SANDIA REPORT

SAND2011-7347

Unlimited Release

Printed September 2011

Final Report for LDRD Project 11-0029: High-Interest Event Detection in Large-Scale Multi-Modal Data Sets: Proof of Concept

Brandon Rohrer

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: <http://www.osti.gov/bridge>

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>

Final Report for LDRD Project 11-0029: High-Interest Event Detection in Large-Scale Multi-Modal Data Sets: Proof of Concept

Brandon Rohrer
Department 6533
Intelligent Systems, Robotics, and Cybernetics Group
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1010
rohrer@sandia.gov

Abstract

Events of interest to data analysts are sometimes difficult to characterize in detail. Rather, they consist of *anomalies*, events that are unpredicted, unusual, or otherwise incongruent. The purpose of this LDRD was to test the hypothesis that a biologically-inspired anomaly detection algorithm could be used to detect contextual, multi-modal anomalies. There currently is no other solution to this problem, but the existence of a solution would have a great national security impact.

This technical focus of this research was the application of a brain-emulating cognition and control architecture (BECCA) to the problem of anomaly detection. One aspect of BECCA in particular was discovered to be critical to improved anomaly detection capabilities: it's feature creator. During the course of this project the feature creator was developed and tested against multiple data types. Development direction was drawn from psychological and neurophysiological measurements. Major technical achievements include the creation of hierarchical feature sets created from both audio and imagery data.

1 Introduction

The problem of processing large amounts of multi-modal data is common to the defense, intelligence, energy, and homeland security communities, and is of particular importance when monitoring nuclear materials and watching for undeclared nuclear development activities. For many of these applications, the data processing problem involves finding rare, high-interest events. General, multi-modal contextual anomaly detection is a problem that has not been addressed in prior work. The strategy taken to find anomalies was to identify unexpected events, i.e. data observations that are not predicted either by other contemporary observations or by preceding observations.

2 Problem Statement

The problem of processing large amounts of multi-modal data is common to the defense, intelligence, energy, and homeland security communities. For many security applications, the problem can be reduced to one of finding rare, high-interest events. For instance, the vast majority of video from security cameras contains no content of interest, but certain rare events, such as a robbery, are of great interest. Rare high-interest events occur in all types of data. Seismic activity may indicate a large detonation. Spikes in radiation detection may point to uncontrolled nuclear material or a rogue nation's undeclared nuclear activities. Subtle changes in satellite imagery may be the result of an adversary's military activity. And inflammatory internet blog postings may presage a bout of terrorist activities. In all these cases, high-interest events are indicated by unusual patterns in the data, but the exact form of those patterns may not be known beforehand. The goal of this project is to test whether anomaly detection can identify such high-interest events. To do this, a proof of concept anomaly detection tool will be developed that is capable of detecting events like these in a wide variety of data types.

Often, the indicators of a high-interest event may be represented across several types of data. For instance a hostile intrusion into a secure facility may be indicated by movement on a surveillance camera, a patterned surge of activity from an array of infrared motion detectors, and rhythmic seismic activity indicating footfalls. The presence of any one of these sensor events in isolation might be due to some other cause, such as a wind-driven tumbleweed. But the simultaneous occurrence of them all increases the confidence that something of interest is happening. Data fusion, the combination of data from multiple sources, can make possible the detection of events that would otherwise go unnoticed. This project is focused on the broader and more difficult problem of multi-modal anomaly detection.

[1] presented a comprehensive review of the field of anomaly detection. They list a number of common applications for anomaly detection, including cyber intrusion, fraud detection, medical anomaly detection, industrial damage detection, image processing, textual anomaly detection, and sensor networks. There is a clear connection between these applications and a number of important security challenges, including national infrastructure monitoring, cyber

defense, site protection for critical assets, battlefield awareness, and electronic intelligence gathering.

Data mining is the application of machine learning to large datasets to answer questions. Various machine learning methods are used, according to the questions one would like to answer. Unsupervised learning (clustering) can be used to answer questions such as “What other websites have similar traffic patterns to Website X?” Supervised learning (including classification and regression) can answer questions like “Is Seismic Event Y natural or man-made?” Rule induction methods are able to answer questions in the form of “Are new computer viruses more likely to be released on Thursdays?” Anomaly detection seeks an answer to the question “Is this event unusual?”

There are a number of challenges inherent in anomaly detection.

- Distinguishing sharply between normal and anomalous events is difficult in general. The boundary between the two can be difficult to define precisely.
- When an adversary is the source of the anomaly, they may try to make it appear as normal as possible, making detection even more difficult
- Many systems vary over time, so the notion what is normal and anomalous must also change over time.
- There is no single definition of an anomaly that is applicable for all domains. Depending on the application, an anomaly might be indicated by an atypical value, rate of change of value, frequency of value occurrence, or sequence of values, or any combination of the above.
- The availability of data for training anomaly detection tools is often limited.
- Data is typically noisy, generating low-interest anomalies.

The combination of these factors makes the creation of a general anomaly detection tool challenging. Most anomaly detection techniques are crafted with a particular application in mind, and make assumptions about the nature of the data that will be handled, such as distributions, constraints, or known relations. The proposed approach is data type agnostic and can identify a wide variety of anomalies without limiting its scope to a single application. It is intended to handle cases where the data sources are not known beforehand, and additional data sources can be added to the data stream during the life of the system.

In the field of anomaly detection, there is a distinction between point anomaly detection and contextual anomaly detection. Point anomaly detection looks at individual data points, such as a measured automobile velocity of 123 miles per hour, and declares it to be atypical. Contextual anomaly detection determines whether the data point is atypical, given the data points the preceded it and co-occurred with it. For instance a daily high temperature in Albuquerque of 45 degrees is not anomalous in itself. However, a temperature of 45 degrees,

given that the temperature in Santa Fe is 76, that the previous day's temperature in Albuquerque was 87, and that the month is July, would definitely be atypical. Multi-modal anomaly detection is inherently contextual. The large majority of the work on anomaly detection has been on point techniques, with only limited exploration of contextual methods, and this has been confined to uni-modal data streams. [2]

3 Technical Accomplishments

The accomplishments of the LDRD are described in detail in a conference publication:

B Rohrer, “**Biologically inspired feature creation for multi-sensory perception,**” Second International Conference on Biologically Inspired Cognitive Architectures, Washington, DC, Nov 5-6, 2011.

This conference paper describes BECCA's feature creator creating hierarchical sets of both audio and visual features.

The BECCA code base is a collection of MATLAB scripts. It consists of all the code for the most recent version of BECCA and the published demonstrations. The code and the paper both can be found at www.sandia.gov/rohrer .

Acknowledgment

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

- [1] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009.
- [2] G. M. Weiss and H. Hirsh. Learning to predict rare events in event sequences. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, pages 359–363, Menlo Park, CA, New York, NY, 1998. AAAI Press.

DISTRIBUTION:

1	MS 1010	Brandon Rohrer, 6533 (electronic copy)
1	MS 1188	Phil Bennett, 1463 (electronic copy)
1	MS 1371	Heidi Smartt, 6832 (electronic copy)
1	MS 1371	Dianna Blair, 6832 (electronic copy)
1	MS 0899	Technical Library, 9536 (electronic copy)