Real Time Categorization of Arbitrary Data Streams as
Encrypted, Scrambled or Compressed, as
Voice, Text or Image

1. COVER SHEET (see attached)

2. IDENTIFICATION & SIGNIFICANCE OF THE OPPORTUNITY

The objective of this proposal is to demonstrate the feasibility of differentiating between arbitrary bit stream types in near real time. The real time categorization of an intercepted bit stream as encrypted data, scrambled data, or compressed data along with a determination of the type of data as voice, text, image or other; is viable and an extremely attractive ability in the intelligence environment. This ability has been marginally demonstrated and this effort will demonstrate the feasibility of making this capability robust and real time.

In the modern digital communication systems the ability to intercept a myriad of data streams has proliferated. In the vast sea of information being broadcast the intercepted signals containing useful intelligence can be lost. This vast amount of data available taxes even the ability to record data for later analysis. A real time ability to categorize a data stream as encrypted or non encrypted, as scrambled or non-scrambled or as voice, text, video or data; would be a tremendous asset for the retrieval of any data stream that contains useful intelligence.

In this effort RBI will refine and demonstrate a pattern recognition algorithm that can perform just such a discrimination on arbitrary bit streams. We will also perform an error analysis of this algorithm's performance on the varying word lengths found in modern communication links. Additionally we will make a comparative analysis of the previous (and ongoing) bit entropy features analysis that demonstrated reasonable promise to perform this bit stream categorization. The results of this effort will demonstrate the feasibility of using this specialized pattern recognition algorithm for completely discriminating arbitrary bit streams in real time, with very minimal amounts of data demodulated and collected. Such a demonstration will lead the way to greatly enhanced and selective digital SIGINT collection systems.

2.1 Background

2.1.1 Intelligence Data Escalation: There is an acute need for enhancements to intelligence gathering systems that enable them to gather, sort, record and report pertinent intelligence in near real time. This necessity is aptly demonstrated by the unchecked terrorist activities of 2001. Further *"Technology advances are enabling new satellite communications systems that combine broad band data rates with small terminals. These novel systems are being designed to provide affordable 'last-mile' network access to home and small business users world wide.*" The new availability of the internett's Transmission Control Protocol (TCP) over the advanced broad band satellite systems distinct from narrow band voice services greatly advances this proliferation of data available for intelligence interception. In the sea of information now available for interception careful selection of data that should undergo more careful scrutinization can greatly increase the probability of a useful intelligence interception.

2.1.2 Arbitrary Bit Stream Categorization.

A primary consideration in filtering through the sea of information available from digital communication channels is the categorization of an intercepted and reconstructed bit stream. Intercepted digital communication may be properly demodulated, properly decoded, and properly unscrambled into several hundred bit streams. These bit streams are arbitrary because of uncertainty in word length, asynchronous timing data, parity data or additional scrambling. Additionally it is uncertain if a bit stream represents encrypted or compressed data and the type of data further contained may be voice, text, image or data. Current software used for analysis of these bit streams

uses a trial and error approach to determine word length and timing/parity aspects followed by manually looking at the results to determine if the data may be usable. These trial and error methods cannot obtain real-time results. An algorithm that makes a determination of the category of bit stream will enable the real time processing and recording of only pertinent data by the appropriate processing tools.

2.1.3 Bit Stream Categorizor Mechanics

Significant research has gone into the problem of categorizing arbitrary bit streams using bit entropy. The usefulness of this method has been limited by the fact that the Shannon entropy calculation is word length and bit stream shift dependent. Although investigation of this dependence may prove useful in determining a bit streams word length or framing information the actual categorization of an arbitrary bit stream is accomplished by introducing additional parameters. In some initial investigation of pertinent parameters that may be useful I explored the following 14 parameters calculated on the whole 4 KByte bit stream sample or on that sample divided into 8 bit word lengths:

1 Shannon Entropy
2 Bit Stream Mean
3 Bit Stream Stdv
4 Bit Stream Min
5 Bit Stream Range
6 Histo Stdv
7 Histo Min
8 Histo Range
9 Sequence of 8
10 Sequence of 7
11 Sequence of 6
12 Number of Reps
13 Number of Reps 3
14 Number of Reps 4

A simple Euclidean distance vector comparison method was used to separate files into categories. This method was able to discriminate how a bit stream had been modified, i.e. compressed, encrypted or scrambled, and remarkable was often able to determine the data type, text, image, voice, etc., through the modification. Analysis of these parameters further demonstrated an intriguing ability to determine the type of scrambling algorithm that had operated on the original data. It is expected that further analysis of this vector of parameters through Eigen vector analysis and weighted Euclidean distance vector analysis will develop an algorithm that is robust enough to compensate for various word lengths and framing (bit slip) error. Remarkably, additional analysis of data from various encryption techniques may demonstrate an ability to discriminate between various encryption algorithms, just like it discriminated between various scrambling algorithms. An ability that would be of great value in the intelligence environment.

3. PHASE I TECHNICAL OBJECTIVES

The overall objective of the proposal is to develop the Euclidean Distance Vector Fitting Algorithm for categorizing arbitrary bit streams and investigate it's robustness against various word lengths and framing errors encountered in arbitrary bit streams. The specific objectives of the program can be enumerated as follows:

1. Investigate the various word lengths normally encountered in modern digital communication systems and develop sample sets of these word lengths with various data types and implementations.
2. Refine the calculations of the 14 vector parameters to provide robustness to investigate the dependence on word length and framing error.
3. Investigate various encryption algorithms normally encountered in modern digital communication systems and develop sample sets of these encryption's with various data types.

4. Refine and optimize the weighting methods used in the Euclidean Distance Vector Fitting Algorithm and explore the benefits of an Eigen value evlautation of the vector..

5. Evaluate the performance of the refined algorithm to discriminate between various data types and data encoding (encryption, compression or scrambling).

A previous objection of this methodology was that it took an empirical approach to the problem rather than an analytical approach. That remains true. This algorithm was developed completly with an empirical analysis of possible bit streams that may be encountered, and a theoretical analysis of what a scrambler or compression algorithm or encryption algorithm does to these 14 parameters has not been fully documented. In truth, this type of analysis borders on the impossible and is left as an exercise for the mathmetician. This approach has demonstrated the ability to differentiate between which of two algorithms scrambled test data. Let the mathmeticians 'go figure.' This algorithm may not solve every possible convergence of data types, but it has been empirically tested to categorize data types that have our keen interest. It is ready for prototype and deployment into operational intelligence platforms.

4. PHASE I - WORK PLAN

Phase I research will be restricted to refining the algorithm and showing feasibility of using this algorithm to categorize arbitrary bit streams.

The Phase I work plan will include the following tasks for achieving the stated objectives:

4.1 Task I - Investigate and Refine the Algorithms versatility

4.2 Task II - Optimize the Algorithm and Evaluate it's Effectiveness

4.3 Reporting

The Phase I work would follow the tentative schedule below:

Task	Days	Hrs	Start	Stop
SBIR Phase I	273.75	720	02/04/2002	11/04/2002
Kick Off	1	3	02/04/2002	02/05/2002
T1 Investigate W L Anomalies	56	100	02/05/2002	04/02/2002
T1 Refine Parameter Calculations	56	100	03/05/2002	04/30/2002
T1 Investigate Encryption Anomalies	56	100	04/02/2002	05/28/2002
T2 Optimize Algorithm	56	150	05/14/2002	07/09/2002
T2 Evaluate Algorithm Performance	56	150	06/11/2002	08/06/2002
Final Report	42	117	08/06/2002	11/04/2002

In addition to day-to-day informal contacts with the program monitor, monthly technical progress reports will be submitted with a complete Technical Progress Report being submitted at the end of each program year, as requested.

5. RELATED WORK

The Principal Investigator has been involved in pursuing categorizing arbitrary bit streams for the past six years. At his previous place of employment he was responsible for the programs that enhance the performance of numerous intelligence collection platforms that would be enhanced by this capability. He is in close contact with other researchers exploring intelligence gathering techniques and digital communications.

6. RELATIONSHIP WITH FUTURE RESEARCH OR RESEARCH AND DEVELOPMENT

Anticipated improvements in sorting and categorizing digital bit streams will be of immediate use where conventional intelligence collection techniques currently have limitations due to data overloading. The proposed Phase I work will determine the fundamental algorithm that best categorizes arbitrary bit streams. The prototyping and deployment of this

algorithm as a software package will be accomplished in the Phase II effort.

7. Commercialization Strategy

The use of digital communications and the necessity to sort through a sea of bit streams is proliferating in the commercial world. The trend is towards seeking minimized data latency and the ability to prioritize transmissions is looking towards accurately categorizing data awaiting transmission. Successful application of these data categorization algorithms may have limited use outside the intelligence collection arena, however there is application, and the need to perform intelligence interception of digital data is fast become a commercial interest.

8. KEY PERSONNEL

Edward G. Rice, Senior Engineer

EDUCATION:

M.S., Electrical Engineering, Air Force Institute of Technology, Wright Patterson AFB Ohio, March 1992.

B.S. Electrical Engineering, Ohio State University, Columbus Ohio, March 1982.

CURRENT POSITION AND RESEARCH:

Edward Rice is a retired USAF officer and the Sole Proprietor of RBI. He has more than 18 years of experience in USAF intelligence systems. RELEVANT EXPERIENCE: Prior to starting RBI Ed Rice was the consultant that developed and tested the basic Euclidean Distance Vector Fitting Technique for Categorizing Arbitrary Bit Streams. From 1989 to 1995 he was assigned to Rome Laboratories where he managed numerous intelligence collection technology initiatives. From 1984 to 1989 as a USAF Electrical Engineer he worked with numerous data collection and weapon delivery systems furthering and rounding out his background in intelligence systems.

He is currently teaching high school math and science part-time, while pursuing a M.Div. Degree.

9. FACILITIES/EQUIPMENT

To support its position as an analysis and algorithm development sole proprietorship, RBI currently maintains an office with several networked computer resources.

10. CONSULTANTS

No consultants are presently foreseen for the Phase I program. If a need should arise, RBI has several consultants available from previous contacts in this technical area.

11. PRIOR, CURRENT OR PENDING SUPPORT

RBI has no prior, current or pending support for a similar proposal.

12. COMPANY COMMERCIALIZATION REPORT (SEE ATTACHED)

13. COST PROPOSAL (SEE ATTACHED)

14. REFERENCES

1)

"SENSOR ACE Capabilities Enhancement Study", Appendix A, "Technology Survey and Recommendations" 6 December 96 submitted to AFRL/IFEC

2) "Technical Notes On Classifying Arbitrary Bit Streams Using Shannon Entropy, Statistical Distributions and Euclidean Distance Vector Fitting Technique" , by Edward G. Rice, Consultant Research Associates of Syracuse Inc. 15 September 1998 submitted to AFRL/IFEC

3) "TCP Performance over Satellite Channels" by Thomas R. Henderson and Randy H. Katz, University of California at Berkeley Dec 1999.

4) "Research Results for Classifying Arbitrary Bit Streams Using Bit Entropy Features" 10 September 1997 by Ed Semplinski, QuesTech, Inc. submitted to AFRL/IFEC

Technical Abstract

In modern digital communication systems the ability to intercept and categorize data is desirable for many disciplines. This effort refines a Euclidean distance vector algorithm which sorts arbitrary bit streams according to various data types and data security/transmission operations. The algorithm calculates 14 statistical parameters from a small sample of a demodulated, un-encoded data stream, and uses this vector of parameters to determine if the bit stream is open data, encrypted data, scrambled data or compressed data. A second evaluation of the same parameters has been able to determine the concealed data type as text, image, voice, or data. Additional evaluation of the same parameter vector has been able to determine which scrambler algorithm was used on the data. This effort is intended to demonstrate the feasibility of using this algorithm to perform multiple real time bit stream categorizations to enable the sorting and selection of digital communication intercepts for intelligence or monitoring purposes.

Anticipated Benefits/Potential Commercial Applications

of the Research or Development.

The potential commercial application of this effort is in the monitoring and statistical compilation of data type usage of a digital communications channel. The monitoring of what type of data is currently being transmitted via a communication channel would be useful in prioritizing communication resources or in evaluating data latency effects. A longer term prioritization of capital investments can be made based on the statistical usage information. The use of digital communications and the necessity to sort through a sea of bit streams is proliferating in the commercial world. The trend is towards seeking minimized data latency and the ability to prioritize transmissions ***th***is requires accurately categorizing data awaiting transmission. Successful application of these data categorization algorithms may have limited use outside the intelligence collection arena, however a current trend for intelligence interception of digital data is fast becoming a commercial requirement as well.

List of 8 Key Words that describe the Project. Digital Communication, Intelligence, Bit Stream, Digital Categorizing, Euclidean Vector, Encryption, Data Scrambling, Data Compression