Content-Based Multimedia Analytics: Rethinking the Speed and Accuracy of Information Retrieval for Threat Detection

Elizabeth K. Bowman
Army Research Lab, APG, MD, United States
Elizabeth.k.bowman.civ@mail.mil

Qingzhe Li
George Mason University, 440 University Drive, Fairfax, VA, United States

Yifeng Gao
George Mason University, 440 University Drive, Fairfax, VA, United States

Jessica Lin
George Mason University, 440 University Drive, Fairfax, VA, United States

Tod Hagan
Securboration, 1050 W NASA Blvd #155, Melbourne, FL, United States

Bob Madahar
Ministry of Defence, Defence Science and Technology Laboratory, United Kingdom

Ranjeev Mittu
Naval Research Lab, 4555 Overlook Avenue, SW, Washington DC, United States

ABSTRACT

Member nations of the NATO Alliance are increasingly threatened by the global spread of terrorism, humanitarian crises/disaster response, and public health emergencies. These threats are informed and/or influenced by the unprecedented rise of information sharing technologies and practices, even in the most underprivileged nations. In this new information environment, agile data algorithms, machine learning software, and threat alert mechanisms must be developed to automatically create alerts and drive quick response. Yet these advanced technologies must be balanced with awareness of the underlying context to accurately interpret machine-processed indicators and warnings and recommendations; human involvement will always remain critical in the decision process. We describe one promising approach to this challenge that brings together information retrieval strategies from heterogeneous media sources and human assessment. These multimedia sources include text, video, and images. Our focus, content based information retrieval and multimedia analytics, involves the exploitation of multiple heterogeneous data sources to deliver timely and accurate synopses of data with information that can be combined with human intuition and understanding to develop a comprehensive worldview.

1.0 INTRODUCTION

The complex nature of problems facing the NATO Alliance today require a comprehensive collection and analysis of many heterogeneous data sources to detect, interpret, and respond to indicators, warnings and threat signals. Multi-source analysis remains a challenging problem for many reasons. This includes the unique nature of the text, image, video signal processes and specialty systems designed for each of those channels. Also, analysts who are trained in one domain may be asked to work across several intelligence areas. In such a case, problems could arise when analysts must correlate causality from multiple input signals.

Additionally, human awareness of the underlying context of the region is critical to interpret the various human- and machine-produced data.

This paper is organized as follows. We first introduce the NATO Research Technology Group (RTG) of the Information Systems Technology (IST) panel, IST-RTG-144, entitled Content Based Multimedia Analytics (CBMA), and describe the mandate and relevance of the group to the topic under discussion in this paper. We then introduce a brief scenario to describe conditions where multi-media streams are needed to add value to the NATO decision making process. Next we will provide exemplar technologies for text and video analysis and describe promising tools emerging from research. We will consider how these could be applied in the NATO context. In the third section we will provide some examples of how these methods can be used together and how that could enhance decision making. To that end we will describe recent demonstration events that have shown emerging capabilities for cross-modal analysis. In the fourth section we will address the cognitive aspects of multimodal analysis and identify challenges with interpreting sensor information within the context of the local operating area. In the final section we will offer concluding remarks and identify future work.

2.0 NATO IST-RTG-144 CONTENT BASED MULTIMEDIA ANALYTICS

IST-RTG-144 was organized in 2015 to conduct a technical review into the current status of theoretical and practical developments of methods, tools and techniques supporting joint exploitation of multimedia heterogeneous data sources. In particular, content-based information retrieval and analytics was considered as a means to allow military experts to exploit multiple data sources in a rapid fashion for sensemaking and knowledge generation. Elements included contextual understanding of complex events through computational/human processing techniques, event prediction through the automated extraction of network features, temporal trends, hidden clusters and resource flows, and the use of machine processing for automated translation, parsing, information extraction, and summarization of unstructured and semi-structured data. The initial conclusions of the study are that important research gaps exist in the coordination of video and text data. Though the research areas and developments are being advanced in the military sector and the civil sector, in particular, they remain at low levels of technical maturity for defense and security system applications. IST-RTG-144 is continuing collaborative research efforts to advance those approaches that are most pertinent to our overall aim of enhancing the contextual understanding of complex events through content based assessment of heterogeneous multimedia streams.

3.0 SCENARIO BACKGROUND

For the purposes of considering how multimedia signals might be analysed and incorporated into a cohesive picture of the operational environment we introduce a notional humanitarian aid scenario within an operational area where civil unrest is occurring due to conflicts in the national leadership. NATO forces were already deployed to the area when a major weather event causes widespread damage to local civilians and infrastructure, but the damage appears to be dispersed across a 1,000 square km area with uncertain awareness of the greatest pockets of need and the related safety and health issues in those regions. Communication links remain intact and the local civilians have a history of using social media technologies to share information. Adversary groups are also active in the area and are threatening to steal humanitarian supplies to sell on the Black Market to resource their violent extremist campaign against the local citizens. The NATO Commander is able to utilize military Intelligence, Surveillance, and Reconnaissance (ISR) resources to establish a visual awareness of the damage existing in the area, to track movements of people and vehicles, and to determine which road infrastructure elements are damaged or intact. The NATO Public Affairs Office (PAO) and Information Operations (IO) cell with Open Source Intelligence (OSINT) support have been actively monitoring the social media space for the area and have acquired necessary language translation software to understand communications among the populace. Immediately after the weather event subsided, the Commander directed the staff to provide updates in the following areas:

- ISR of the entire affected area to determine areas hardest hit, numbers of civilians affected in each area, and extent of urgent need.

- Determine the road infrastructure status to evaluate NATO's ability to reach affected regions.
- Identify movements of adversary forces in the region to identify threats to NATO forces traveling to hardest hit areas.

4.0 TECHNOLOGIES FOR CONTENT BASED MULTIMEDIA ANALYTICS

In this section we introduce emerging video and text technologies currently in development that will mature the ISR domain for rapid awareness and assessment in complex terrain and scenarios such as outlined above.

4.1 Kinematic Motifs

Spatial trajectory mining has received a lot of attention from researchers in the last decades, as the discovery of previously unknown movement patterns provides insights into the motives and nature of human behaviors. This research effort focuses on the fundamental challenges of identifying meaningful features or patterns that characterize trajectory data (e.g. the Wide Area Motion Surveillance (WAMI) data, GPS data, etc.) and the establishment of normality in the tracks (cf. patterns of life) for the purpose of anomaly detection. Trajectory data are intrinsically complex to explore since patterns of movement are associated with unknown environments, and objects often move at varying speeds, making it difficult to capture long-term correlations. Towards that end, we study the problem of characterizing user behavior and detecting novel patterns in trajectories based on frequently occurring patterns which we refer to as (kinematic) motifs [6]. Frequent patterns in temporal or spatiotemporal data provide useful high-level information about the data. For example, one could potentially summarize a set of trajectory data by the frequently taken routes, frequently visited locations or a sequence of locations.

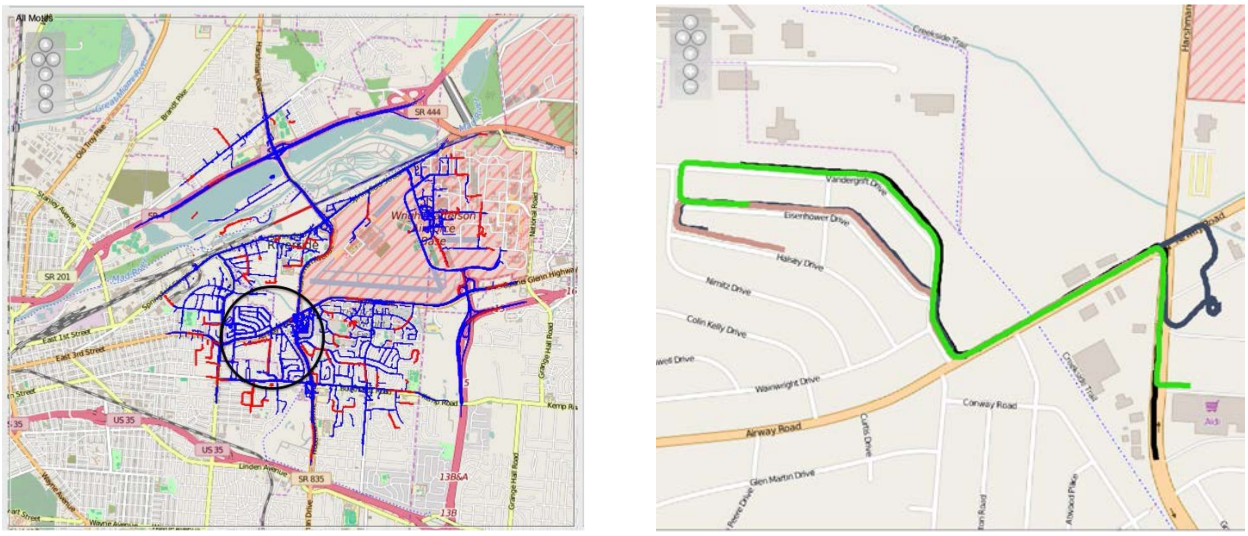
Our goal is thus to build a model that describes normal behavior using patterns that frequently occur in the data. We define a spatiotemporal trajectory T as a time indexed and ordered set of (noisy) location points ({latitude, longitude, time stamp}) sampled from an object moving in 2-dimensional geographic space. Given a dataset ST containing n noisy trajectories, we identify a set of trajectory motifs $M = \{M_1, M_2, ..., M_m\}$, where each M_i consists of a set of sub-trajectories extracted from ST, and the distance between each pair of sub-trajectories is within some error threshold ε. Notice that the sub-trajectories can be of any length, and that we do not have temporal constraint on the discovered motifs. As a result, our motif discovery is speed-invariant – that is, it has the ability to detect sub-trajectories following similar routes even if they have significantly varied speeds.

In recent work [2][7], we show that grammar induction, the process of learning rules of a formal language from a set of observations, allows the discovery of hierarchical structures and regularities from temporal and spatiotemporal data. The current work differs from the previous work in that we proposed a new discretization strategy to handle high level of noise and to enable the discovery of speed-invariant kinematic motifs. Similar to the previous work, our approach is built on a three-phase process: spatiotemporal trajectory discretization, context-free grammar induction, and motif discovery. The first step represents the trajectories as discretized elements by converting them into symbols. We employ a simple strategy and divide the region of interest into equal-sized blocks. Since trajectory data are typically noisy and high-dimensional, we applied some noise reduction and dimensionality reduction techniques to ensure good quality results [6]. The second step parses the symbolic series and decomposes it into a set of context-free grammar rules [1][2]. Since rules of a context-free grammar are hierarchically organized, and each grammar rule represents a discretized subsequence pattern extracted from the input trajectories, frequently used rules are likely to correspond to recurrent subsequences, while infrequently used rules are likely to correspond to rare subsequences [1][2][6][7].

4.1.1 Experimental Results

As reported in [6], in our experiments, we use the publicly released LAIR (Large Area Image Recorder) dataset from Air Force Research Laboratory (AFRL). LAIR is a WAMI (Wide-Area Motion Imagery) dataset, which is extracted from 7 minutes of video, containing 3,615 trajectories. Two motifs are shown in

Figure 1 and Figure 2, respectively. The first motif shows a route between a residential area and a shopping center. The second motif shows a route inside the Air Force Base. We can see that the primary roads (frequently travelled routes) are covered with blue lines, and infrequently travelled routes are in red. The preliminary results [6] demonstrate the ability of the proposed algorithm to identify kinematic motifs, as well as its potential to identify anomalous (infrequently visited) routes.



(a) WAMI tracks (b) A repeated route

Figure 1: Motif example in the WAMI dataset, showing the route from a residential area to some shopping center just outside of the Air Force Base [6].

(a) WAMI tracks (b) A repeated route

Figure 2: Motif example in the WAMI dataset, showing a repeated route in the base [6].

4.2 Text Analytics – Sociocultural Reasoning Framework

Unfortunately, in today's world, understanding the indicators present in the digital footprints of terrorist perpetrators or wide-spread social movements that impact the global landscape, is done after the fact. After an unforeseen event, analysts comb through massive data sets of social media searching for clues and indicators to dig deeper into the event for more intelligence. Examples include identifying potential cells in the case of a terrorist event or to determine what clues might have indicated that a social protest would evolve into a large-scale revolution. Central to the problem of early detection is the shortage of skilled analysts combined with a lack of tools for processing overwhelming heterogeneous datasets – specifically OSINT and social media. Consequently, the ratio of data to actionable intelligence continues to increase to the point where the issue is not finding the *needle in a haystack* but even initially locating the *right haystack* that might have that proverbial needle in the first place. As a result, decision makers are often not provided with actionable intelligence within useful spatial and temporal lines in order to make the most informed tactical and strategic decisions.

An emerging text analytics tool that can help analysts extract knowledge from social media and open source data is Social Cultural Reasoning Framework (referred to as SURF) [4]. As depicted in Figure 3, the vision is for SURF to reside as a suite of cloud based services which automatically alerts analysts to social networks that SURF discerns as 'interesting' in terms of their potential to forecast an event of interest.

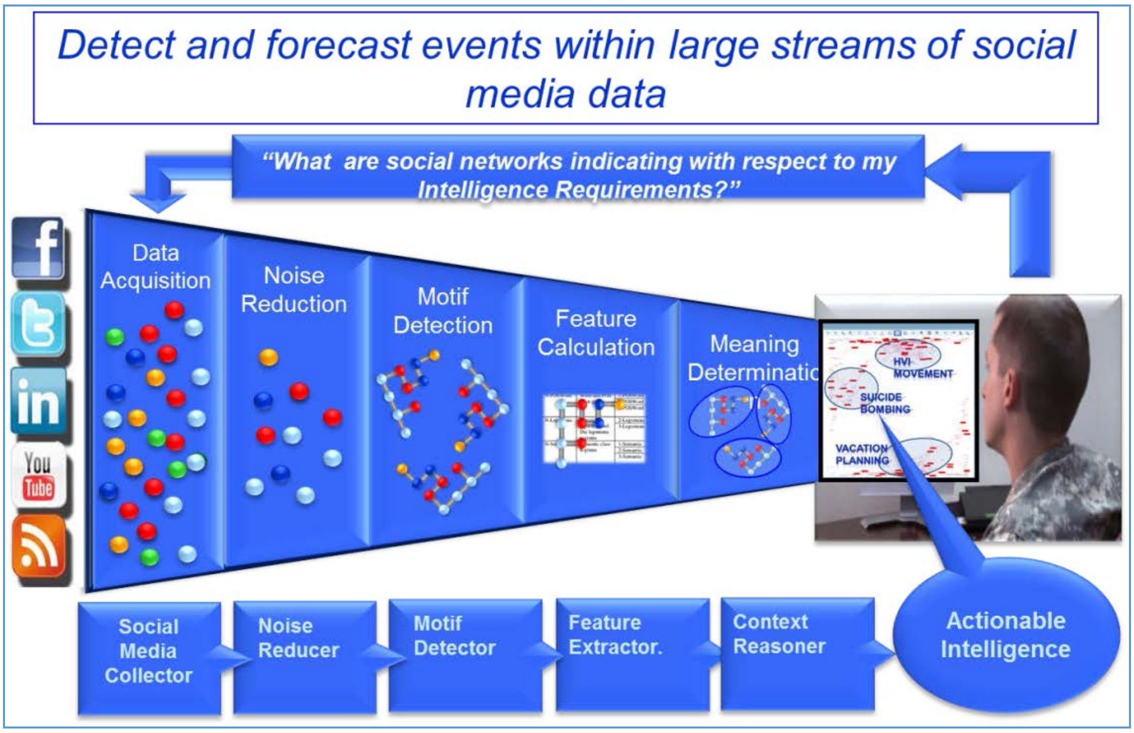


Figure 3. SURF Concept of Operations

In a military setting, such as the scenario outlined earlier, intelligence is generally collected, processed, exploited, and disseminated based on Information Requirements (or Commander's Critical Information Requirements, CCIRs). Based on the operational context, these range from high-level intelligence requirements based on strategic goals to tactical goals. These serve as a contextual starting point from which to apply the SURF capability.

1. Data Acquisition. Data flows into SURF from the social media global data landscape.
2. Noise Reduction. The ingested data is filtered through the noise reduction algorithms, applied within the context of the intelligence requirements.
3. Motif Detection. Repeatable subgraph patterns within the social network are detected.

4. Feature Extraction. Relevant features among the motifs are extracted. These features are distinguishing characteristics which include elements such as sentiment, metrics (e.g. tweet frequency), and demographics.
5. Meaning Determination. Based on the motifs and their feature sets, a probabilistic determination is made with respect to the likelihood that it forecast an anomalous or event of interest related to the intelligence requirement.
6. Actionable Intelligence. The culmination of SURF processing, providing a valuable tool for the analyst to use in conjunction with their expertise to focus and refine their intelligence gathering and ultimately take appropriate action.

4.2.1 Using Motif Models to Determine Meaning

SURF uses motif models to increase algorithmic accuracy while lowering computational requirements. Once motifs have been detected and features gathered, SURF constructs a model from past experiences in order to infer what future events might occur if similar motifs and features are encountered. Using a database of past events allows more accurate model creation and the ability to predict new events in a timely manner. Since the model is nonlinear in nature we do not know the function that represents each of the features to give us a correct classification. Thus the model builder's job is to construct an appropriate function from these observations. This is unlike linear or nonlinear regression model building found in statistics because of the soft margins to which each of the classifiers adhere. These soft margins mean that the model can output probabilities that a data point belongs to a certain class. This probability can be used to calculate the total posterior probability by using Bayes theorem and taking into account past inferences in order to more accurately predict the solution.

The first step in model creation is creation of a decision plane which intersects all data points found during training, which is shown in Figure 4 below. The black line represents the decision boundary and the classifier takes this boundary as the point at which one class becomes another. The support vector machine learns this function through a simple equation based on solving multiple Lagrange multipliers in some infinite dimensional space. The equation shown in Equation 1 shows the learning function of a Support Vector Machine (SVM) which is used to minimize the W component in order to learn the black line in the figure. To learn this function during SURF, we used a program called Vowpal Wabbit that produces the same function, but with simplifications for efficiency purposes[1].

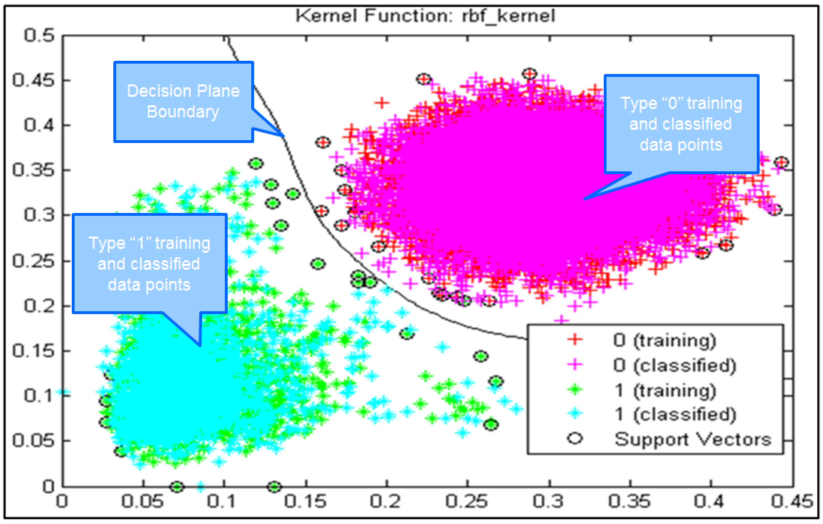


Figure 4. Typical decision plane output of a Support Vector Machine

[1] L. J., L. L. and A. Strehl, "Vowpal wabbit online learning," Yahoo Research, 2007.

$$\min_{\mathbf{w},\xi,b}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left\{\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x_i}-b)-1+\xi_i]-\sum_{i=1}^{n}\beta_i\xi_i\right\}$$

Equation 1: SVM soft margin classification

4.2.2 Evaluation

The SURF tool demonstrated accurate event prediction for two difficult challenges for which large volumes of social media data were timely, readily, and freely available. For both challenges, data acquired a large volume of Twitter data and trained it to derive control (i.e. baseline) motifs/features related to movie ranking and stock pricing. The development team used tweets from February 18th 2013 to March 31st 2013 as the training set. The control motifs allow SURF to detect how close newly detected motifs and features align, and to identify anomalies – i.e. tracking of how a detected motif deviates from the control group. For both challenges, the SURF algorithms outperformed previous Twitter based movie prediction rankings and the results were consistent regardless of changes to the training set.

4.3 Applying Content Based Multimedia Analytics to the Scenario Problem

In the context of the scenario described previously in section 3.0, we envision applications of these video and text tools to the Commander's staff. Using airborne platforms as a beginning point, we would expect video to identify anomalous patterns, such as disrupted routes, unusually large gatherings of people, flooded areas, and other like characteristics of disastrous weather events. However, using the kinematic motifs described in section 4, we might expect to detect unusual movement indicative of adversary activity, to alert the Commander's staff to unexpected movement of the adversary group. In the case where an adversary's headquarters might not have been known to the NATO staff, movement in this particular condition might signal a new organizing site that could have useful purposes.

If we include the SURF capability in the Command staff's toolkit, we could take advantage of social media text signals sent by the civilians affected by the event. In these situations, it is expected that people will share information related to their health and safety and provide some indication as to the emergency nature and extent of the problems being experienced. However, it is well known that social media is a very noisy medium and caution must be exercised to detect spurious information. The SURF capability for motif detection, feature calculation, and meaning determination would guide the analysts in filtering the various categories of information being shared in the region. We note that deconflicting adversary messages from those of civilians requesting emergency assistance is of significant importance, and each should be used to its unique advantage and purpose.

5.0 DEMONSTRATING MILITARY RELEVANCE OF CONTENT BASED MULTIMEDIA ANALYTICS

Demonstrations of emerging technology capabilities follows a continuum of technology readiness levels (TRLs) with the more mature tools being integrated into military systems at large-scale military events and exercises. Early capability tools are featured in the US Naval Postgraduate School (NPS) Joint Interagency Field Experimentation Program (JIFX). Begun in 2012 under the sponsorship of the Office of the Secretary of Defense (OSD) and the Department of Homeland Security (DHS), JIFX events are held quarterly. The JIFX setting offers government, private sector, academic, and non-profit technologists opportunity to engage directly with DHS, DOD, federal civilian agencies, and civilian first responders in a semistructured learning environment promoting collaboration and innovation while fusing interactive community networking and knowledge-sharing activities.

The JIFX environment is a unique opportunity for developers to engage with end users in new or emerging concepts and solutions. By interacting directly with end users in a replicated operational environment, technologists are able to leverage the knowledge gained to improve or enhance their products, steer development to requirements, and reduce the time required to transition a technology, often saving millions of dollars in research and development. Each quarter, a JIFX event uses different methods of interaction, all of which focus on end user input. This reflects and addresses the most complex challenges identified by those directly engaged in homeland defense and security.

The NPS JIFX setting has been used successfully by government, industry, and academic Research and Development (R&D) teams [5]. To prepare, the team organizes a relevant scenario and collects data for the performers to train models, algorithms, and prepare visualizations of data products. In [5] we describe a scenario and dataset that was used to detect threat signals from text data that was then used to cue aerial sensors, and vice-versa. The team included seven text analytics performers. The text analytics team provided the following services: early warning system using discourse and sentiment analysis, single and multi-document summary, relationship analysis, and language translation and social network analysis. The video analytics team provided advanced tracking algorithms that allowed detection of anomalous driving behaviour, heat maps of unusual amounts of activity, and advanced tracking features.

We plan to demonstrate the technologies identified in this paper to the NPS JIFX event 17-1, to be held in November 2016 for a similar demonstration. The demonstration dataset includes California Department of Transportation (Caltrans) Performance Measurement System (PeMS) vehicle detector station data, California Highway Patrol (CHP) Traffic Incidents data and Twitter text messaging data obtained from Twitter's public API. The data was collected over several weeks with a 10 to 15 mile radius about Los Angeles Dodger stadium on days both with and without events (e.g., baseball games, concerts). These locations were selected due to the historical difficulty in obtaining coincident video and text data from military settings.

Regarding the complimentary features of video and text analytics, we propose that automated machine learning tools can provide useful situational awareness capabilities for big data sources, such as social media or kinematic data (typically derived from ISR sensors – e.g., tracks). We also expect that geospatial context in social media can be leveraged to provide useful situational awareness information. Finally, we hypothesize that the fusion of text social media and kinematic data can provide more utility than each sensor in isolation. We will assess the ability of the user to understand and summarize the behavior in a large geospatial region with and without the prototype tools. The particular challenge from a user perspective involves the ability to productively ingest and understand a large quantity of social media and kinematic data.

In the experiment, measures of performance and effectiveness will include the quantifiable computational requirements for the algorithms automatically assessing the data. We will compare the ability of a user to develop a situational awareness summary with and without the prototype tools. A simple user survey and measured task times will allow us to develop an assessment of the potential of the prototype capabilities. The kinematic motif approach for video analysis, described in Section 4.1, characterizes geospatial regions based on the behavior that takes place in those areas. The kinematic motif approach is general, and can fuse data from multiple sources to characterize and summarize the behavior in the geospatial regions. We will investigate the fusion of social media and traffic monitoring data to characterize the behavior in an urban region.

6.0 CONCLUSIONS AND FUTURE WORK

The two exemplar video and text tools described in this paper provide an excellent point of departure for bringing the analysis of video and text closer together to realize rapid cross-cueing and alerting features. The kinematic motif capability characterizes user behaviour to detect novel patterns in trajectories based on frequently occurring patterns, or motifs. These frequency patterns have shown how they can be used to summarize trajectory data and potentially signal anomalous activity. The SURF tool uses motif and message sentiment as an important feature for social network analysis and feature identification. Rapid and accurate detection of motifs that can reliably signal threat activity, when combined with other sensor products and in the context of an event, could produce actionable intelligence and allow proactive military response. Both of these tools will be demonstrated in the NPS JIFX 17-1 event in November 2016 with subsequent evaluation results documented.

The NATO IST-RTG-144 group will explore collaborations arising from the demonstration event and will attempt to incorporate learning from these activities in addition to those through research of the participating group team members. The following areas of exploitation for video and text analysis are expected to produce additional capabilities to support NATO decision makers, at headquarters (e.g. OSINT) and in operations, in the future:

- Capture and index motion imagery; further investigate intelligent capturing and initial processing by sensor systems, to include initial video indexing and key frame information produced in audio and metadata entries.
- Exploit imagery indexing through hierarchical methods using semantic identifiers and human evaluations of exploitation results.
- Explore motion-based index generation to generate rapid and robust retrieval of context. Types of motion include background motion of static structures related with sensor flight, background motion generated by normal patterns such as traffic flow, an explosion and after effects at a location, etc.
- Expand the Machine Learning/Deep Learning (ML/DL) approach for semantic video analytics through a semantic hierarchy of Full Motion Video. Long-term impact is the provision of optimal semantic information to users in rapid fashion while adapting to dynamically varying computational resources.
- Explore the mechanisms by which text analysis results can be used to drive/exploit video and imagery indexing and retrieval.
- Explore frameworks for optimizing multimedia analytics via systems engineering and architectural design concepts.

NATO and coalition military leaders, commanders, and intelligence analysts need interoperable tools that cross-cue knowledge obtained from one method to generate taskings in another. The research focus identified by IST-RTG-144 should help build the cross-cued solution from advances in content-based multimedia data analytics. If the research is successful, it will significantly improve NATO abilities to generate knowledge from extremely large stores of text, imagery, and video caches to speed situational awareness and decision making.

References

[1] Nevill-Manning, C. G. and Witten, I. H., “Identifying hierarchical structure in sequences: A linear-time algorithm," J. Artif. Intell. Res.(JAIR) 7, 67{82 (1997).

[2] Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A. P., Chen, C., and Frankenstein, S., “Time series anomaly discovery with grammar-based compression," in [International Conference on Extending Database Technology], (2015).

[3] Patel, P., Keogh, E., Lin, J., and Lonardi, S., “Mining motifs in massive time series databases," in [Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on], 370-377, IEEE (2002).

[4] Securboration, (2013). “Securboration to Research Link Analysis of Knowledge from Social Media for Air Force”. Available on the Internet at: http://www.securboration.com/2013/07/24/securboration-to-research-link-analysis-of-knowledge-from-social-media-for-air-force/. Retrieved September 14, 2016.

[5] Bowman, E.K. & Zimmerman, R. J. (2015). US Army Research Laboratory Joint Interagency Field Experimentation 15-2 Final Report. Adelphi: ARL

[6] Mittu, R., Lin, J., Li, Q., Gao, Y., Rangwala, H., Shargo, P., Robinson, J., Rose, C., Tunison, P., Turek, M., Thomas, S. & Hanselman, P. (2016). Foundations for context-aware information retrieval for proactive decision support. Proc. SPIE 9851, Next-Generation Analyst IV.

[7] Oates, T., Boedihardjo, A., Lin J., Chen, C., Frankenstein, S. and Gandhi, S. (2013). Motif discovery in spatial trajectories using grammar inference. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*. San Francisco, CA. Oct 27-Nov 1.