# Technical Evaluation Report

**Drs. B. K. Madahar[1], M. Harries[1], E. Bowman[2], R. Rao[2], G. Burghouts[3], L. Overlier[4], D. Gustafsson[5]**

[1]UK Dstl, [2]US ARL, [3]NLD TNO, [4]NOR FFI, [5]SWE FOI

bkmadahar, mharries@dstl.gov.uk; elizabeth.k.bowman.civ@mail.mil; raghuveer.m.rao@mail.mil; gertjan.burghouts@tno.nl; lasse.overlier@ffi.no; david.gustafsson@foi.se

## 1. INTRODUCTION

The specialist meeting on "Content Based Real-Time Analytics of Multi-media streams" (CBRAM), NATO-IST-158-RSM-10, was held from the 6th to 8th September, 2017, at Middlesex University, London, United Kingdom. There were just over 30 registrations and about 25 attendees across coalition Nations (e.g. citizens of UK, USA, NLD, NOR, SWE, ITA, ESP, DEU) on the first two days and about half that number on the final day which was primarily for the technical committee and work group leads to summarise the findings for this report. There was good representation from Government, Industry, and Academia.

It was one of the outputs of current research task group (IST-RTG-144) of the NATO Information Systems Technology (IST) on "Content-Based Multi-Media Analytics". This is a three year RTG which started in April 2016 and is undertaking collaborative research to achieve the following outcome for defence:

- **Better automated integration of text and video information into decision support environments.**

In pursuing this aim, the RTG wanted to identify leading developments and emerging Science and Technology (S&T) that could help advance the RTG's work in the following key areas:

- Intelligent Capture and indexing of motion imagery;
- Expand the Deep Learning approach for semantic video analytics;
- Explore the mechanisms by which text analysis results can be used to drive/exploit video and imagery indexing and retrieval; and
- Explore frameworks for optimizing multi-media analytics via systems engineering and architectural design concepts.

A specialist meeting was considered to be a suitable instrument to progress this aim.

## 2. SCOPE

The RSM was to bring together experts and practitioners from NATO member military agencies along with industry leaders and academic visionaries to present and discuss the state-of-the-art developments and hard challenges in content analysis of multi-media and the application and exploration of exploiting text and video for rapid understanding.

A technical committee was established (active members are listed in Annexe 1 and most are co-authors of this report) to define scope and structure.

The meeting scope was limited to the technical topics below and to elicit technical contributions, and invite key notes, to address leading developments in one or more of these areas. The overall context however was defence and security environments with coalitions, or inter-agencies, to support rapid decision making enabled by intelligent analytics of heterogeneous multi-media streams.

RSM topics:

- Capture and indexing of motion imagery
- Imagery exploitation
- Human evaluations of exploitation results
- Index generation for content retrieval
- Deep Learning for semantic video analysis
- Test analytics to exploit video indexing
- Architectural design concepts
- Integrated text and video indexing

## 3. APPROACH

The RSM was structured to have a key note each day, three sessions (Semantic Multi-Media Analysis, Imagery Exploitation, and Emerging Issues) for technical presentations over the two days and a specific session for small inter-disciplinary working groups to discuss debate and agree on key technical areas and challenges that should be addressed in future research. In addition ‘ice-breaking’ sessions and a visit around research facilities at Middlesex University were organised to stimulate creativity and innovation. The final programme is shown in Annexe 2.

The meeting was fortunate to obtain two keynote speakers, one addressing detection of persons of interest in social-media using motifs and the other on digital forensics. An important part of RSMs are the working groups and three were established for the RSM with approximately 8 members each. They formed during the ice-breaking session, elected a rapporteur, and carried out the work in earnest on the second day. The questions they were asked to address, following a brief on context and suggested ways of working, were:

- What: Identify the key technical areas, including emerging areas, in analytics and related S&T that are likely to make the most difference to defence and security in the near to long term;
- Why? Articulate the value of capabilities, and identify where and why they can help make a difference, such as challenges/capabilities being addressed, and possible benefits/value they may bring to nations and coalitions.

A summary of the key outputs from the technical presentations and the working group sessions are reported in the following sections.

## 4. SUMMARY OF PRESENTATIONS

A summary of the work presented during the three sessions (Semantic Multi-Media Analysis, Imagery Exploitation, and Emerging Issues) can be found in Annexe 3.

# 5. WORKING GROUP OUTPUTS

Drs. Raghuveer Rao, Lasse Overlier, and Gertjan Burghouts developed an overview that captured the three discussion groups' conclusions regarding thoughts on ***what*** the key technical areas in analytics and related S&T are likely to make the most difference to NATO partners in terms of defence and security in the near-to-long term, and ***why*** these capabilities can make a difference in terms of possible benefits they bring to the NATO Alliance. Prior to identifying specific technologies, the groups agreed that executing any recommendation should be conducted with the following considerations: 1) using an agile methodology that includes 2) an iterative process with smaller closed loops or spirals, that 3) addresses basic research, applications and the end user, and 4) is built with a full understanding of the workflow process of the intelligence operators.

Four overarching categories were identified to frame a unified construct for multimedia analytics. These are: Capture, Analysis, Dissemination, and Training. Cutting across each of those frames is the need to take into account consideration of bias, confidence, and ethics. Ideas from the discussion topics are presented below within each frame.

## 5.1 Capture

The ability to collect relevant data and analytics methodologies for the NATO alliance will rely on the use of community datasets that provide a common foundation and frame of reference for research efforts. These will be supported by an assembled group of use or case folders that supply the operational meaning behind the analytics (or the ***why*** factor). Each dataset must be curated, which includes crowdsourced tasks such as labelling and the creation/simulation of operationally relevant datasets. It is also recommended that role of metadata be specified.

An example of how the *capture* frame could be understood in an analyst's role is the issue of propaganda campaigns in the current information environment using global reach of social media platforms and mobile technologies. Analytics that would be needed here include issues of how to detect various levels of false or misleading information, an understanding of how messages spread, and correlations between messages and platforms (various platforms utilize different types of information, e.g., text, images, videos). Basic and applied research needs in this area include an understanding of how influence spreads in online networks, approaches to verifying truth, how marketers choose target audiences and how these features are identified, how are information campaign effectiveness measured, and what is the inference mechanism used to target audiences (as these are not selected at random).

The discussion about the need for common datasets led to several sources for potential acquisition or support, one of which resides with the Visual Analytics Community (VAC) and the annual Visual Analytics Science and Technology (VAST) Challenges. These are an annual contest with the goal of advancing the field of visual analytics through competition. The VAST Challenge is designed to help researchers understand how their software would be used in a novel analytic task and determine if their data transformations, visualizations, and interactions would be beneficial for particular analytic tasks. VAST Challenge problems provide researchers with realistic tasks and data sets for evaluating their software, as well as an opportunity to advance the field by solving more complex problems. Researchers and software providers have repeatedly used the data sets from throughout the life of the VAST Challenge as benchmarks to demonstrate and test the capabilities of their systems. The ground truth embedded in the data sets has helped researchers evaluate and strengthen the utility of their visualizations. Organizations in the UK (Dstl) and the US (Pacific Northwest National Lab (PNNL)) contribute to the VAST dataset development and access. (See http://www.vacommunity.org/HomePage).

It was noted that a first step for using a dataset like the VAST corpora would be to create a methodology for adding to an existing dataset with an established standard for academia, government, and industry to add data from multiple sources (video, text, audio, images, social media, etc.).

## 5.2 Analysis

The second step in the multimedia analytics workflow is to transform data to indexable forms to enable the use of data mining tools, e.g., video detections to text labels, or online translation of audio. It is recommended that processing occur at the edge with high performance tools, e.g., neuromorphic chips, GPUs & field processing units. Such tools have become very flexible nowadays; configurations and models can be modified easily. This would allow the labelling of objects and actions in real time at point of need. This would improve the adaptibility of models and the ability to perform context aware analysis. These capabilities together would allow the auditability and assurance of artificial intelligence (AI) decisions. Further needs are the anticipatory and predictive intelligence (machine sentience), e.g., the ability to exploit patterns of life from multimodal data to inform adversary action awareness

When combining multi-modal data, discussion centered around what it means to combine data and at what level? How are data integrated and why? Which communications are the most interesting? If we use social media as an example, companies tried to use analytic pipelines that use multidimensional scaling techniques with embeddings for signatures. These are then used to match signatures to allow for recommender system applications. In a military sense, we need to help analysts identify and understand causal links in the data, much like the commercial signature algorithms. It was suggested that the Joint Directors of Laboratories (JDL) fusion levels may provide inspiration for a process flow.

Further discussion illuminated the nature of spatio-temporal (ST) data within the realm of multi-modal information. These data exist with a matter of scale, and not all data may link to ST analysis methods. It will be important to understand how spatial uncertainty can be communicated to analysts, given the reality that many different types of analysis functions exist within military settings that traditionally rely on geospatial representations of the world. For example, it was mentioned that many interpretations of the same map would be made by Navy and Army analysts. The former would focus on the littorals and the latter on the land formations. In creating visualizations of analysis products, it is important to comply with the modality of data for which the intended audience is aware. Using a medical analogy, it would be senseless to provide a general practitioner of medicine with a 3D model of an MRI; they need to use information in ways that they can comprehend and utilize. As researchers develop technologies they need to understand the process of hypothesis creation and should include iterative experiments with teams of analysts to produce an agile product.

## 5.3 Dissemination

When considering how to propagate information to relevant users, it is recommended that we investigate methods for efficient and context-relevant spread. These would nominally be Warfighters close to tactical edge with the analyst ‘in the back’ but available for reachback and support. Transparency of machine algorithms is important so that the analyst can explain and understand results and question results that appear as outliers. The (cognitive) workflow of the analyst should be taken into account (see Annex section A3.5).

In addition to the multi-nation experiment in data indexing discussed by Dr. Lasse Overlier in his presentation (see section 4.4.2), it was suggested that a human subjects’ experiment be undertaken by interested nations to explore more fully the human aspects of incorporating multi-modal data into the analysis process. Such an undertaking would support knowledge creation for the NATO alliance. The suggested plan is to generate a shared experiment design, scenario, and multi-modal data for presentation to human subjects. The trials would be conducted in national labs with initial results analysed accordingly. Results of data analysis would be shared among partners with a final report coalescing findings from each national study. In some cases, if national analysis tools cannot be shared with other nations participating in the experiment, screenshots of analysis products might be used to simulate the workings of the particular tool. Such a study is a recommended strategy for the last 18 months of the IST-RTG-144.

### 5.4 Training

The full realization of content-based multimedia analytics for defense applications will rely on the literacy of intelligence analysts in dealing with autonomous decision systems and augmented data vs raw data. It is important that the analysts understand biases and limitations of decision systems, and at the same time embrace new ways of human machine interactions. As the science of multimedia analysis continues to race forward, the relative roles of the human and the machine are likely to change over time. Human analysts must be adaptive and ensure that machine technologies continue to support their sensemaking roles.

## 6. CONCLUSIONS/RECOMMENDATIONS

The specialist meeting on “Content Based Real-Time Analytics of Multi-media streams” (CBRAM), NATO-IST-158-RSM-10 provided insightful lectures and discussion from two keynote and eight subject matter experts in text and video analytics. The findings of the meetings will contribute to the successful results of the sponsoring research task group (IST-RTG-144) of the NATO Information Systems Technology (IST) on “Content-Based Multi-Media Analytics”.

The knowledge captured at this meeting was summarized in this report within the existing IST-RTG-144 objectives:

- Intelligent Capture and indexing of motion imagery;
- Expand the Deep Learning approach for semantic video analytics;
- Explore the mechanisms by which text analysis results can be used to drive/exploit video and imagery indexing and retrieval; and
- Explore frameworks for optimizing multi-media analytics via systems engineering and architectural design concepts.

Two major recommendations follow from the IST-158-RSM-10; 1) a continuation of the shared video indexing experiment and 2) a follow-on human subjects experiment to explore the human analysts’ ability to rapidly and accurately incorporate multi-modal information with a variety of visual displays. With respect to the former, the incorporation of existing open source datasets could be a potential addition to the indexed corpora and could be used to support the human subjects’ experiment(s). The execution of a multi-nation set of experiments using shared resources will serve to enhance the understanding of the analysis workflow and cognitive demands. Furthermore it could serve as a concept technology demonstrator to NATO stakeholders.

RTG-144 is to consider these recommendations and submit technical activity proposals to the IST panel for research activities outside of its current scope (e.g. experiment, community datasets and demonstrator) before the spring 2018 panel business meeting.

# REFERENCES

[1] R. Brackin, R. (2016). Incident Management Information Sharing (IMIS) Internet of Things (IoT) Extension Engineering Report, OGC 16-092, Open Geospatial Consortium, available from: https://www.dhs.gov/publication/incident-management-information-sharing-imis-internetthings-iot-extension-engineering

[2] Goudos, S.K., Kalialakis, C. and Mittra, R. (2016). Evolutionary Algorithms Applied to Antennas and Propagation: A Review of State of the Art, International Journal of Antennas and Propagation, 12 pages.

[3] Health and Safety Executive (2017). Hazard pictograms, available from: http://www.hse.gov.uk/chemical-classification/labelling-packaging/hazard-symbols-hazard-pictograms.htm

[4] Robinson, Neil, Emma Disley, Dimitris Potoglou, Anais Reding, Deirdre May Culley, Maryse Penny, Maarten Botterman, Gwendolyn Carpenter, Colin Blackman and Jeremy Millard. Feasibility Study for a European Cybercrime Centre. Santa Monica, CA: RAND Corporation, 2012. https://www.rand.org/pubs/technical_reports/TR1218.html.

[5] Van Baar, R.B., van Beek, H.M.A., and Eijk, E.J. (2014). Digital Forensics as a Service: A game changer. Digital Investigation, Vol 11, Supplement 1, p. S54-S62. Available online: http://ac.els-cdn.com/S1742287614000127/1-s2.0-S1742287614000127-main.pdf?_tid=35725cd8-9d1f-11e7-8143-00000aacb360&acdnat=1505814482_864ecbb453915051f385ee697b9b6ff8

# ANNEX 1: SUMMARY OF PRESENTATIONS

## SESSION 1: SEMANTIC MULTIMEDIA ANALYSIS

### A3.1 USING MOTIF DETECTION TO PREDICT GROUP MEMBERSHIP FROM MULTI-SOURCE DATA

Mr. Tod Hagan, US, presented the first keynote address to discuss a new capability to classify individuals active in social media platforms. This tool is the Social Understanding Reasoning Framework (SURF) that was developed for the US Department of Defense (DOD) under the Small Business Innovative Research (SBIR) program. SURF uses features to classify a motif as a specific type of event. For example, motif detection might show a social movement occurring in a region. Features of the motif can inform contextual details of the activity (e.g., whether the gathering represents a protest, revolution, etc.). This is analogous to knowing the shape of a puzzle piece (motif detection) and using feature extraction to know what is pictured on the piece. Features tell us distinguishing characteristics about the event in the same way that skin attributes or hair color can be used to identify a person’s age or ethnicity. The SURF tool allows discovery of features that inform interesting details about the discovered motifs and what they mean in relation to each other. The unifying theory is that motifs of similar class structure will have different attributes associated with a specific event. Possible motif features are well documented in previous publications and in the network science field in general. These features include clustering coefficient, average path length, and average number of common neighbours. The parameters which are more specific to the SURF use cases such as detecting probable social movements must be added in order increase prediction accuracy. One primary feature of SURF is what topics are being discussed in the motifs, and the prevailing sentiments of exchange. In the situation of a revolution, one could assume that the prevailing sentiment would be that of anger while topics such as past revolutions, uprisings and government would be expressed more than the baseline.

### A3.2 AD-HOC SEARCH IN VIDEO DATA

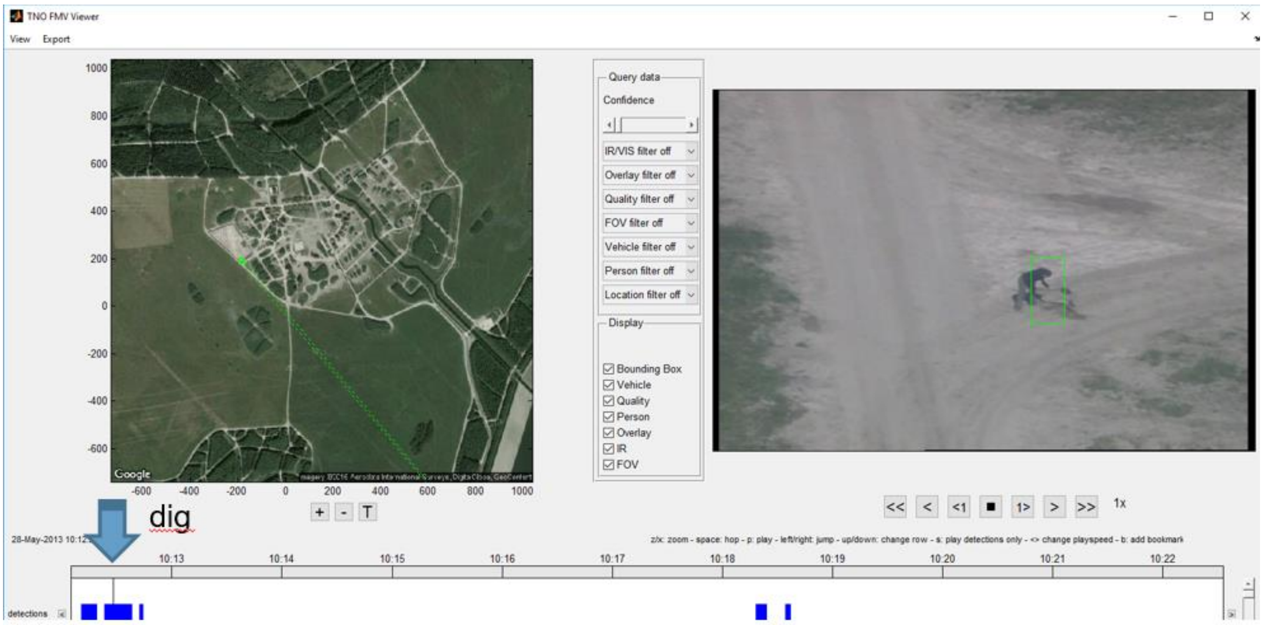
Maaike de Boer, NL, was the first speaker in session 1, Semantic Multimedia Analysis. As an analyst with NL’s TNO, her research focuses on capturing activities from the world around us in real-time. However, with the increasing amount of video data, it becomes unfeasible to watch all data. Computer vision techniques make it possible to automatically detect concepts in videos, such as people, objects and activities. It is, however, not possible to train all possible concepts, especially higher-level concepts, such as a terrorist attack or a robbery. These higher-level concepts often have not enough data to reliably detect them. In her presentation, she showed how we can still search in the video stream with queries that do not contain trained concept detectors (ad-hoc queries), such as the high-level concepts. The focus was on the topics: 1) query interpretation: how to translate a natural language query (robbery) into a combination the things that can be detected in the video stream (person + running + bag); 2) fusion: how to combine information from multiple sources, such as detections on image level, motion, audio and text, and; 3) feedback interpretation: how to improve the system while using it. Through her research, de Boer has improved visual search effectiveness by using a combination of i) query-to-concept mapping based on semantic word embeddings, ii) exploiting user feedback and iii) fusion of different modalities (data sources). Also, she has made some advances to provide a (real time) search capability that handles ad-hoc textual queries (i.e. contains non pre-trained concepts). As future research, de Boer plans to conduct evaluations of this approach with security operators and/or analysts to verify that the capability is effective. In conclusion, she noted that the capability is dependent on the pre-trained concepts. For example, without a concept *fire* it is hard to detect an event *extinguishing a fire.*

## A3.3 COMPRESSED-DOMAIN DEEP LEARNING FOR SEMANTIC VIDEO ANALYTICS

Dr. Ioannis Andreopoulos, UK, (with colleagues Aaron Chadha and Alhabib Abbas) was the second speaker of session 1. He noted that video comprises the major communications and entertainment media asset, occupying more than 60% of today's Internet traffic. Video signals also comprise the cornerstone of surveillance and security systems. Yet, video remains the least-manageable element of the big data ecosystem. This is because all state-of-the-art methods for high-level semantic description in video require either manual annotation or compute-intensive video decoding and processing with very deep neural network architectures. In their recent work, Dr. Andreopoulos and his team have begun investigating video content classification and retrieval via deep learning systems that directly ingest compressed bitstream information. Their first results show that such approaches may lead to more than 100-fold increase in content parsing and processing speed while being competitive or superior to state-of-the-art deep learning systems based on video frame decoding and pixel-domain processing. This idea is based on the observation that video macroblock (MB) prediction modes (that are very compact and readily available from the compressed bitstream with minimal decoding) are inherently capturing local spatio-temporal changes in each video scene. Therefore, a deep learning system that is trained primarily based on MB prediction modes can allow for tremendous acceleration in comparison to frame decoding and pixel-domain processing. This can unlock the rich structure preserved within compressed video bitstreams, thereby allowing for large-scale deployment in content delivery networks, online video crawlers, real-time video tagging and alert generation systems for surveillance, video recommendation systems based on content similarity and content type, etc. Dr. Andreopoulos summarised his framework for training a temporal stream of MB motion vectors extracted directly from the video bitstream and a spatial stream comprising selective (motion-dependent) MB RGB texture decoding, and considered how the two streams can be fused during testing. He presented recent results with a 3D convolutional neural network (CNN) architecture for video classification that utilizes compressed-domain motion vector information for record-breaking speed. A key aspect of his team's approach is the fusion of the 3D CNN with a spatial stream that ingests selectively-decoded frames, determined by the motion vector activity. He reported results showing that he is able to classify videos up to an order of magnitude faster than recent proposals, whilst maintaining competitive classification accuracy. Given the observed performance within the two standard benchmark datasets, further work in this area (and the utilization of higher-resolution video datasets) may provide for even further advantages for our approach and previous methods. Such systems may find important applications in video data classification and retrieval systems.

## A3.4 COMBINATION OF VIDEO AND TEXT ANALYTICS FOR MULTI-SOURCE INTELLIGENCE

Dr. Gertjan Burghouts, NLD, presented the final talk in Session 1. He discussed results from a small-scale experiment to explore possibilities for intelligence gathering based on text and video. Text and pictures were obtained from social media and tracks of people and video were labelled in aerial footage by video analytics. From social media, some messages were collected by text search. For example, using keywords related to insurgent activity, the analyst using this system might search for 'jeep', 'weapon', or 'suspicious'. In this Dr. Burghouts' use case, the analyst's search for 'jeep' returned a Tweet with three persons standing near a jeep, one of them was a woman. The analyst then used the image of those persons to search for other instances of those individuals within the available video footage. The browser returned detections in an intuitive manner (i.e., grouping similar appearances of these people in close proximity). A timeline viewer was developed to allow the analyst to view the activities of these persons of interest to determine potential threat levels (see Figure 1).



IR/VIS filter off
Overlay filter off
Quality filter off
FOV filter off
Vehicle filter off
Person filter off
Location filter off

☑ Bounding Box
☑ Vehicle
☑ Quality
☑ Person
☑ Overlay
☑ IR
☑ FOV

z/x: zoom - space: hop - p: play - left/right: jump - up/down: change row - s: play detections only - <> change playspeed - b: add bookmark

**Figure 1 - Person of interest on timeline showing activities to determine potential threats.**

This small-scale experiment was used as a proof-of-concept to explore possibilities for intelligence gathering using text and video. Text and pictures were obtained from social media and tracks of people and vehicles from video analytics. Various viewers enabled the analyst to interact with the metadata to pinpoint persons or objects of interest in social media, to select appearances of persons or objects, and to show a timeline of behaviour to aid analyst understanding of threats. Dr. Burghouts concluded by saying that this experiment was about a social media search as a cue to video search.  However, video search may also cue social media search, such as looking for persons in pictures contained in messages, and to explore messages in a particular time period and area. Moving forward, Dr. Burghouts stressed that military applications will require more than cross-cueing; fusion using interactive two-way search will be needed. Using the use case in the study, this might involve the analyst viewing a social media picture of the Jeep but not the woman, where the woman was seen in the video but not in the Jeep.  Using fusion capabilities, the analyst would associate the woman and the Jeep based on the time and place of their appearance in the video stream. Dr. Burghouts concluded by noting that the study demonstrated the search for relevant content across modalities, thereby providing tools to

## A3.5 DEMONSTRATION OF VISUAL ANALYTICS FOR SENSE-MAKING IN CRIMINAL INTELLIGENCE ANALYSIS (VALCRI)

Drs. William Wong and Neesha Kodagoda, UK, provided an overview of the VALCRI system. This tool for analysing and investigating criminal activity is being developed under a European Commission grant to Middlesex University. In this effort, Middlesex University leads a consortium of 18 international organizations from 8 countries. The system is a system that facilitates human reasoning and analytic discourse, tightly coupled with semi-automated human-mediated semantic knowledge extraction. It uses machine learning to process vast quantities of data and to identify semantically similar reports (from command and control, intel reports, witness statements) and suspicious patterns of behaviour. The VALCRI system is designed to provide early warnings of impending criminal activity and support the complex work of police and intelligence agencies in the age of the internet and global terrorism. VALCRI will be used by analysts using powerful analytics software and data organized in an interactive and graphical manner designed to lessen cognitive workload often experienced by humans while encouraging imagination, enabling insight, ensuring transparency and engaging with fluidity and rigour. The entire project is designing the technology from cognitive, legal, ethical and privacy perspectives to protect the rights of the individual to security and liberty while ensuring the good of society. VALCRI will also enable law enforcement agencies to make their processes more transparent so the process by which their conclusions are reached are made easier to inspect. During September and October 2017, the VALCRI team will be making presentations to the Los Angeles County, CA and the Pasco County, FL Sheriff's Departments, the IEEE Visualization, and the Human Factors and Ergonomics Society conferences in Phoenix, AZ and Austin, TX, respectively.

## SESSION TWO: IMAGERY EXPLOITATION

## A3.6 FORGERY DETECTION IN IMAGES, DR. KATRIN FRANKE, NTNI

Dr. Katrin Franke, Professor of Computer Science, Centre for Cyber and Information Security (CCIS), Norwegian University of Science and Technology (NTNI)[1], gave the second keynote presentation on Image Forgery Detection. The CCIS is a joint force in Norway, opened in August 2014 and includes 13 law enforcement, industry, and academic organizations at the national and local levels. The CCIS uses a 2Centre Model (Cybercrime Centres of Excellence Network for Training Research and Education) (Robinson et al., 2012) and builds upon previous Cybercrime Training projects funded by the European Commission and DG Home. The primary objective of the Centre is to enhance the capability of combating cybercrime in the European Union (EU) and beyond. It was initiated between Ireland and France, with additional nodes in Belgium, Bulgaria, England, Estonia, Greece, Lithuania, Romania, and Spain.

The NTNU Digital Forensics Group (DRG) includes three CCIS funded positions through the Norwegian Police Directorate. They sponsor a Masters of Science (MSc) track in digital forensics and cybercrime investigation. Joint research projects are pursued with law enforcement, to include Dark Net training for the International Police Organization (INTERPOL) in Norway.

Dr. Franke outlined her perspectives on digital investigations: Legal (regulations, policies, rule of law), Technological (security, archival), Organizational (information management, procedures, governance) and Knowledge (capacity building, training public awareness through pedagogical methods). While she acknowledged that all aspects of these investigations are important, her talk focused on the technological approaches. As depicted in Figure 2, the goals of recognition accuracy and feature complexity are inter-related.

[1] The Norwegian University of Science and Technology, when translated into Norwegian, is NTNI.

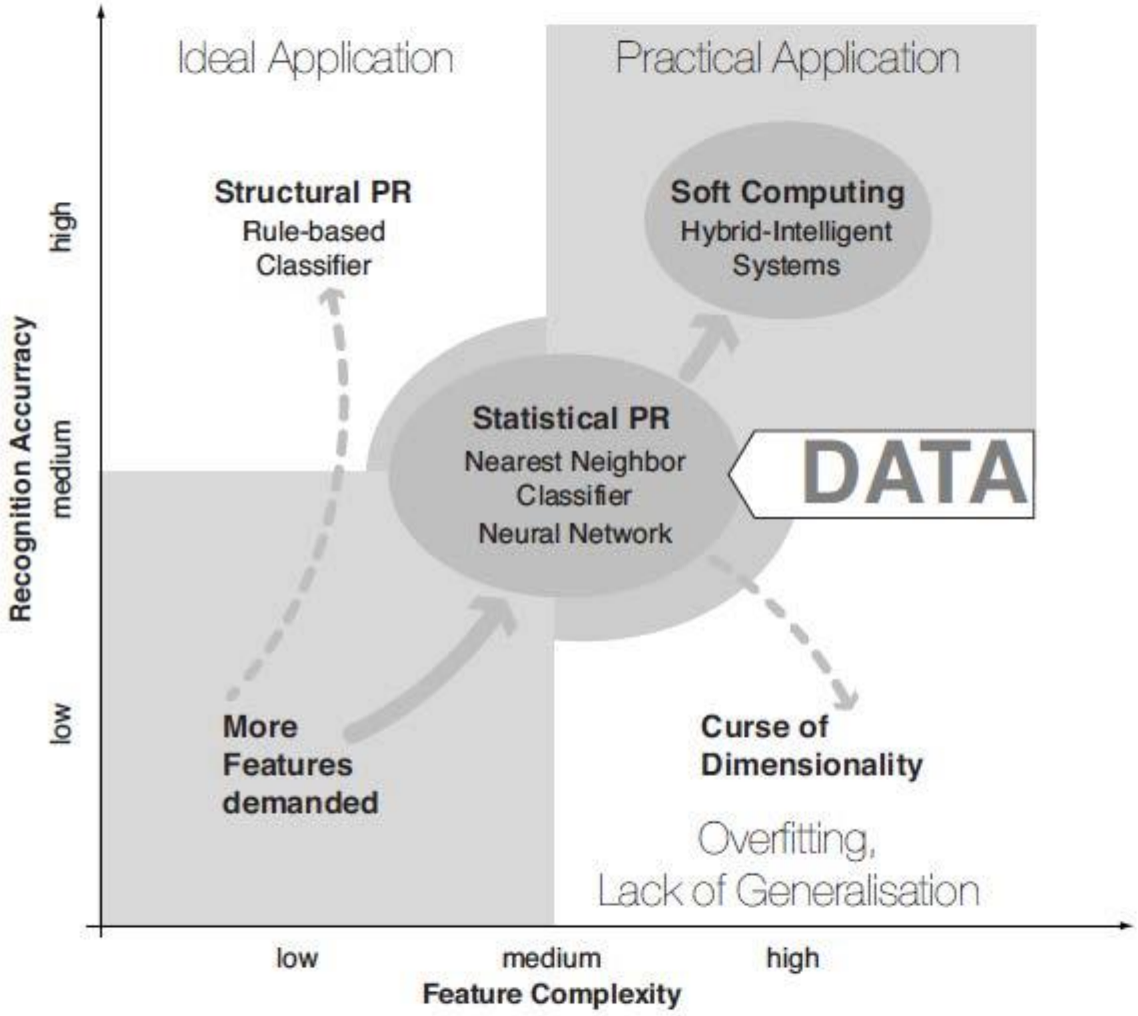


**Figure 2 - Inter-relation of feature complexity and expected recognition accuracy**

For a practical application example, Dr. Franke discussed a large-scale forensic computing infrastructure feasibility study conducted over 18 months involving CCIS researchers. This project, conducted collaboratively with the Netherlands Forensic Institute (NFI), developed the forensics research engine Hansken.2 The Dutch Police use Hansken to conduct digital investigations using huge amounts of data that are made searchable through a Digital Forensics as a Service (DFaaS) (van Baar, Beek & Eijk, 2014). This process has become a standard in the Netherlands for hundreds of cases and over a thousand detectives.

Before concluding with many examples of digital forensics applications (see below), Dr. Franke summarized her thoughts by cautioning the audience to bridge theory and practice, understand results, and talk to experts fully gain relevant domain knowledge. Her applied examples covered the following topics:

- Malicious Code Detection

[2] The name 'Hansken' originates from an elephant that toured the Netherlands in the 17th century and that was sketched by Rembrandt van Rijn in 1637. As the largest land animal in the world, the elephant stands for the huge amount of data that can be processed by the software platform. Besides, elephants are noted for their good memory. See http://www.forensicinstitute.nl/about_nfi/news/2015/nfi-developed-forensic-search-engine-for-digital-investigation.aspx?cp=34&cs=578 for a video about Hansken.

- Network Intrusion Detection
- Malware Analysis
- Email Analysis and Author Identification from Text-based Communications
- Data-driven Threat Intelligence
- Economic Crime Investigation
- Blockchain Technology
- Questioned Document Analysis (historical documents, signature analysis, secure border control documents, reconstruction of torn documents, stamp analysis, large-scale document analysis, and robotic signature simulation).
- Forensic Handwriting Examination

### A.3.7 A DEEP CONVOLUTIONAL NETWORK FOR TRAFFIC CONGESTION CLASSIFICATION, DR. CHRIS WILLIS ET AL., BAE SYSTEMS

Dr. Willis presented the first talk of Session 2, Imagery Exploitation. In this talk, he described the development of an image classification process for the recognition of traffic congestion. As background, he explained that ‘Transport for London’ manages a network of surveillance cameras which are used to monitor road junctions in the city. Images and video sequences (of a few seconds duration) from over 1,200 cameras are collected and made available for third party use in the development of new applications. The images and video sequences are released with a five-minute refresh rate. This paper focuses on processing the still images; these are 352x288 pixels.

Imagery was collected over an extended time period and includes examples at all times of the day and night, in different location types and in the range of traffic and environmental conditions prevailing at collection time. Typical examples of images collected and used in the presentation are shown below. These illustrate just some of the variation in lighting, traffic and road configuration which is present in the dataset.

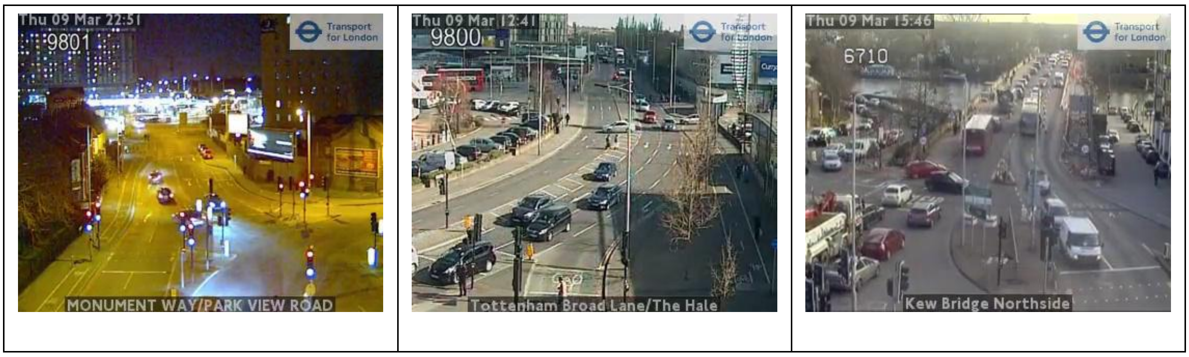


**Figure 3 - Examples of Imagery Configurations in Traffic for London dataset**

A subset of the imagery covering five locations and a 24 hour time period was selected for ground truth labelling. In order to carry out this step an internet-accessible image annotation tool was developed which allows the labelling of each image as one of: uncongested, congested, unknown or broken. These allow both hard and soft (probabilistic) labels to be derived for the images, enabling their use for training and testing a classifier.

Deep learned classifiers have been designed based on: GoogLeNet with transfer learning with a follow-on subnet, and; as a bespoke network trained from scratch. For the former the subnet is a five-layer fully-connected deep network which takes its inputs from the final convolutional layer of GoogLeNet. The subnet transforms the image features into the classification result we seek. The bespoke network is comprised of convolutional, max-pooling and fully-connected network layers. In both cases an Adam optimizer is used on stochastic mini-batches until the cross-entropy on a held-out validation image set stops improving.

Results show classification performances of over 95% correct are achievable, and examination of the misclassified cases suggests many of these examples are borderline cases in which it is difficult to decide whether they are congested or not. The images above show examples of an image classified as: uncongested (left), congested (right) and one of the borderline cases (centre). Extracting intermediate outputs from the deep networks allows them to be used as a regression model enabling the ordering of scenes, from low congestion to high congestion to be carried out.

The classifier is designed to be one of several inputs to a higher-level capability which combines data from different sources and modalities for the extraction of situational awareness. Future activities will add interpretability to our classification model. This will help address the issue of trust of deep-learned classifiers when sharing intelligence with other processes, and may help share justifications for decisions without having to release restricted input data.

The method examined here assesses the broad state of imagery content, combining information from all over the scene into a conclusion, rather than in the recognition of a specific object. Such an approach might be applicable in many areas of military significance, for example in the recognition of crowd formation. The method presented here is clearly directly useful in the routing of military vehicles in contested or possibly congested urban environments.

### A.3.8 COLLABORATIVE AUTONOMY USING DEEP LEARNING FOR EMERGENCY SCENE ASSESSMENT AND RESPONSE, BY DR. STEFANO CAVAZZI ET AL., ENVITIA

Dr. Cavazzi and his team, Gobe Hobona and Roger Brackin, presented results from a project using collaborative autonomy using deep learning for emergency scene assessment and response. This project was motivated by the problems that emergency responders might have upon arriving at a contaminated scene and their need to identify and process crucial information to plan a response. Within a short period of time, responders have to assimilate as much information about the scene to arrive at an understanding of the hazard. To collect and assimilate this information can present significant safety risks to human responders if they have to personally enter and survey the contaminated scene. Even more so, if the hazard includes Chemical, Biological, Radioactive or Nuclear (CBRN) materials.

Dr. Cavazzi and his team’s products offer faster scene assessment and improved situational awareness to support decision making. The aim of the project was to design and implement a proof-of-concept autonomous decision-making support system that allowed data from multiple sensors and other sources to be continuously analysed, tagged and prioritised to identify the most appropriate courses of action.

The idea behind collaborative autonomy using deep learning was to enable intelligent teaming between individual platforms or software agents such that the data collected by each component enabled the system to learn and adapt. The system made use of an interactive geospatial map display, Deep Learning analysis, Semantic Web technologies, feature detection through computer vision, and a modular open system architecture. Deep Learning is a machine learning approach that attempts to model high-level abstractions in data by using a deep graph with multiple processing layers, composed of several linear and non-linear transformations (Goudos et al., 2016). Deep Learning can adaptively mine features from raw sensor data by transforming original sensor data into a highly abstract expression through the stack of non-linear models. The project also employed Semantic Web technologies to enable the system to make inferences about courses of action. Information from multiple sensors was transmitted to a coordinating metadata registry through the use of service interfaces based on standards from the Open Geospatial Consortium (OGC). The use of standards-based interfaces applies lessons learnt from a recent OGC project “Incident Management Information Sharing (IMIS) IoT Pilot” (Brackin, 2016) which prototyped and demonstrated open system sensor integration for emergency and disaster response.

The developed concept and its implementation demonstrate the feasibility of collaborative autonomy using deep learning to enable emergency responders to assess a hazardous scene by providing real-time sensor data, creating a digital representation of the scene and assisting the commander in decision making. The benefits of the presented research include:

- Improved situational awareness across emergency responders attending a contaminated scene;
- Reduction in the risk to initial responders as they can reach assessments quicker;
- Increased speed of carrying out scene assessment as historical assessments are used to train the Deep Learning algorithms of the system;
- Greater capacity for reach-back to specialist scientific personnel and disposal units.

The research, which was commissioned by Dstl, will be of interest to a number of emergency response and defence agencies interested in collaborative autonomy as a way to and accelerate situational awareness, decision making and scene assessment.

## SESSION THREE: EMERGING ISSUES

### A.3.9 EXPLORING THE VALUE OF TARGET MOTION DEPENDENCY, DRS. SIMON JULIER, & RAGHUVEER RAO

Dr. Simon Julier presented the first talk of Session 3. He noted that situational awareness in urban environments often entails being able to track vehicles to determine their number, location, and possibly even their behaviour. Many approaches, such as Persistent Wide Area Surveillance, have been developed to take streams of multimedia data, process these to detect vehicles, and carry out multi-target tracking. However, the constraints of these environments can make reliable tracking difficult. Challenges include the occlusions caused by buildings and trees, huge variations in illumination and light levels, and the fact that some vehicles, by their nature, can simply be difficult to detect. As a result, many multi-target tracking systems might not be able to initialise tracks or lose tracks of targets over time.

One of the reasons for this difficulty is that many multi-target tracking systems assume that all the targets move independently of one another. In other words, the motion of one target does not influence the movement of any other target. This assumption is mathematically convenient and is a very good approximation for systems in which there is little interaction between targets. However, this is not the case in dense traffic in which any vehicle influences the pattern of behaviour of the vehicles around it. For example, because vehicles cannot drive through one another, a gap in a flow of traffic might indicate the presence of an undetectable vehicle. Because vehicles tend to drive at a desired minimum distance from one another, if a vehicle leaves a lane of traffic, the lane will typically close up.

Dr. Julier reported on an ongoing project which is investigating the use of interactions between targets for vehicle monitoring and situation awareness. In particular, he focused on the issue of how to formulate the estimation problem which efficiently models the interaction of large numbers of targets with nonlinear process models. He presented an approach based on Bayesian networks and evaluate its performance in terms of accurate of estimates and ability to predict the existence of unobservable targets.

### A.3.10 A COLLABORATIVE EXPERIMENT FOR IMAGE INDEXING, DR. LASSE OVERLIER, NOR

Dr. Lasse Overlier provided an overview of the NATO Information Systems Technology (IST) Panel Research Technology Group (RTG) 144, Content-Based Multimedia Analytics (CBMA) (IST-RTG-144) effort to establish an online experiment to demonstrate the potential of activities undertaken by the group in text and video analytics. He explained that by setting up a common shared experimental/research system for machine learning based on current state-of-the-art technologies we can provide data to all these topics, get resource sharing and cooperation demonstrated, and in addition get a new resource for all countries to use. Dr. Overlier explained the motivation is to create a bigger image analysis database to include images with military content, like military planes/trucks/tanks, boats, submarines, guns, weapon systems, people in uniforms/etc. Therefore we want to take the fast track and extend an existing database with (pre-) annotated images containing these objects. He anticipates the task group can do this as a common experiment to which all can contribute to and benefit from. The countries should be able to test the new extended model with their own data without releasing the image content. This could also be done by downloading the (new) model and testing/using it off-line, e.g. on classified images/videos.

A starting point for sharing information among the participating NATO nations would be to establish basic requirements, such as storage space, bandwidth, restricted access, and sharing among members of internal data. Because the NATO shared site was not a candidate for this purpose, a private cloud solution was adopted by Dr. Overlier. He created a shared rtg144.zone.no with access control for members only. This provides a secure upload and download capacity. Currently, existing data has been uploaded to the site that include datasets for machine learning. Examples of type include ImageNET, MNIST, Cifar, UCF, etc. Dr. Gertjan Burghouts, TNO, shared video data sets with group. Specifically ignored for this particular effort are the following: distributed solution, system integration, analysis platform, and real time input & storage.

Dr. Overlier provided an overview of his thoughts on architecture requirements, summarized below.

- Input: Regarding sensor information input, we should allow an unlimited number of sources, from multimedia sources (text, images, videos, etc.) based on bandwidth capacity and easy to manage sensors. Open source data input can be obtained from news reports, social media texts/images/videos, blog postings, and other easy to manage sources.
- Storage: storage and database solutions should be developed with redundancy, the database must be capable of handling huge amounts of data, must work well with machine learning applications and tools, must have fast and easy access from all sites, with secure access, simultaneous write and read from multiple agents/clients, high speed access to recent data & "reasonable" access to old data, and support "real time" recent data queries and extraction.
- Multimedia: Data should support streams of text, images and video and provide indexing and annotations online (automatic) or offline (manual/tools) through:
  - Semantic identifiers (text/images/videos)
  - Object identification (text/images/videos)
  - Person identification (text/images/videos)
  - Activity identification (text/images/videos)
  - Location detection (text/images/videos)
  - Manual tagging (text/images/videos)
- Machine learning: Would be used to enhance sensor input, for example:
  - To identify in images/video:
    - Date and time, people, objects, activities, locations, authenticity
    - Identify falsified information of same types
  - To identify in texts and social media:
    - People, activities, language, fake content (identity, news, claims, …)
  - Cluster analysis
    - Use date/time, people, location, activities, objects,
  - Feedback to indexing
- Analysis: Needed capabilities include fast and simple access to all data and automatic analysis results, easy to add analysis tools and/or results, especially easy to integrate with machine learning tools, and background processes working 24/7.
- System: Requirements include ease of management and maintenance, deployment, testing of new algorithms, models, applications, and scaling considerations. Distributed system requirements include redundancy and the expectation that all installations must have potential for on-site access to data.
- Software candidates were identified by Dr. Overlier for consideration by the larger audience. By type of application, these include:
  - Gathering data from sensors and sources
    - Apache Kafka
    - Apache Spark Streaming
  - Storage -- no public cloud solutions
    - Elasticsearch
    - Splunk
    - HDFS - Hadoop Distributed File System
    - IPFS

- Databases
    - Apache Kafka, Apache Spark
    - MapD Core
    - MongoDB
- Analysis backend
    - Apache Spark
    - Apache Kafka Streams
    - Apache Flink
- Machine learning toolkits
    - Tensorflow
    - Apache Spark MLlib
    - Or: Caffe, Theano, Torch
- Image and video parsing tools
    - FFmpeg
    - MEncoder
- Integration and management
    - Python & Jupyter
    - Docker, Kubernetes
- Visualization and front end
    - Kibana
    - MapD Immerse
    -

Dr. Overlier outlined the experiment process in the following way. First, the contributing nations would provide resources for running a computer in their national lab with a distributed machine learning library available to all partners. Resources could be 1) a simple GPU system in a couple of countries, 2) access to some resources at any HPC facility that will provide them, or 3) renting GPU-resources in the cloud. Second, the members would take one of the large image databases already pre-trained in an ML-model. If using TensorFlow we might use the pre-trained Inception-v3 (http://arxiv.org/abs/1512.00567). Third, partners would look at the categories in the existing model, look through our images-to-add, and decide which categories we must add to the model. We create a common list for all countries, and identify a method to modify the pre-trained model to allow more categories for new images without causing problems with the old data. Fourth, all countries add (categorized) images to the system for extended training of the model. The number of training images in each new category should be close to the number of images in the "old" categories. Fifth, after a training period, all countries test to see if the model is able to categorize new images correctly, meaning having close to the same accuracy as the "old" pre-trained model had. Here the countries should also try to use the model for video analytics as well. Sixth, if successful everyone can test/implement for internal use and we distribute the new model to all NATO countries expressing interest for testing the system. If unsuccessful we go back to step 3 or possibly 2, but the next round will at least have the new categorized images ready for extending the model and will hopefully take a shorter time. If results are good we can retrain the entire database of images, both old and new with the extended amount of categories. This would be expected to result in a better model, which will only be verified by testing. Once established, the system can be used in different ways to document results in all topic areas through: indexing and searching using text analytics, automatic triggering, and video analytics.

## A.3.11 INTELLIGENT OUTREACH TO AI CAPABILITIES, MR. PHIL GIBSON ET AL.

Dr Russell McKinlay presented the third talk of Session 3. He and his colleagues, Mr. Benjamin Howes, Dr. Michael Harries, Mr Phil Gibson, and Dr. Bob Madahar, provided an overview of the Defense Science and Technology Laboratory (Dstl) effort to reach out to external partners in the area of Analytics and Artificial Intelligence (AAI). He explained that one of current challenges in AAI enabled by Machine

Intelligence/Deep Learning, especially in government, is the availability of sufficient Suitably Qualified Experienced Personnel (SQEP). As a result it is proving challenging to build sufficient and enduring capability that can readily meet the increasing demands in defence and security for data science and AI experts to research and address ‘big data’ problems. Hitherto it has been possible to address such capability challenges by contracting the research to external capability providers in industry and academia working in the military and civil sectors (the traditional approach). However this too is also becoming challenging. Promising applications of machine intelligence developments, such as deep convolutional neural networks, with record breaking performance for text, audio and image processing has stimulated significant growth in the civil sector thereby stressing the capability pool further. Hence different approaches and alternative instruments need to be considered and assessed which is the subject of this paper.

Dr. McKinlay outlined how Dstl has gone beyond her traditional sphere of engagement to reach out to others with capabilities in AAI. Science and Technology (S&T) is evolving rapidly in these fields and UK Government must demonstrate its ability to respond. Such an environment demands partners who are capable of working with dynamic requirements where innovation and collaboration bring success. This requires Dstl to push how she conducts research and development to challenge the current processes and practices and to reach out beyond traditional suppliers to ensure access to the most skilled and experienced data scientists. Dstl sought to meet this requirement and completed a number of new approaches to delivery which involved improving relationships with non-traditional suppliers, closer collaboration cross government, more transparent ways of working and greater trust with partners.

Dr. McKinlay presented the approach taken and the results from the delivery of three different initiatives addressing specific defence and security challenges. These included 1) the Kaggle Feature Detection Challenge, development of one of the UK Governments first ‘crowd-sourcing’ platform, 2) the Data Science Challenge – a re-design of traditional contracted research, and 3) the Digital Catapult Pitstop, a Sandpit type event on autonomy and partnering with the Alan Turing Institute.

The results will show that for AAI Dstl were successful in – a) broadening our engagement with suppliers beyond the traditional, b) building datasets, c) accessing innovations for image feature detection, c) linking to new capabilities and SQEP networks, d) raising complex problems defence and security is wrestling to new audiences, and e) demonstrating that these new instruments are cost effective and efficient in delivering results. Such approaches are not limited to AAI but can be used across defence and security S&T to access new ideas and capabilities.