously (see 2-of-2 Field Detector) is for a 2-of-2 detector. This paper also defined an objective function for a 1-2 detector. The 1-2 detector requires an initial detection from the magnetic sensor on one node followed by a confirmed detection from the acoustic sensor on a second node. Results for both the 2-of-2 detector and the 1-2 detector are reported below.

Simulation Results

The results from the simulation are given in Table 1. The results are provided in units of days.

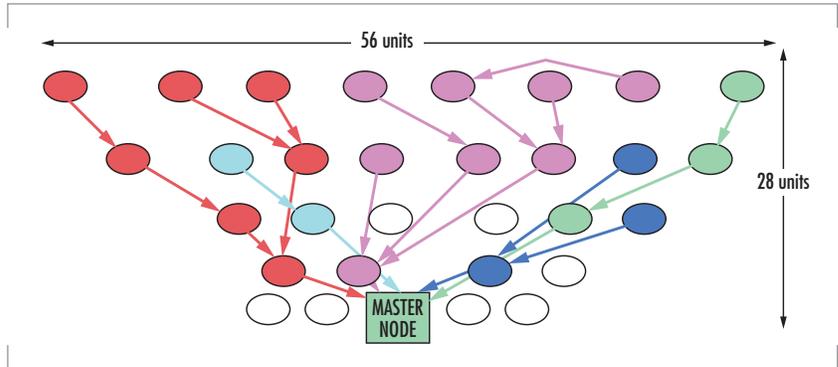


FIGURE 7. Fixed-field routes.

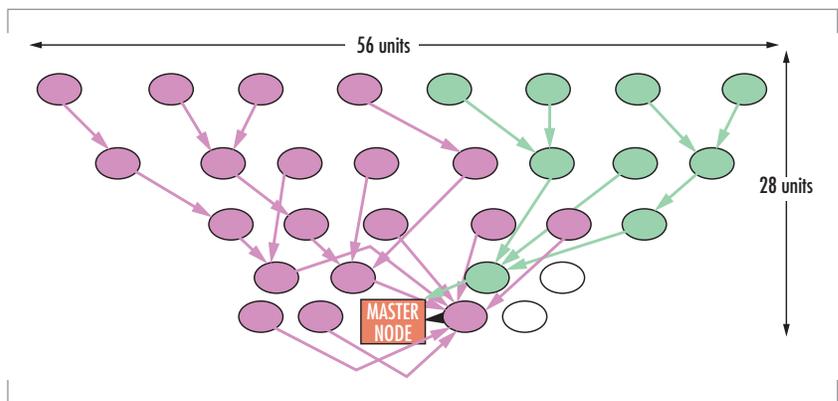


FIGURE 8. Single optimized field routes.

Figure 9 shows the results from running the fixed-field simulation. In the fixed field, the routing assignment was performed by using the minimum number of hops between the master node and each node in the field. This result is for the 2-of-2 detector processing for the second field laydown. It shows that running no optimization algorithm and just a greedy algorithm to assign a route for the field only yields a field life of 74 days. As shown in Figure 9, one single node begins to lose its power immediately. This node is the main communication node to the master node. Once one node in the field loses all of its power, the field is considered to be dead.

Figure 10 shows the results from the single optimized field simulation. The routes for this result were calculated by running the EP algorithm once for the whole life of the field. This optimization result yielded a field life of 106 days for the 2-of-2 detector for the second field laydown. As shown in this figure, a single node still drives the field to death, but there are several other sensor nodes that are also losing power at a similar rate.

The field life was extended over the fixed-field implementation by using at least one planned optimal route for the whole simulation.

Figure 11 shows the results from the vector-optimized field simulation. This result has its routes recalculated each day by running the EP optimization algorithm. This optimization result yielded a field life of 154 days for the 2-of-2 detector for the second field laydown. As shown in this figure, a group of sensor nodes all lose power similarly at the same rate. Approximately one-third of the sensor nodes in the field died on day 154. This result more than doubled the life of the field over the fixed-field result of Figure 9. It also increased the life of the field from 106 days for the single optimized solution shown in Figure 10 to 154 days for the vector-optimized solution.

Observations

The following observations are made regarding the simulation results:

1. The vector-optimized solution more than doubled field life as compared to the fixed-field solution.
2. The 2-of-2 detector has a longer life than the 1-2 detector. This is because the 2-of-2 detector has stringent initial detection rules, which translates to fewer reports and less communication as shown in Table 1.

TABLE 1. Simulation results in days.

Field Laydown	Detector	Fixed Field	Single Optimized	Vector-Optimized
1	1-2	21	32	45
	2-of-2	40	70	118
2	1-2	26	45	55
	2-of-2	74	106	154

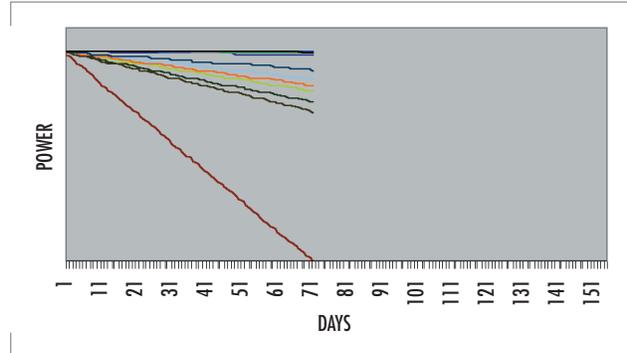


FIGURE 9. Fixed-field life.

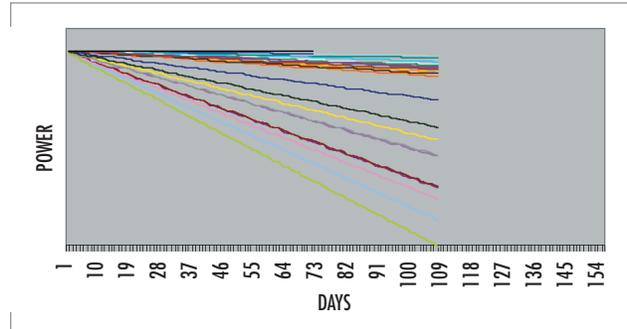


FIGURE 10. Single optimized field life.

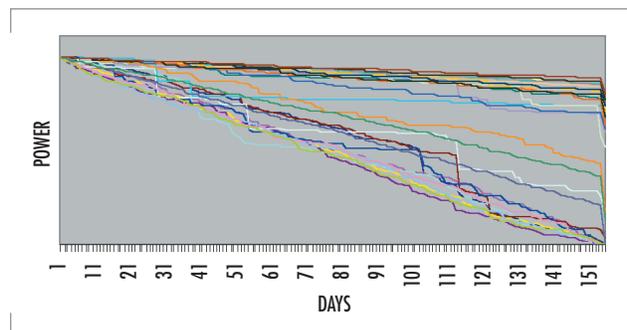


FIGURE 11. Vector-optimized field life.

3. Field life increased when the master node was moved from the edge of the field to the center of the field for the second field laydown. This is because when the master node is in the center of the field, there are more direct routes to the master node, which spreads out battery drain.
4. The vector-optimized solution has a longer field life than the single optimized solution. This is because changing the routes every 24 hours allows the battery drain to be spread more evenly across the field. With the vector-optimized solution, approximately one-third of the field will die on the same day.

CONCLUSIONS

In this paper, we have applied a stochastic optimization technique to adapt the thresholds of an autonomous sensor field and plan the communication routes. This stochastic optimization algorithm is known as evolutionary programming. The evolutionary program adapted the thresholds of a 2-of-2 detector for a set of sensors as well as a 1-2 detector. The algorithm is an evolutionary computation technique where an analytic solution is not attainable mathematically. Each sensor node in the 2-of-2 detector contained two thresholds to adapt, yielding four total thresholds to compute. The four thresholds are combined to meet a field-level probability of detection constraint and extend the life of a field of sensor nodes. Results show the benefits of adaptive threshold control in an autonomous sensor field by reducing communication costs and extending the life of the field by two.

AUTHORS

Dale M. Klammer

MS in Mathematics, San Diego State University, 1972.

Current Research: Adaptive data fusion; advanced tracking; information assurance.

Barbara Dean

BS in Electrical Engineering, Widener University, 1989

Current Research: Software engineering for Navy sonar systems.

REFERENCES

1. Hatch, M., M. Owen, et al. 1998. "Data Fusion Methodologies in the Deployable Autonomous Distributed System (DADS) Project," *Proceedings of the International Conference on Multisource-Multisensor Information Fusion*, (July), Las Vegas, NV, pp. 470-477.
2. Jahn, E., J. Kaina, and M. Hatch. 1999. "Fusion of Multi-Sensor Information from an Autonomous Undersea Distributed Field of Sensors," *Proceedings of the Second International Conference on Multisource-Multisensor Information Fusion*, (July), Sunnyvale, CA, pp. 4-11.
3. Shea, P. and M. Owen. 1999. "Fuzzy Control in the Deployable Autonomous Distributed System," *Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition VIII 1999*, (Ivan Kadar, ed.), vol. 3720, (April), Orlando, FL.



Mark W. Owen

MS in Electrical Engineering,
California State University
Long Beach, 1997

Current Research: Data fusion;
signal processing; autonomous
control.

4. Helstrom, C. W. 1968. *Statistical Theory of Signal Detection*, Pergamon Press, New York.
5. Klamer, D. and M. Owen. 2000. "Adaptive Threshold Control in an Autonomous Sensor Field," *Proceedings of SPIE: Signal and Data Processing of Small Targets 2000*, (Oliver Drummond, ed.), vol. 4048, (April), Orlando, FL.
6. Fogel, D. B. 1995. *Evolutionary Computation*, IEEE Press, New York.



Use of One-Point Coverage Representations, Product Space Conditional Event Algebra, and Second-Order Probability Theory for Constructing and Using Probability-Compatible Inference Rules in Data-Fusion Problems

I. R. Goodman
SSC San Diego

INTRODUCTION

Programmatics

This paper documents one aspect of the ongoing FY 01 In-house Laboratory Independent Research Project CRANOF (a Complexity-Reducing Algorithm for Near-Optimal Fusion), Project ZU014, with Principal Investigator, Dr. D. Bamber, and co-investigator, Dr. I. R. Goodman (both SSC San Diego), and with associate support from Dr. W. C. Torrez (SSC San Diego) and Prof. H. T. Nguyen (Department of Mathematical Sciences, New Mexico State University and U.S. Navy American Society for Engineering Education Fellow during summers at SSC San Diego). A preliminary version of this paper can be found in [1, section 3.3].

Background on Underconstrained Conditional Probability Problems

Philosophy of Approach and General Motivations

To improve the timeliness and accuracy of decision-supported human decision-making, one is faced with an array of crucial problems, including how to handle large amounts of incoming and uncertain information from disparate sources. These sources can be human-based or mechanical-based, and the information can arrive in different forms, such as qualitative and linguistic, numerical and statistical-probabilistic, or some mixture of both. At SSC San Diego, the CRANOF project addresses such crucial issues solely within the realm of statistics and probability. The issue of underconstrained or underspecified probabilities is treated by a novel use of second-order probabilities (i.e., probabilities of probabilities) in Bayesian framework. Underconstrained probabilities arise in a wide variety of problems, including quantitatively formulated rule-based systems, tracking and correlation, assessment of network intrusions, information retrieval, and simulation of human behavior in war games. This paper serves as a beginning extension of the capabilities of CRANOF to include linguistic-based information.

ABSTRACT

This paper covers issues relating to the establishment of a sound and conditional probability-compatible rationale for generating linguistic-based inference rules concerning a population. By extending previous preliminary results, we detail, in a fully rigorous manner and within the confines of traditional probability theory, that a comprehensive technique can be derived that converts linguistic-based conditional information, couched only in fuzzy-logic terms, into naturally corresponding conditional probabilities. In turn, we demonstrate how such typically underconstrained conditional probabilities can be combined for suitable conclusions and decision-making, via a new use of second-order probability logic. This research is part of the ongoing SSC San Diego In-house Laboratory Independent Research FY 01 project CRANOF (a Complexity-Reducing Algorithm for Near-Optimal Fusion).

Quantitatively Formulated Rule-Based Systems

Consider quantitatively formulated rule-based systems, with the rules or conditional relations symbolized typically as $(a_1 | b_1), (a_2 | b_2), \dots$ —read "if b_1 , then a_1 " (or equivalently, " a_1 , given b_1 ," etc.), "if b_2 , then a_2 ," ..., where events or sets $a_1, b_1, a_2, b_2, \dots$ may themselves represent quite complicated logical combinations of simpler events or sets, and where it may or may not be known what logical relations exist among such events. Each such rule is also assigned quantitative reliability in the form of naturally corresponding conditional probabilities. Thus, for some otherwise unspecified probability measure P , rule $(a|b)$ is assigned value $P(a|b) = P(ab)/P(b)$, the conditional probability of a given b , using standard Boolean and probability notation and assuming antecedent probability $P(b) > 0$. Because typical rule $(a|b)$ is not perfect, in general $P(a|b) < 1$, but, on the other hand, one would expect $P(a|b)$ to be reasonably high. A common problem that such rule-based systems address is: Consider incoming information in the form of events, d_1, \dots, d_n , possibly gleaned from different sources, such as $d_1 =$ "visibility is up to 1 mile," $d_2 =$ "winds between 15 mph and 30 mph," $d_3 =$ "enemy movement detected last night in Sector C," ..., $d_n =$ "political situation with enemy country Q at level R ," and a collection of reasonably related rules, such as $(a_1|b_1), (a_2|b_2), \dots, (a_m|b_m)$, where the a_j, b_j involve not only parts or all of the d_j (or various logical combinations of them), but possibly other related events (or logical combinations of such). Then, one wishes to test for viability of possible decisions, based upon this information, such as $c_1 =$ "fully successful attack by us can be accomplished by attacking in Sectors C or D," $c_2 =$ "partially successful attack by us can be accomplished by attacking Sectors D or H," Symbolically, one is considering the validity or degree of validity of the *entailment schemes* $G_i = [(a_1|b_1), \dots, (a_n|b_n); (c_i|d)]$, $i = 1, 2, \dots$, where $d = d_1 \& \dots \& d_n$ (conjunction of all data), and where $((a_1|b_1), \dots, (a_n|b_n))$ can be considered the *premise set* of G_i and $(c_i|d)$ its *potential conclusion*. Ideally, one would like to know just what each $P(c_i|d)$ would be, based on having either, say, the *exact threshold situation* holding, i.e., $P(a_j|b_j) = t_j$, $j = 1, \dots, n$, or, the *lower bound threshold situation* holding, i.e., having $P(a_j|b_j) \geq t_j$, where all the thresholds t_j are known or estimable in either situation. However, in general, it is readily demonstrated that the n equalities (or inequalities) are not enough to determine P and/or $P(c_i|d)$ completely. Thus, one is faced with the problem of best estimating, in some sense, just what P and/or $P(c_i|d)$ should be.

Adams' Approach to Analyzing Quantitatively Formulated Rule-Based Systems

In a series of papers [2, 3], E. W. Adams proposed, in effect, the estimate of $P(c_i|d)$ to be a pessimistic one in the form of his "minimum conclusion" function, using multivariable abbreviation t_j for $(t_j)_{j \text{ in } J}$, $(a|b)_J$ for $(a_j|b_j)_{j \text{ in } J}$, $P(a|b)_J \geq t_j$ for $P(a_j|b_j) \geq t_j$, $j \text{ in } J$, 1_J for column vector of all 1's indexed by J , etc.,

$$\begin{aligned} & \text{estimate}_{\text{HPL}} \text{ of } (P(c_i|d) \text{ from } G_i) \\ & = \text{minconc}(G_i)(t_j) = \inf\{P(c_i|d) : \text{for all possible probability measures } P \text{ such that } P(a|b)_J \geq t_j\}, \quad (1) \end{aligned}$$

with $P(c_i|d)$ for the exact threshold situation analogously estimated. The subscript $(\)_{\text{HPL}}$ is used to indicate "High Probability Logic," since Adams also introduced the idea of an entailment scheme being *HP-valid* or *HP-invalid*, which, in the case of any G_i here simply means for the former that

$$G_i \text{ is HPL-valid} \quad \text{iff} \quad \lim_{(t_j \uparrow 1_j)} (\text{minconc}(G_i)(t_j)) = 1. \quad (2)$$

But, unfortunately, both the minconc function and its limiting forms to test for HPL-validity/invalidity produce a number of results very much at odds with commonsense reasoning, including the fact that three very fundamental entailment schemes, *transitivity* (or *hypothetical syllogism*) [(a|b), (b|c); (a|c)] (the heart of any rule-based system); *contraposition* [(a|b); (b'|a')]; and *strengthening of antecedent* [(a|b); (a|bc)] are all HPL-invalid. In fact, one can find P's that satisfy their premise thresholds for any choice of t_j close to (but not exactly equal to) 1_j , but for which the corresponding conclusion probabilities are arbitrarily close to (or actually equal to) 0. Moreover, more generally, Eq. (2) can be complemented by the fact that any

$$G_i \text{ is HPL-invalid} \quad \text{iff} \quad \lim_{(t_j \uparrow 1_j)} (\text{minconc}(G_i)(t_j)) = 0. \quad (3)$$

Finally, Adams pointed out another type of validity, CPL (Certainty Probability Logic), that, although still based on the minconc function, can be characterized as "too optimistic" in contrast with HPL, whereby the criterion is

$$G_i \text{ is CPL-valid} \quad \text{iff} \quad \text{minconc}(G_i)(1_j) = 1. \quad (4)$$

Close connections exist between CPL validity/invalidity (the latter satisfying a relation analogous to that of Eq. (3)) and that of CL (classical logic) validity or invalidity, noting

$$G_i \text{ is CL-valid} \quad \text{iff} \quad \&(b' \vee ab)_j \leq d' \vee c_i d. \quad (5)$$

(For further analysis, criticism, and extension of Adams' ideas, see [3].)

CRANOF Approach to Analyzing Quantitative Rule-Based Systems and Other Underconstrained Probability Problems

The previous conclusions show that the minconc function is not a reasonable measure (for reasonably high thresholds) of the degree of validity/invalidity of an entailment scheme and also show that the HP-validity/invalidity test is too stringent. Therefore, it seemed natural to replace the extremal minconc function by the more moderating *meanconc* function (well-justified from decision analysis in the form of conditional expectation and justified as always admissible, least-squares error, etc.—see any standard texts such as Rao [4] or Wilks [5]) within a Bayesian framework, where the unknown probability measure P here is treated as a random quantity with some appropriately assigned prior distribution, subject to the given premise set threshold constraints. Utilizing additional new theoretical results [6], an "optimal" choice of prior or priors essentially must come from the well-known Dirichlet family of distributions. It should be noted that, unlike the minconc function, the meanconc function in the

unity-limiting threshold case can take on nontrivial values and, in a natural sense, at any fixed threshold level, provides a reasonable measure of degree of validity of that entailment scheme under consideration. In particular, in full agreement with commonsense reasoning, transitivity, contraposition, and strengthening of antecedent are all SOPL-valid, where SOPL stands for Second-Order Probability Logic and where one defines validity of any G_j as

$$G_j \text{ is SOPL-valid} \quad \text{iff} \quad \lim_{(t_j \uparrow 1_j)} (\text{meanconc}(G_j)(t_j)) = 1, \quad (6)$$

SOPL-validity depending on some degree, of course, on the particular choice of prior for P . However, it has been pointed out (Bamber [7] and personal communications) that the limit in Eq. (4) remains the same as if the prior of P is a uniform distributional one, when the corresponding probability density function is bounded uniformly above and below (from zero) over its *natural* domain (again, see references).

Also, see [8] for additional background on both the theoretical structure of the meanconc function and its practical implementational form CRANOF—whereby a significant reduction in the complexity of computing $\text{meanconc}(G_j)(t_j)$ is achieved by, in effect, reducing the premise set of G_j to a single constraint, also taking into account the unity-limiting threshold behavior of meanconc ([7]). Finally, Table 1 is presented below to illustrate a few typical evaluations of $\text{meanconc}(G)$ for relatively simple entailment schemes G with P assigned a uniform prior distribution [8].

TABLE 1. Abridged table of calculations of degree-of-entailment functions, minconc and meanconc, for fixed threshold levels, and a comparison of CPL-, SOPL-, and HPL-validities for different types of entailment schemes.

Name of Entailment Scheme $D = [(a b)_j; (c d)]$	Given Levels of Premises: $P(a b)_j = t_j$, for otherwise arbitrary prob. meas. P	minconc(D)(t_j) (inequality threshold form)	meanconc(D)(t_j), assuming uniform prior for P 's (exact threshold form)	D is CPL-valid?	D is SOPL-valid?	D is HPL-valid?
Cautious Monotonicity: [(a b), (c b); (a bc)]	$P(a b) = s$, $P(c b) = t$	$\geq \max(s+t-1, 0)$	$\geq \max(s+t-1, 0)$	YES	YES	YES
Transitivity: [(a b), (b c); (a c)]	$P(a b) = s$, $P(b c) = t$	0	$= st + (1-t)/2 - p(s,t)/q(s,t)$, $p(s,t) = s(1-s)(2s-1)t(1-t^2)$, $q(s,t) = t+2t^2 + (s(1-s)(1-t)(2+3t-t^2))$	YES	YES	NO
Contraposition: [(a b); (b' a')]	$P(a b) = t$	0	$1/t + \frac{(1-t)\log(1-t)}{t^2}$	YES	YES	NO
Positive Conjunction: [(a b), (a c); (a bc)]	$P(a b) = t$, $P(a c) = t$	0	$(1+t)/3 + [((1+t)(2-t)/(3t)) \theta(t)]$, $\theta(t) = (t^2/4)[\log((2-t)/t)]/(1-t) - ((1-t)^2/4) \cdot \log((1+t)/(1-t))$	YES	YES	NO
Nixon Diamond: [(ab c), (d a), (d' b); (d c)]	$P(ab c) = s$, $P(d a) = t$, $P(d' b) = t$	0	1/2	YES	NO	NO
Abduction: [(a b), a; b]	$P(a b) = s$, $P(a) = t$	0	If $s \geq t$: $t/(2s)$, If $s < t$: $\frac{t^3 s(1-t)^2}{2(t^2 - 2st + s)}$	NO	NO	NO

EXTENDING APPLICABILITY OF CRANOF TO LINGUISTIC-BASED SYSTEMS

In considering linguistic-based information in rule-based systems and in formulating the linguistic analogue of the underconstrained conditional (including unconditional) probability problem, the role of fuzzy logic comes immediately to mind. This is based in part on the great practical success of fuzzy logic in running systems such as elevators, washing machines, etc., and on the now very large body of scientific literature supporting the modeling of linguistic information, relations, and decision processes via fuzzy logic. (See, e.g., past *Proceedings of IEEE International Conferences on Fuzzy Systems* or the *Proceedings of the Joint Conference on Information Sciences*, as well as basic texts, such as Dubois & Prade's now classic treatise [9] and Nguyen & Walker's [10].)

On the other hand, there still exists a lively controversy considering the merits of using probability theory and techniques in place of fuzzy logic and vice versa. (See Goodman's summary and listing of literature papers directly involved in this controversy [11].) This leads to the following area in which this author and H. T. Nguyen have played some role over the past several years: *the issue of the possible direct connection between fuzzy logic and probability theory* [12, 13, and 1]. Until this is completely resolved, it is this author's opinion that a comprehensive view of data fusion, which both theoretically and practically integrates linguistic-based information with probabilistic-based information, will not be achieved. In particular, this applies to rule-based systems, where the fuzzy logic community has developed a common approach that is claimed to be more satisfactory than any probability approach.

This paper once again points out the existence of deep, but tractable, relations among fuzzy logic, linguistic-based principles, probability theory, and commonsense reasoning mainly through the use of two basic mathematical tools: SOPL/CRANOF (as briefly described in the first section), and the representation theory of fuzzy sets by the one-point coverages of random sets (see [12, 13]) in conjunction with other recently developed mathematical tools (*conditional and relational event algebra* [14; 15, section 3]). In particular, homomorphic-like relations were established, connecting fuzzy-logic concepts and corresponding random-set concepts, where each fuzzy-set membership function involved is, in effect, interpreted as the weakest way to specify any of a class of corresponding random subsets of the fuzzy set's domain. These relations include natural random-set interpretations of various combinations of fuzzy-logic operators and Zadeh's well-known "extension theorem." This time, these connections are extended to include the formulation and use of inference rules obtained from a population of interest. The results presented here extend preliminary efforts provided in Goodman & Nguyen [1], where it was demonstrated that one type of fuzzy-logic approach to the modeling of inference rules for a population, relative to a given collection of attributes, using the ratio of fuzzy cardinalities or averaged membership level of the attributes, could also be interpreted in a probability framework. In addition, by using similar techniques, it is shown how other fuzzy-logic

concepts, commonly thought of as not directly relating to probability, may now also be put into a complete probabilistic setting, including the illustration for normalization of membership functions.

MATHEMATICAL RESULTS ESTABLISHING GENERAL FUZZY LOGIC POPULATION CONDITIONING PROBLEM AS AN UNDERCONSTRAINED CONDITIONAL PROBABILITY PROBLEM TREATABLE VIA SOPL/CRAFNOF

As in the previous sections, standard Boolean algebra and probability theory notation will be employed, with $[0,1]$ indicating unit interval; $\{0,1\}$ indicating the two element set containing 0, 1; \mathbf{R} indicating the real (or Euclidean) line and \mathbf{R}^m indicating the real (or Euclidean m -space), $P(D)$ indicating the power class of D (sometimes written 2^D —the class of all subsets of D), etc. "Equal by definition" is denoted as $=_d$. For background on copulas, see Schweizer & Sklar [16] and the recent excellent monograph by Nelsen [17]. Recall that copulas are any joint cdf's (cumulative probability distribution functions), all of whose one-dimensional marginal cdf's correspond to identical uniformly distributed rv's (random variables) over $[0,1]$.

Theorem 1. Modification of Goodman [18]

Let D be a finite set, $f, g: D \rightarrow [0,1]$ any two fuzzy-set membership functions, and $\text{cop}: [0,1]^{D \times D} \rightarrow [0,1]$ any copula with that domain, with (x,y) -marginal copulas indicated by, e.g., $\text{cop}_{x,y}$, x, y in D , etc. Then:

(i) There is a probability space (Ω, B, P) and a joint collection of 0-1-valued rv's, $Z_{f,x}, Z_{g,y}: \Omega \rightarrow \{0,1\}$, for all x, y in D with overall joint cdf $F_{f,g,\text{cop}} = \text{cop}_o((F_{f,x})_{x \in D}, (F_{g,y})_{y \in D}): \mathbf{R}^{D \times D} \rightarrow [0,1]$ (via Sklar's Theorem [16]), and, indicating the joint marginal (x,y) -components of cop , as $\text{cop}_{x,y}$, the joint cdf of $(Z_{f,x}, Z_{g,y})$ is, correspondingly, $F_{f,g,\text{cop},x,y}(\cdot, \cdot) = \text{cop}_{x,y} \circ (F_{f,x}(\cdot), F_{g,y}(\cdot))$, where \circ indicates functional composition and $F_{f,x}, F_{g,y}$ are each one-dimensional cdf's corresponding to mass-point probability functions $h_{f,x}, h_{g,y}$, respectively, where

$$\begin{aligned} P(Z_{f,x} = 1) &= h_{f,x}(1) = f(x); & P(Z_{f,x} = 0) &= h_{f,x}(0) = 1-f(x); \\ P(Z_{g,y} = 1) &= h_{g,y}(1) = g(y); & P(Z_{g,y} = 0) &= h_{g,y}(0) = 1-g(y); \end{aligned} \quad (7)$$

whence

$$F_{f,x}(s) = \begin{cases} 0, & \text{if } s < 0, \\ 1-f(x), & \text{if } 0 \leq s < 1, \\ 1, & \text{if } 1 \leq s; \end{cases} \quad F_{g,y}(s) = \begin{cases} 0, & \text{if } s < 0, \\ 1-g(y), & \text{if } 0 \leq s < 1, \\ 1, & \text{if } 1 \leq s; \end{cases} \quad \text{all } x, y \text{ in } D \quad (8)$$

(ii) Define random sets $S(f, \text{cop}), S(g, \text{cop}): \Omega \rightarrow P(D)$, $S(f, g, \text{cop}): \Omega \rightarrow P(D) \times P(D)$ as follows, for each ω in Ω :

$$\begin{aligned} S(f, g, \text{cop})(\omega) &= S(f, \text{cop})(\omega) \times S(g, \text{cop})(\omega) = \{(x,y): x, y \text{ in } D, Z_{f,x}(\omega) Z_{g,y}(\omega) = 1\}; \\ S(f, \text{cop})(\omega) &= \{x: x \text{ in } D, Z_{f,x}(\omega) = 1\}; \quad S(g, \text{cop})(\omega) = \{y: y \text{ in } D, Z_{g,y}(\omega) = 1\}; \end{aligned} \quad (9)$$

whence, by straightforward combinatoric considerations, the entire probability distributions of the marginal random subsets of D , $S(f, \text{cop}), S(g, \text{cop})$, as well as the joint random subset of $D \times D$, $S(f, g, \text{cop})$, are completely determined.

(iii) For any x, y in D , the following equality of one-point coverage events hold:

$$(x \text{ in } S(f, \text{cop})) = (Z_{f,x} = 1) ; (y \text{ in } S(g, \text{cop})) = (Z_{g,y} = 1); \quad (10)$$

$$((x,y) \text{ in } S(f, g, \text{cop})) = (x \text{ in } S(f, \text{cop})) \& (y \text{ in } S(g, \text{cop})) = (Z_{f,x} = 1) \& (Z_{g,y} = 1). \quad (11)$$

(iv) For any x, y in D , the following *one-point coverage representations* for f, g hold:

$$P(x \text{ in } S(f, \text{cop})) = P(Z_{f,x} = 1) = f(x) ; P(y \text{ in } S(g, \text{cop})) = P(Z_{g,y} = 1) = g(y); \quad (12)$$

$$\begin{aligned} P((x \text{ in } S(f, \text{cop})) \& (y \text{ in } S(g, \text{cop}))) &= P((Z_{f,x} = 1) \& (Z_{g,y} = 1)) \\ &= 1 - P(Z_{f,x} = 0) - P(Z_{g,y} = 0) + P((Z_{f,x} = 0) \& (Z_{g,y} = 0)) \\ &= 1 - P(Z_{f,x} = 0) - P(Z_{g,y} = 0) + P((Z_{f,x} \leq 0) \& (Z_{g,y} \leq 0)) \\ &= 1 - (1-f(x)) - (1-g(y)) + F_{f,g,\text{cop}_{x,y}}(0, 0) \\ &= f(x) + g(y) - 1 + \text{cop}_{x,y}(1-f(x), 1-g(y)) \\ &= f(x) + g(y) - \text{cocop}_{x,y}(f(x), g(y)) \\ &= {}_d \text{cop}_{x,y}^{\wedge}(f(x), g(y)), \end{aligned} \quad (13)$$

where we use the relation

$$\begin{aligned} F_{f,g,\text{cop}_{x,y}}(0, 0) &= \text{cop}_{x,y}^{\circ}(F_{f,x}(0), F_{g,y}(0)) \\ &= \text{cop}_{x,y}^{\circ}(b_{f,x}(0), b_{g,y}(0)) \\ &= \text{cop}_{x,y}^{\circ}(1-f(x), 1-g(y)) \end{aligned}$$

and where the functions cocop , cop^{\wedge} are called the *cocopula*, *survival copula*, respectively, of cop (the latter apparently being the special designation of Nelsen for modular transform [17, section 2.6]), where, for any s, t in $[0,1]$:

$$\text{cocop}(s, t) = {}_d 1 - \text{cop}(1-s, 1-t) ; \text{cop}^{\wedge}(s, t) = {}_d s+t - \text{cocop}(s,t). \quad (14)$$

(v) Specializing (iv) for $x = y$ in D arbitrary,

$$P(x \text{ in } S(f, \text{cop}) \cap S(g, \text{cop})) = P((x,x) \text{ in } S(f, g, \text{cop})) = \text{cop}_{x,y}^{\wedge}(f(x), g(x)). \quad (15)$$

(vi) As copula cop is allowed to vary arbitrarily, the full solution set of distribution-distinct random subsets of D that are one-point coverage equivalent to f, g , respectively in the sense of Eq. (12), is exhausted. ■

Remark 1. Note first that cocop is the DeMorgan transform of cop —so that if one thinks of cop as a generalized conjunction or "and" operator—as in fuzzy logic (with the usual desirable properties of being nondecreasing in its arguments and having appropriate boundary properties when one of the arguments is 0 or 1), then, naturally, cocop can be thought of as a general disjunction or "or" operator. Nelsen [17, section 2], shows that the survival copula is always a legitimate copula and shows the characterization

$$\text{cop}^{\wedge} = \text{cop} \text{ iff } \text{cop} \text{ is } \textit{radially} \text{ symmetric}, \quad (16)$$

where the latter means that the joint r.v. Y represented by cop is such that $Y - (1/2, 1/2)$ and $(1/2, 1/2) - Y$ have the same distribution. In particular, radial symmetry—and hence the validity of Eq. (16)—holds for all Gaussian copulas Ψ_{ρ} ($\Psi-1(\cdot)$, $\Psi-1(\cdot)$), where Ψ_{ρ} is the joint cdf of distribution Gaussian $(0_2, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$ and Ψ is the cdf of the standardized

one-dimensional Gaussian distribution Gaussian (0,1) and all of Frank's Archimedean copula family [17], [19] (i.e., associative, commutative with $\text{cop}(s,s) < s$, for $0 < s < 1$)—which includes the copulas prod and minsum, as well as the special copula min, where for all s, t in $[0,1]$, min, prod are the usual arithmetic minimum and product of s, t , respectively, while minsum(s,t) is given as

$$\text{minsum}(s,t) = \min(s+t-1, 0). \quad (17)$$

■

Theorem 2. Extension of Goodman & Nguyen [13]

Suppose that D is a finite set, $f, g: D \rightarrow [0,1]$ are any two fuzzy set membership functions, $\text{cop}: [0,1]^{D \times D} \rightarrow [0,1]$ is any copula with that domain, and $w: D \rightarrow [0,1]$ is a probability function. Define

$$((f|g)_{\text{cop},w}) = \frac{\sum_{x \in D} (w(x) \cdot \text{cop}^{\wedge}(f(x), g(x)))}{\sum_{x \in D} (w(x) \cdot g(x))}. \quad (18)$$

Then, in the sense of Theorem 1, there is a probability space (Ω, B, P) and random sets $S(f, \text{cop}), S(g, \text{cop}): \Omega \rightarrow P(D)$, $S(f, g, \text{cop})$: with the one-point coverage relations holding as in Eqs. (12), and, without loss of generality, there exists a random variable $V: \Omega \rightarrow D$, independent of $S(f, g, \text{cop})$, and hence of $S(f, \text{cop}), S(g, \text{cop})$, such that the probability function of V is w , so that

$$(f|g)_{\text{cop},w} = P(a_{f, \text{cop}} | b_{g, \text{cop}}), \quad (19)$$

an ordinary conditional probability, where events $a_{f, \text{cop}}, b_{g, \text{cop}}$ in B are defined as the two-stage randomization events

$$a_{f, \text{cop}} =_d (V \text{ in } S(f, \text{cop})), \quad b_{g, \text{cop}} = (V \text{ in } S(g, \text{cop})), \quad (20)$$

so that in reduced form,

$$P(a_{f, \text{cop}} | b_{g, \text{cop}}) = P(a_{f, \text{cop}} \& b_{g, \text{cop}} | b_{g, \text{cop}}) = P(V \text{ in } S(f, \text{cop}) \cap S(g, \text{cop})) / P(V \text{ in } S(g, \text{cop})). \quad (21)$$

Proof: Use the usual conditioning property of probabilities, independence of V , and Eq. (11) at each outcome of r.v. V ,

$$\begin{aligned} P(V \text{ in } S(f, \text{cop}) \text{ and } V \text{ in } S(g, \text{cop})) &= E_V(P(V \text{ in } S(f, \text{cop}) \text{ and } V \text{ in } S(g, \text{cop}) | V)) \\ &= E_V(\text{cop}^{\wedge}(f(V), g(V))) = \sum_{x \in D} (w(x) \cdot \text{cop}^{\wedge}(f(x), g(x))). \end{aligned} \quad (22)$$

Similarly (and more simply), now using Eq. (12) in place of Eq. (13),

$$P(V \text{ in } S(g, \text{cop})) = E_V(P(V \text{ in } S(g, \text{cop}) | V)) = E_V(g(V)) = \sum_{x \in D} (w(x) \cdot g(x)). \quad (23)$$

The desired results hold by dividing Eq. (22) by Eq. (23). ■

Remark 2 and an Example. In Theorem 2, for the special case of w corresponding to a uniform distribution over *population* D , canceling the $1/\text{card}(D)$ factor, and usually—but not always choosing cop to be either min or prod—the numerator of the quantity $(f|g)_{\text{cop},w}$ reduces to the popular fuzzy-logic concept of the *fuzzy cardinality* of f "and" g for population D , i.e., to what extent the entire population D has characteristics described by f "and" g , while, similarly, the denominator represents the fuzzy cardinality of g (by itself) for population D . In turn, the arithmetic

division of these, i.e., the quantity $(fg)_{\text{cop},w}$ becomes the *relative fuzzy cardinality* of f "and" g for D compared to fuzzy cardinality of g for D, i.e., the overall fuzzy conditioning of f to g with respect to population D. The latter, beginning with Zadeh's ideas [20, 21], followed by Dubois & Prade's modifications [22], and Kosko's related concept of *fuzzy subsethood* [23], are used ubiquitously in the fuzzy-logic community for reasoning. In this process, one considers the premise set of a particular linguistic entailment of interest, the latter being formally the same as the probability-framed previous $G_i = [(a|b)_j; (c_i|d)]$, but now where each $(a_j|b_j)$ is replaced by a fuzzy conditional—in its general form *the same as* $(f_j|g_j)_{\text{cop},w}$ —formed as in Eq. (18), now with f replaced by f_j , g by g_j (for possibly pre-logically compounded fuzzy-set membership functions), j in J; and with similar remarks applicable to the potential conclusion $(c_i|d)$ replaced by $(f_{o,i}|g_o)_{\text{cop},w}$, for some fuzzy sets $f_{o,i}$, g_o , etc. But, Theorem 2 (with suitable modifications, where required) essentially shows that any such $(f_j|g_j)_{\text{cop},w} = P(a_{f_j, \text{cop}} | b_{g_j, \text{cop}})$, with a similar relation holding the potential conclusion. Moreover, the variability of P subject to whatever arbitrary but fixed levels t_j are set for the premise collection holds in the same meaningful manner as in the case where one began the problem in a probability framework, i.e., for typical entailment schemes of the form G_i . As an application of this, suppose one considers the transitivity scheme, which Zadeh has also considered and modeled his premise set as indicated above, but has used a method solely developed within fuzzy logic for determining what the appropriate conclusion should be [21]. Thus, three attributes are present, where, e.g., population D here is the set of all enemy ships in area A, "ships with type 1 weapons onboard" corresponds to known or estimated fuzzy-set membership function f over D; "ships with elongated hulls" corresponds to known or estimated fuzzy-set membership function g over D; "ships with signature pattern Q" corresponding to known or estimated fuzzy-set membership function h over D. Moreover, other truth modifiers may be present, such as "it is mostly true," "it is somewhat true," etc. Here, for simplicity, suppose for the premise set, one actually has "it is highly true that the enemy ships in A with signature pattern Q have elongated hulls," "it is moderately likely that an enemy ship in A with an elongated hull has type 1 weapons onboard." Can one conclude "it is x-likely that an enemy ship in A with signature pattern Q has type 1 weapons onboard," where the degree of truth x is to be determined? Assume that "it is highly true" is represented by a known or estimated fuzzy-set membership function M over [0,1], which is monotone increasing, "it is moderately likely" is also represented by a (different—not as steep toward 1 as M, etc.) known or estimated fuzzy-set membership function N over [0,1], where $M(r) = N(r) = r$, for $r = 0$ or 1. Hence, for any arbitrary levels s, t in [0,1], the conditional fuzzy relations here are, for some choice of copula and population weighting function w,

$$\begin{aligned} M((f|g)_{\text{cop},w}) = s, N((g|h)_{\text{cop},w}) = t \quad & \text{iff } (f|g)_{\text{cop},w} = M^{-1}(s), (g|h)_{\text{cop},w} = N^{-1}(t) \\ & \text{iff, using Theorem 2, } P(a_{f, \text{cop}} | b_{g, \text{cop}}) = M^{-1}(s), \\ & P(b_{g, \text{cop}} | c_{h, \text{cop}}) = N^{-1}(t). \end{aligned} \tag{24}$$

Thus, for any given levels s, t , one can now consider the SOPL-estimate of the potential conclusion for transitivity, $P(a_{f, \text{cop}} | b_{g, \text{cop}})$, with respect to the premise set above at thresholds s, t , where the entire entailment scheme is

$$G = [(a_{f, \text{cop}} | b_{g, \text{cop}}), (b_{g, \text{cop}} | c_{h, \text{cop}}); (a_{f, \text{cop}} | c_{h, \text{cop}})]; \quad (25)$$

$$\text{meanconc}(G)(M^{-1}(s), N^{-1}(t)) = E_P(P(a_{f, \text{cop}} | c_{h, \text{cop}}) |$$

$$P(a_{f, \text{cop}} | b_{g, \text{cop}}) = M^{-1}(s), P(b_{g, \text{cop}} | c_{h, \text{cop}}) = N^{-1}(t)). \quad (26)$$

In turn, Table 1 shows that under a uniform distributional assumption on what P could be, subject to its constraints in the premise set of G , for any given s, t in $[1/2, 1]$

$$\text{meanconc}(G)(M^{-1}(s), N^{-1}(t)) = \rho(M^{-1}(s), N^{-1}(t)),$$

where, for any s, t in $[1/2, 1]$,

$$\begin{aligned} \rho(s, t) &= {}_d st + (1-t)/2 - p(s, t)/q(s, t); \quad p(s, t) = {}_d s(1-s)(2s-1)t(1-t^2); \\ q(s, t) &= {}_d t+2t^2 + (s(1-s)(1-t)(2+3t-t^2)), \end{aligned} \quad (27)$$

where,

$$\rho(s, t) \approx \rho_0(s, t) = {}_d st + (1-t)/2, \text{ for values of } s, t \text{ sufficiently close to } 1. \quad (28)$$

Hence, the posterior conditional (given the premise constraints for any s, t) is approximately equal to $\rho_0(M^{-1}(s), N^{-1}(t))$, which can be interpreted also as a truth modifier with respect to two variables, noting its limit is unity as s, t approach unity, etc. Of course, all of the above applies to any fuzzy-logic entailment scheme relative to the original premise sets utilizing overall fuzzy conditioning for some population D .

Remark 3. In the same spirit of Theorem 2, other fuzzy-logic concepts can now be fully interpreted. Due to space limitations, only the example of fuzzy normalization will be considered here. In this situation, a fuzzy membership function, say, $f: D \rightarrow [0, 1]$ is given, followed by its normalization function $\text{norm}(f): D \rightarrow [0, 1]$, which is now obviously a legitimate probability function over finite population D , where

$$\text{norm}(f) = \left(1 / \sum_{x \in D} (f(x))\right) \cdot f. \quad (29)$$

But, if one considers, à la Theorem 1, for any choice of copula cop , a probability space (Ω, B, P) , for which, without loss of generality, there is both a random set $S(f, \text{cop}): \Omega \rightarrow P(D)$ and an independent random variable $V: \Omega \rightarrow D$ uniformly distributed over D , with the one-point coverage relation holding

$$P(x \text{ in } S(f, \text{cop})) = f(x), \text{ all } x \text{ in } D, \quad (30)$$

for any x in D , specializing Eq. (23) with g replaced by f ,

$$\begin{aligned} P(V = x | V \text{ in } S(f, \text{cop})) &= P(V = x \text{ and } x \text{ in } S(f, \text{cop})) / P(V \text{ in } S(f, \text{cop})) \\ &= ((1/\text{card}(D)) \cdot f(x)) / \sum_{x \in D} (1/\text{card}(D)) \cdot g(x) = \text{norm}(f)(x), \end{aligned} \quad (31)$$

showing fuzzy normalization is actually a simple conditional probability restriction of the two-stage randomization for one-point coverages. A future paper will deal with related issues.

REFERENCES

1. Goodman, I. R. and H. T. Nguyen. 1999. "Application of Conditional and Relational Event Algebra to the Defining of Fuzzy Logic Concepts," *Proceedings of Signal Processing, Sensor Fusion & Target Recognition VIII*, Society of Photo-Optical Instrumentation Engineers (SPIE), vol. 3720, pp. 37–46.
2. Adams, E. W. 1986. "On the Logic of High Probability," *Journal of Philosophical Logic*, vol. 15, pp. 255–279.
3. Adams, E. W. 1996. "Four Probability-Preserving Properties of Inferences," *Journal of Philosophical Logic*, vol. 25, pp. 1–24.
4. Rao, C. R. 1973. *Linear Statistical Inference & Its Applications*, 2nd Ed., Wiley, New York, NY.
5. Wilks, S. S. 1963. *Mathematical Statistics*, Wiley, New York, NY.
6. Goodman, I. R. and H. T. Nguyen 1999. "Probability Updating Using Second-Order Probabilities and Conditional Event Algebra," *Information Sciences*, vol. 121, pp. 295–347.
7. Bamber, D. 2000. "Entailment with Near Surety of Scaled Assertions of High Conditional Probability," *Journal of Philosophical Logic*, vol. 29, pp. 1–74.
8. Bamber, D. and I. R. Goodman. 2000. "New Uses of Second-Order Probability Techniques in Estimating Critical Probabilities in Command and Control Decision-Making," *Proceedings of the 2000 Command & Control Research & Technology Symposium*, Naval Postgraduate School, http://www.dodccrp.org/2000CCRTS/cd/html/pdf_papers/Track_4/124.pdf.
9. Dubois, D. and H. Prade. 1980. *Fuzzy Sets & Systems: Theory and Applications*, Academic Press, New York, NY.
10. Nguyen, H. T. and E. A. Walker. 1997. *A First Course in Fuzzy Logic*, CRC Press, New York, NY.
11. Goodman, I. R. 1998. "Random Sets and Fuzzy Sets: a Special Connection," *Proceedings of the International Conference on Multisource-Multisensor Information Fusion (Fusion'98)*, vol. 1, pp. 93–100.
12. Goodman, I. R. and H. T. Nguyen. 1985. *Uncertainty Models for Knowledge-Based Systems*, North-Holland Press, Amsterdam.
13. Goodman, I. R. and G. F. Kramer. 1997. "Extension of Relational and Conditional Event Algebra to Random Sets with Applications to Data Fusion," in *Random Sets: Theory & Applications* (J. Goutsias, R. P. Mahler, and H. T. Nguyen, eds.), Springer, New York, NY, pp. 209–242.
14. Goodman, I. R. and H. T. Nguyen. 1995. "Mathematical Foundations of Conditionals and Their Probabilistic Assignments," *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems*, vol. 3, no. 3 (September), pp. 247–339.
15. Goodman, I. R., R. P. Mahler, and H. T. Nguyen. 1997. *Mathematics of Data Fusion*, Kluwer Academic, Dordrecht, Holland.
16. Schweizer, B. and A. Sklar. 1983. *Probabilistic Metric Spaces*, North-Holland, Amsterdam.
17. Nelsen, R. B. 1999. *An Introduction to Copulas* (Lecture Notes in Statistics, no. 139), Springer, New York, NY.



I. R. Goodman

Ph.D. in Mathematics, Temple University, 1972

Current Research: Mathematical foundations of data fusion via conditional probabilistic logic; Boolean conditional event algebra; one-point random set representations of fuzzy logic.

18. Goodman, I. R. 1994. "A New Characterization of Fuzzy Logic Operators Producing Homomorphic-Like Relations with One Point Coverage of Random Sets," in *Advances in Fuzzy Theory & Technology*, (P. P. Wang, ed.), Duke University, Durham, NC, vol. 2, pp. 133–159.
19. Frank, M. J. 1979. "On the Simultaneous Associativity of $F(x,y)$ and $x+y-F(x,y)$," *Aequationes Mathematicae*, vol. 19, pp. 194–226.
20. Zadeh, L. A. 1985. "Syllogistic Reasoning as a Basis for Combination of Evidence in Expert Systems," *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-85)*, vol. 1, pp. 417–419.
21. Zadeh, L. A. 1978. "PRUF: a Meaning Representation Language for Natural Languages," *International Journal of Man–Machine Studies*, vol. 10, pp. 395–460.
22. Dubois, D. and H. Prade. 1988. "On Fuzzy Syllogisms," *Computational Intelligence*, vol. 4, pp. 171–179.
23. Kosko, B. 1992. *Neural Networks and Fuzzy Systems*, Prentice-Hall, Englewood Cliffs, NJ.



On Knowledge Amplification by Structured Expert Randomization (KASER)

Stuart H. Rubin
SSC San Diego

INTRODUCTION TO RANDOMIZATION

The theory of randomization was first published by Chaitin and Kolmogorov [1] in 1975. Their work may be seen as a consequence of Gödel's Incompleteness Theorem [2] in that it shows were it not for essential incompleteness, a universal knowledge base could, in principle, be constructed—one that need employ no search other than referential search. Lin and Vitter [3] proved that learning must be domain-specific to be tractable. The fundamental need for domain-specific knowledge is in keeping with Rubin's proof of the Unsolvability of the Randomization Problem [4]. This paper went on to introduce the concept of knowledge amplification. Production rules are expressed in the form of situation action pairs. Such rules, once discovered to be in error, are corrected through acquisition. Conventionally, a new rule must be acquired for each correction. This is linear learning.

The acknowledged key to breakthroughs in the creation of intelligent software is cracking the knowledge acquisition bottleneck [5]. Learning how to learn is fundamentally dependent on representing the knowledge in the form of a society of experts. Minsky's seminal work here led to the development of intelligent agent architectures [6]. Furthermore, Minsky [7] and Rubin [4] independently provided compelling evidence that the representational formalism itself must be included in the definition of domain-specific learning if it is to be scalable.

A KASER is defined to be a knowledge amplifier that is based on the principle of structured expert randomization. A Type I KASER is one where the user supplies declarative knowledge in the form of a semantic tree using single inheritance.

A Type II KASER can automatically induce this tree through the use of randomization and set operations on property lists, which are acquired by way of database query and user-interaction. An overview of a Type II KASER is provided below. Unlike conventional intelligent systems, KASERs are capable of accelerated learning in symmetric domains.

Figure 1 plots the knowledge acquired by an intelligent system vs. the cost of acquisition. Conventional expert systems will generate the curve below break-even. That is, with conventional expert systems, cost increases with scale and is never better than linear. Compare this with KASERs where cost decreases with scale and is always better than linear unless the domain has no symmetries (i.e., it is random). Note that such

ABSTRACT

We define Knowledge Amplification by Structured Expert Randomization (KASER). A KASER can automatically acquire a virtual rule space exponentially larger than the actual rule space and with an exponentially decreasing nonzero likelihood of error. The KASER cracks the knowledge acquisition bottleneck in intelligent systems by amplifying user-supplied knowledge. This enables the construction of an intelligent system, which is creative, fail-soft, learns over a network, and otherwise has enormous potential for automated decision-making.

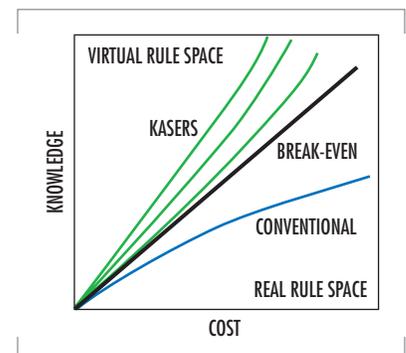


FIGURE 1. The comparative costs of knowledge acquisition.

domains do not exist with scale in practice. Similarly, purely symmetric domains do not exist with scale in practice either. The more symmetric the operational domain, the less the cost of knowledge acquisition and the higher the curve appears in the graph. It is always the case that the virtual rule space \gg the real rule space.

INDUCING PROPERTY LISTS

We will define a production system that can automatically acquire a virtual rule space that is exponentially larger than the actual rule space with an exponentially decreasing non-zero likelihood of error. Moreover, the generalization mechanism will not only be bounded in its error, but other than for a straightforward user-query process, it will operate without any *a priori* knowledge supplied by the user.

To begin, define a production rule (e.g., using ANSI Common LISP) to be an ordered pair—the first member of which is a set of antecedent predicates, and the second member of which is an ordered list of consequent predicates. Predicates can be numbers (e.g., [1..2] \vee [10..20]) or words [8].

Previously unknown words or phrases can be recursively defined in terms of known ones. For example, the moves of a Queen in chess (i.e., unknown) can be defined in terms of the move for a Bishop (i.e., known) union those for a Rook (i.e., known). This is a union of property lists. Other basic set operations may likewise be used (e.g., intersection, difference, not, etc.). The use of fuzzy set operators here (e.g., "almost the same as") pertains to computing with words [8].

In a Type I KASER, words and phrases are entered through the use of pull-down menus. In that manner, semantically identical concepts (e.g., Hello and Hi) are not ascribed a distinct syntax, which would otherwise serve to dilute the efficiency of the learning mechanism. In a Type II KASER, distinct syntax may be equated to yield the equivalent normalized semantics. To better visualize this, think of a child who may ask, "What is a bird?" to which the reply is, "It is an animal that flies," to which the question is, "What is an animal?" to which the reply is, "It is a living thing," to which the question is, "What is a living thing?" to which the reply (often) is, "Eat your soup!" (i.e., a Type I delimiter, or stop marker gene).

Two sample rules and their representation follow.

Hydrogen \wedge Oxygen \wedge Spark \rightarrow Steam

R1: ({Hydrogen, Oxygen, Spark} (Steam))

Hydrogen \wedge Oxygen \wedge Match \rightarrow Steam

R2: ({Hydrogen, Oxygen, Match} (Steam))

R1 and *R2* may be generalized, since the consequent predicates are identical (i.e., the right-hand sides [RHSs] are equivalent) and the antecedent terms differ in exactly one predicate. This is termed a level-1 generalization because it is one level removed from ground truth. In a level-*i* generalization, *i* is the maximum level of generalization for any antecedent predicate. The need for a generalization squelch arises because contexts may be presented for which there is no matching rule in the real space. Generalizations can be recursively defined.

The advocated approach captures an arbitrary rule's context—something that cannot be accomplished through the use of property lists alone. If veristic terms such as "Warm" are generalized to such terms as "Heat" for example, then qualitative fuzziness will be captured.

A1: ({Heat} {Spark, Match})(X001 Explosive-Gas-Igniter))

Generalization, *A1*, tells us that antecedent predicate, "Heat" is more general than either a Spark or a Match. We may also write this as $\text{Heat} > \{\text{Spark, Match}\}$. Note that the relation ">" is used to denote ancestral generalizations (and vice versa). The general predicate is initially specified as *X00i*, but this is replaced after interactive query with the user, where possible. Otherwise, the next-level expansions will need to be printed for the user to read. Also, "redundant, at-least-as-specific" rules are always expunged.

The common property list follows the set of instances. Here, the list informs us that a spark or a match may be generalized to Heat because both share the property of being an Explosive-Gas-Igniter. Properties are dynamic. They must be capable of being hierarchically represented, augmented, and randomized. In addition, property lists are subject to set operations (e.g., intersection). Properties can be acquired by way of database and/or user query.

User-queries can be preprocessed by a companion veristic mining system. Similarly, system-generated queries can be post-processed by companion systems. Companion systems can also play a role in imparting tractability to the inference engine.

Consequent terms, being sequences, are taken to be immutable. The idea here is to automatically create a hierarchy of consequent definitions to maximize the potential for rule reuse. Begin by selecting a pair of rules having identical left-hand sides (LHSs), where possible. Consider:

R3: ({Hydrogen, Oxygen, Heat} (Steam))

R4: ({Hydrogen, Oxygen, Heat} (Light, Heat))

Next, an attempt is made to generalize the consequent sequences with the following result.

C1: ((Energy) ((Steam) (Light Heat))(X002 Power-Source))

Here, the properties of Steam intersect those of Light and Heat to yield the property, Power-Source. Thus, a property of Energy, in the current context at least, is that it is a Power Source. Rules *R3* and *R4* are now replaced by their valid generalization, *R5*:

R5: ({Hydrogen, Oxygen, Heat} (Energy))

A key concept is that further learning can serve to correct any latent errors. In addition, notice that as the level of randomization increases on the LHS and RHS, the potential for matching rules, and thus inducing further generalizations, increases by way of feedback. Consequent randomization brings the consequents into a normal form, which then serves to increase the possibility of getting antecedent generalizations, since more RHSs can be equated. Antecedent randomization is similar.

Next, consider *R5*, where *R6* is acquired and appears as follows after substitution using *C1*.

R6: ({Candle, Match} (Energy))

The system always attempts to randomize the knowledge as much as possible. Using *A1* and *C1* leads to the level-1 conjecture, *R7*, which replaces *R6*.

R7: ({Candle, Heat} (Energy))

R7 is not to be generalized with *R6*. This is because {Match, Heat} is the same as {Match, Spark, Match}, which of course reduces to Heat and is already captured by *R7*.

At this point, learning by the system can be demonstrated. Suppose the user asks the system what will happen if a spark is applied to a candle. While this is a plausible method to light a candle, this method will not usually be successful. Thus, the user must report to the system the correct consequent for this action:

R8: ({Candle, Spark} (No-Light))

R8 is a more-specific rule than is *R7* because the former is a level-1 generalization, while the latter is at level-0. Thus, *R8* will be preferentially fired when possible by using a most-specific agenda mechanism. It, too, will be subject to subsequent generalization. Notice that the new consequent will protect against similar error.

The learning process has not completed. We still need to correct the properties list so that Matches and Sparks can be differentiated in the context of lighting a candle. The following property (i.e., LISP) list is obtained.

P1: (Match Explosive-Gas-Igniter Wick-Lighter)

P2: (Spark Explosive-Gas-Igniter)

Now, since Heat is a superclass of Match, its property list is unioned with the new property(s): Wick-Lighter. Suppose, at this point, the user poses the same question, "What will happen if a spark is applied to a candle?" Rule *R7* informs us that it will light; whereas, *R8* informs us that it will not. Again, the inference engine can readily select the appropriate rule to fire because of specialization. However, here there is yet more to learn. Here is what is known: *R7* and *R8* differ on the LHS in exactly one predicate and $prop(Heat) \cap prop(No - Light) = \emptyset$. The reason that the candle lights for a match, but not for a spark can be delimited by computing, $prop(Match) - prop(Spark) = prop(P1) - prop(P2) = (Wick-Lighter)$. Rule *R7* is now replaced by *R7'*:

R7': ({Candle, (X003 Wick-Lighter)} (Energy))

that is, a property list named X003 has been substituted for Heat. Notice that X003 is necessarily a subclass of Heat. Then, anything that has (all) the properties on the property list (i.e., X003) can presumably light a candle (e.g., a torch). Observe that the human in the loop need not know why a list of properties is relevant, since the reasons will be automatically discovered. Notice that a Spark can no longer light a candle and only those items having at least Wick-Lighter in their property classes can light a candle. Observe the nonlinear learning that has been enabled here!

Consider now the rule:

R9: ({Candle, Match} (Energy))

Clearly, this rule is correct as written. Candles do indeed produce steam, light, and heat. The usefulness of induction follows from the fact that the

system has no knowledge that a candle is a hydrocarbon and hydrocarbons produce steam as a byproduct of combustion.

Antecedent predicate generalizations can be rendered more class-specific as necessary to correct overgeneralizations by increasing the number of levels of available generalization. The rule consequents will not be affected. For example:

A2: ({Car} {Ford, Fiat})

yields:

A3: ({Car} {Family-Car, Sports-Car})

A4: ({Family-Car} {Ford})

A5: ({Sports-Car} {Fiat})

Property lists can be automatically organized into a hierarchical configuration through the use of simple set operations. This means that rules can be generalized or specialized through the use of the disjunctive or conjunctive operators, respectively. Such property lists can be associatively retrieved through the use of a grammatical randomization process [9]. Moreover, matching operations then need to incorporate searching subclasses and superclasses as necessary.

Finally, we note that this system can incorporate fuzzy programming [10]. Fuzzy programming will enable the system to explore a space of alternative contexts as delimited by optional consequent filters and ranked by the level of generalization used to obtain a contextual match (see below).

GRAMMATICAL RANDOMIZATION

Consider the following three property lists:

P3: (Ice A B C)

P4: (Water B C D)

P5: (Steam C E)

Here, ice, water, and steam share a common property, C, which, for example, might be that they are all composed of H₂O. Also, only ice has property A (e.g., frozen); only water has property D (e.g., liquid); and only steam has property E (e.g., gaseous). Observe that only ice and water share property B (e.g., heavier-than-air).

The use of a hierarchical object-representation is fundamental to the specification of property lists, antecedent sets, or consequent sequences. For example, when one specifies the object, "aircraft carrier," one implicitly includes all of its capabilities, subsystems, and the like. One cannot and should not have to specify each subsystem individually. We proceed to develop a randomization for the sample property lists; although, it should be clear that the same approach will work equally well for the antecedent and consequent predicates. Perhaps the most relevant distinction is that one needs to distinguish object sequence dependence from independence in the notation. Of course, property lists are sequence-independent.

As the example stands now, to specify the properties of ice or water, one need state the three properties of each (in any order). This may not seem

too difficult, but this is only because the list-size is small. Consider now the randomized version of the property lists:

P3: (Ice A Precipitation)

P4: (Water Precipitation D)

P5: (Steam C E)

P6: (Precipitation B C)

Here, the property of precipitation has been randomized from the property data. Observe, that if the user states property B, then the system will offer the user exactly three choices (e.g., by way of a dynamic pull-down menu): B, Precipitation, or Random. The Random choice allows the user to complete the specification using arbitrary objects. In other words, an associative memory has been defined. Similarly, if the user selects Precipitation, then the system will offer the user exactly four choices (e.g., again by way of a dynamic pull-down menu): Precipitation, Ice, Water, or Random.

Suppose that in keeping with the previously described nomenclature conventions, we had the following property list specifications:

P3': (X004 A Precipitation)

P4': (X005 Precipitation D)

In this case, if the user selects Precipitation, then the system will offer the user the following four choices: Precipitation, A, D, or Random. In other words, it attempts to pattern-match and extrapolate the set.

In practice, randomization is based on known classifications—not arbitrary ones. Thus, in the previous example, the randomization of *P3* and *P4* requires that *P6* be known *a priori*. Again, this still allows for the use of integer identifiers.

Next, it can be seen that the usefulness of randomization is a function of its degree. The relevant question then pertains to how to realize the maximal degree of randomization. First, recall that as rules are generalized, the possibilities for further predicate generalization are increased. This, in turn, implies that the substitution and subsequent refinement of property lists for predicates is increased. Finally, as a result, the virtual space of properly mapped contexts (i.e., conjectures) grows at a rapid rate. Experimental evidence to date indicates that this rate may be exponential for symmetric domains.

Next, we turn our attention to the inference engine, which is common to Type I and II KASERs. Basically, in a Type I KASER, conflict resolution is accomplished through the use of a hierarchical tree of objects evolved by a knowledge engineer, which define generalization and specialization (see below); whereas, in a Type II KASER, conflict resolution is the same as in a Type I KASER, but where the system, instead of the knowledge engineer, evolves hierarchical property lists, which serve to increase the size of the virtual contextual space—without sacrificing convergence in the quality of the response. In effect, declarative knowledge is randomized to yield procedural knowledge.

ACTIVE RANDOMIZATION

Active randomization is a symbiosis of property lists and grammatical randomization. Property lists are really just predicates that are subject to

grammatical randomization. Moreover, randomized predicates allow the user to specify contexts and associated actions by using minimal effort [9]. Next, suppose that we had:

(A B C D) (i.e., the properties of A are B, C, D)

Here, A is the randomization of B, C, D. Similarly, we may have

(B E F)

and the two rules (i.e., antecedent differentiation):

$R10: A S \rightarrow W$

$R11: X S \rightarrow W$

Then, we can create a randomization:

(Q A X)

which, since valid, leads to the following replacement of $R10$ and $R11$:

$R12: Q S \rightarrow W$

This replacement allows for the possibility of new rule pairings and the desired process then iterates. Thus, we have

(Q: A \cup X) {expanding A, X}

These are *active transforms* [9] in the sense that whenever A or X change their membership, the properties of Q may change. Evidently, this is a converging process. However, if subsequently we had

$R13: A S \rightarrow T$

$R14: X S \rightarrow G$

where, T and G have no properties in common (i.e., neither is a subsequence of the other), then it becomes clear that A cannot substitute for X and vice versa. In other words,

$R13: (A-X) S \rightarrow T$

$R14: (X-A) S \rightarrow G$

Thus, we have

(A: A - X) {contracting A}

(X: X - A) {contracting X}

These are active transforms, and again, this is a converging process. Next, suppose that T and G are such that G is a subsequence of T without loss of generality. Then, it follows that A is a subset of X and

(A: A \cap X) {contracting A}

(X: X \cup A) {expanding X}

These are active transforms. This is not, however, necessarily a converging process. That is not to say that it will diverge without bounds. It is just not stable. We do not view this as a problem. It is to be viewed as an oscillatory system that, in some ways, may mimic brain waves. The complexity of interaction will increase as the system is scaled up. The eventual need for high-speed parallel/distributed processing is apparent. The case for consequent differentiation is similar. Here though, one is processing sequences instead of sets.

OBJECT-ORIENTED TRANSLATION MENUS

The Type I KASER requires that declarative knowledge be (dynamically) compiled in the form of object-oriented hierarchical phrase translation

menus. Each class (i.e., antecedent and consequent) of bipartite predicates can be interrelated through their relative positions in an object-oriented semantic tree. A declarative knowledge of interrelatedness provides a basis for commonsense reasoning, as will be detailed in the next section. The subject of this section pertains to the creation, maintenance, and use of the object-oriented trees as follows.

1. The phrase-translation menus serve as an intermediate code (as in a compiler) where English sentences can be compiled into menu commands by using rule-based compiler bootstraps. KASERs can be arranged in a network configuration where each KASER can add (post) to or delete from the context of another. This will greatly expand the intelligence of the network with scale and serves to define Minsky's "Society of Mind" [6]. Furthermore, the very-high-level domain-specific language(s) used to define each predicate can be compiled through a network of expert compilers. Alternatively, neural networks can be used to supply symbolic tokens at the front end.
2. Each antecedent or consequent phrase can be associated with a textual explanation, Microsoft's Text-to-Speech engine (Version 4.0), an audio file, a photo, and/or a video file. Images may be photographs, screen captures, scans, drawings, etc. They may also be annotated with arrows, numbers, etc. Voice navigation may be added at a later date.
3. Antecedents and consequents can be captured by using an object-oriented approach. The idea is to place descriptive phrases in an object-oriented hierarchy such that subclasses inherit all of their properties from a unique superclass and may include additional properties as well. Menus can beget submenus, and new phrases can be acquired at any level.

Consider the partial path, office supply, paper clip and the partial path, conductor, paper clip. Here, any subclass of paper clip will have very different constraints depending on its derivation. For example, anything true of paper clips in the context of their use as conductors must hold for every subclass of paper clips on that path. Unique antecedent integers can be set up to be triggered by external rules. Similarly, unique consequent integers can be set up to fire external procedures. All we need do is facilitate such hooks for future expansion (e.g., the radar-mining application domain).

Each project is saved as a distinct file, which consists of the antecedent and consequent trees, the associated rule base, and possibly the multimedia attachments.

4. A tree structure and not a graph structure is appropriate because the structure needs to be readily capable of dynamic acquisition (i.e., relatively random phrases) and deletion, which cannot be accomplished in the presence of cycles due to side effects. Note that entering a new phrase in a menu implies that it is semantically distinct from the existing phrases, if any, in that menu.
5. A tree structure is mapped to a context-free grammar (CFG), where the mapping process needs to be incremental in view of the large size of the trees. Each node or phrase is assigned a unique number, which serves to uniquely identify the path.
6. Each phrase may be tagged with a help file, which also serves the purposes of the explanation subsystem. This implies that conjuncts are not necessary to the purpose of the antecedent or consequent trees.

7. Each menu should be limited to on the order of one screen of items (e.g., 22). Toward this end, objects should be dynamically subdivided into distinct classes. That is, new submenus can be dynamically created and objects moved to or from them.
8. Three contiguous levels of hierarchy should be displayed on the graphical user interface (GUI) at any time, if available.
9. A marker gene or bookmark concept allows the user to set mark points for navigational purposes.
10. A list of recently visited menus serves to cache navigational paths for reuse.
11. A global find mechanism allows the user to enter a phrase and search the tree from the root or present location and find all matches for the phrase up to a prespecified depth. The path, which includes the phrase, if matched, is returned.
12. Entered phrases (i.e., including pathnames) can be automatically extrapolated where possible. This "intellisense" feature facilitates keyboard entry. It can also assist with the extrapolation of pathnames to facilitate finding or entering a phrase. Pathname components may be truncated to facilitate presentation.
13. A major problem in populating a tree structure is the amount of typing involved. In view of this, copy, paste, edit, and delete functions are available to copy phrases from one or more menus to another through the use of place-holding markers. Phrase submenus are not copied over because distinct paths tend to invalidate submenu contents in proportion to their depth. Again, new integers are generated for all phrases. Note that the returned list of objects still needs to be manually edited for error and/or omissions. This follows from randomization theory. This maps well to natural language translation.
14. Disjuncts in a menu serve as analogs and superclasses serve as generalizations for an explanation subsystem. In addition, help files and pathnames will also serve for explanative purposes.
15. An "intellassist" feature allows the system to predict the next node in a contextual, antecedent, or consequent tree. Each node in a tree locally stores the address (number) of the node to be visited next in sequence. If a node has not been trained, or if the pointed-to address has been deleted without update, then a text box stating "No Suggestion" pops up, and no navigation is effected if requested. Otherwise, potentially three contiguous menus are brought up on the screen, where the farthest right menu contains the addressed node. Navigation is accomplished by clicking on a "Suggest" button. Otherwise, all navigation is manually performed by default. The user can hop from node to node by using just the suggest button without registering an entry. The use of any form of manual navigation enables a "Remember" button immediately after the next term, if any is entered. Clicking on this enabled button will result in setting the address pointed to by the *previously entered* node to that of the *newly entered* node. The old pointer is thus overwritten. Note that this allows for changing the item selected within the same menu. Note, too, that if a node (e.g., Toyota) is deleted, then all pointers to it may be updated to the parent class (e.g., car menu) and so on up the tree (e.g., vehicle type menu).

A pull-down menu will enable one of two options: (1) Always Remember (by default) and (2) Remember when Told. The Remember button is not displayed under option (1), but the effect under this option is to click it whenever it would have otherwise been enabled. The system always starts at the root node of the relevant tree.

16. It does not make sense to retain a historical prefix for use by the intellassist feature. That is, there is no need to look at where you were to determine where you want to go. While potentially more accurate, this increase in accuracy is more than offset by the extra training time required, the extra space required, and the fact that it will take a relatively long time to reliably retrain the nodes in response to a dynamic domain environment.

AN A* ORDERED SEARCH ALGORITHM

Expert compilers apply knowledge bases to the effective translation of user-specified semantics [11]. The problem with expert compilers is that they use conventional expert systems to realize their knowledge bases. A KASER is advocated because it can amplify a knowledge base by using an inductively extensible representational formalism.

Here, we present a relatively high-level view of the KASER algorithm. We claim that it represents a great advance in the design of intelligent systems by reason of its capability for symbolic learning and qualitative fuzziness:

1. Click on antecedent menus to specify a contextual conjunct. Alternatively, a manual "hot button" will bring up the immediately preceding context for reuse or update. Renormalization is only necessary if a generalization was made—not for term deletion (see below). Iteratively normalize the context (i.e., reduce it to the fewest terms) by using the tree grammar. Note that contextual normalization can be realized in linear time in the number of conjuncts and the depth of search. Here are the reduction rules, which are iteratively applied in any order—allowing for concurrent processing:
 - a. $S \rightarrow A \mid B \mid C \dots$ then replace $A, B, C \dots$ with S just in case all of the RHS is present in the context. This step should be iteratively applied before moving on to the next one.
 - b. $S \rightarrow A \dots$ and $A \rightarrow B \dots$ and $B \rightarrow C \dots$ then if S, A, B, C are all present in the context, then remove A, B, C since they are subsumed by S . It is never necessary to repeat the first step after conclusion of the second.
2. Compute the specific stochastic measure. Note that the specific stochastic measure does not refer to validity—only to the creative novelty relative to the existing rules while retaining validity. For example, given the antecedent grammar: $C5 \rightarrow C3 \mid C4$; $C4 \rightarrow C1 \mid C2$:
 - a. $\{C3 \ C1\} \{\{C3\}, \{C2 \ C3\},\}$ covers and matches the first $\{C3\}$ at level 0. Note that the first covered match, if any, that does not have a covered superset is the one to be fired—a result that follows from the method of transposition.
 - b. $\{C3 \ C1\} \{\{C5\}, \{C2 \ C3\},\}$ matches nothing at the level 0 expansion, so we expand the RHS with the result, $\{C3 \ C1\} \{\{C5 \ (C3 \ C4)\},\}$

- $\{ *C2 *C3 \}$, where the $C2 C3$ are both primitives and $*Ci$ can be matched, but not expanded again. (...) is used to denote disjunction. Here, $\{C5\}$ is matched at level 1. Note that at any level, only one term inside the parentheses (e.g., $C3$) need be covered to get a match of any one disjunct.
- c. $\{C3 C6\} \{ \{C2\}, \{C5 C6\} \}$ matches nothing at the level 0 expansion, so we expand the RHS with the result, $\{C3 C6\} \{ \{ *C2 \}, \{ *C5 (C3 C4), *C6 \} \}$, which matches at level 1 because we matched $(C3 \text{ OR } C4) \text{ AND } C6$. Note that $C6$ was never expanded because it was pre-matched by the existing context. This economy is possible as a result of pre-normalizing the context.
 - d. The result of applying the method of transposition to the above step is $\{ \{C5 C6\}, \{C2\} \}$.
 - e. Each matched {...} fires a consequent, which, if not primitive, matches exactly one row header (i.e., a unique integer) and step (2) iterates.
 - f. Maintain a global sum of the number of levels of expansion for each row for each consequent term. The specific stochastic measure is taken as the maximum of the number of levels of expansion used for each consequent term.
3. Exit the matching process with success (i.e., for a row) or failure based on reaching the primitive levels, a timer-interrupt, a forced interrupt, and/or by using the maximum allocated memory.
 4. If a sequence of consequent actions has been attached, then the sequence is pushed onto a stack in reverse order such that each item on the stack is expanded in a depth-first manner. A parenthesized sequence of actions will make clear the hierarchy. For example, ((Hold Writing Instrument (Hold Pencil with Eraser)) (Press Instrument to Medium (Write Neatly on Paper))). Here, the subclasses are nested. Such a representation also serves explanative purposes. Thus, here one has, Hold Writing Instrument, Press Instrument to Medium, at the general level, and Hold Pencil with Eraser, Write Neatly on Paper, at the specific level. A companion intelligent system could transform the conceptual sequences into smooth natural language (e.g., Pick up a pencil with an eraser and write neatly on a sheet of paper.) Set the general stochastic measure (GSM) to zero. Note that the stochastic measures for each predicate are computed and held in a data structure. The data will be used by the inference engine.
 5. If a match is not found, then since we already have an expanded antecedent {...}, we proceed to expand the context in a breadth-first manner (i.e., if enabled by the level of permitted generalization). Compute the general stochastic measure. Initialize the general stochastic measure to GSM. Note that the general stochastic measure is a measure of validity. Set the starting context to the context.
 - a. A specialized match was sought in step (2), and a generalized match is sought here. Expanding the context can lead to redundancies. For example, $\{ *C1 *C2 *C3 *C4 C1 C2 \}$. Here, the solution is to simply not include any term that is already in the (expanded) context. Stochastic accuracy is thus preserved. Any method that does not preserve stochastic accuracy is not to be used.

- b. $\{C5\} \{\{ *C3\}, \{ *C2 *C3\},\}$ failed to be matched in step (2), so a level 1 expansion of the context is taken:
 $\{ *C5 C3 C4\} \{\{ *C3\}, \{ *C2 *C3\},\}$ where C3 is matched at level 1.
- c. $\{C5 C6\} \{\{ *C1\}, \{ *C2 *C3\},\}$ matches nothing at level 0, so a level 1 expansion of the context is taken:
 $\{ *C5 C3 C4 *C6\} \{\{ *C1\}, \{ *C2 *C3\},\}$ matches nothing at level 1, so a level 2 expansion of the context is taken:
 $\{ *C5 *C3 *C4 C1 C2 *C6\} \{\{ *C1\}, \{ *C2 *C3\},\}$ matches C1 OR C2 AND C3 at level 2. The first covered set is the one to be fired (i.e., even though both sets are covered), since it does not have a covered superset. Next, the method of transposition is trivially executed with no resulting change in the logical ordering.
- d. Each matched $\{...\}$ fires a consequent, which, if not primitive, matches exactly one row header (i.e., a unique integer) and step (2) iterates.
- e. One should maintain a count of the maximum number of levels of expansion for the context below the initial level. The general stochastic measure is defined by GSM plus the maximum number of levels that the context minimally needs to be expanded to get the "first" (i.e., method of transposition) match. This stochastic is represented by the maximum depth for any expansion.
- f. If the context fails to be matched, then generalize each term in the starting context one level up in the tree. Remove any redundancies from the resulting generalization. If the generalized context differs from the starting context, then add one to GSM and go to step (5). Otherwise, go to step (6). For example, the starting context $\{C2 C3\}$ is generalized to yield $\{C4 C5\}$. If this now covers a $\{..\}$, then the general stochastic measure is one. Otherwise, it is subsequently expanded to yield $\{ *C4 C1 C2 *C5 C3 C4\}$ at the first level. Notice that the second C4 has a longer derivation, is redundant, and would never have been added here. Note, too, that C4 is also a sibling or analog node. If this now covers a $\{..\}$, then the GSM remains one, but the specific stochastic measure is incremented by one to reflect the additional level of specialization.
- For another example, Toyota and Ford are instances of the class car. If Toyota is generalized to obtain car, which is subsequently instantiated to obtain Ford (i.e., an analog), then the general and specific stochastic measures would both be one. The general stochastic measure represents the number of levels of expansion for a term in one direction, and the specific stochastic measure represents the number of levels of expansion from this extrema in the opposite direction needed to get a match. The final general (specific) stochastic is taken as the maximum general (specific) stochastic over all terms.
- g. Conflict resolution cannot be a deterministic process as is the case with conventional expert systems. This is because the number of predicates in any match must be balanced against the degree of specialization and/or generalization needed to obtain a match. Thus, a heuristic approach is required. The agenda mechanism will order the rules by their size, general stochastic, and specific stochastic with recommended weights of 3, 2, and 1 respectively.

6. Exit the matching process with success (i.e., for the entire current context for a row) or failure based on reaching the primitive levels, a timer-interrupt, a forced interrupt, and/or by using the maximum allocated memory. Note that a memory or primitive interrupt will invoke step (5f). This enables a creative search until a solution is found or a timer-interrupt occurs. Note, too, that it is perfectly permissible to have a concept appear more than once for reasons of economy of reference, or to minimize the stochastic measures (i.e., provide confirming feedback). The stochastic measures also reflect the rapidity with which a concept can be retrieved.
7. Knowledge acquisition:
 - a. Note that new rules are added at the head.
 - b. If exit occurs with failure, or the user deems a selected consequent (e.g., in a sequence of consequents) in error (i.e., trace mode on), then the user navigates the consequent menus to select an attached consequent sequence, which is appropriate for the currently normalized context.
 - c. If the user deems that the selected "primitive" consequent at this point needs to be rendered more specific, then a new row is opened in the grammar, and the user navigates the consequent menus to select an attached consequent sequence.
 - d. A consequent sequence can pose a question, which serves to direct the user to enter a more specific context for the next iteration (i.e., conversational learning). Questions should usually only add to the context to prevent the possibility of add/delete cycles.
 - e. Ask the user to eliminate as many specific terms (more general terms will tend to match more future contexts) from the context as possible (i.e., and still properly fire the selected consequent sequence given the assumptions implied by the current row). A context usually consists of a conjunct of terms. This tends to delimit the generality of each term as it contributes to the firing of the consequent. However, once those antecedent terms become fewer in number for use in a subsequent row, then it becomes possible to generalize them while retaining validity. The advantage of generalization is that it greatly increases reusability. Thus, we need to afford the user the capability to substitute a superclass for one or more terms. Note that this implies that perfectly valid rules that were entered can be replayed with specific (not general) stochastics greater than zero. This is proper, since the specific stochastic preserves validity in theory. Thus, the user may opt to generalize one or more contextual terms by backtracking their derivational paths. If and only if this is the case, step (1) is applied to normalize the result. An undo/redo capability is provided. Validated rule firings are only saved in the rule base if the associated generalization stochastic is greater than zero. The underlying assumption is that rule instances are valid. If a pure rule instance proves to be incorrect, then the incorrect rule needs to be updated or purged, and the relevant object class menu(s) may be in need of repair. For example, what is the minimal context to take FIX_CAR to FIX_TIRE? A companion intelligent system could learn to eliminate and otherwise generalize specific terms (e.g., randomization theory).

- f. The system should verify for the user all the other {...} in the current row that would fire or be fired by the possibly over-generalized {...} if matched. (Note that this could lead to a sequence of UNDOs.)

For example, {{C5} A2} {{{C5} A1} {{C5 C6} A2} {{C5 C7} A2, A3}} informs the user that if the new C5 acquisition is made, then A2 and not A1 is proper to fire. If correct, then the result is {{C5} A2 {C5 C7} A2, A3}. {C5 C6} A2 has been eliminated because it is redundant. Also, {C5 C7} A2, A3 is fired just in case C5 AND C7 are true—in which case, it represents the most specific selection since it is a superset of the first set. If the elimination of one or more specific terms causes one or more {...} to become proper supersets, then warning message(s) may be issued to enable the user to reflect on the proposed change(s). If the elimination and/or generalization of one or more specific terms enables the firing of another rule in the same row in preference to the generalized rule, then the generalization is rejected as being too general. Note that there is no need to normalize the results, as they would remain in normal form. Also, any further normalization would neutralize any necessary speedup.

- g. A selected consequent number may not have appeared on the trace path with respect to the expansion of each consequent element taken individually. Checking here prevents cycle formation.
- h. It should never be necessary to delete the least frequently used (LFU) consequent {...} in view of reuse, domain specificity, processor speed, and available memory relative to processor speed. Nevertheless, should memory space become a premium, then a hierarchy of caches should be used to avoid deletions.
8. A metaphorical explanation subsystem can use the antecedent/consequent trees to provide analogs and generalizations for explanative purposes. The antecedent/consequent paths (e.g., ROOT, FIX_CAR, FIX_TIRE, etc.) serve to explain the recommended action in a way similar to the use of the antecedent and consequent menus. The antecedent/consequent menus will provide disjunction and "user-help" to explain any level of action on the path. Note that the system inherently performs a fuzzy logic known as computing with words [4] (i.e., based on the use of conjuncts, descriptive phrases, and tree structures). The virtual rule base is exponentially larger than the real one and only limited by the number of levels in the trees, as well as by space-time limitations on breadth-first search imposed by the hardware.
9. A consequent element could be a "do-nothing" element if need be (i.e., a Stop Expansion). The provision for a sequence of consequents balances the provision for multiple antecedents. The selected consequent(s) need to be as general class objects as can be to maximize the number of levels and, thus, the potential for reuse at each level. The consequent grammar is polymorphic since many such grammars can act (in parallel via the Internet) on a single context with distinct, although complementary results. Results can be fused as in a multi-level, multicategory associative memory. Multiple context-matched rules may not be expanded in parallel because there can be no way to ascribe *probabilities* to partially order the competing rules and

because any advantage would be lost to an exponential number of context-induced firings. The consequent {...}s cannot be ranked by the number of matching terms (i.e., for firing the most specific first) because the most specific terms are generally incomparable. However, a covered superset is always more specific than any of its proper subsets. Thus, the first covered set that does not have a covered superset in the same row is the one to be fired. If it does have a covered superset, then the superset is fired only if it is the next covered one to be tested in order. It is not appropriate to tag nodes with their level, use a monotonically increasing numbering system, or any equivalent mechanism to prevent the unnecessary breadth-first expansion of a node(s) because the menus are dynamic, and it would be prohibitively costly to renumber, for example, a terabyte of memory. Note that node traversal here is not synonymous with node visitation. Even if parallel processors could render the update operation tractable, the search limit would necessarily be set to the depth of the deepest unmatched node. Here, the likelihood of speedup decreases with scale. The contextual terms should only be *'d if this does not interfere with their expansion—even if normalized. Let the context be given as {C5 C6} and the RHS be {C5 C7}, {C1 ...}. Clearly, if the context had *C5, then the C1 might never be matched.

10. Unlike the case for conventional expert systems, a KASER cannot be used to backtrack consequents (i.e., goal states) to find multiple candidate antecedents (i.e., start states). The problem is that the preimage of a typical goal state cannot be effectively constrained (i.e., other than for the case where the general and specific stochastics are both zero) in as much as the system is qualitatively fuzzy. Our answer is to use fuzzy programming in the forward-chained solution. This best allows the user to enter the constraint knowledge that he/she has into the search. For example, if the antecedent menus are used to specify CAR and FUEL for the context and the consequent is left unconstrained for the moment, then the system will search through all instances, if any, of CAR crossed with all instances of FUEL (i.e., to some limiting depth) to yield a list of fully expanded consequents. Generalization-induced system queries, or consequents that pose questions, if any, will need to be answered to enable this process to proceed. Thus, in view of the large number of contexts that are likely to be generated, all interactive learning mechanisms should be disabled or bypassed whenever fuzzy programming is used. Note that CAR and FUEL are themselves included in the search. Each predicate can also be instantiated as the empty predicate in the case of the antecedent menus, if user-enabled. If the only match occurs for the case of zero conjuncts, then the consequent tree is necessarily empty. A method for fuzzy programming is to simply allow the user to split each conjunct into a set of disjuncts and expand all combinations of these to some fixed depth to obtain a list of contexts. This use of a keyword filter, described below, is optional. For example, the specification $(A \vee A' \vee !A'') \wedge (B \vee B') \wedge (C)$ yields 23 candidate contexts—including the empty predicate (i.e., if one assumes that A'' is primitive and allows for redundancy), which excludes the empty context. The exclamation mark, "!", directs the system to expand the nonterminal that follows it to include (i.e., in addition to itself) all of the next-level instances of its class. For example, !CAR would yield (CAR TOYOTA FORD MAZDA HONDA ... λ). Here, lambda denotes the empty predicate and is included as a user option.

A capability for expanding to two or more levels if possible (e.g., "!!") is deemed to be nonessential but permissible (e.g., for use with relatively few conjuncts). This follows because the combinatorics grow exponentially. One can always take the most successful context(s) produced by a previous trial, expand predicates to another level by using "!"s where desired, and rerun the system. Note that, in this manner, the user can insert knowledge at each stage—allowing for a far more informed, and thus, deeper search than would otherwise be possible. Moreover, the fuzzy specialization engine will stochastically rank the generalized searches to enable an accurate selection among contexts for possible rerun.

The search may be manually terminated by a user interrupt at any time. The search is not to be automatically terminated subsequent to the production of some limit of contexts because to do so would leave a necessarily skewed distribution of contexts—thereby giving the user a false sense of completeness. We would rather have the user enter a manual interrupt and modify the query subsequently. A terminated search means that the user either needs to use a faster computer, or more likely, just narrow down the search space further and resubmit. For example, if we have the antecedent class definitions:

```
(CAR (FORD TOYOTA)) (FUEL (REGULAR_GAS HIGH_TEST
DIESEL)) (AGE (OLD (TIRES ...)) (NEW (TIRES ...)))
```

and the contextual specification:

```
(!CAR) ^ (!FUEL) ^ (NEW),
```

then we would have the following 35 contexts allowing for the empty predicate. Note that the use of the empty predicate is excluded by default, since its use is associated with an increase in the size of the search space and since it may not be used with the consequent menus (see below).

```
CAR
DIESEL
FORD
FUEL
HIGH_TEST
NEW
REGULAR_GAS
TOYOTA
CAR DIESEL
CAR FUEL
CAR HIGH_TEST
CAR NEW
CAR REGULAR_GAS
DIESEL NEW
FORD NEW
FUEL NEW
HIGH_TEST NEW
REGULAR_GAS NEW
TOYOTA FUEL
TOYOTA DIESEL
TOYOTA HIGH_TEST
TOYOTA NEW
TOYOTA REGULAR_GAS
CAR DIESEL NEW
```

```

CAR FUEL NEW
CAR HIGH_TEST NEW
CAR REGULAR_GAS NEW
FORD DIESEL NEW
FORD FUEL NEW
FORD HIGH_TEST NEW
FORD REGULAR_GAS NEW
TOYOTA DIESEL NEW
TOYOTA FUEL NEW
TOYOTA HIGH_TEST NEW
TOYOTA REGULAR_GAS NEW

```

The user may also have used the consequent menus to specify an optional conjunctive list of key phrases, which must be contained in any generated consequent. Those generated consequents, which contain the appropriate keywords or phrases, are presented to the user in rank order—sorted first in order of increasing generalization stochastic and within each level of generalization stochastic in order of increasing specialization stochastic (i.e., best-first). For example, (general, specific) (0, 0) (0, 1) (1, 0) (1, 1) ... Recall that only the specific stochastic preserves validity.

The specified antecedent and consequent classes should be as specific as possible to minimize the search space. Neither the antecedent nor consequent terms specified by the user are ever generalized. For example, if we have the consequent class definitions:

```

(COST_PER_MILE (CHEAP MODERATE EXPENSIVE))
(MPG (LOW MEDIUM HIGH))

```

then we can constrain the space of generated consequents in a manner similar to the way in which we constrained the space of generated antecedents. Thus, for example we can write:

```

(!CAR) ^ (!FUEL) ^ (NEW) -> (!COST_PER_MILE) ^ (!MPG)

```

This is orthogonal programming; that is, reusing previous paradigms unless there is good reason not to reuse them. Each candidate solution has been constrained so that it must contain at least one phrase from the four in the COST_PER_MILE class *and* at least one phrase from the four in the MPG class—including the class name, but excluding the empty predicate of course. IF an asterisk, "*" is placed after the arrow, then the compiler is directed not to filter the produced consequents in any way.

The user can make changes wherever (i.e., to the antecedents, the consequents, or both) and whenever (e.g., interactively) appropriate and rerun the system query. This represents *computing with words* because fuzziness occurs at the qualitative level. It is not really possible for distinct classes to produce syntactically identical phrases because pathnames are captured using unique identifiers. That is, the identifiers are always unique even if the represented syntax is not.

It is not necessary to weight the consequent phrases because instance classes preserve validity (i.e., at least in theory) and because it would be otherwise impossible to ascribe weights to combinations of words or phrases. For example, "greased" and "lightning" might be synonymous with fast, but taken together (i.e., "greased lightning"), an appropriate weight should be considerably greater than the sum of the partial weights. The degree to which the conjunctive weight should be increased does not lend itself to practical determination. Moreover, one is then faced with the indeterminable question (i.e., for ranking) as to which is

the more significant metric: the weight or the two stochastics. Besides, if one follows the dictates of quantum mechanics or veristic computing, it suffices to rank consequent phrases by group as opposed to individually.

Feedback produced, in the form of implausible generalizations, serves to direct the knowledge engineer to modify the involved declarative class structures by regrouping them into new subclasses so as to prevent the formation of the erroneous generalizations. This, too, is how the system learns. The iterative pseudocode for accomplishing the combinatorial expansion follows.

1. Initialize the list of Candidate Contexts to λ .
2. Each conjunct—e.g., $(A \vee A' \vee !A'')$ —in the starting list—e.g., $(A \vee A' \vee !A'') \wedge (B \vee B') \wedge (C)$ will be processed sequentially.
3. Note that $!A''$ means to expand the disjunct to include all members of its immediate subclass, if any. Similarly, $!!A''$ means to expand the disjunct to a depth of two. The provision for multilevel expansion is implementation-dependent and is thus optional. Each expanded conjunct is to be augmented with exactly one λ if and only if the user has enabled the λ -option. This option is disabled by default.
4. Expand the first conjunct while polling for a manual interrupt. Here, the result is
 $(A \vee A' \vee A'' \vee A''.a \vee A''.b \vee \lambda)$.
5. Note that the fully expanded list of conjuncts for illustrative purposes appears:
 $(A \vee A' \vee A'' \vee A''.a \vee A''.b \vee \lambda) \wedge$
 $(B \vee B' \vee \lambda) \wedge (C \vee \lambda)$
6. Initialize a buffer with the disjuncts in the first conjunct. Here, the first six buffer rows are populated.
7. Copy the contents of the buffer to the top of the list of Candidate Contexts;
8. Current Conjunct = 2;
9. Note that there are three conjuncts in this example.
10. WHILE (Current Conjunct <= Number of Conjuncts) and NOT Interrupt DO
 {
11. Expand the Current Conjunct while polling for a manual interrupt.
12. Let d = the number of disjuncts in the Current Conjunct;
13. Using a second buffer, duplicate the disjuncts already in the first buffer d times. For example, here, the second conjunct has three disjuncts and would thus result in the buffer: $A, A, A, A', A', A', A'', \dots, \lambda, \lambda, \lambda$.
14. FOR each element i in the buffer WHILE NOT Interrupt DO
15. FOR each Disjunct j in the Current Conjunct WHILE NOT Interrupt DO
 {
16. Buffer $[i] =$ Buffer $[i] \parallel$ Current Disjunct $[j]$.
17. (For example, $AB, AB', A\lambda, A'B, A'B', A'\lambda, \dots, \lambda B, \lambda B', \lambda\lambda$.)
- }
18. IF the λ -option has been enabled THEN
 Append the contents of the buffer to the bottom of the list of Candidate Contexts while polling for a manual interrupt.

19. Current Conjunct++
}
20. An interrupt may be safely ignored for the next two steps.
21. IF the λ -option has been enabled THEN
 Final Contexts = Candidate Contexts - λ
22. ELSE
 Final Contexts = contents of the buffer.
23. Duplicate contexts are possible due to the use of λ and possible duplicate entries by the user. Searching to remove duplicate rows is an $O(n^2)$ process. Thus, it should never be mandated, but rather offered as an interruptible user-enabled option.

The iterative pseudocode for constraining the generated consequents follows.

1. Expand each conjunct—e.g., $(A \vee A' \vee !A'')$ —in the starting list—e.g., $(A \vee A' \vee !A'') \wedge (B \vee B') \wedge (C)$. Note that the λ -option is disabled.
 2. Here, the result is
 $(A \vee A' \vee A'' \vee A'' . a \vee A'' . b) \wedge (B \vee B') \wedge (C)$.
 3. FOR each consequent sequence (i.e., rule) WHILE NOT Interrupt DO
 {
 4. match = FALSE;
 5. FOR each expanded conjunct (i.e., required key concept) WHILE NOT Interrupt DO
 {
 6. FOR each predicate in an expanded conjunct (i.e., PEC) WHILE NOT Interrupt DO
 {
 7. FOR each predicate in a consequent sequence (i.e., PICS) WHILE NOT Interrupt DO
 {
 8. IF PEC = PICS THEN
 {
 - match = TRUE;
 - BREAK;
 - BREAK;
 - (Each BREAK transfers control to the next statement outside of the current loop.)
9. IF NOT match THEN BREAK
 }
10. IF NOT match THEN remove current rule from the candidate list
11. ELSE the rule is saved to the set of candidate rules, which is sorted as previously described.
 }

SUGGESTED NAVAL APPLICATIONS

Figure 2 presents a screen capture of a Type I KASER for diagnosing faults in a jet engine. Observe that the general and specific stochastics are both one. This means, in the case of the general stochastic, that the KASER needed to use a maximum of one level of inductive inference to arrive at the prescribed action. Similarly, the specific stochastic indicates that a maximum of one level of deduction was necessarily employed to arrive at this prescribed action. Contemporary expert systems would not have been able to make a diagnosis and prescribe a course of action, since they need to be explicitly programmed with the necessary details. In other words, the KASER is offering a suggestion here that is open under deductive process. Simply put, it created new and presumably correct knowledge. Here are the two level-0 rules, supplied by the knowledge engineer (i.e., *R15* and *R16*), that were used in conjunction with the declarative object trees to arrive at the new knowledge, *R18*:

R15: If Exhaust Flaming and Sound Low-Pitched Then Check Fuel Injector for Carbonization

R16: If Exhaust Smokey and Sound High-Pitched Then Check Fuel Pump for Failure

R17: If Exhaust Smokey and Sound Low-Pitched Then Check Fuel Pump for Failure

Upon confirmation of *R17*, *R16* and *R17* are unified as follows.

R18: If Exhaust Smokey and Sound Not Normal Then Check Fuel Pump for Failure

The KASER finds declarative antecedent knowledge, which informs the system that the three sounds that an engine might make, subject to dynamic modification, are high-pitched, low-pitched, and normal. By generalizing high-pitched sounds one level to SOUNDS (see Figure 3) and then specializing it one level, one arrives at the first-level analogy: low-pitched sounds. This analogy enables the user context to be matched and leads to the creation of new knowledge. Figure 4 depicts the consequent tree and is similar to the antecedent tree shown in Figure 3. The consequent tree is used to generalize rule consequents so as to maximize reusability. Object reuse may simultaneously occur at many levels, even though this example depicts only one level for the sake of clarity. There are many more algorithms, settings, and screens that may be detailed.

Another application is the automatic classification of radar signatures. Basically, the radar data are assigned a feature set in consultation with an expert. Next, a commercial data-mining tool is applied to the resulting very large database to yield a set of rules and associated statistics. These rules are manually fed into the Type I KASER, which interacts with the knowledge engineer to create the antecedent and consequent trees, as well as a fully generalized rule base and miscellaneous sundry. Upon completion of the manual acquisition, the KASER is given a procedure

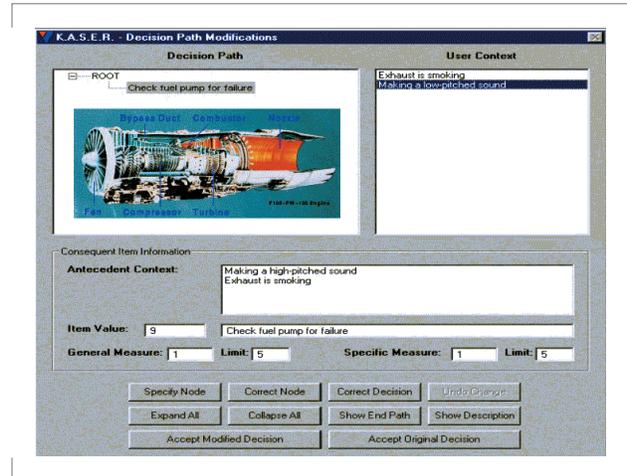


FIGURE 2. Screen capture of an operational Type I kaser.

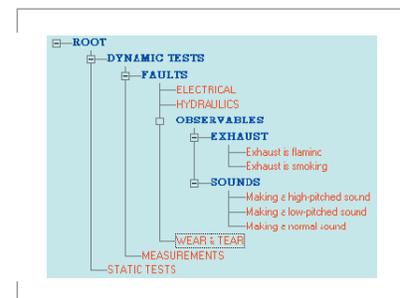


FIGURE 3. Screen capture of an antecedent tree.

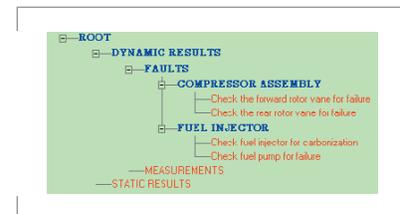


FIGURE 4. Screen capture of a consequent tree.

to link it through open database connectivity (ODBC) to an external electronic intelligence (ELINT) database. This database supplies the radar signatures in approximately real time. The signatures are then automatically classified by the KASER's virtual rule space and the generated stochastics provide an indication of reliability. The KASER, having a virtual rule space >> real rule space can produce erroneous advice if the general stochastic is greater than zero. In this event, the user is requested to supply a corrective consequent(s), which may be "radioed" to the base computer for subsequent update on a daily basis, followed by uploading the more learned KASER. The main benefit here is that the KASER can supply solutions to complex signature-identification problems that would not be cost-effective to supply otherwise (see Figure 1). A Type II KASER should be able to automatically acquire the feature set.

CONCLUSIONS

This project seeks to demonstrate (1) a strong capability for symbolic learning, (2) an accelerating capability to learn, (3) conversational learning (i.e., learning by asking appropriate questions), (4) a metaphorical explanation subsystem, (5) probabilistically ranked alternative courses of action that can be fused to arrive at a consensus that is less sensitive to occasional errors in training, and (6) a capability to enunciate responses. It is argued that the intelligent components of any Command Center of the Future (CCOF) cannot be realized in the absence of a strong capability for symbolic learning.

Randomization theory holds that the human should supply novel knowledge exactly once (i.e., random input), and the machine should extend that knowledge by way of capitalizing on domain symmetries (i.e., expert compilation). In the limit, novel knowledge can only be furnished by chance itself. This means that, in the future, programming will become more creative and less detailed, and thus, the cost per line of code will rapidly decrease. According to Bob Manning [12]: "Processing knowledge is abstract and dynamic. As future knowledge management applications attempt to mimic the human decision-making process, a language is needed that can provide developers with the tools to achieve these goals. LISP enables programmers to provide a level of intelligence to knowledge-management applications, thus enabling ongoing learning and adaptation similar to the actual thought patterns of the human mind."

Moreover, according to Erann Gat at the Jet Propulsion Laboratory, California Institute of Technology, working under a contract with the National Aeronautics and Space Administration [13]: "Prechelt concluded that 'as of JDK 1.2, Java programs are typically much slower than programs written in C or C++. They also consume much more memory.' "

Gat states that "We repeated Prechelt's study by using Franz Inc.'s Allegro Common LISP 4.3 as the implementation language. Our results show that LISP's performance is comparable to or better than C++ in execution speed; it also has significantly lower variability, which translates into reduced project risk. The runtime performance of the LISP programs in the aggregate was substantially better than C and C++ (and vastly better than Java). The mean runtime was 41 seconds versus 165 for C and C++. Furthermore, development time is significantly lower and less variable than either C++ or Java. This last item is particularly significant because it translates directly into reduced risk for software development.

Memory consumption is comparable to Java. LISP thus presents a viable alternative to Java for dynamic applications where performance is important."

In conclusion, the solution to the software bottleneck will be cracking the knowledge-acquisition bottleneck in expert systems (compilers).

ACKNOWLEDGMENTS

I would like to thank Robert Rush, Jr., and James Boerke for their technical programming support in the implementation of the Type I KASER. The Office of Naval Research sponsored this In-house Laboratory Independent Research project.

REFERENCES

1. Chaitin, G. J. 1975. "Randomness and Mathematical Proof," *Scientific American*, vol. 232, no. 5, pp. 47-52.
2. Uspenskii, V. A. 1987. *Gödel's Incompleteness Theorem*, translated from Russian. Ves Mir Publishers, Moscow, Russia.
3. Lin, J-H. and J. S. Vitter. 1991. "Complexity Results on Learning by Neural Nets," *Machine Learning*, vol. 6, no. 3, pp. 211-230.
4. Rubin, S. H. 1999. "Computing with Words," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 29, no. 4, pp. 518-524.
5. Feigenbaum, E. A. and P. McCorduck. 1983. *The Fifth Generation*, Addison-Wesley Publishing Company, Reading, MA.
6. Minsky, M. 1987. *The Society of Mind*. Simon and Schuster, Inc., New York, NY.
7. Clark, C. T. 2000. "An Interview with Marvin Minsky," *Knowledge Management*, (June) pp. 26-28.
8. Zadeh, L. A. 1999. "From Computing with Numbers to Computing with Words--From Manipulation of Measurements to Manipulation of Perceptions," *IEEE Transactions on Circuits and Systems*, vol. 45, no. 1, pp. 105-119.
9. Rubin, S. H. 1999. "The Role of Computational Intelligence in the New Millennium," Plenary Speech, *Proceedings of the 3rd World Multiconference on Systemics, Cybernetics, and Informatics (SCI '99) and 5th International Conference on Information Systems Analysis and Synthesis (ISAS '99)*, pp. 3-13.
10. Rubin, S. H. 1998. "A Fuzzy Approach Towards Inferential Data Mining," *Computers and Industrial Engineering*, vol. 35, nos. 1-2, pp. 267-270.
11. Hindin, J. 1986. "Intelligent Tools Automate High-Level Language Programming," *Computer Design*, vol. 25, pp. 45-56.
12. Manning, B. 2000. "Smarter Knowledge Management Applications: LISP," *PC AI*, vol. 14, no. 4, pp. 28-31.
13. Gat, E. 2000. "LISP as an Alternative to Java," *Intelligence* (winter), pp. 21-24.



Stuart H. Rubin

Ph.D. in Computer and Information Science, Lehigh University, 1988

Current Research: Intelligent systems; knowledge management.

Establishing a Data-Mining Environment for Wartime Event Prediction with an Object-Oriented Command and Control Database

Marion G. Ceruti

SSC San Diego

S. Joe McCarthy

Space and Naval Warfare Systems Command

INTRODUCTION

The ability to predict attacks and other hostile events during times of conflict is important to military commanders from the standpoint of readiness. The more advanced the notice and the more widespread the notification, the better able all echelons are to respond to threats efficiently and with the correct combination of forces.

The literature is replete with recent research results on data mining and data classification. (See, for example, [1, 2, 3, and 4].) Data mining, data classification, and data correlation are related to data fusion. As these techniques mature, better tools become available to model and to correlate data from complex operational scenarios. The purpose of this research is to create and extend a method to predict attacks on the U.S. Marine Corps using an object-oriented command and control database and data-mining techniques [5].

Data Mining

Data mining is the search for and extraction of hidden and useful patterns, structures, and trends in large, multidimensional, and heterogeneous data sets that were collected originally for another purpose. (See, for example, [4].) Data mining is an art that is supported by a considerable body of science, engineering, and technology. For example, data mining uses techniques from such diverse areas as data management, statistics, artificial intelligence, machine learning, pattern recognition, data visualization, and parallel and distributed computing. Data mining is possible today because of advances in these many fields; however, this multidisciplinary characteristic also makes data mining a difficult subject to teach and learn. Whereas the Structured Query Language (SQL) is inadequate to answer many complex queries, data mining can support searches for patterns in temporal and spatial databases in a more efficient manner. Data mining is important to the military because commanders and the analysts who support them cannot anticipate all future uses of information at the time of data collection.

Limitations of Data Mining

Whereas the goal of data mining is to identify hidden patterns, the search algorithms chosen for the particular task may miss an important and interesting pattern or even a class of similar patterns. A systematic method to preclude this problem is not available.

ABSTRACT

This paper documents progress to date on a research project, the goal of which is wartime event prediction. The paper describes the operational concept, the data-mining environment, and the data-mining techniques that use Bayesian networks for classification. Key steps in the research plan are (1) implement machine learning, (2) test the trained networks, and (3) use the technique to support a battlefield commander by predicting enemy attacks. Data for training and testing the technique can be extracted from the object-oriented database that supports the Integrated Marine Multi-Agent Command and Control System (IMMACCS). The class structure in the IMMACCS data model is especially well suited to support attack classification.

Similarly, there is no guarantee that any given data-mining effort will yield something new and useful, regardless of how many well-designed data-mining tools are used. This is because the data may not contain the desired patterns. Data mining is a search for observational data and the relationships between them, rather than the measurement of experimental data.

CONCEPT OF OPERATIONS

The concept of operations for a future system based on this research is (1) to use data-mining and data-classification algorithms to detect patterns associated with attacks (e.g., to identify factors that indicate an imminent attack) and (2) to correlate these patterns with current events with a view toward supplying military commanders with a prediction of the next attack and a confidence level that pertains to that prediction. A considerable amount of data associated with events that have preceded known attacks is required to model attacks, to search for common features, and to find these patterns in new data.

Success in this effort depends on a characterization of the circumstances that translate to well-defined observables that preceded past attacks. The more detailed the available knowledge, the better the resulting model, and the greater the probability that data instantiating critical variables can be collected. We expect that such detailed data for all variables will not be available prior to future attacks and that all available data may not be useful in predicting attacks (i.e., will function as "noise" in the analysis). Thus, the task involves identification of algorithms that can detect pre-attack features in clutter and the use of pattern recognition. Modern methods of statistical pattern recognition are sufficiently computationally oriented to use a larger dimensional space and are less sensitive to noise than older methods. Success in attack prediction will depend, at least in part, on how well these methods can be implemented with the available data.

GENERAL APPROACH

Hostile events can be characterized with respect to as many relevant variables as are deemed necessary and available to predict future attacks. An object-oriented message-traffic database can be analyzed for the occurrence of telltale signs of pending attacks. Our objective is to generate an event prediction (in terms of a probability) with a confidence value associated with it. Therefore, it is necessary to determine the combinations of events and observations that will have a higher probability of indicating a future attack. A baseline can be modeled from normal operational scenarios and from military events during times of conflict that do not constitute attacks per se.

The attack alarm-generation process and the reduction of false positives can be approached using constraints from models of known attacks. The identification of the appropriate features (and groups of features) that can flag imminent attacks is the most challenging part of the process. One approach is to explore the generation of a knowledge base encoded in Bayesian networks.

A literature search was conducted for publications on various subjects that relate to data mining, including algorithms and their applications. Data-mining algorithms can be used to identify complex patterns in the

data that correlate well to hostile events. Criteria can be developed for sufficient correlation and confidence levels in data associations. For example, one metric that could be used is correlation strength, which is the ratio of the joint probability to the individual probability of observing a pattern [1].

BAYESIAN NETWORKS

Bayesian networks can be used to classify data into categories. Bayesian networks are:

- probabilistic networks,
- directed acyclic graphs that encode certain dependences between nodes that represent random variables,
- knowledge bases with knowledge in the network's structure and in its conditional probability table, and
- structures that can be used to infer causality.

Naive Bayesian Networks

A naive Bayesian network is a very simple structure in which all random variables representing observable data have a single, common parent node—the class variable. The naive Bayesian classifier has been used extensively for classification because of its simplicity, and because it embodies the strong independence assumption that, given the value of the class, the attributes are independent of each other.

Naive Bayesian networks work remarkably well considering that this independence assumption may not be valid from a logical standpoint. The performance of a naive Bayesian network can be improved with the addition of trees that provide augmenting edges to a naive Bayesian network by representing correlations between the attributes.

Tree Augmented Naive (TAN) Bayesian Classification Algorithm

SSC San Diego has access to SRI International's classifier algorithms developed under the Defense Advanced Research Projects Agency's High Performance Knowledge Base Program. For example, SRI's Tree Augmented Naive (TAN) Bayesian Classification Algorithm is a classifier algorithm based on Bayesian networks with the advantages of robustness and polynomial computational complexity [2 and 3].

Bayesian networks have some drawbacks that SRI has addressed in the TAN algorithm. In ordinary naive Bayesian networks, the variables (data) are assumed to be conditionally independent given the class. Logically, this is not always true. For example, suppose enemy troops are observed at location X and enemy tanks are observed at location Y. When using naive Bayesian networks, one assumes that these events are independent. However, both events may be part of the overall enemy battle plan. In the TAN algorithm, the trees provide edges that represent correlation between the variables.

Bayesian networks, especially with tree augmentation, are a suitable technology for data-mining classification and event prediction for the following reasons:

- First, one need not provide all joint probability values to specify a probability distribution for collections of independent variables [6].
- Second, one could mix modeling (e.g., explicit knowledge engineering for knowledge elicited from experts) with statistical data induction and

adaptivity. This mix would require fewer data values to induce better quality models.

- Third, one could use these models to compute the value of information. For example, having seen signs "A" and "B" of an imminent attack, what is the best information to collect next to confirm that hypothesis?
- Fourth, one could characterize explicitly the kinds of attacks. For example, given an attack of type "air attack," what are the most likely signals? These signals could be collected regularly to fill the database used as input into the TAN algorithm.

The TAN algorithm makes some tradeoffs between accuracy and computation. It approximates a probability distribution using some constraints on the complexity of the representation; however, it is extremely fast (low polynomial), efficient (one pass over the data), and robust (low-order statistics).

The TAN algorithm accepts data sets as input and induces Bayesian networks as output. Specifically, the TAN algorithm is intended to be used as a classification algorithm, which means that the input would be a file with tuples of the form $\{x_1, x_2, x_3, \dots, x_n, c\}$ where the x_i are values that variable X_i takes and c is the value that a class (C) variable can take. To set the range of each variable, the TAN algorithm needs an auxiliary file that contains a description of each variable, including the range of values representing the degree of intensity.

The TAN algorithm's output is a Bayesian network encoding of $P(C, X_n, \dots, X_1)$ in an efficient manner. To use TAN as a classifier, one simply computes $P(C|x'_n, \dots, x'_1)$. Given a new vector X'_n, \dots, X'_1 and having a probability distribution over c , one can select the event with highest probability as the one to classify. To compute the confidence in this value, the bootstrap method can be used [7].

The TAN algorithm outperforms naive Bayesian networks while maintaining its robustness and computational simplicity (polynomial vs. exponential complexity).

The TAN algorithm captures the best of both discrete and continuous attributes. Therefore, the TAN algorithm achieves classification performance that is at least as good as, and in some cases better than, models that use purely discrete or purely continuous variables. Studies at SRI have demonstrated that the TAN algorithm performs competitively with other state-of-the-art methods.

TAN, and similar algorithms, can be made to perform the classification of certain battlefield situations for the Marine Corps. Much work needs to be done in this area, particularly with regard to data-set selection, data cleansing, and the refinement of the algorithm to meet specific needs.

In addition to the TAN algorithm, SRI has more general algorithms for inducing Bayesian networks that do not make the compromises that the TAN algorithm does. These algorithms try to fit the best distribution possible with no constraints. The disadvantage is that the computation of these models is slower; however, this may be acceptable and desirable in some cases. Algorithms can be implemented with the same data and the results compared.

GaussMeasurePredict Program

The GaussMeasurePredict program was developed by Nir Friedman to measure the performance of an induced TAN model. (See, for example, [2]).

The input of GaussMeasurePredict consists of the following items: (1) an induced Naive Bayesian network from TAN, (2) the name of the variable to predict, and (3) a test data set that contains instance information.

When testing the Bayesian network model, the variable to predict is specified and known to be correct. Usually this will be the outcome of the class variable.

GaussMeasurePredict also has the option to calculate and display the probability of each class value for each instance in the input file. This feature is particularly useful for receiver operating characteristic (ROC) curves as well as for determining other statistics [8]. Thus, with this option, GaussMeasurePredict can output the probability distribution for each instance in addition to a summary.

The output of GaussMeasurePredict is a prediction of the accuracy of the network in the TAN Bayesian network .bn file. It can be used to predict the accuracy of other classifier algorithms as long as the output file matches the format of TAN's Bayesian network file.

GaussMeasurePredict is intended to be used to measure the accuracy of predictions and not to generate predictions for unlabeled instances. Unlike the TAN algorithm, GaussMeasurePredict does not accept instances with "?" for missing values in an instance input file. All variables must have filled values in each instance. However, because GaussMeasurePredict compares the induced Bayesian network to the test data set, it also can be used to infer the class of an unknown instance by filling in the class (Outcome) variable with a guessed value. Using the option described above, GaussMeasurePredict can output a predicted class probability for each class value. The class with the highest probability is the predicted class for that instance.

Fortunately, in the simplest case of attack predictions, only two values are possible for the class variable: ATTACK_LIKELY and ATTACK_NOT_LIKELY. In more detailed cases of attack predictions in which specific attack types are listed in the data-definition input file, the class variables may assume $2N$ values where N is the total number of attack types considered in the class. (The $2N$ arises from including the negation of the likelihood of an attack of each type.)

SOFTWARE IMPLEMENTATION AND PLANS

Data-mining software was tested for correct operation with clean data sets designed specifically for testing. The programs described below are included in the research environment. The software includes the TAN algorithm and the GaussMeasurePredict that uses the output of the TAN algorithm. Inputs to GaussMeasurePredict must be complete. Plans include the acquisition of additional algorithms that are designed to operate on incomplete data sets.

TAN 2.1 Availability

The TAN version 2.1 software and user's manual are available for download via file transfer protocol (FTP) from SRI's Web site: <http://edi.erg.sri.com/tan/TANintro.htm>. The user is required to register with a name and password. To obtain the TAN algorithm, Netscape is recommended and may be required. The Solaris CDE Web browser, HotJava, is not recommended to download TAN. The TAN user manual is included with the software (See, for example, [8]).

The TAN software was downloaded from SRI's Web site onto a Solaris SPARC Station 20 computer running the Solaris 2.7 UNIX operating system and using the Common Desktop Environment (CDE).

TAN 2.1 constitutes the main data-mining tool in the research environment of this project. TAN can be used as a base classifier and also as a method to fuse the output of other data-mining and classification algorithms. When algorithms have been tested and programmed, data visualization tools can be identified, tested, and used to view the data and to continue the pattern-recognition process.

GaussMeasurePredict Availability

The GaussMeasurePredict program is available along with the TAN software from SRI's Web site. The program is included with the TAN package and can be executed when files are "unzipped" and when the appropriate input files are available.

OBJECT-ORIENTED DATA IMPLEMENTATION

The object model, on which the Integrated Marine Multi-Agent Command and Control System (IMMACCS) database is based, is a detailed representation of the battlespace with objects derived from the March 1998 Urban Warrior Advanced Warfighting Exercise [9 and 10]. Object attributes and their associations, as well as class inheritance, are also described in [10]. The IMMACCS database uses the Unified Modeling Language symbolic representation method [10].

The IMMACCS database includes in its structure the following topics of interest to the Marine Corps: aircraft; ground vehicles; sea-surface vehicles; weapons and weapon systems; electronic devices of many kinds; terrain; bodies of water; logistics information; transportation infrastructure; various specialized units; personnel data; and most importantly for this application, military events. Class inheritance paths and allowed values are specified [10]. The use of an object-oriented database and the representation of military entities in object form provide a degree of interoperability and extensibility that allows multiple services to use and add to this common tactical picture [9].

The data sets for this data-mining effort will come from IMMACCS. The class structure in the IMMACCS data model is especially well-designed for adaptation to the attack/non-attack classification task. When data fill becomes available, especially for the attributes and object classes of interest, the IMMACCS database will be a very desirable data source for reasons described in the next subsection.

CONSTRUCTION OF TRAINING DATA SETS

The following discussion illustrates the strategy for constructing training data sets using certain IMMACCS object-oriented data classes as examples. The data-mining classification task is to identify the value of the Bayesian-network class variable of an unknown data set. Initially, two Bayesian-network class variables will be considered, "imminent attack likely" or "imminent attack not likely." To train the TAN algorithm, the value of the Bayesian-network class variable will be identified in the training data sets for both classes.

Various types of attacks and defenses are listed as allowed values (among others) in the MILITARY_EVENT object class in the IMMACCS database.

These are AIR_ATTACK, GROUND_ATTACK, AIR_DEFENSE, GROUND_DEFENSE, and SMALL_SCALE_ATTACK. Only instances that correspond to attacks from hostile forces on the Marine Corps will be considered. Any attack launched by the Marine Corps on hostile forces will not be counted in the "attack" category. In contrast, defenses by the Marine Corps against hostile attacks, whether the attacks are launched from the air or the ground, are likely to play a role in the over-all model when they influence subsequent enemy attacks. For example, enemy commanders may select a battle plan that does not involve an air attack on an area with a strong Marine Corps air defense.

Several naive Bayesian networks can be induced, one for each attack type and one for the combined data for all attack types. For the combined attacks, the class variable can take multiple values, corresponding to the likelihood of a particular attack type and the likelihood that this attack type will NOT occur. Initially, all attack types will be assumed to be independent, although this is rarely true in actual battles. For example, ground attacks are more likely to follow air attacks at the same location than vice versa.

For the non-attack training instances, data associated with the other values of the MILITARY_EVENT object class will be used, such as WITHDRAWAL_EVENT, DELAYING_ACTION, AIR_REINFORCEMENT, or DRILL_EVENT. Other non-attack training instances also can be derived, for example, from the AIR_DEFENSE and GROUND_DEFENSE values, provided the instances pertain to events associated with enemy air defenses and ground defenses.

The date-time groups (DTGs) associated with each instance, both of attack and non-attack situations, will be noted and other data objects with the same DTGs (and with DTGs just prior to the event) will be included in the training data sets. The training data also could include objects present in the same vicinity as the attack or non-attack event that do not have DTGs. This will provide as comprehensive a description of the battlespace at the time and place of the attack as is possible, given the level of data granularity. This method of formulating training data sets can be extended by including in each data set the data that pertain to DTGs several days prior to the event to ascertain whether this will yield better results. The exact time span that each data set should cover is an open research issue.

Design Considerations in the Construction of Test Data Sets

Changes can be made in the test data sets, depending on the desired outcome of the test. For example, to determine how far in advance an attack can be predicted, the instances that pertain to an entire day immediately prior to the attack can be omitted systematically from test data sets. If the algorithm still makes the correct prediction, one can conclude, at least as far as that test data set is concerned, that an attack can be predicted 24 hours in advance. Similarly, if 2-days worth of data immediately preceding the attack can be omitted without a significant decline in the prediction accuracy, this is an indication that attacks can be predicted 48 hours in advance.

We expect, however, that omitting more and more data that pertain to the days just prior to an attack will cause the attack-prediction accuracy to degrade. The exact functionality of this degradation (linear, exponential,

logarithmic, etc.) is another open research question. This type of testing can enable researchers to determine the number of days to include in the data collection and the specific data elements to be collected necessary to formulate as accurate a prediction as possible.

Test and training data sets will be formulated according to an n-fold cross-validation procedure. For example, to implement the first cycle of a five-fold cross validation with a data set consisting of 1,000 records, the first 800 records can be selected for training, with the last 200 records being reserved for testing. During the second phase of training and testing, the first 600 records and the last 200 records together will comprise the test data set, and the remaining records will be used for testing. In the third phase, the first and last 400 records will be used for training and the middle 200 for testing, etc. The advantage of this procedure is that it can be used to identify anomalies in the testing and training so that if the results are comparable for all five tests, a higher level of confidence in the method is obtained.

CONCLUSION

This paper describes a data-mining environment designed to support wartime event prediction using Bayesian networks to perform a data-classification task. The TAN algorithm was selected to induce a network using data extracted from an object-oriented database that contains information from exercise message traffic. Future work could include a user-friendly interface designed on top of the algorithms to provide automated input of selected data sets to the algorithm of choice. Success in this research project will pave the way for a more precise indication-and-warning system for the U.S. Marine Corps.

ACKNOWLEDGMENTS

The authors thank SSC San Diego's Science and Technology Initiative and the Defense Advanced Research Projects Agency for their financial support.

REFERENCES

1. Clifton, C. and R. Steinheiser. 1998. "Data Mining on Text," *Proceedings of the 22nd Annual IEEE International Computer Software and Applications Conference, COMPSAC'98*, pp. 630-635.
2. Friedman, N., D. Geiger, and M. Goldszmidt. 1997. "Bayesian Network Classifiers," *Machine Learning*, vol. 29, no. 2/3, November/December, pp. 131-163.
3. Friedman, N., M. Goldszmidt, and T. J. Lee. 1998. "Bayesian Network Classification with Continuous Attributes: Getting the Best of Both Discretization and Parametric Fitting," *Proceedings of the International Conference on Machine Learning '98, ITAD-1632-MS-98-043*.
4. Thuraisingham, B. M. 1999. *Data Mining: Technologies, Techniques, Tools and Trends*, CRC Press, Boca Raton, FL.
5. McCarthy, S. J. and M. G. Ceruti. 1999. "Advanced Data Fusion for Wartime Event Correlation and Prediction," *Proceedings of the 16th Annual AFCEA Federal Database Colloquium and Exposition, AFCEA*, pp. 243-249.
6. Charniak, E. 1991. "Bayesian Networks without Tears," *AI Magazine*, pp. 50-63.



Marion G. Ceruti

Ph.D. in Chemistry, University of California at Los Angeles, 1979

Current Research: Information systems analysis, including database and knowledge-base systems, artificial intelligence, data mining, cognitive reasoning, software scheduling and real-time systems; chemistry; acoustics.

S. Joe McCarthy

Ph.D. in Solid-State Electronics, University of Washington, 1973

Current Work: Assistant Program Manager for Processing and Analysis, Space and Naval Warfare Systems Command.

7. Friedman, N., M. Goldszmidt, and A. Wyner. 1999. "On the Application of the Bootstrap for Computing Confidence Measures on Features of Induced Bayesian Networks," *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*.
8. Lee, T. J. and M. Goldszmidt. 1998. "TAN Tree Augmented Naive Bayesian Network Classifier Version 2.1 User Manual," <http://edi.erg.sri.com/tan/TANintro.htm>, pp. 1-27.
9. Alderson, S. L. 1999. "Urban Warrior Advanced Warfighting Experiment: Information Dominance in the Battlefield," *Proceedings of the 16th Annual AFCEA Federal Database Colloquium and Exposition*, AFCEA, pp. 213-228.
10. Leighton, R. and J. Pohl. 1998. The IMMACCS Object Model and Database, OBDATA00, November, IOM Version 1.5, Cal Poly, San Luis Obispo, CA.



Thermal Pixel Array Characterization for Thermal Imager Test Set Applications

Ike Bendall, Ted Michno, Don Williams, Matthew Holck,
and Richard Bates

SSC San Diego

José Manuel López-Alonso

Laboratorio de Termovision, Madrid, Spain

Robert J. Giannaris

Applied Technology Associates

Gordon Perkins and H. Ronald Marlin

The Titan Corporation

INTRODUCTION

Infrared scene projection (IRSP) technology has advanced rapidly in the last few years in an effort to support testing of missiles and other munitions that use infrared seekers. Existing infrared scene generation technology is very expensive, with available scene generators falling in the million-dollar price range. These infrared scene projectors are prohibitively expensive for most infrared (IR) sensor test and evaluation applications. Low-cost alternative technologies would open the door to a much greater range of test applications.

A thermal imager test set using IRSP technology would have several advantages over traditional test sets consisting of a blackbody source and target wheel. Portable thermal imager test sets have a small number of target wheel positions. Test patterns must be installed to match sensor test requirements. IRSP technology eliminates the need for physical test patterns and allows the operator to generate test patterns appropriate for each sensor. Target wheels are generally too large to be effectively cooled. Blackbody sources can be controlled to maintain a constant temperature difference, but changes in the ambient temperature produce temperature changes that lie outside the camera's dynamic range. The IRSP arrays under investigation in this study are small and can be cooled with thermoelectric coolers. The use of IRSP technology in place of the blackbody/target wheel allows control of both the source and background temperatures and guarantees that the scene lies within the camera's dynamic range. The thermal imager testing community is developing improved methods of testing thermal imagers that do not use traditional test patterns. IRSP technology provides the tester with the flexibility to generate the test patterns appropriate to these alternative test procedures.

SSC San Diego has been funded through the Office of Naval Research to develop a low-cost thermal pixel array (TPA) for portable test set applications that provides a path to built-in test applications. The Real-Time Infrared (RTIR) TPA is a micro-electromechanical systems (MEMS) device consisting of a two-dimensional array of miniature IR heater elements (thermal pixels). In contrast to other IRSP technologies, the RTIR TPA is a silicon-based, micro-machined Complementary Metal Oxide Semiconductor (CMOS) array. This process yields a single chip device that is significantly less expensive than alternative approaches. Each IR

ABSTRACT

An array of thermal emitters has been developed for use in a portable test set to enable field-testing of low-performance infrared imaging systems and seekers. It is not known if this technology can be used to evaluate the performance of state-of-the-art thermal imagers. This paper describes the preliminary measurements of thermal pixel array (TPA) performance. The radiant output of TPA was measured as a function of pattern size and drive voltage. Simple models were developed that agree with many aspects of the experimental data. Spatial and temporal noise characteristics of the TPA have been ascertained through three-dimensional noise analysis. Detection algorithms were used to compare images of test patterns produced by the TPA to images of similar test patterns produced by a standard blackbody.

heater is suspended over a micro-machined cavity and surrounded by pixel-specific electronics that allow rapid loading and retention of the image data. The micro-machined cavity thermally isolates the heater from the parent substrate, allowing each pixel to be individually set and maintained at a temperature different from that of its neighbors. Four heater elements are shown in Figure 1(A). Each heater element can be addressed independently of any other heater element. This allows the operator to vary both the shape and location of test patterns. This capability is shown in Figures 1(B), 1(C), and 1(D).

The RTIR TPA specifications were selected to meet dynamic scene requirements for missile testing and were not intended for use in a thermal imager test set. Minimum resolvable temperature difference (MRTD) is an important thermal imager figure of merit and is routinely measured during sensor evaluations. State-of-the-art thermal imagers have MRTDs of a few tens of milliKelvin at low spatial frequencies. Characterization of these imagers requires blackbodies with temperature resolutions that exceed those of the imager. Temperature resolutions of this scale exceed the RTIR TPA design specifications by at least an order of magnitude. In spite of the drawbacks of the RTIR TPA design, it was felt that it would be beneficial to compare the performance of this technology to a traditional thermal imager test set. This approach would provide insight into the feasibility of the RTIR TPA technology, help identify unknown problems, and provide a basis for developing thermal imager test set TPA performance specifications.

INSTRUMENTATION

The standard blackbody used in this comparison was furnished by Santa Barbara Infrared (SBIR). The telescope has a 6-inch aperture and a 30-inch focal length. Differential temperature resolution is ± 3 milliKelvin when the unit is operated in the temperature difference mode. The thermal pixel array test set is shown in Figure 2. The RTIR TPA is a 128 by 128 array with pixel pitch of 88.6 microns.

The temperature range of the TPA is approximately 250°C with a thermal resolution of 0.250°C. The TPA area fill factor is 15%, and its emissivity is approximately 60%. The collimating telescope has a total transmission of 91% in the 3- to 5-micron band, a focal length of 233 mm, and a 50-mm aperture. The losses due to the fill-factor, emissivity, and telescope transmission result in an efficiency of 0.082 and an effective temperature resolution of 20 milliKelvin. A pixel non-uniformity correction capability is planned but is not currently available.

An Amber Galileo thermal imager with a 75-mm focal-length lens was used for these measurements. The Galileo is capable of extremely high frame rates; however, for this analysis, images were acquired at

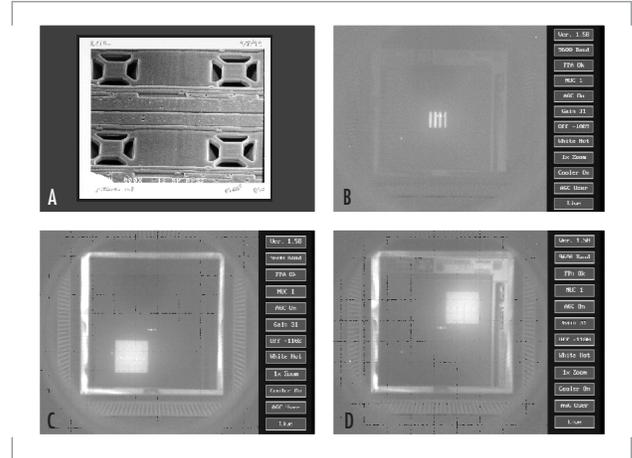


FIGURE 1. (A) Four micro-machined heater elements, (B) four-bar pattern in center of array, (C) square in lower left-hand corner, and (D) square moved to upper right-hand corner.

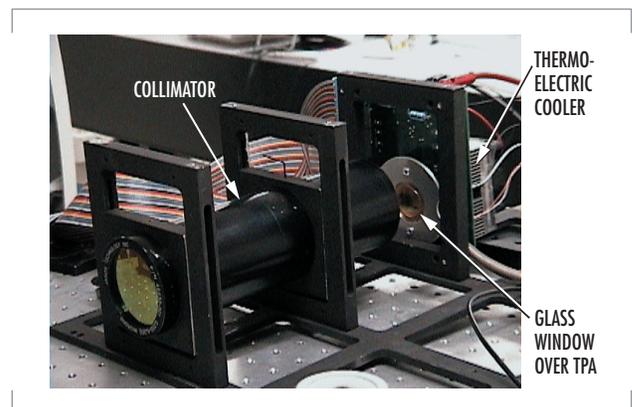


FIGURE 2. TPA blackbody.

the standard 30 Hz. The thermal imager was mounted on a rotation stage as shown in Figure 3. The thermal imager focus was adjusted to image the bar patterns from the SBIR blackbody. The imager was then rotated 90 degrees to view the TPA blackbody. The focus of the TPA bar patterns was achieved by adjusting the position of the TPA. This configuration allowed rapid collection of both TPA and SBIR images. Images were digitized with a Matrox Pulsar frame grabber with 8 bits of resolution.

PATTERN SIZE AND VOLTAGE EFFECTS

Traditional test sets consist of a thermal source, a collimator, and a target wheel that holds the test patterns or masks. The wheel is physically separated from the blackbody source and its temperature is unaffected by changes in the temperature of the source. Changing the wheel's position does not affect the temperature difference between the blackbody and mask; therefore, temperature differences are independent of the pattern size. This is not necessarily true for a TPA blackbody. The thermal insulation provided by the micro-machined cavity does not completely isolate the heater from the parent substrate. Thermal conduction through the substrate affects the background temperature of the array and decreases the effective temperature difference (Figure 4). This effect may depend on both pattern size and control voltage.

The first characterization task was to examine the relationship between pattern size and radiometric temperature. Three test patterns (two squares and a four-bar pattern) were selected for the analysis. The squares were generated by heating 30-pixel by 30-pixel and 6-pixel by 6-pixel regions on the array. The bar pattern consisted of four bars each 21 pixels long by 3 pixels wide. This pattern is consistent with the 7:1 aspect ratio of bar patterns used in MRTD measurements. Two measurements were

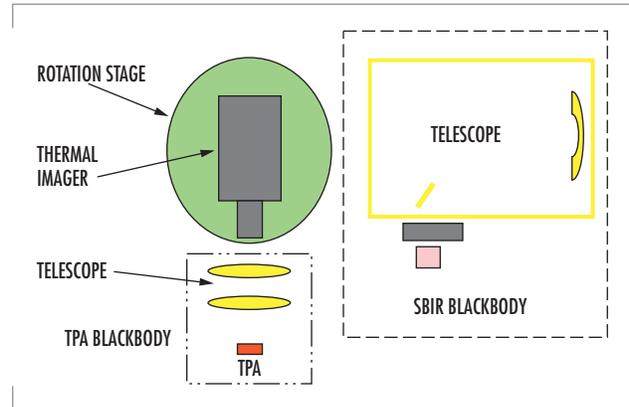


FIGURE 3. Diagram of experimental apparatus.

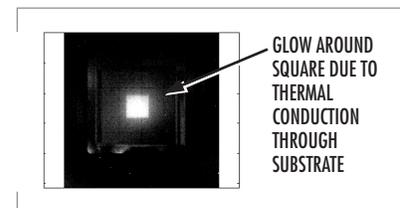


FIGURE 4. 40-pixel by 40-pixel heated area. It is apparent that the heat is not confined to the pattern area but is conducted into the surrounding area.

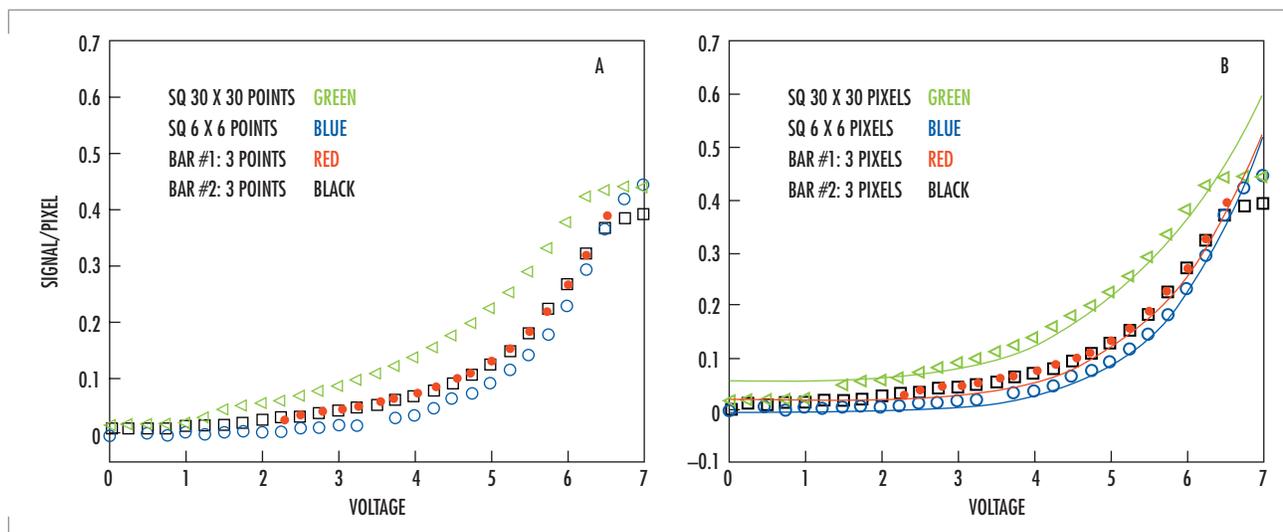


FIGURE 5. (A) Signal/pixel for three patterns (30 x 30 square, 6 x 6 square, and 21 x 3-bar pattern). Pattern size effects are evident. (B) Fit to data based on simplistic heating model. Model has three parameters and a nonlinear fit is used to achieve best fit. Good agreement is achieved between 2.5 and 6 V.

made with the bar pattern on separate days to evaluate TPA temporal stability. Images of each pattern were obtained as a function of voltage. The values of the pixels in each pattern were summed to produce a total signal. The signal was divided by the number of pixels comprising the pattern to produce a signal/pixel value. The signal/pixel values were compared for the three patterns. The results are plotted in Figure 5(A). A temperature dependence on pattern size is clearly evident. In contrast, the four curves would overlap if a traditional blackbody/target wheel test set had been used. It was encouraging that the two four-bar pattern curves (red and black) were in excellent agreement and that the shapes of the curves were similar for all three patterns. A simplistic model, relating the radiometric energy measured to voltage, was developed and used to fit the data. The model, which had three unknown parameters, was in excellent agreement with the data from 2.5 to 6 V (Figure 5(B)).

Thermal imagers suffer from blurring due to a reduction of the modulation transfer function with an increase in pattern spatial frequency. A traditional blackbody does not affect the pattern fidelity; therefore, any loss of fidelity can be attributed to the thermal imager. The blurring due to thermal conduction in the TPA test set results in a loss of pattern fidelity that must be separated from the degradation in image quality due to the thermal imager.

The investigation of the impact of thermal conduction on pattern fidelity was continued by examining the shape of square test patterns as a function of size and voltage. The results are summarized in Figure 6. Horizontal line profiles were taken through the center of the heated area. Line profiles of a 40 by 40 square as a function of voltage are plotted in Figure 6(A). Figure 6(B) compares line profiles for squares with sides of 10, 15, 20, 30, and 40 pixels at a constant 6 V. A parabolic curve, described by the equation below, was plotted through the peak of each curve.

$$S = S_{x_c} - a_{p,V}(V - V_T)(x - x_c)^2$$

where S is the pixel value, S_{x_c} is the peak pixel value, x_c is the pixel location at which the peak pixel value occurs, V_T is a threshold voltage (~ 3 V), and $a_{p,V}$ is a coefficient that can depend on pattern size and voltage. The curves shown in Figure 6 are generated by setting $a_{p,V}$ equal to a constant independent of pattern size or voltage. The curves appear to represent a reasonable fit to the data. This relatively simple relationship was unexpected and suggests that thermal conduction distortions can be readily understood, which is encouraging given the complexity of the TPA structure.

THERMAL MODELS

The results from the previous section suggested that a simple thermal conduction model might predict the effects of pattern size and control voltage on the array's temperature distributions. A finite-element analysis model was used to predict array temperature distributions. The TPA is a very complex structure, but for the first attempt, a simplistic model of the TPA was constructed. The TPA was assumed to be a homogeneous, isotropic material with a constant thermal conductivity and emissivity. It was further assumed that cooling occurs only through the bottom surface of the array and that the thermoelectric cooler maintains this surface at a fixed temperature. The objective of this analysis was to generate curves with trends similar to those shown in Figure 6. In particular, four features

in Figure 6 were of interest: (1) the increase in peak temperature with pattern size, (2) the long tail in the unheated region, (3) the sharp transition between the heated area and the tail, and (4) the flat tops of the small squares. A typical result is shown in Figure 7.

The model results do show the increase in peak temperature with pattern size and the long tail in the unheated areas. The transition between the heated and unheated areas is not as sharp, and the tops of the small squares are more rounded than experimentally measured. A more complex model of the TPA is being constructed that should replicate these features.

NOISE BEHAVIOR

Noise is an important factor in thermal imager performance especially for tasks involving detection threshold measurements such as MRTD. The three-dimensional (3-D) noise model [1] provides an effective method of determining the noise characteristics. Image sequences of 30 frames were obtained from both the TPA and the reference blackbody. Thermal conduction through the TPA substrate distributes heat throughout the entire array. This low-frequency background is not apparent in the blackbody. For this reason, the low-frequency noise components were suppressed by means of a polynomial fit prior to the 3-D noise analysis. The results are summarized in Table 1.

The intrinsic noise of the blackbody should be small compared to that of the Galileo, and it is safe to attribute blackbody noise components in Table 1 to the Galileo. Inherent TPA noise is indicated by the increase between blackbody and TPA noise components. The magnitudes of the TPA and blackbody noise components are remarkably similar, with σ_{tvh} and σ_{vh} being the most significant noise components for both sources. This behavior is typical of staring thermal imagers, such as the Galileo.

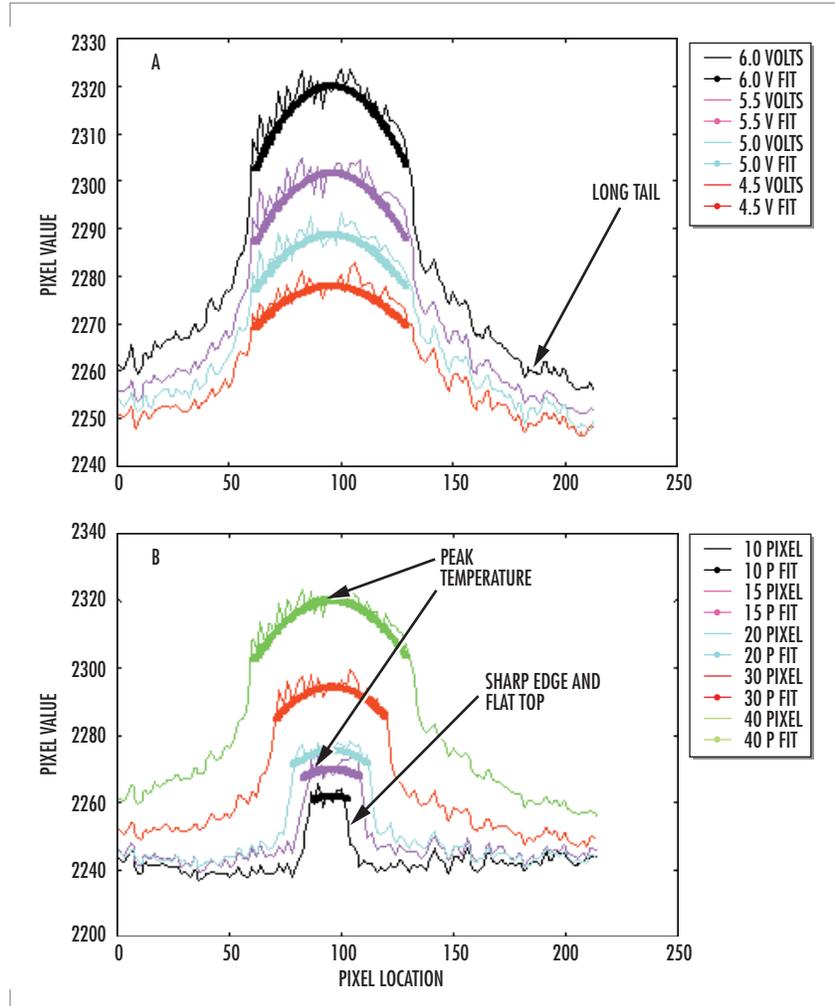


FIGURE 6. (A) Horizontal line profiles for a 40-pixel by 40-pixel pattern over a range of control voltages. (B) Horizontal line profiles at a fixed 6 V for square patterns of 10, 15, 20, 30, and 40 pixels. In both (A) and (B), the solid curve is parabolic fit.

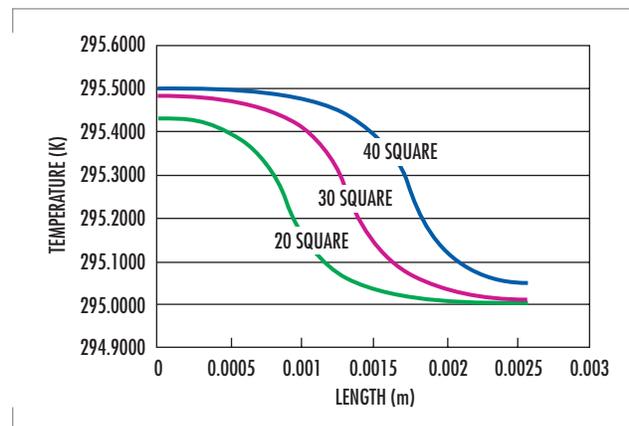


FIGURE 7. Finite-element analysis of TPA temperature profiles.

HUMAN VISION MODELS

In recent years, significant advances have occurred in the field of vision research to model the response of the human visual cortex. While human vision is far from solved, the principal mechanisms are understood. The visual cortex can be modeled as a collection of filters each sensitive to a restricted spatial frequency bandwidth. The model can be extended to compare filtered responses of two similar images and compute probabilities that a human observer will detect differences between the images. A visual cortex model developed by one of the authors [2] was used to compare high- and low-contrast bar patterns from a traditional blackbody and the TPA. A full description of the model and the analysis is beyond the scope of this paper; however, the model indicated that at low contrast the TPA and traditional blackbody images were indistinguishable to a human observer.

TABLE 1. 3-D noise analysis results.

	Blackbody (counts)	TPA (counts)
Sigma tvh	1.43	1.48
Sigma tv	0.23	0.24
Sigma th	0.16	0.22
Sigma vh	1.21	1.28
Sigma v	0.37	0.29
Sigma h	0.29	0.32
TOTAL	1.96	2.04

CONCLUSIONS

An assortment of measurements has been performed during the initial phase of the TPA characterization. In general, the results were extremely promising. Noise characteristics were in better agreement with a traditional blackbody than expected. Use of human vision models provided a novel characterization tool and indicated that TPA and blackbody images are very similar at low contrast. Crude estimates based on low-contrast images yield TPA MRTD measurements two to four times higher than MRTD measurements made with a traditional blackbody. This was better than expected, considering the poor temperature resolution of the RTIR TPA. Pattern blurring from thermal conduction is an important difference between TPA and traditional blackbodies. The effects of thermal conduction on pattern contrast must be understood or eliminated before a TPA-based thermal imager test set will be achievable. Simple thermal conduction models reproduce some of the experimentally measured features, but a more complete model is needed. Understanding the important factors affecting thermal conduction will help develop TPAs less susceptible to thermal distortions. Further investigation and development of the TPA is required, but the results are extremely promising and indicate that the TPA technology is a potential candidate for use in a thermal imager test set.

This technology may be the subject of one or more invention disclosures assignable to the U.S. Government, including N.C. #82901. Licensing inquiries may be directed to:

Harvey Fendelman
 Office of Patent Counsel D0012
 SSC SAN DIEGO
 53510 Silvergate Avenue Room 103
 San Diego, CA 92151-5765
 (619) 553-3001

AUTHORS

Ted Michno

BA in Engineering Physics, Point Loma Nazarene University, 1996
Current Research: Microradiometry; infrared technology development.

Don Williams

BS in Electrical Engineering, Northrup Institute of Technology, 1963
Current Research: MEMS for thermal infrared sources and for RF power measurement; passive millimeter-wave surveillance.

Matthew Holck

MS in Physics, San Diego State University, 1999
Current Research: Signal processing; image processing; computer automation.

Richard Bates

BA in Physics, Loma Linda University, 1960
Current Research: Radiation-induced Irtran 2 absorption; polarization independent narrow channel (PINC) wavelength division multiplexing (WDM) fiber coupler fusing.

José Manuel López-Alonso

Graduate in Physics, Universidad Complutense de Madrid, 1994
Current Research: Characterization of thermal imagers, image quality evaluation.

Robert J. Giannaris

Ph.D. in Mechanical Engineering, Purdue University, 1972
Current Research: Infrared displays; microwave sensors; and hyperspectral sensors.

Gordon Perkins

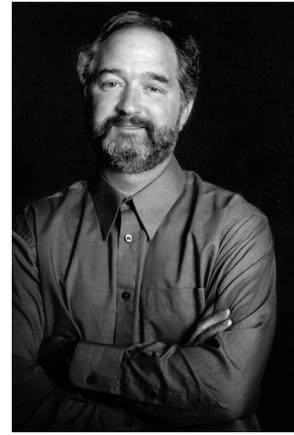
BA in Physics, University of California at San Diego
Current Research: RF MEMS devices; millimeter-wave remote atmospheric sensing.

H. Ronald Marlin

BA in Physics, La Sierra University, 1959
Current Research: Microradiometry.

REFERENCES

1. D'Agostino, J. and C. Webb. "3-D Analysis Framework and Measurement Methodology for Imaging System Noise," *SPIE*, vol. 1488, pp. 110–120.
2. Alonso, J. and I. Bendall. 2000. "The Use of Vision Models for the Characterization of Scene Projector Devices," NATO TG12 Panel Group, (September), and private communication.



Ike Bendall

Ph.D. in Physics, Arizona State University, 1984
Current Research: Infrared and hyperspectral sensor characterization.

Hyperspectral Imaging for Intelligence, Surveillance, and Reconnaissance

David Stein, Jon Schoonmaker, and Eric Coolbaugh
SSC San Diego

INTRODUCTION

The optical spectrum is generally considered to include the ultraviolet (200 to 400 nm), the visible (400 to 700 nm), the near infrared (700 to 1100 nm), and the short-wave infrared (1100 to 2500 nm). Sensors operating in these bands detect reflected light which is used to discriminate an object from its background and to classify it based on spectral characteristics. Spectral sensors capitalize on the color difference between objects and the background. A color video camera that divides the reflected light into red, green, and blue components is thus a simple spectral sensor. More complicated sensors break the spectrum into finer and finer bands and/or selectively tune to bands appropriate for a specific object or background. In general, a multispectral sensor, illustrated in Figure 1, is defined as a sensor using two to tens of bands, while a hyperspectral sensor, illustrated in Figure 2, is defined as a sensor using tens to hundreds of bands. Spectral sensors are divided into four types or approaches. Currently, the most common type is the "pushbroom" hyperspectral sensor. In this approach (Figure 2), a single line is imaged through a dispersing element so that the line is imaged in many different bands (colors) simultaneously. A second spatial dimension is realized through sensor motion. A second type is a multispectral filter wheel system in which a scene is imaged consecutively in multiple bands. A third type images multiple bands simultaneously using multiple chips (or multiple areas on the same chip). This approach uses multiple apertures or a splitting technique, such as a series of dichroic prisms or a tetrahedral mirror or lens. The fourth approach is the use of a Fourier transform spectrometer. The product of any of these sensors is an image cube as illustrated in Figure 3.

Hyperspectral Imaging at SSC San Diego

SSC San Diego has supported a number of hyperspectral programs over the last several years for a variety of government agencies, including the Defense Advanced Research Projects Agency (DARPA), the Spectral Information and Technology Assessment Center (SITAC), the Office of Naval Research (ONR), the Office of the Secretary of Defense (OSD), and the High Performance Computing Management Office (HPCMO). We have worked on DARPA's Adaptive Spectral Reconnaissance Program (ASRP), the goal of which was to demonstrate the detection of concealed terrestrial military targets and the cueing of a high-resolution imager. For ONR, we have been involved with maritime applications of

ABSTRACT

This paper highlights SSC San Diego contributions to the research and development of hyperspectral technology. SSC San Diego developed the real-time, onboard hyperspectral data processor for automated cueing of high-resolution imagery as part of the Adaptive Spectral Reconnaissance Program (ASRP), which demonstrated a practical solution to broad area search by leveraging hyperspectral phenomenology. SSC San Diego is now implementing the ASRP algorithm suite on parallel processors, using a portable, scalable architecture that will be remotely accessible. SSC San Diego performed the initial demonstrations that led to the Littoral Airborne Sensor Hyperspectral (LASH) program, which applies hyperspectral imaging to the problem of submarine detection in the littoral zone. Under the In-house Laboratory Independent Research (ILIR) program, SSC San Diego has developed new and enhanced methods for hyperspectral analysis and exploitation.

hyperspectral sensors. Under OSD sponsorship, we have demonstrated the capabilities of hyperspectral remote sensing for search and rescue applications. For SITAC, we have provided ground truth measurements of ocean optical properties and illumination required for controlled experiments, and we have analyzed the bands required for optimal ocean imaging. The HPCMO is sponsoring our work to develop scalable and portable implementations of the ASRP algorithms. Under ONR and SSC San Diego In-house Laboratory Independent Research (ILIR) funding, we have developed new and enhanced methods for hyperspectral analysis and exploitation. Highlights of these efforts are described in more detail below.

Terrestrial Hyperspectral Remote Sensing

The DARPA ASRP successfully demonstrated the capability to detect military targets of interest in real time by using an airborne hyperspectral system to cue high-resolution images for ground analysis. SSC San Diego led all research, development, coding, and implementation of the end-to-end processing and critical hyperspectral detection and recognition algorithms. The algorithms and processing architecture developed are applicable to a broad scope of missions, targets of interest, and platform architectures. ASRP pushed the state of the art beyond simple detection of targets in the open, making detection of difficult, realistically positioned targets possible at low false alarm rates. Figure 4 shows the difficult environment, used by ASRP for real-time hyperspectral system demonstrations, that may be encountered during military operations. The variety of natural and man-made materials and the

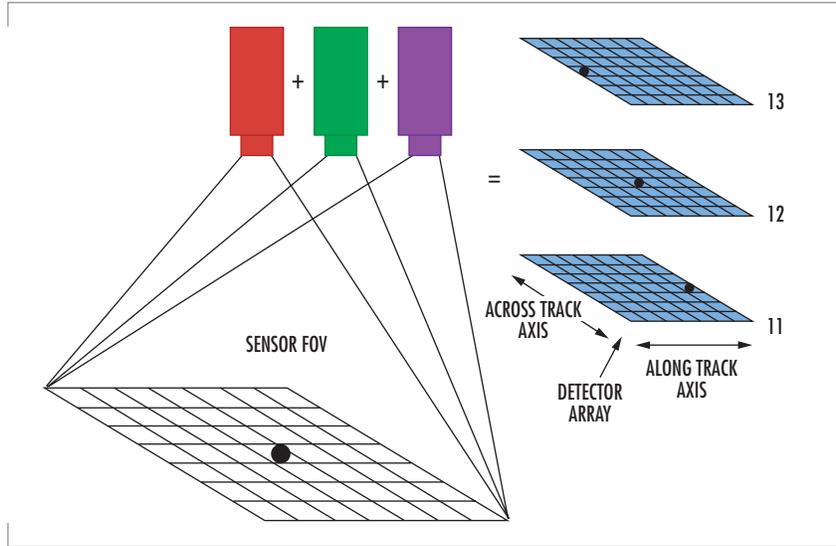


FIGURE 1. Schematic of three-band multispectral imaging camera.

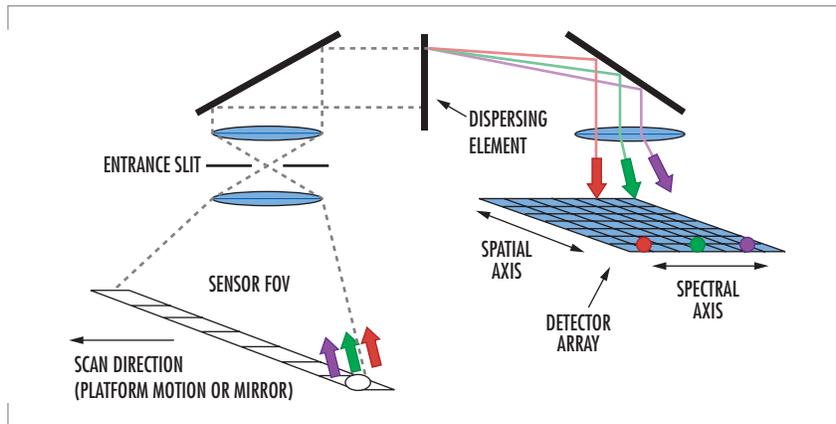


FIGURE 2. Schematic of a pushbroom dispersive hyperspectral sensor.

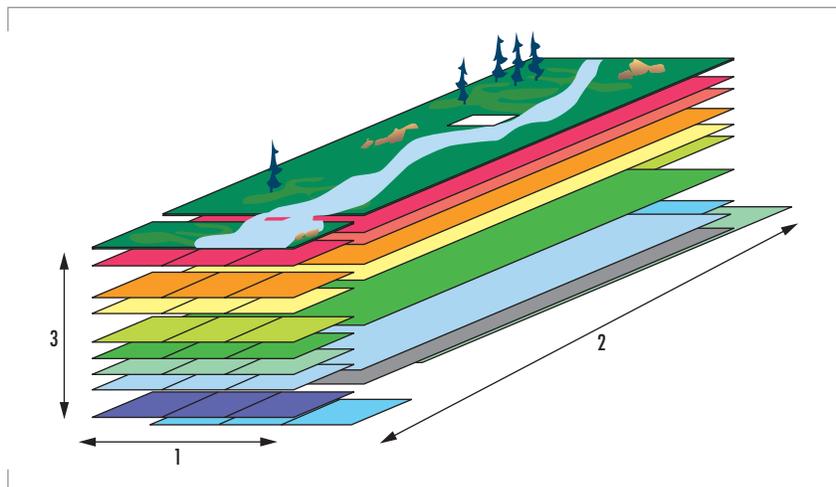


FIGURE 3. Hyperspectral image cube's cross-track, 1; along track, 2; and spectral dimension, 3.

variability of illumination combine to form a highly complex spectral detection challenge. Figure 5 compares the visibility of two targets in high-resolution imagery (top), in a red-green-blue (RGB) image (middle), and in the output of a detection statistic (bottom). These detections exemplify the ability of the hyperspectral system to identify target positions even when they may not be evident in traditional high-resolution imagery.

The High Performance Computing Management Office (HPCMO) has funded SSC San Diego, as part of the Hyperspectral Information Exploitation Project, to implement the ASRP hyperspectral algorithm suite and end-to-end processing on high-performance computer (HPC) platforms in a portable, scalable architecture accessible by a wide variety of Government users. Parallel processing capabilities will provide a new dimension for hyperspectral processing, allowing multiple hyperspectral algorithms to optimize target detection and recognition on massive data sets.

Maritime Sensor Systems

SSC San Diego has been instrumental in initiating and demonstrating the use of hyperspectral imagery for surveillance of the littoral. In 1996, SETS Technology, working with SSC San Diego, flew the SETS Technology Advanced Airborne Hyperspectral Imaging System (AAHIS) over submarines at the Pacific Missile Range Facility northwest of Kauai. The results of these flights led to the Littoral Airborne Sensor Hyperspectral (LASH) program.

LASH is an integrated optical sensor system that uses pushbroom scanning for the detection of submarines in the littoral environment. The LASH system consists of a



FIGURE 4. Three-color image of an ASRP hyperspectral image.

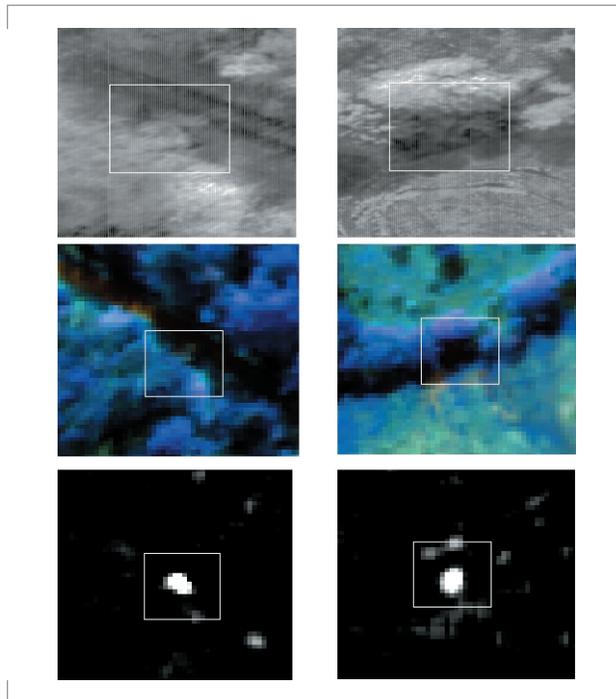


FIGURE 5. These figures show a high-resolution panchromatic imager (6-inch ground sample distance [GSD]) [top], and RGB image created from three hyperspectral bands (1-meter GSD) [middle], and one hyperspectral algorithm detection statistic image [bottom] for two different targets hidden along tree lines in shadow.

passive hyperspectral imager (HSI) assembly, an image processor, a data storage (archival) unit, a data display unit for operator use that incorporates the system monitoring, and control functions. The system is integrated into a modified ALE-43 (chaff cutter and dispenser pod) and mounted on a standard pylon at wing station 12 (Figure 6). All principal elements of the LASH system are contained within the pod. The units installed within the aircraft itself are limited to the system display processor, the power interface to the aircraft, the operator controls, and a global positioning system (GPS) antenna. This design was established to provide a system that could be considered independent of the individual aircraft tail number. It is estimated that all of the internal aircraft mounted units could be installed in less than 2 hours if necessary.

The passive and stabilized hyperspectral sensor collects both spatial (770 pixels) and spectral data (up to 288 pixels) on each instantaneous image increment. The data are binned by 2 spatially and 6 spectrally to give 385 spatial and 48 spectral channels. This imaged data is framed at 50 Hz, with each frame covering a 40-degree lateral field of view and approximately a 0.06-degree (1 milli-radian) field of view in the direction of flight. The data are simultaneously recorded in the archival storage system, processed by the image processor, and presented in a pseudo-color waterfall display to the operator. The processing system evaluates the data sensed in near real time using both spectral and spatial processing, and it provides a "frozen" display of the target along with its position in longitude and latitude. A stabilization system automatically adjusts the sensor so that it compensates for aircraft roll, pitch, and yaw. A "point to track" option forces the stabilization system to point the sensor along a predetermined track (otherwise the sensor points directly down).

These sensors can perform a wide range of ocean sensing tasks. Targets range from submarines and sea mines for military applications, to chlorophyll and sediment load in physical oceanographic applications, to schools of dolphins and whales in marine biology applications. Figure 7 demonstrates the ability of the sensor to image a pod of humpback whales. In these applications and others, a common goal is to detect an extremely low-contrast target in a high-clutter background.

Ocean Environmental Measurements

Hyperspectral systems such as LASH are being developed that use spectral and spatial processing algorithms to discern objects and organisms below the sea surface. The performance of such systems depends on environmental and optical properties of the sea. An instrument suite, the Portable Profiling Oceanographic Instrument System (PorPOIS), was developed to ascertain and quantify these environmental and hydro-optic conditions. Profiling of the downwelling irradiance leads to a value of the diffuse attenuation coefficient, k_d , for the water column. Measurements of the beam absorption, a , and attenuation, c , provide information about the non-pure water absorption and scattering characteristics of the water. Measurement of the backscatter at different wavelengths determines what fraction of the downwelling photons is scattered back toward space. These and a number of other measurements made by PorPOIS allow for a thorough characterization of the water body. These data are used in the LASH program to optimize parameters of the processing algorithms and to predict the performance of the sensor by using modeling software that requires these oceanographic data as inputs.

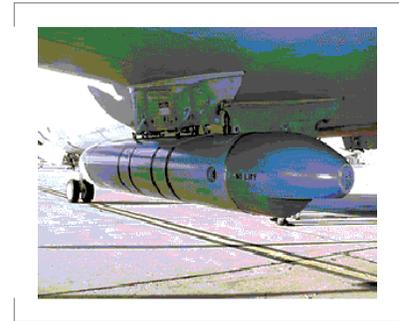


FIGURE 6. LASH pushbroom hyperspectral imager mounted on the wing of a P3 aircraft.

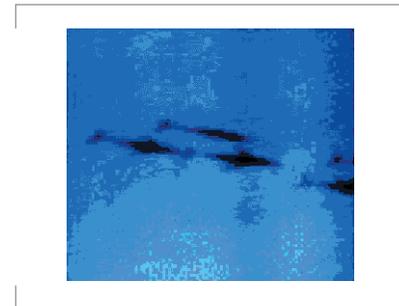


FIGURE 7. A pod of humpback whales imaged using the AAHIS sensor, a precursor to LASH.

The PorPOIS system is deployed on two submersible cages and a surface data-gathering station. The instruments are controlled and the data collected on a laptop computer running a Windows-based control and data acquisition software package, the Sensor Interface Display (SID), developed at SSC San Diego. The instruments (Figure 8) used to measure surface conditions and ship location include a wind transducer (anemometer), a magnetic compass, a surface irradiometer, and a GPS receiver. There are currently seven instruments used to measure optical and environmental conditions below the sea surface. These instruments include a downwelling and upwelling irradiometer (Biospherical Instruments PER600 and PER700), an upwelling radiometer (PER600), a transmissometer (Seatech), an absorption and attenuation meter (WETLabs ac-9), a conductivity-temperature-depth (CTD) (SeaBird Electronics SBE-19), a fluorometer (WETStar), and a backscattering meter (HobiScat-6). The devices are bundled in a single beehive-type stainless-steel profiling cage as shown in Figure 9. The cage is suspended from a davit via the underwater cable. The SeaBird SBE-32 carousel water sampler (Figure 10) holds twelve 2.5-liter bottles and the SBE-19 CTD. It uses the same underwater cable as the profiling cage. Deployment of the cage is nearly identical to that of the instrument cage. A deck unit mounted in the control rack translates the CTD information from the carousel and transfers the data to SID. This allows the user to capture water samples from target depths by monitoring the position of the carousel as it travels through the water column. New instruments can be added to the configuration as required.

Sample PorPOIS products are shown in Figures 11 and 12. Figure 11 shows downwelling irradiance at 490 nm measured off San Clemente Island, CA. These data are used to determine the rate of attenuation of irradiance at 490 nm, k_{490} , as shown in Figure 12. Optical depth, $1/k_{\lambda}$, is defined as the depth at which surface irradiance of wavelength λ diminishes by $1/e$. System performance is parameterized in terms of optical depth.

SSC San Diego ILIR and ONR-sponsored Research on Hyperspectral Algorithms

Pre-processing transforms are a common initial step in the processing of hyperspectral imagery that is performed in order to determine spectra of the fundamental constituents of the scene or for data compression. The principal component transform is based on minimizing loss in mean-square error, and the vector quantization (VQ) transform is based on minimizing the worst-case angle error between a datum and its projection onto a subspace. These transforms may have unintended consequences on the signal-to-noise ratio (SNR) of a target of interest. We have evaluated the loss in SNR that may result from applying a linear transform and developed several new transforms that use different knowledge of the signals of interest to reduce the loss in SNR in comparison with commonly applied transforms. Figures 13 and 14 illustrate the detectability of an underwater target in data that has been transformed using vector quantization and one of the newly defined transforms, whitened vector quantization (WVQ), that uses no signal information. Clearly, the WVQ algorithm can reduce the dimension of the data and preserve the target SNR for these



FIGURE 8. The Biospherical Instruments PRR-610 surface irradiometer, the NEXUS wind transducer, and the NEXUS magnetic compass are used to measure surface conditions.

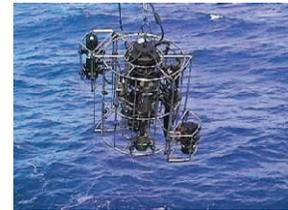


FIGURE 9. Submersible cage containing instruments used to measure ocean optical properties.



FIGURE 10. Submersible cage containing a CTD and water collection bottles used to measure absorption and scattering as a function of depth.

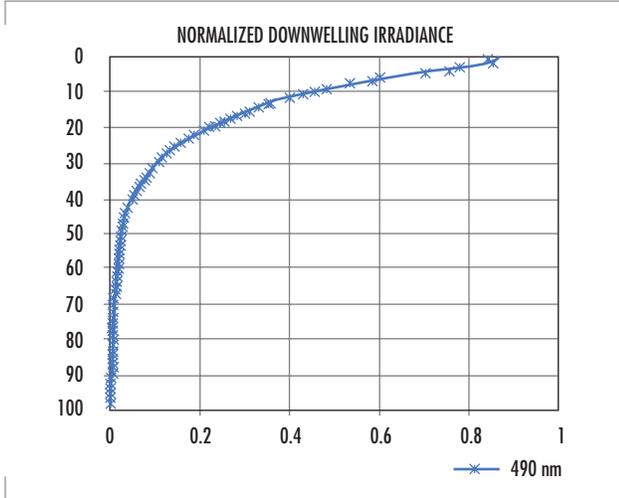


FIGURE 11. Plot of downwelling irradiance at 490 nm as a function of depth as measured using PorPOIS in waters off San Clemente Island, CA.

data. The transformed data are processed here with the Reed-Xiaoali (RX) quadratic anomaly detector. The enhanced discrimination of the target at lower dimension using the WVQ algorithm arises from the fact that the performance of quadratic detectors improves for a given SNR if the dimension is reduced.

Linear unmixing and image segmentation are common means of analyzing hyperspectral imagery. Linear unmixing models the observed spectra as

$$y^{ij} = \sum_{k=1}^d a_k^{ij} e_k, \text{ such that } \sum_{k=1}^d a_k^{ij} \leq 1 \text{ and } 0 \leq a_k^{ij} \leq 1.$$

The spectral vectors, e_k are known as endmembers, and a_k^{ij} is the abundance of the k^{th} material at pixel (i,j) . There are several means available for estimating the endmembers. The abundances are usually estimated by solving the constrained least-squares problem.

Image segmentation typically models the observation vector as arising from one of several classes, such that each class has a multi-variate normal distribution. The number of classes, d , is selected and the mean and covariance of the classes $\{(\mu_{\ell k}, \Sigma_k) \mid 1 \leq k \leq d\}$ are estimated from the hyperspectral data. The expectation maximization and the stochastic expectation maximization algorithm are two methods of estimating these parameters. Given the parameters and the probability of each class, the data may be classified by assigning y^{ij} to the class that, conditioned on the observation, is most likely. This computation is carried out using Bayes Law.

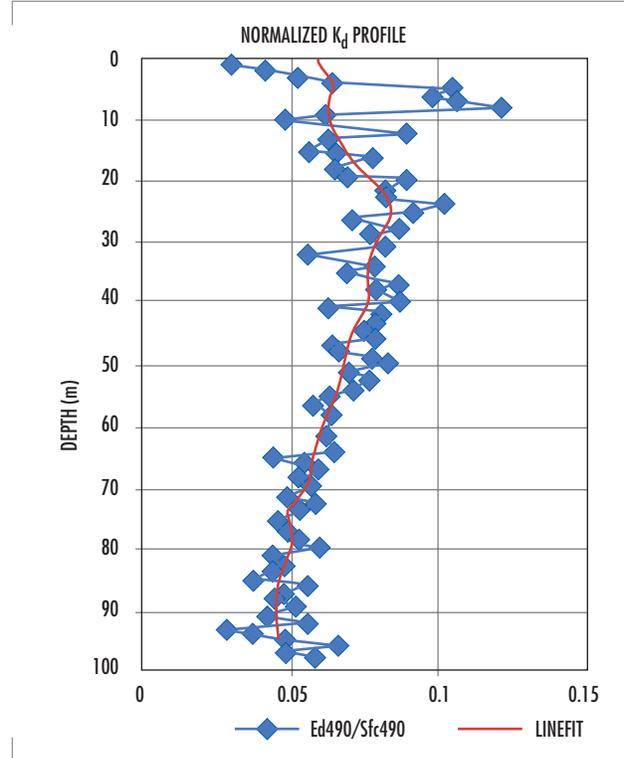


FIGURE 12. Rate of attenuation of downwelling irradiance at 490 nm derived from PorPOIS measurements of downwelling irradiance.

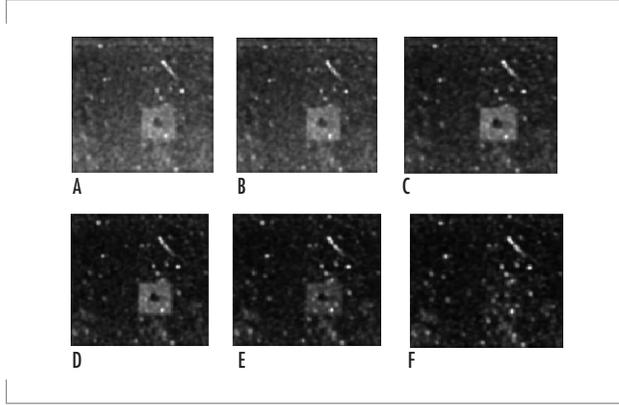


FIGURE 13. The RX algorithm applied to VQ-transformed 48-band hyperspectral imagery transformed to 48, 36, 20, 12, 9, and 7 dimensions (A through F, respectively).

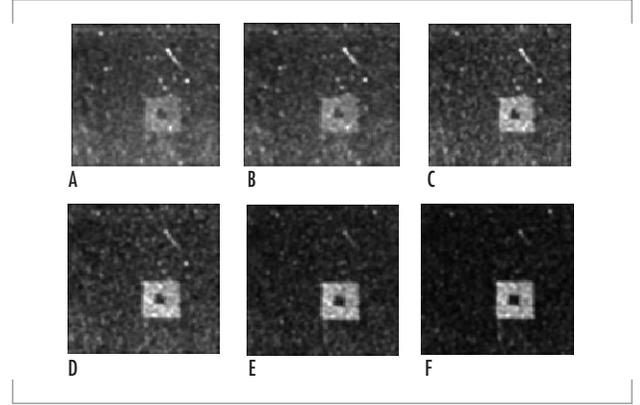


FIGURE 14. The RX algorithm applied to WVQ-transformed 48-band hyperspectral imagery transformed to 48, 36, 24, 8, 4, and 2 dimensions (A through F, respectively).

We have generalized the linear unmixing and image segmentation approaches in the development of the stochastic compositional model. We assume an $A \times B$ image of multivariate data: $y^{ij} \in \mathbb{R}^n$, $1 \leq i \leq A$, $1 \leq j \leq B$. The stochastic compositional approach models each observation vector as a constrained linear combination of normally distributed random variables. Let d be the number of classes, and let $N(\mu_k, \Sigma_k)$, $1 \leq k \leq d$ denote the normal distribution with mean μ_k and covariance Σ_k then

$$y^{ij} = \sum_{k=1}^d a_k^{ij} x_k^{ij} \text{ such that } x_k^{ij} \sim N(\mu_k, \Sigma_k), 0 \leq a_k^{ij} \leq 1, \text{ and } \sum_{k=1}^d a_k^{ij} = 1. \quad (1)$$

To account for scalar variation in the illumination, we also consider the model that uses an inequality constraint:

$$y^{ij} = \sum_{k=1}^d a_k^{ij} x_k^{ij} \text{ such that } x_k^{ij} \sim N(\mu_k, \Sigma_k), 0 \leq a_k^{ij} \leq 1, \text{ and } \sum_{k=1}^d a_k^{ij} \leq 1. \quad (2)$$

For given parameters (μ_k, Σ_k) , $1 \leq k \leq d$, and given abundances

$\alpha = (a_1, \dots, a_d)$, let (dropping the pixel indices) $\mu(\alpha) = \sum_{k=1}^d a_k \mu_k$, and $\Sigma(\alpha) = \sum_{k=1}^d a_k^2 \Sigma_k$. Then, $y^{ij} \sim N(\mu(\alpha), \Sigma(\alpha))$. Maximum likelihood abundance estimates are thus obtained by solving

$$\hat{\alpha}^{ij} = \arg(\max(\frac{1}{|\Sigma(\alpha)|^{0.5} (2\pi)^{n/2}} \exp\left(-\frac{1}{2}(y^{ij} - \mu(\alpha))\Sigma(\alpha)^{-1}(y^{ij} - \mu(\alpha))\right)). \quad (3)$$

Let $X = (x_1, \dots, x_d)$; the maximum likelihood estimates of the decomposition of the observation into contributions, x_k from the classes is obtained by solving

$$\begin{aligned} \hat{X} &= \arg(\max(p(X | y, \alpha, \mu_k, \Sigma_k)) \\ &= \arg\left(\max\left(\prod_{k=1}^d \frac{1}{(2\pi)^{n/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x_k - \mu_k)' \Sigma_k^{-1} (x_k - \mu_k)\right)\right)\right) \\ \text{such that } y &= \sum_{k=1}^d a_k x_k. \end{aligned} \quad (4)$$

The stochastic compositional model and deterministic linear unmixing have been compared by using simulated hyperspectral imagery. Class statistics were estimated from hyperspectral imagery by using the stochastic expectation maximization algorithm. Using these parameters, a set of simulated hyperspectral imagery was generated so that the mixing proportions of the classes were known. The test data were then unmixed by using both deterministic unmixing (with the class means as endmembers) and by stochastic compositional modeling, such that the class parameters were estimated using the expectation maximization algorithm. Figure 15 compares the error in the abundance estimates of one of the classes using the two methods. In this example, the stochastic compositional model reduces the abundance estimation error by a factor of two to three. Work is ongoing to compare the performance of detection algorithms emanating from the segmentation, linear unmixing, and stochastic compositional models.

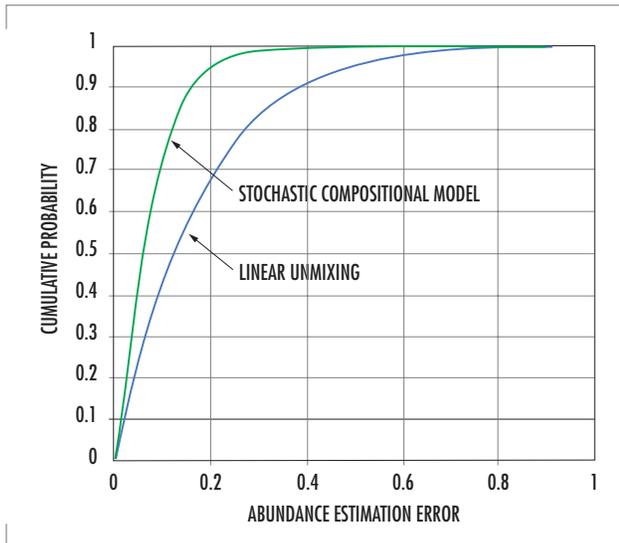
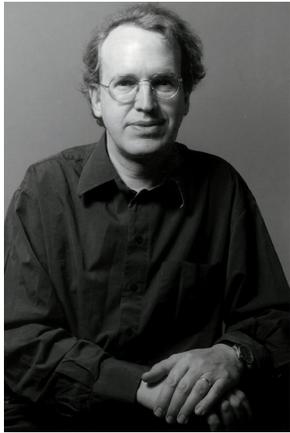


FIGURE 15. A comparison of the absolute error in the abundance estimate using linear unmixing and stochastic compositional modeling.

SUMMARY

SSC San Diego has been involved in many aspects of hyperspectral imaging. We are making important contributions in the areas of real-time processing implementations, system design for a variety of missions, environmental characterization, and the development of new models and methods. SSC San Diego is continuing to work across the Department of Defense (DoD)/Intelligence communities to bring mature hyperspectral technologies to the warfighter, making this unique source of critical information more widely available and user friendly.



David Stein

Ph.D. in Mathematics, Brandeis University, 1986

Current Research: Multidimensional statistics; detection theory; hyperspectral algorithms; remote sensing of littoral processes.



Jon Schoonmaker

BS in Physics/Mathematics, University of Oregon, 1985

Current Research: Hyperspectral systems; data analysis and algorithm development; remote sensing of littoral processes.



Eric Coolbaugh

MS in Oceanography and Meteorology, Naval Postgraduate School, 1989

Current Research: Hyperspectral imaging systems; high-performance computing; hyperspectral algorithms.

Surface Plasmon Tunable Filter for Multiband Spectral Imaging

Stephen D. Russell, Randy L. Shimabukuro,
Ayax D. Ramirez, and Michael G. Lovern

SSC San Diego

Yu Wang

Jet Propulsion Laboratory

BACKGROUND

An important aspect of theater missile defense is the multiband spectral characterization of plume radiation during the boost phase of a missile. Current Ballistic Missile Defense Organization (BMDO) plans call for study of the utility of a dual-mode ultraviolet (UV) and mid-wave infrared (MWIR) seeker. Combining the conventional MWIR sensor with shorter wavelengths provides increased information content for the image and can aid in optical target characterization. However, even dual-mode seekers have potential problems. Onboard optical seekers are subject to some vehicle self-interference. Sources of optical interference include out-gassing of vehicle contaminants, and by-products of the vehicle plume and attitude control systems, especially if solid aluminized propellants are used. Carbon particles are commonly present in the exhaust plume of kerosene liquid-oxygen (LOX) motors used by Atlas-type rockets. Once formed, carbon may contribute a continuum-like feature to the optical radiation of a rocket exhaust plume, especially in the near-UV [1]. A carbon monoxide–oxygen chemiluminescence mechanism may also be a source of radiation for the Atlas propellant because carbon dioxide is a large plume exhaust species and atomic oxygen is formed in the shear layer of the plume where the ambient oxygen molecules are dissociated [2]. Such optical interference effects lead to an increased background radiation level for the seeker in all spectral bands, but are most problematic in the infrared. Sensor confusion may also be caused by deliberate countermeasures. Therefore, multi-spectral imaging is important for ground-based imagery for optical signature characterization and onboard seekers.

One approach for multi-spectral imaging uses an imaging spectrometer that acquires images in many contiguous spectral bands simultaneously over a given spectral range. By adding wavelength to the image as a third dimension, the spectrum of any pixel in the scene can be calculated. These images can be used to obtain the spectrum for each image pixel, which can identify components in the target. The most common method of image spectroscopy is changing fixed dichroic filters. Existing systems suffer from large size and weight and operate slowly (approximately a millisecond). Several tunable filters have been proposed, but they all have severe problems. For example, the acousto-optic tunable filter is power-hungry (in kilowatts), while the liquid crystal tunable filter is slow (approximately tens of milliseconds for nematic liquid crystals) and has low efficiency.

ABSTRACT

The SSC San Diego Advanced Technology Branch and the Jet Propulsion Laboratory have been developing a novel technology that can be applied to multiband imaging. The surface plasmon tunable filter (SPTF) uses color-selective absorption by a surface plasmon at a metal-dielectric interface to achieve its optical selectivity. If an electro-optic material is used as the dielectric and a voltage is applied to change the surface plasmon resonance, the reflected light can be modulated, i.e., the photons at surface plasmon resonance will be absorbed and the photons out of the resonance will be totally reflected. Therefore, the applied voltage controls the reflection spectrum, and an electrically tunable color filter is formed. This paper details progress in developing SPTF technology as a replacement for discrete filters. This technology will allow multi-band or hyperspectral imaging with a single filter/camera system.

The Advanced Technology Branch at SSC San Diego and the Jet Propulsion Laboratory (JPL) have been developing a novel technology that can be applied to BMDO's needs for multi-spectral imaging. The surface plasmon tunable filter (SPTF) described in this paper uses color-selective absorption by a surface plasmon at a metal-dielectric interface to achieve its optical selectivity. If an electro-optic (EO) material is used, an applied voltage can control the resonant frequency of the surface plasmon, and an electrically tunable color filter is formed [3, 4, 5, and 6]. The technology may replace discrete filters and allow for multi-spectral or hyperspectral imaging with a single filter/camera system. This feature is particularly important if minimal payload weight and volume is desired for imager or seeker systems on rockets or missiles.

SURFACE PLASMON TUNABLE FILTER

The surface plasmon (SP) has been studied since the 1960s. It is a collective oscillation in electron density at the interface of a metal and a dielectric [7]. At SP resonance, the reflected light vanishes. This resonance is attenuated total reflection and depends on the dielectric constants of the metal and the dielectric. If an EO material is used as the dielectric and a voltage is applied to change the SP resonance condition, the reflected light can be modulated [8 and 9]. Using this principle, an SP spatial laser light modulator with a contrast ratio greater than 100 has been reported [10]. If we consider the SP light modulator in frequency space, the photons at the SP resonance frequency will be absorbed by the free electrons in the metal, and the photons away from the SP resonance will be totally reflected. If a voltage is applied to the EO material, the resonance frequency will change, and a tunable filter is formed. The SP tunable notch filter was invented based on this voltage-induced color-selective absorption [11 and 12]. Figure 1 schematically shows a reflective-mode SPTF.

The structure of the SPTF in Figure 1 shows white light incident on the metal-EO interface using a high-index prism (SF57 glass) for coupling. The color of the reflected light is determined by the SP resonance that is a function of the dielectric properties of the materials. Using a thin (55-nm) layer of silver and a liquid crystal (Merck E49) as the EO material, a narrowband SP resonance is obtained (Figure 2). Note that as the applied voltage is increased from 0 to 30 V, the SP absorption shifts from red to violet.

Figure 3 shows a symmetric geometry of metal/EO/metal used to form a transmissive filter. Two high-index glass prisms are used for the coupling with a thin metal film evaporated on each prism, and an EO material sandwiched between the two prisms. The thickness of EO material layer is less than 1 wavelength. When an SP wave is excited on one side of the metal/EO material interface by the incident

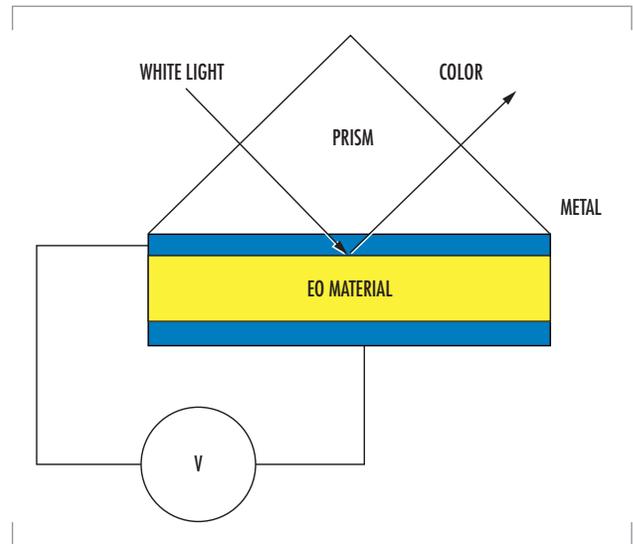


FIGURE 1. Reflective SPTF.

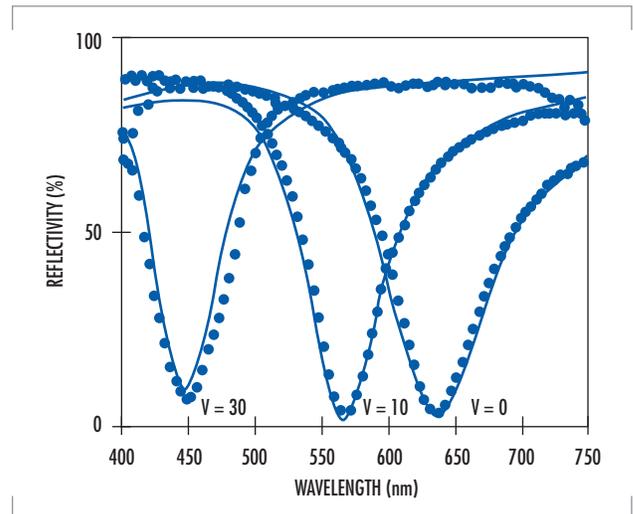


FIGURE 2. SPTF reflection spectra.

photons, the energy of the resonance photons convert into the motion of free electrons of the metal film. The optical field penetrates the thin EO layer and excites another SP wave with the same frequency at the other EO/metal interface because of the symmetric structure. The resonance photons will then re-radiate out as transmitted light. When a voltage is applied to the EO material, the index of the EO material changes, leading to a change of the SP resonance frequency and the transmission spectrum. Theoretical calculation shows that for two silver films separated by a 150-nm EO material layer (Merck E49), a change in the index of the EO layer from 1.5 to 2.0 leads to transmission peak shifts from 450 to 650 nm.

Varying the thickness of the dielectric layer between the two metal films can also change the coupling mechanics. Using a symmetric geometry similar to what was used in Figure 3, a SPTF can be constructed using a changeable air gap to select the spectrum.

Figure 4 shows the theoretical calculation of reflectivity vs. wavelength of the Air Gap SPTF and its effective tuning ability. Using silver as the metal films, when the thickness of the air gap changes from 300 to 5000 nm, the peak transmission shifts from 400 to 1600 nm. Though the structure of the Air Gap SPTF is schematically similar to the Fabry-Perot filter, the physics is totally different. The photons are incident at an angle greater than the critical angle, and two metal films must be used to generate the SP resonance. Furthermore, the tunable range runs from 400 to 1600 nm and is not limited by 2X as the Fabry-Perot filter requires. The SPTF can also be configured to operate based on angle of incidence [13].

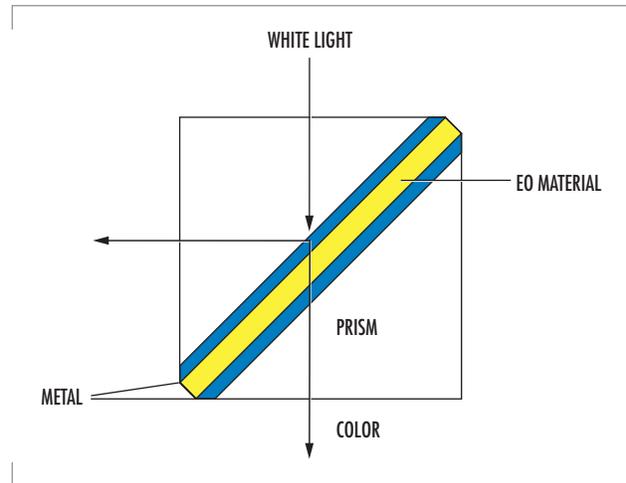


FIGURE 3. Transmissive SPTF.

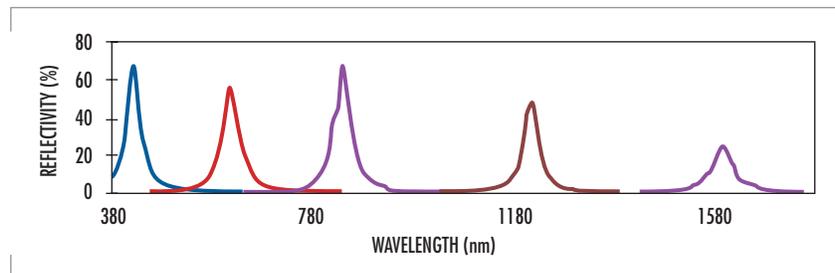


FIGURE 4. Tuning of the Air Gap SPTF.

FUTURE ADVANCES

A major advantage of SPFT technology is the ability to integrate it with various optical sensors and detectors. These products include state-of-the-art miniature photo-multiplier tubes available commercially (e.g., Hamamatsu R5600), microelectronic photo-multipliers [14 and 15], and solid-state detectors such as charge-coupled devices (CCDs) and active pixel sensors [16]. Compared with an acoustic-optic tunable filter and liquid crystal tunable filter, the SPTF is lightweight, low-power, and works in a wide temperature range. If the glass material is chosen so that its thermal expansion matches the thermal expansion of the EO material, this device works in a wide temperature range (-200 to +200°C. Though liquid crystal material was used in these experiments, the liquid crystal material can be replaced by solid-state EO materials such as potassium di-hydrogen phosphate (KDP), potassium titanyl phosphate (KTP), ethylene oxide (EO) polymers, organic crystals, and organic salts. If a solid-state material is used, the SP modulator can reach very fast (less

than 1- μ s) modulation speeds. Materials optimized for near-infrared (IR) and mid-IR can also optimize the device for specific applications. Such devices can be used for multi-spectral and hyperspectral imaging, for chemical analysis, and in surveillance and reconnaissance.

ACKNOWLEDGMENTS

The research described in this paper was supported by the SSC San Diego In-house Laboratory Independent Research Program and in part by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA).

AUTHORS

Randy L. Shimabukuro

Ph.D. in Applied Physics, University of California at San Diego, 1992

Current Research: Microsensors; photonic devices; optoelectronics; microelectromechanical systems (MEMS).

Ayax D. Ramirez

MS in Physics, San Diego State University, 1991

Current Research: Microsensors; photonic devices; optoelectronics; microelectromechanical systems (MEMS); laser applications.

Michael G. Lovern

BS in Electrical Engineering, University of Arizona, 1985

Current Research: Optical target characterization; advanced optics and detectors; laser systems and applications.

Yu Wang

Ph.D. in Physics, University of Toledo, 1992

Current Research: Optoelectronic devices.

REFERENCES

1. Levin, D. A. 1997. "Modeling of Optical Target Characterization from High Temperature Hypersonic Flows," unpublished proposal.
2. Slack, M. and A. Grillo. 1985. "High Temperature Rate Coefficient Measurements of CO+O Chemiluminescence," *Combustion and Flame*, vol. 59, p. 189.
3. Wang, Y. 1995. "Voltage-Induced Color-Selective Absorption with Surface Plasmons," *Applied Physics Letters*, vol. 67, p. 2759.
4. Wang, Y., S. D. Russell, and R. L. Shimabukuro. 1997. "Surface Plasmon Tunable Filter and Spectrometer-on-a-Chip," *Proceedings of SPIE*, vol. 3118, p. 288.
5. Wang, Y., S. D. Russell, and R. L. Shimabukuro. 1998. "Electronically Tunable Mirror with Surface Plasmons," *Proceedings of SPIE*, vol. 3292, p. 103.
6. Wang, Y. 1977. "Electronically Tunable Color Filter with Surface Plasmon Waves," *Proceedings of SPIE*, vol. 3013, p. 224.
7. Raether, H. 1980. *Excitation of Plasmons and Interband Transitions by Electrons*, monograph in series: *Springer Tracts in Modern Physics*, vol. 88, Springer-Verlag, Berlin, Germany.
8. Wang, Y. and H. J. Simon. 1993. "Electrooptic Reflection with Surface Plasmons," *Optical and Quantum Electronics*, vol. 25, p. S925.



Stephen D. Russell

Ph.D. in Physics, University of Michigan, 1986

Current Research: Microsensors; photonic devices; optoelectronics; microelectromechanical systems (MEMS); laser applications.

9. Schildkraut, J. S. 1988. "Long-Range Surface Plasmon Electrooptic Modulator," *Applied Optics*, vol. 20, p. 1491.
10. Caldwell, M. E. and E. M. Yeatman. 1992. "Surface Plasmon Spatial Light Modulators Based on Liquid Crystal," *Applied Optics*, vol. 31, p. 3880.
11. Wang, Y. 1996. "Surface Plasmon High Efficiency HDTV Projector," U.S. Patent #5,570,139.
12. Russell, S. D., R. L. Shimabukuro, and Y. Wang. 1998. "Transmissive Surface Plasmon Light Valve," U.S. Patent #6,122,091.
13. Ramirez, A. D., S. D. Russell, and R. L. Shimabukuro. "Resonance Tunable Optical Filter," Patent Pending, Navy Case No. 79,095.
14. Shimabukuro, R. L. and S. D. Russell. 1993. "Microelectronic Photomultiplier Device with Integrated Circuitry," U.S. Patent #5,264,693.
15. Shimabukuro, R. L. and S. D. Russell. 1994. "Multilayer Microelectronic Photo-multiplier Device," U.S. Patent #5,306,904.
16. Fossum, E. R. 1995. "Low Power Camera-on-a-Chip Using CMOS Active Pixel Sensor Technology," *1995 Symposium on Low Power Electronics*, 9 to 10 October, San Jose, CA.



Knowledge Base Formation Using Integrated Complex Information

Douglas S. Lange
SSC San Diego

INTRODUCTION

Command and control involves three fundamental processes that fit together in a tight cycle. Situation analysis provides the context on which to act. Decisions are made based on analysis results. These decisions constitute planned movements, engagement orders, and many other possible actions. Decisions must be communicated to those who are to carry out the actions. The results of these actions are observed as part of a new situation analysis.

As command, control, communications, computers, intelligence, surveillance, and reconnaissance (C⁴ISR) systems have evolved, system integration has been the general theme. Stand-alone systems, each with its own database, were first interfaced to allow some data transfer. Data management schemes provide some consistency among databases and operational units. System federation gradually allowed multiple applications to run on users' workstations, preventing the need for specialized hardware and support software for large numbers of individual systems. The current state of system integration not only allows multiple applications to share hardware, operating system, and network platforms, but also uses a layered service architecture that eliminates redundancy of some capabilities.

The evolution of system integration has broadened the stovepipes that were so narrow in previous system generations. The resulting view is of a few broad systems made up of many small applications, any of which may be accessible through the user's workstation. Some applications work on common data managed through centralized services. Many data categories still form separate stovepipes since they are maintained in separate data repositories because of their differing technical natures and programmatic backgrounds. Users must associate the tactical situation shown in one application with the results of a logistical query conducted through another application.

Information Complexity

The focus on systems integration ignores the true goal in decision support. Information is of ultimate value to the decision-makers. Integrating the information is the next step. Unlike data-warehousing applications, military information is not just collecting and crunching sales and inventory figures from various branch offices. The military environment is complex. The variety of concepts, events, and situations that can be

ABSTRACT

An intelligence support system has been developed using open hypermedia architecture. This approach integrates information from distributed disparate sources into a knowledge base. A public interface supports access by external applications. Filtering and change detection functions have also been implemented. The approach has shown promise in multiple domains, indicating possible wide application. This paper discusses the principles of the hypermedia framework for this system and how these principles may influence command, control, communications, computers, intelligence, surveillance, and reconnaissance (C⁴ISR) systems in general.

described subjectively or measured and reported objectively is probably limitless. No ontological study can *a priori* determine all of the possible data types needed to describe the military environment. Therefore, bringing all data into a relational or object database will not completely accomplish information integration.

Pattern of Analysis

In researching the requirements for an intelligence support system for the U.S. Defense Intelligence Agency (DIA), a pattern of analysis was uncovered that was common to those used in some other domains. The primary feature of this pattern is that an analyst's role is to create associations among existing data. Analysts rarely create data, but search, filter, and review all available information. As they do, they form networks of related information [1].

DIA intelligence analysts spend some of their time building up a private model of their area of expertise. They spend the rest of their time responding to queries from DIA's various customers. The responses are typically linear essays. Analysts also periodically produce background reports on particular matters of interest. These reports also take a strictly linear, book-like form, even when delivered over a computer network.

Analysis of the current approach yielded the following problems:

- Products were static or updated using a paper publishing schedule.
- Customers with local information have no mechanism to share it with others.
- Only a particular question was answered, even if it was not the correct question.
- Analyst turnover causes a large loss of knowledge.

As a result of these insights, work was initiated to find a way of recording the knowledge built by the intelligence analyst and communicating this knowledge to intelligence consumers. The goal was to move away from the linear essay to a more collaborative communications method. This method would allow for continuous update of the knowledge jointly held between the intelligence agency and its customers.

Recording Decisions

Decisions also take the form of associations among data or information elements. A classic example may be the order for a surface combatant to engage a hostile aircraft. The decision-maker did not create the aircraft or the positional and attribute data known about that aircraft. Likewise, the decision-maker did not generate the information related to the surface combatant. The value added by the decision-maker is that an engagement relationship (perhaps with other amplifying information) should exist between the two.

As the data on the two combatants changes, the association must be reviewed, but is not necessarily invalidated. Likewise, a reversal of the decision changes the relationship among the combatants, but does not change any individual data. This fundamental distinction between the structural representation of the associations among concepts or real-world objects and the content that describes them is common between the knowledge created by analysts and decision-makers.

HYPERMEDIA ARCHITECTURES

Hypermedia systems automate the management of information that is structured as described previously. Such systems provide the capability to work with a wide variety of data, while using the powerful information available through the structures created by the connections made among the various data items [2]. Hypermedia accurately records information, but its non-linearity allows the reader to access information in ways that the author did not necessarily expect. Users of analysis results can make new discoveries from the same body of data [3]. Likewise, distribution of responsibilities in a large command and control environment is aided by ensuring that not all uses of the data must be preconceived, though accurate representation of constraints is essential.

The basic features of most hypermedia systems are as follows:

- *Node*. A node is an object that represents a document or some other media element.
- *Link*. Links create relationships among nodes.
- *Anchor*. Anchors connect nodes to the actual media that make up their content.

Open Hypermedia

From 1987 to 1991, researchers noted that the hypermedia systems did not support the needs of collaborative work groups and could not be integrated into computing environments used in large enterprises [4 and 5]. Requirements were found for hypermedia systems that were not addressed. These requirements included the following:

- Interoperability to access and link information across arbitrary platforms, applications, and data sources.
- Link and node attributes to record the author of a link, what the permissions are for the particular link or node, and other management information.
- Anchors that allow attachment to the exact data desired.
- Link types to provide more information about the meaning of a particular link and what functions the link is intended to support.
- Public and private links to support collaborative environments.
- Templates for automating routine analysis tasks.
- Navigational aids that can act as filters and supply powerful querying mechanisms.
- Configuration control so that information important to the analysis effort can be developed and managed in hypertext.

To address these requirements, open hypermedia systems evolved. Open hypermedia systems have been defined as those that exhibit the following characteristics [6]:

- A system that does not impose any markup on the data. By marking up data to create hyperlinks, the data are changed, making the data inaccessible to systems that cannot handle the markup.
- A system that can be integrated with any tool that runs under the host operating system. This can be extended to mean a system that can be integrated with distributed object environments.
- A system in which data and processes may be distributed across a network and hardware platforms.

- A system in which there is no artificial distinction between readers and authors. This requirement is quite important for systems supporting analysis.
- A system to which new functionality can be easily added.

Since analysts and decision-makers are simultaneously readers and authors of node contents and links, these characteristics are vital in an information support environment. Likewise, the ability to link objects without changing them is critical. The information linked together by the analysts may be coming from other applications and databases integrated with the hypermedia system. These applications will not understand changes imposed on the data to support linking. The links must be separated from the content. This separation is the basic premise of an open hypermedia system. It has been demonstrated in many research systems [7].

The prototypical open hypermedia system is structured as shown in Figure 1.

Graph-Based Hypermedia

Several other hypermedia system types contribute capabilities necessary to support analysis functions. Chief among these is graph-based hypermedia. Graph-based hypermedia are based on set and graph theory, providing mathematically defined filter, search, and navigation methods. This category of hypermedia also includes human-computer interaction methods featuring graphical depictions of the hypermedia.

The idea of a schema made of node and link types provides the basis for much of this method's power [8]. The relationships among schema types and between schema entries and the instances created in the hypermedia closely mirror the relationships in object-oriented design.

One result of the typing found in graph-based hypermedia systems is that the resulting hypermedia forms a semantic network. Semantic networks are used to model concepts and real-world situations, making them a natural tool for modeling a tactical situation or the results of intelligence analysis.

Another result is that sophisticated filtering mechanisms can be defined. Graph-based hypermedia provide the concept of a perspective. A perspective contains three elements. The first element is the perspective pattern. A perspective pattern is a hypergraph that is a subset of the schema hypergraph. The second element is a filter, which is a constraint on the instance set. The filter may constrain either through the node and link attributes, or the content attached through the anchors. Finally, a subset of the instance set satisfies both of the constraints.

HYPEROBJECT PROCESSING SYSTEM

The design of the HyperObject Processing System (HOPS) inherits features from both open hypermedia systems and graph-based hypermedia systems. Some modification to the established research architectures was

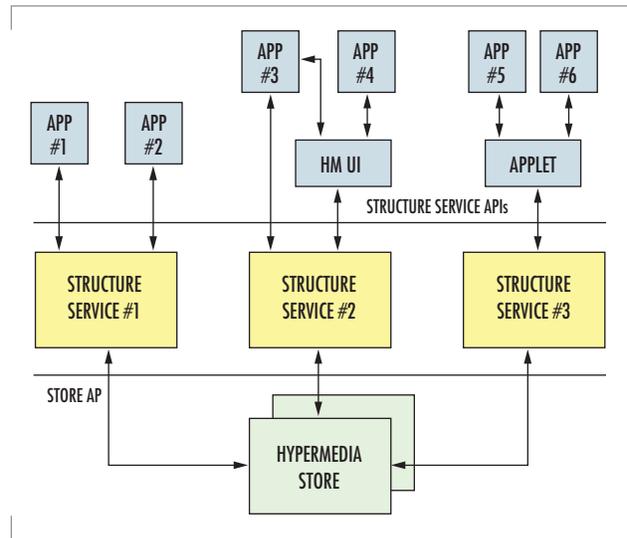


FIGURE 1. Open hypermedia architecture [9].

required to support analysis of the kind performed by DIA. These same modifications would appear to be important for related C⁴ISR systems.

General Architecture

HOPS follows the open hypermedia form with the architecture shown in Figure 2.

In this figure, circles labeled "RT" represent runtime applications supporting a user directly or automated processing. "HOMIS" (HyperObject Multimedia Information Systems) are modified multimedia information systems [8] that can handle hypergraphs rather than simple graphs [10]. HOMIS function as structure servers, as called for in generic open hypermedia systems; however, they provide graph-based hypermedia functions. Each HOMIS has a schema and instance set. Perspectives ("P" in Figure 2) and filters can be defined, and graph-based navigation interactions are possible. "ORB" represents an object request broker, in our case, supporting Java Remote Method Invocation. Object request brokers allow the system to be distributed over multiple platforms.

Unique Hypermedia Features

Most hypermedia systems found in research literature work with information spaces constrained by either the level of diversity and quantity of the information, or by restrictions on the structure of information, or by limited change of the underlying data. Several aspects of HOPS are unique among hypermedia systems. The features are necessary to allow HOPS to handle the dynamic unbounded nature of military information integration.

Multiple Anchors

The middle layer of HOPS holds the semantic network. Classical hypermedia systems use a node to represent a piece of media and anchor to a single media element to provide content. A semantic network forms that describes the relationships among media elements rather than the tactical situation. To remedy this problem, HOPS uses multiple anchors per atomic node. Use of multiple anchors allows the nodes to define concepts or real-world objects and allows the links to represent relationships among them rather than relationships among the content elements.

Large Open-Ended Schema

Schemas imply an ability to predict all the types of information to be used and the entire range of associations that will exist among the elements. In some domains this is possible, but not in the military information domain [1]. An example can be demonstrated in terms of exercise plans. During Tandem Thrust 97, one of the primary requirements concerned protecting the Great Barrier Reef. Environmental mitigation strategies and environmental reports are not typically found in the command and control systems of our armed forces. There will always be unpredicted situations in warfare and military exercises. Information systems must adapt on the fly to allow analysts and decision-makers to see

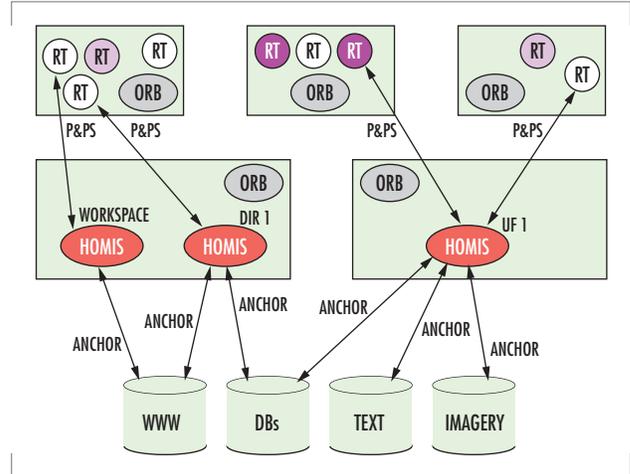


FIGURE 2. HyperObject Processing System.

and interpret information and record and inform regarding decisions. The HOPS design allows users to include information not accounted for in the schema through the object-oriented method of deriving all nodes and links from common ancestors. This allows users to bypass rules in the schema and connect nodes and links in ways not previously predicted. The user or an administrator can then update the schema on the fly to allow autonomous tools to process the information more easily.

Analysis schemas and instance sets can become quite large. The problems modeled are quite complex. The size of the schema represents the complexity of the model while the size of the instance set represents the quantity of information. Consumers of the analysis model must filter both in terms of the complexity and in terms of the size of the knowledge base that they work with to avoid being overwhelmed. HOPS allows this capability through adaptations of the graph-based hypermedia concepts of perspective patterns and filters [10]. Perspective patterns allow the user to limit the kinds of information being worked with, while filters focus attention on information with particular content.

Link and Anchor Integrity

When important decisions are being made based on the information presented, error is less tolerable than in our daily workings with the World Wide Web. Anchored content must not disappear unexpectedly.

Likewise, if content changes, the model must be re-evaluated to determine if it is still valid. The typed links of the storage layer must also be carefully managed to prevent dangling links. HOPS accomplishes these goals by caching anchored content and providing periodic checks using an autonomous change detection agent. Agents used for this purpose can use whatever rules suit the application.

Link Equality

Although hypermedia relies on associations between elements for its character, many interaction techniques found in research literature still focus on the content (e.g., string matching filters and searches, searches on images). Links are primarily used for navigation. This may be because, in many applications, links are addresses used to point to more information, or typed paths to get to related nodes. Since the primary value added by intelligence analysts and decision-makers is found in associations among elements, authors and readers of the products will want the ability to interact with typed links in ways other than simply using them for navigation. They themselves provide critical information. HOPS handles this by making links special types of nodes, allowing all the mathematics of filtering, searching, and browsing to work on links. [10].

Framework

HOPS itself is not a command and control system or an analysis system. HOPS is a hypermedia framework designed to support analysis and to provide some generic applications for interacting with the hypermedia. HOPS is intended to be used by adding domain-specific applications along with an initial schema to create an analysis system of the type needed. Such work is in support of DIA's mission.

In the Military Operations in the Built-up Areas project, HOPS was integrated with the Lightweight Extensible Information Framework (LEIF) to provide geographic and temporal views of the hypermedia.

An intelligence product creation wizard and intelligence-specific anchors were also used. Together with the generic applications within the framework, users have a variety of ways to work with the information.

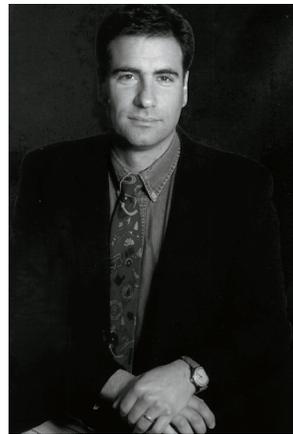
PROSPECTS FOR INFORMATION INTEGRATION

Hypermedia systems hold promise for information integration. Any number of decision-support tools can access the semantic network formed of the associations and nodes. Decision-makers can have access to all the information they need because the hypermedia can be made from information elements from all available systems. While the semantic network is serving higher level decision tools, the content is left untouched, and is still accessible by those tools that interact directly with content databases.

Beyond executing applications from a single workstation, integrated information could provide decision-makers with a competitive advantage. An integration method that brings the information into a semantic network can allow meaningful access to human beings and autonomous agents. The goal of command and control systems should be to integrate information rather than just the applications. Architecture such as that used for HOPS, centered on the structure of information, can accomplish this goal. Military plans, tactical situations, and their interaction can be described using hypermedia-induced semantic networks.

REFERENCES

1. Lange, D. 1999. "Hypermedia Potentials for Analysis Support Tools," *Proceedings of Hypertext '99*, Association for Computing Machinery (ACM), pp. 165-166.
2. Nurnberg, P., J. Leggett, and E. Schneider. 1997. "As We Should Have Thought," *Proceedings of Hypertext '97*, ACM, pp. 96-101.
3. Nielson, J. 1990. *Hypertext and Hypermedia*, Academic Press, San Diego, CA.
4. Halasz, F., 1987 "Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems," *Proceedings of the ACM Conference on Hypertext*.
5. Malcom, K., S. Poltrock, and D. Schuler, 1991. "Industrial Strength Hypermedia: Requirements for a Large Engineering Enterprise," *Proceedings of the Third ACM Conference on Hypertext*.
6. Davis, H., W. Hall, I. Heath, G. Hill and R. Wilkins. 1992. "Towards an Integrated Information Environment with Open Hypermedia Systems," *Proceedings of Hypertext 1992*, ACM, pp. 181-190.
7. Wiil, U. 1997. "Message from the OHSWG Chair," Open Hypermedia Systems Working Group Web Site, <http://www.ohswg.org/intro/chair.html> (December).
8. Lucarrela, D. and A. Zanzi. 1996. "A Visual Retrieval Environment for Hypermedia Information Systems," *ACM Transactions on Information Systems*, vol. 14, no. 1 (January), pp. 3-29.
9. Wiil, U. and P. Nurnberg. 1999. "Evolving Hypermedia Middleware Services: Lessons and Observations," *Proceedings of the 1999 ACM Symposium on Applied Computing*, pp. 427-436.
10. Lange, D. 1997. "Hypermedia Analysis and Navigation of Domains," Master's Thesis, Computer Science Department, Naval Postgraduate School, Monterey, CA.



Douglas S. Lange

MS in Software Engineering, Naval Postgraduate School, 1997
Current Research: Software generation; knowledge bases; enterprise architectures.

A Real-Time Infrared Scene Simulator in CMOS/SOI MEMS

Jeremy D. Popp, Bruce Offord, and Richard Bates
SSC San Diego

H. Ronald Marlin and Chris Hutchens
Titan Systems Corporation

Derek Huang
Advanced Analog VLSI Design Center

INTRODUCTION

The objective of the real-time infrared (RTIR) project is to develop a reliable prototype infrared (IR) test set for use in calibration and testing of IR systems, including built-in-test to ensure the real-time reliability of IR sensing systems. The potential of RTIR as built-in-test equipment (BITE) is to improve the reliability of IR sensors, thus lowering the overall system cost of operation. Infrared scene simulators that use bulk complementary metal-oxide semiconductor (CMOS)/micro-electromechanical systems (MEMS) have been reported previously [1]; however, this work uses silicon-on-insulator (SOI) as the starting material. The MEMS area is scaled down to create higher density pixel arrays, with low leakage at higher temperatures.

DESIGN

The integrated circuit (IC) consists of a data input block, address write control, and pixel-specific electronics including a microheater suspended over a micromachined cavity in the silicon substrate. The display IC consists of an array of 64 x 128 thermally isolated, resistive emitters. The thermal pixel array (TPA) elements have response times less than a millisecond, making them suitable for real-time scene simulation. The pixel cell contains a resistive heater element (or infrared emitter), a storage capacitor, pixel drive transistors, and switches (Figure 1). The user digitally specifies a specific row and column and then writes a pixel voltage to the desired cell via the analog multiplexer (MUX). The infrared pixel array IC is designed for use with a computer or an electronic controller to service or update the real-time images. The computer sends gray-scale scene data to the pixel array in the form of voltages, which the TPA displays as a gray-scale image. The computer controls digital row and column address lines and writes the analog inputs via a digital-to-analog converter (DAC) to the RTIR IC. The voltage magnitude reflects the desired IR intensity of the pixel element, thereby achieving the gray-scale levels. After writing to the pixel, the desired voltage is stored dynamically by Chold, producing the desired IR pixel intensity while the remaining pixels are updated. The pixel electronics of the array are designed to exploit the low leakage properties of SOI during high-temperature operation. Voltage droop is the greatest problem affecting pixel dynamic range and accuracy. Droop is primarily a result of the leakage currents through

ABSTRACT

A 64 x 128 real-time infrared (RTIR) complementary metal-oxide semiconductor (CMOS)/silicon-on-insulator (SOI) scene generation integrated circuit (IC) is described. The RTIR IC offers real-time dynamic thermal scene generation. This system is a mixed-mode design, with analog scene information written and stored into a thermal pixel array. The design uses micro-electromechanical sensors (MEMS) in conjunction with SSC San Diego's 0.8- μm CMOS/SOI process to develop a RTIR IC scene generator.

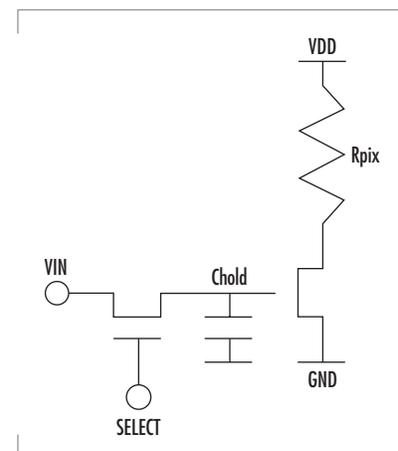


FIGURE 1. Pixel schematic: the drive transistor is a BTS device; the access transistor is an HGATE device.

the pn junctions of the sampling switch and secondarily a result of excessive channel leakage. As an option to further reduce droop, the designer can place a compensation pn junction by using half a negative-channel metal-oxide semiconductor (NMOS) transistor at the hold capacitor node.

FABRICATION

CMOS/MEMS technology is used as a technique to thermally isolate the infrared emitter microstructures from the substrate after the CMOS processing is completed. SSC San Diego's 0.8- μm CMOS partially depleted SOI process was selected to fabricate the array of electronically addressable 20 x 20 micron emitter elements (Figure 2). The process is a single poly, double metal, salicided process with a high-value resistor option of up to 1 Kohm/square. This allows modest density arrays, and, together with the high-value silicon resistor available in the 0.8- μm process, provides lower pixel current operation. The micromachined cavity is constructed by using a silicon etchant that undercuts the desired pattern in the silicon substrate, while leaving it electrically connected to create a suspended structure/microheater (Figure 3). This pattern is created by patterning and plasma etching silicon dioxide after the CMOS passivation, thereby exposing the substrate silicon of the CMOS chip. The exposed silicon is then exposed to a tetra-methyl ammonium hydroxide (TMAH) solution, an aniso-tropic silicon etchant. The TMAH etchant was chosen because, with the addition of silicic acid, it does not attack the exposed aluminum bonding pads [2].

RESULTS

The thermally isolated resistor emitter has been characterized using a calibrated blackbody and adjusting for fill factor using a method described in [3]. The temperature of the emitter as a function of voltage across the resistor is plotted in Figure 4, together with the current through the resistor. A maximum temperature of 262°C is achieved at a voltage of 8.25 V and a current of 0.85 mA.

SUMMARY

A 64 x 128 scene generator RTIR IC architecture has been described with each key component discussed. A MEMS device, the TPA, is produced using CMOS/SOI technology with post CMOS process etching.

ACKNOWLEDGMENTS

This work was funded by the Office of Naval Research, Physical Science Division, Dr. Phil Abraham.

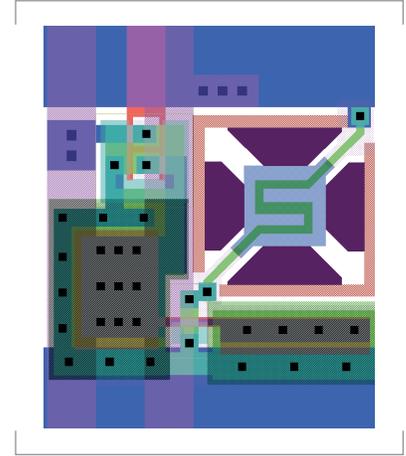


FIGURE 2. The heater element and pixel electronics layout.

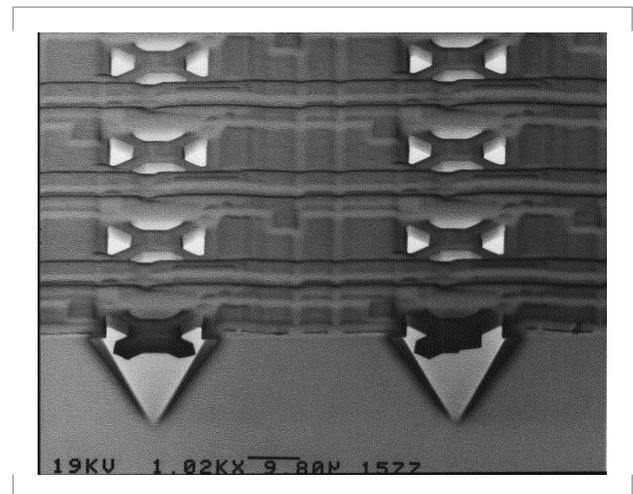


FIGURE 3. Scanning electron microscopy (SEM) of a cross-sectioned sample of suspended microheaters.

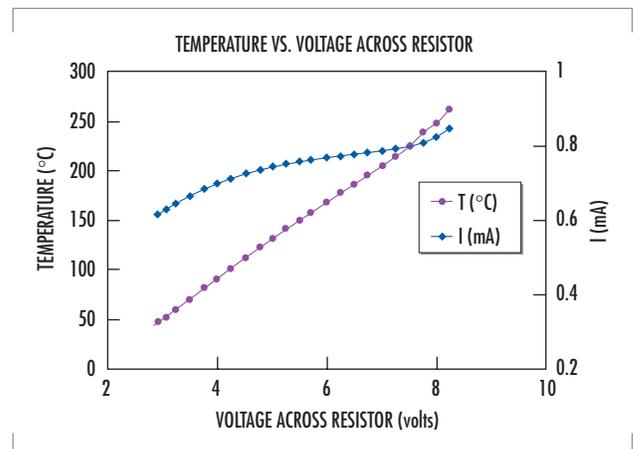


FIGURE 4. Pixel thermal response to applied voltage across resistor.

AUTHORS

Bruce Offord

BS in Engineering Physics, University of San Diego, 1985

Current Research: Very large-scale integrated (VLSI) SOI process development; novel IC design.

Richard Bates

BA in Physics, Loma Linda University, 1960

Current Research: Radiation-induced Irtran 2 absorption; polarization independent narrow channel (PINC) wavelength division multiplexing (WDM) fiber coupler fusing.

H. Ronald Marlin

BA in Physics, La Sierra University, 1959

Current Research: Infrared radiometry.

Chris Hutchens

Ph.D. in Electrical Engineering, University of Missouri, 1979

Current Research: Low-power, mixed-signal SOI CMOS and analog CMOS.

Derek Huang

MS in Electrical Engineering, Oklahoma State University, 2001

Current Research: Low-power, mixed-signal SOI CMOS.



Jeremy D. Popp

BS in Electrical Engineering, Portland State University, 1998

Current Research: Low-power, mixed-signal application-specific integrated circuit (ASIC) design; novel systems on a chip; reconfigurable computing.

REFERENCES

1. Parameswaran, M., A. M. Robinson, D. L. Blackburn, M. Gaitan, and J. Geist. 1991. "Micromachined Thermal Radiation Emitter from a Commercial CMOS Process," *IEEE Electron Device Letters*, vol. 12, no. 2 (February), pp. 57-59.
2. Tabata, O., R. Asahi, H. Funabashi, K. Shimaoka, and S. Sugitama. 1992. "Anisotropic Etching of Silicon in TMAH Solutions," *Sensors and Actuators, A.34*, pp. 51-57.
3. Marlin, A. H. R., R. L. Bates, M. H. Sweet, R. M. Carlson, R. B. Johnson, D. H. Martin, R. Chung, J. C. Geist, M. Gaitan, C. D. Mulford, E. S. Zakar, R. J. Zeto, R. Olson, and G. C. Perkins. 1997. "Real-time Infrared Test Set: Assessment and Characterization," *SPIE*, vol. 3084.

