Content-based multimedia analytics: US and NATO research

Elizabeth K. Bowman, Gertjan Burghouts, Lasse Overlier, Sue E. Kase, Randal J. Zimmerman, et al.

Elizabeth K. Bowman, Gertjan Burghouts, Lasse Overlier, Sue E. Kase, Randal J. Zimmerman, Serena Oggero, "Content-based multimedia analytics: US and NATO research," Proc. SPIE 10653, Next-Generation Analyst VI, 106530H (27 April 2018); doi: 10.1117/12.2307756

SPIE.

Event: SPIE Defense + Security, 2018, Orlando, Florida, United States

Content-Based Multimedia Analytics: US and NATO Research

Elizabeth K. Bowman*[a], Gertjan Burghouts[b], Lasse Overlier[c], Sue E. Kase[a], Randal J. Zimmerman[d], Serena Oggero[b]
[a]Army Research Laboratory, 321 Johnson St, APG, MD, USA 21005; [b]TNO Intelligent Imaging Oude Waalsdorperweg 63 2597 AK The Hague The Netherlands; [c]FFI Instituttveien 20, 2007 Kjeller, Norway; [d]Zimmerman Consulting Group, LLC Leavenworth, KS, USA

ABSTRACT

The US and 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 influenced by the unprecedented rise of information sharing technologies and practices, where mobile access to social networking sites is ubiquitous. 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. US science and technology investments in Artificial Intelligence and Machine Learning (AI/ML) and Human Agent Teaming (HAT) are increasingly focused on developing capabilities toward that end. A critical foundation of these technologies is the awareness of the underlying context to accurately interpret machine-processed warnings and recommendations. In this sense, *context* can be a dynamic characteristic of the operating environment and demands a multi-analytic approach. In this paper, we describe US doctrine that formulates capability requirements for operations in the information environment. We then describe a promising social computing approach that brings together information retrieval strategies using multimedia sources that include text, video, and imagery. Social computing is used in this case to increase awareness of societal dynamics at various scales that influence and impact military operations in both the physical and information domains. Our focus, content based information retrieval and multimedia analytics, involves the exploitation of multiple data sources to deliver timely and accurate synopses of data that can be combined with human intuition and understanding to develop a comprehensive worldview.

Keywords: social computing, multimedia analytics, information environment, text and video analytics

1. INTRODUCTION

Social Computing technologies that harvest forensic social and digital media will boost agility of military operations both in the physical and information spaces through deep understanding of adversary perspectives, intent, and threats. The pooling of social science theory, aggregated data on social phenomena, and computational science are enabled by the nature of human behavior in the new Information Environment (IE). In this paper we describe recent doctrine describing the concept of 'Operations in the Information Environment' and associated data analytics that characterize key trends and potential threat activities. This is followed by a social computing paradigm used to formulate research strategies. We then describe a short scenario that was used to demonstrate text, video, and imagery exploitation capabilities developed by members of the NATO Information Systems Technology Research Technology Group 144, Content-Based Multimedia Analytics (NATO IST-RTG-144). The overall concept is to correlate video and text analytics for cross-cueing between data types, which results in a significant increase in the Commander's and Staff's ability to understand adversary perspectives, intent and threats.

1.1 Operations in the Information Environment

Joint and Army doctrine are evolving with respect to identifying and assessing dynamic threats formed and executed in the IE. Information was recently declared the 7th Joint function in the US Department of Defense (DOD) [1], which recognizes it an instrument of national power. Commanders at all levels must understand the impact of an information-rich world on the planning and conduct of military operations [2]. This is a complex environment where state and non-state actors exert influence through combinations of truthful, biased, and false messaging from human and machine sources (e.g., trolls and bots). [3] Commanders and staffs must also understand the multifaceted nature of human networks

* Elizabeth.k.bowman.civ@mail.mil, 410-278-5924

Next-Generation Analyst VI, edited by Timothy P. Hanratty, James Llinas, Proc. of SPIE Vol. 10653, 106530H · © 2018 SPIE · CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2307756

including those expressing adversary, friendly, and neutral affiliations because each impacts military operations (which often shift). [1, 3, 4] The complexities of the operational environment (OE) argue for a systems perspective that allows a holistic view across the political, military, economic, social, information, and infrastructure (PMESII) systems [5]. Such an integration of PMESII system parameters is troublesome due to 1) difficulty capturing relevant information (especially S and I), 2) uncertainty about how to combine data elements when the systems are not independent, and 3) lack of agile algorithms capable of accounting for shifting system characteristics.

Within the PMESII paradigm, it can be argued that the Social and Information elements are the most dynamic and difficult to characterize, prior to the rise of mobile devices, enhanced internet availability, and social networking platforms. New technologies have evolved to identify human and machine influencers, capture topic trends and associated sentiment, expose false claims or images, and characterize social group norms and activities. These technologies can enable the formulation of a framework based on human-domain-centric theories and analysis for planning and executing operations in the IE (OIE). The framework can be incorporated into the joint process of intelligence preparation of the operational environment (JIPOE) [4] and iteratively developed and refined by harvesting "forensic" social and digital media data while applying new research contributed by academic partners (See Fig.1). For example, social scientists refer to "cognitive vulnerabilities" as weak points in individuals' critical thinking in which they are willing to accept ideas and information confirming their existing sentiments and beliefs. "Targeted biases" is the strategy of targeting previously held biases that have deep emotional value and little critical evaluation. From this perspective, the formulation of an effective OIE strategy depends on a combination of topics of genuine interest to the public with a distorted, distracting, dismaying or highly entertaining dismissive content so as to appeal directly to emotions and pre-existing biases.

The goal of the framework is to determine the advantages and disadvantages of OIE approaches and combinations of approaches in order to increase understanding and quantify capabilities, limitations, uncertainly, and error associated with each approach under different IE conditions. In the next section we look at OIE capabilities and tasks to identify enabling data analytic technologies.

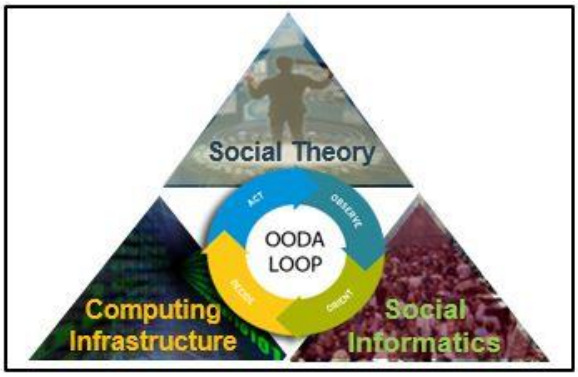

Fig. 1 Conceptualization of framework using computational social science technologies to integrate innovative human-domain-centric theories and approaches providing military commanders and staff with the information necessary to develop courses of action and make decisions.

We consider the range of trends that impact OIE across the world order, human/geography, and science/technology domains as articulated in Table 1. Each of these themes relate to the co-evolution of mobile computing technologies and social networking behavior that witnessed the rise of information as a powerful social force. Since the world first witnessed the power of social networking in the Arab Spring revolution of 2011, information has become multifaceted. It is pervasive and allows people to organize faster and influence socio-cultural norms. It provides insight into worldviews that frame perceptions and attitudes that drive behavior. Technology has elevated information to an instrument of power that can be wielded by individuals in groups, politics, economics, and warfare. As such, it has changed the way information is generated, shared, and interpreted and allows non-state actors to wield influence around the globe. This diffusion of power over information is shifting cultural norms around the world.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the authors'' employers or their governments.

Table 1 Themes and trends that impact Operations in the IE [5]

Thematic Area		
World Order States with the political will, economic capacity, and military capabilities may compel change at the expense of others	**Human Geography** Social, economic, environmental, and political pressures will push states past the breaking point and creating wide-ranging international problems	**Science, Technology, & Engineering** Technological parity and innovative mixes of high and low technology may allow adversaries to more effectively challenge U.S. interests
Trends that Impact the How, What, and Why of Operations in the IE		
• Shifting Strategic Relationships • Pursuit of Regional Primacy • Regional Powers with Global Reach • Evolving Roles of International Institutions • Consequences of Fragile/Failing States	• Intensifying Consequences of Population Growth & Migration • Urban Concerns as Global Security Issues • Evolving Ideological Conflict • Alternative Hubs of Authority • Rise of Privatized Violence	• Multidisciplinary Scientific Research • Systems and Systems Integrations • Emerging Measure / Countermeasure Competitive Spaces • Proliferated Information Technologies • Emergence of New High-End, Capital Intensive Capabilities

1.2 Social Computing for Context

Our interests in using social computing to derive context from multi-domain data center around these factors: Human, Information, Interpretation, and Influence. In the Human domain, we are concerned with individuals, social groups, organizations, analysts, and decision makers. With respect to Information, computer-mediated messaging, text, images, videos, geo-locations, and networks are driving concerns. Topic modeling, social network analysis, narrative identification and exploration, pattern of life, and relationship linking are modes of analysis that allow a user to identify messaging variations, such as deception, fake news, disinformation, misinformation, bots, distortion, contagion, and spread. Influence is considered to foster tailored truthful messaging with relevant platforms and emic (insider) perspectives that are supported by Tactics, Techniques, and Procedures (TTPs) for countering adversary messaging.

US doctrine is developing rapidly to establish common understanding and procedural guidelines for operating in the dynamic information environment and interacting with adversary, neutral, and friendly groups [3, 4, 5, 6]. In [5] four capabilities are identified that will provide Joint and Coalition Forces with dominance across OIE. A fundamental requirement for this capability is to characterize and assess the informational and human aspects of the operating environment. These aspects include understanding perceptions and attitudes that drive behavior, how these can change over time, and the ability to share contextual understanding.

1.3 Multi-Media Analytics Tools to Drive Context

In the context of this paper, a variety of video and text analytic tools are used together to drive rapid determination and monitoring of context relative to the military objectives. We briefly review them here but their capabilities will be described, in turn, in the following sections.

1.3.1 Combining Video and Text Analytics for Multi-Source Intelligence

[7] provides insight gained from a small-scale experiment into how combining text and video analytics can support intelligence analysts. Text and pictures were obtained from social media and tracks of people and video were labelled in aerial footage by video analytics algorithms. 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 experiment, the analyst’s search for ‘jeep’ returned a Tweet with three persons standing near a jeep and 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.

1.3.2 Live Image and Video Detection

The huge number of Tweets in a given period of time makes it impossible for an analyst to view and mark the content relevant or not. Using a simplified tool for dynamic development and testing, Jupyter[8], we have designed a capability to shorten the list with flexible triggers for deeper analysis. This 'Demonstrator' is used by an analyst to specify the relevant tags, users and keywords. The monitoring of the active social media platform(s) will then result in a filtered list of the relevant postings and links for options and possible actions for further analysis. The Demonstrator works with several aspects of Tweets, to include text, image/video clips, geolocation markers, names, and locations. One capability of Demonstrator is to identify if an image/video clip has been observed online before, and if so, by whom, when, and where? This enables the analyst to rapidly identify disinformation or false messaging. A capability is to automatically index objects in the image/video clips (e.g., tanks, weapons, vehicles, drone strikes, etc.) for automatic match/detection and more in-depth human processing.

1.3.3 Text: Social-media Understanding and Reasoning Framework (SURF)

SURF [9] is a text analytics platform that determines threat levels of individuals based on their social media interactions using machine learning and bioinformatics-inspired algorithms to find and accurately classify group membership. The tool currently operates with Twitter messages (Tweets) and is language agnostic because classification is based on network patterns. SURF has been successfully tested on Arabic, English, French, and Spanish. The classification algorithm is based on a proven formula used to classify biological entities from their protein interaction network. The SURF adaptation to social media is >85% accuracy using cross-folded validation with ten random K-folds. Group models are built from examples of true positive and true negative messages; currently four models exist. These are businessmen, hackers, Islamic State of Iraq and Syria (ISIS) 1 and ISIS 2 (e.g., earlier and later ISIS group adaptations). SURF provides two categories of group membership: nodes and ego networks. The first identifies individuals active in the Twitter network under study and the second considers friends and followers of individuals. The second category is important to consider because of the combined influence friends and followers can have on individuals.

2. ANALYTICS WORKFLOW

The concept for analytics is built around the workflow illustrated in Figure 1 [10]. In this depiction of an intelligence analysis process, analysts are searching for specific vehicle targets related to insurgent activity in full motion video (FMV). Text inputs in the form of Twitter messages indicate that 'jeeps (are) on the move' and 'jeeps are leaving the crossroad' and these are used to cross-cue aerial assets for verification and tracking. This type of iterative cueing between text and video has been demonstrated to speed analyst awareness of threats [11].

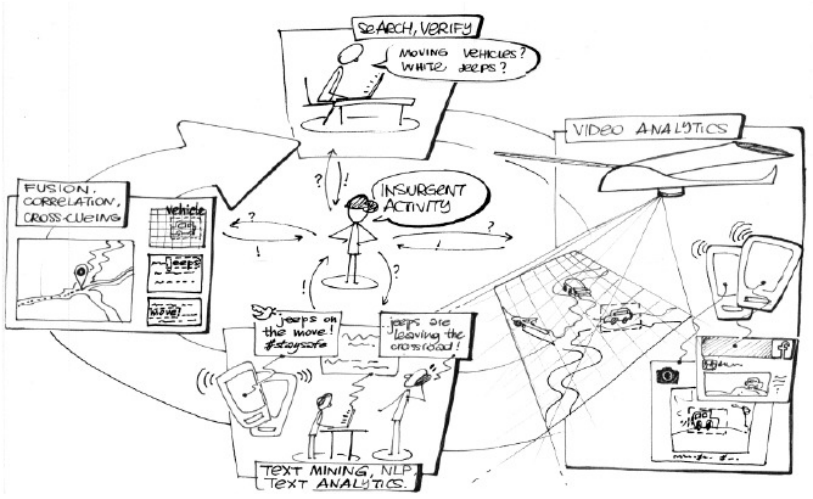

Figure 2: Intelligence Gathering Based on Text and Video [10]

3. ACQUIRING INTELLIGENCE FROM TEXT AND VIDEO

3.1 Text and Video Data Analytics

In this section we discuss how text and video analytics are integrated by the analyst cell to rapidly form situational awareness. The starting point for the analyst's search is a simulated Twitter message containing text and an image, displayed in Figure 3. The image displays the jeep with the woman and the two men. Analysts report that the incoming videos from social media and surveillance assets show increasing volume of vehicles and people. They need to determine types of activity that might be occurring to indicate intentionality. Twitter reports suggest that a woman in a jeep with two men might be of particular interest. The analyst wants to cross-cue from the Twitter reports to surveillance data from other sources including UAVs.

Figure 3. The Twitter message that suggests that a woman in a jeep with two men might be of particular interest.

The analyst will search for the woman in surveillance data, consisting of full-motion video, recorded by a UAV. In total the video spans 32 minutes, including infrared and daylight imagery, in various field of views. The recordings are accompanied by metadata about the position of the UAV and its field of view, as shown in Figure 4.

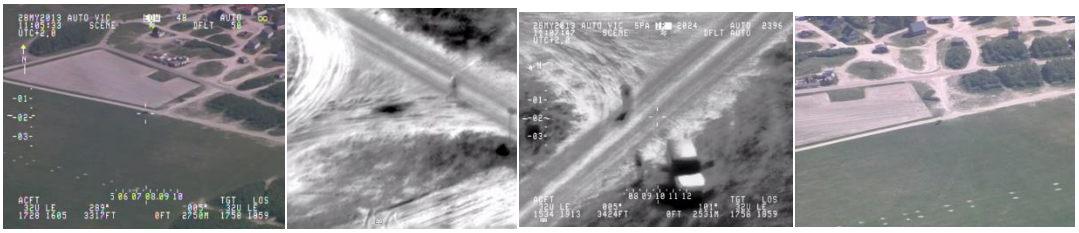

Figure 4. Impressions of the full-motion videos recorded by the UAV.

3.2 Algorithm

The analyst searches the woman in the Twitter image in the UAV imagery. The analyst indicates the woman in the image, by selecting a bounding box, from which a 'snippet' is extracted to show the picture of the woman, as depicted in Figure 5.

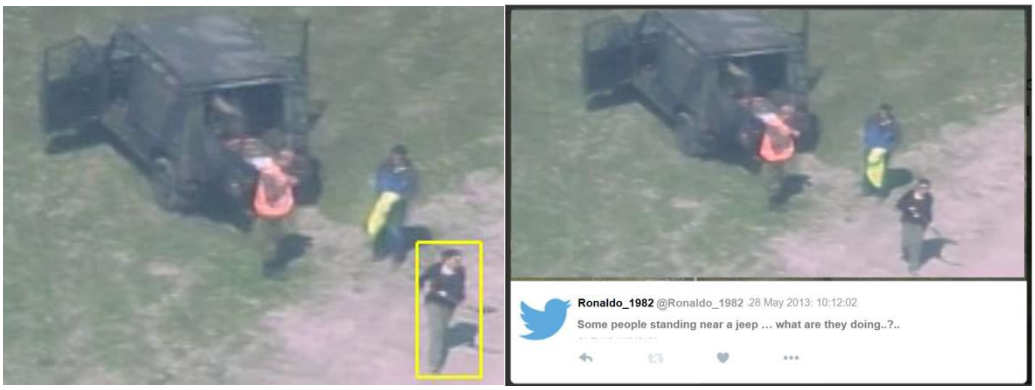

Figure 5. The analyst selects the woman as a starting point for the search.

The analyst wants to use this snippet to cross-cue to the UAV videos. To limit the search, persons and vehicles are detected in the video. Persons are detected by the ACF detector [12], which provides bounding boxes in video frames. The boxes are checked against the expected size of persons, which is derived from the metadata about the video. The ACF detector is limited to persons that appear relatively large within the image, typically larger than a height of 50 pixels. Persons that appear smaller in the image, are detected when they move, by a moving object detection (MOD) method. For MOD, the method in [13] was adopted and modified to search specifically for moving persons by considering the expected size of persons as derived from the metadata. The boxes provided by ACF and MOD are tracked to obtain temporal consistency and to remove erroneous boxes. The tracker is based on kinematics of the boxes, where speed and orientation in world coordinates are derived from the metadata. The association of boxes is performed by considering the consecutive overlap between the projected previous box and the current box.

The snippet of the person of interest, i.e., the woman, is now searched within all tracks in the UAV videos. From the tracks, snippets are extracted. The snippet of the woman (i.e., the query snippet), is compared with all other snippets (i.e., the candidate snippets). For comparison, the re-identification algorithm from [14] is deployed. It provides a score between two snippets of 1 (similar) and 0 (dissimilar). This score demonstrated a large degree of robustness for a change of camera, viewpoint and illumination [14]. This is beneficial for searching in a large volume of surveillance data, potentially recorded by various cameras. Pair-wise scores are computed between the query snippet and the candidate snippets.

The similarity scores are presented to the analyst on an intuitive two-dimensional canvas. For the projection onto the canvas, the t-SNE embedding [15] is applied. This embedding groups similar snippets together, with the rationale that local structure is preserved, and dissimilarities are less important. The grouping of similar snippets aids the analyst to search for the woman in the surveillance video data.

A graphical user interface (GUI) has been developed to sift through the canvas of snippets. In practice, there may be a large amount of snippets. These will be projected on top of each other. The GUI enables the analyst to show more or less snippets on the canvas, and to navigate through the canvas. A single snippet often does not provide a good impression of the person's activity. When a snippet is clicked by the analyst, the accompanying part of the video is showed. While the analyst is searching, snippets can be confirmed and added to the search results. The confirmed search results can be showed on a timeline viewer, such that the analyst can playback the related video parts and analyze intentionality.

3.3 Results

The graphical user interface with the image snippets of persons is shown in Figure 6. The interface is a canvas of snippets, where their position is based on the similarity embedding (see previous section). On the left of Figure 6, all snippets are shown. On the right, the zoom-in mode is shown, which are the snippets inside the red rectangle which is shown on the left. When a snippet is clicked by the analyst, the video part plays, which is shown in Figure 6 inside the pop-up window. This enables the analyst to explore the persons in the imagery data and their activities. Figure 6 illustrates this by the search for the person of interest (the woman). She was found in a UAV video, while carrying a shovel.

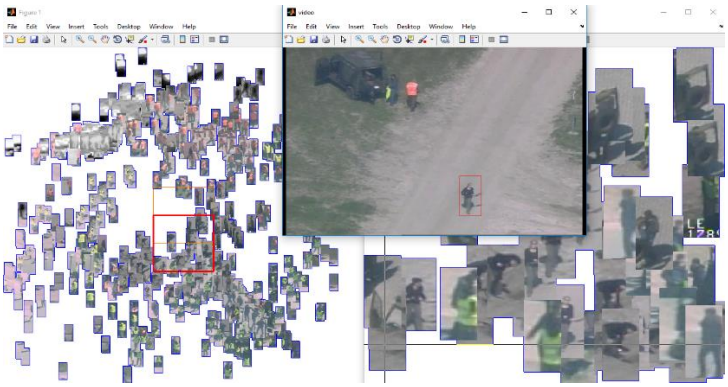

Figure 6. The graphical user interface with the image snippets of persons and their local grouping by the embedding based on the similarity measure (see text).

The person of interest (the woman) was found multiple times in the imagery data. This is shown in Figure 7 on the right side. The analyst has confirmed these appearances by double-clicking the snippets. In Figure 7 this is depicted by the green contours. The analyst has given these appearances the label ‘woman at jeep’. This tag enables the analyst to trace-back to the original imagery data based on a text query.

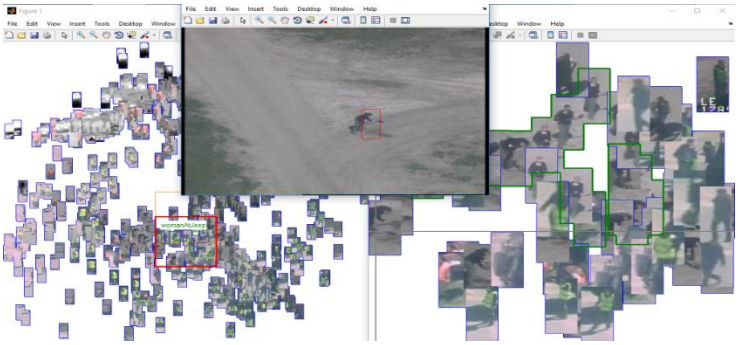

Figure 7. Multiple appearances of the woman have been found by the analyst (in green).

To determine intentionality, the analyst needs to analyze what happened where and when. To aid the analyst with this task, the tagged appearances of the person of interest (the woman) are shown on a timeline (when), map (where), accompanied by the imagery (what). Figure 8 shows the timeline viewer, with its timeline (bottom), map (left) and imagery (right). On the timeline, the analyst can mark a time point (one image) or interval (part of the video). The interesting behaviours can be indicated by a bookmark with a tag. Figure 8 shows the moments that the woman steps out of the jeep, then digs, and finally leaves the area.

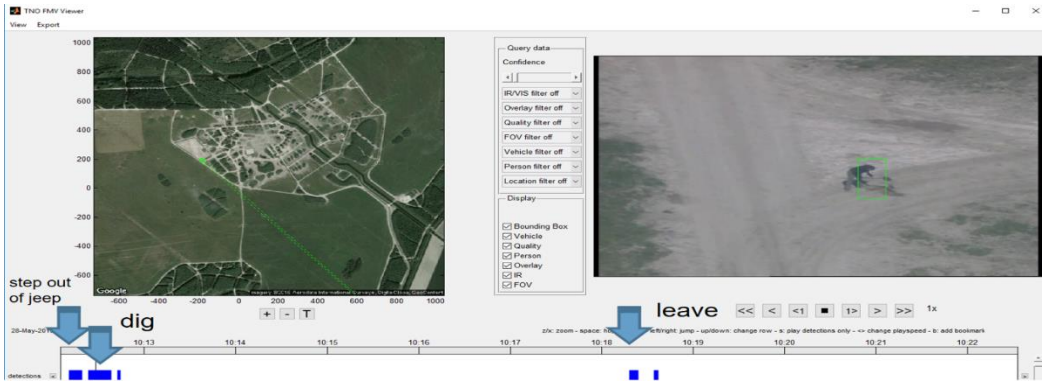

Figure 8. Timeline view of the found appearances of the person of interest and her activities.

4. LIVE IMAGE AND VIDEO DETECTION

4.1 Demonstrator

The sheer number of tweets in a given period of time makes it impossible for an analyst to view and mark the content relevant or not. [16] has designed a capability to shorten the list with flexible triggers for deeper analysis. This 'Demonstrator' is used by an analyst to specify the relevant tags, users and keywords. The monitoring of the active social media platform(s) will then result in a filtered list of the relevant postings and links for options and possible actions for further analysis as shown in Figure 9. The Demonstrator works with several aspects of Tweets, to include text, image/video clips, geolocation markers, user names, relevant links, and locations. The capability of Demonstrator, an early version of which is shown in Figure 9, is to identify if an image/video clip has been observed online before, and if so, by whom, when, and where.

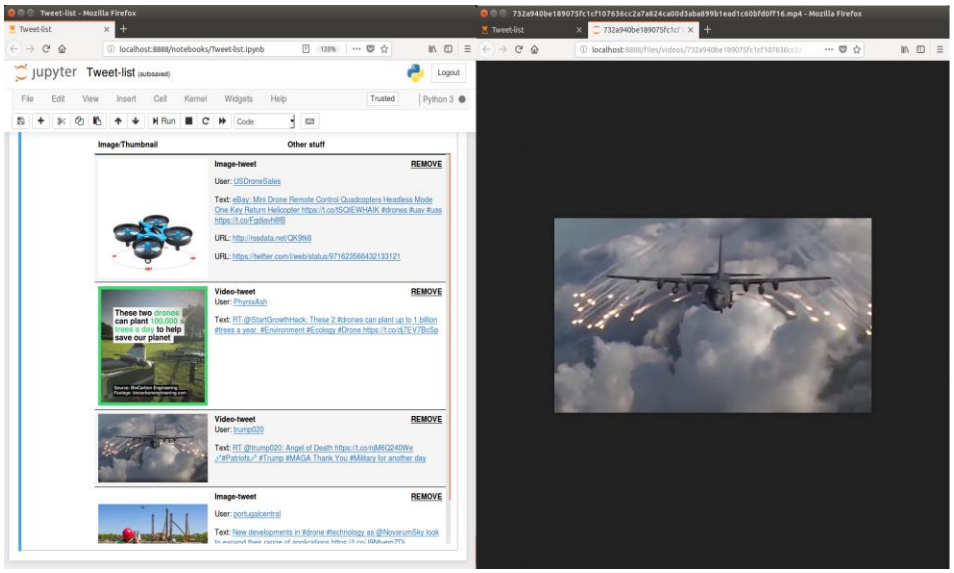

Figure 9: Stream of relevant and filtered images and videos shown in list to the left. On the right manual inspection of item in list from local repository

The user is presented with the option to manually specify the relevant tags, users and keywords. The monitoring of the active social media platform(s) will then result in a filtered list of the relevant postings and links for options and possible actions for further analysis as shown in the left part of Figure 9. This list will enable the analyst to rapidly address the detailed content of the posting and look for more relevant material through the user behind the posting. Using one of the items in Figure 9 as an example, the analyst can on the left choose to more deeply examine the marked video for accuracy as shown to the right in the figure. This in-depth analysis will happen from the locally stored copy of the video identified in the posting, and will not go on-line again to look for postings that may have been changed or deleted.

To go more in depth of how the Demonstrator works we will first describe the different elements it consists of and then we will take a deeper look into each of these elements. The demonstrator/system is configured to follow a set of tags, users and/or keywords. It enables automatic pull down of the complete content of related messages, together with any content the messages link to from the current social media stream. Some information may also be fetched from last two weeks, but if earlier historical data is needed this must be purchased from commercial services and is not a part of our Demonstrator. Any image or video linked/referenced in a social media posting on-line which matches our set of criteria will now be downloaded and automatically analyzed in-depth with regards to:

- Does the posting have identifying marks, like geolocation markers, specific content, names, groups, etc.?

- Does the included or referenced image/video contain marked objects like tanks, weapons, drone strikes, etc.?

- Has the image/video clip been observed online before? When, where, by whom?

- Does the image/video contain text like posters, airplane identifiers, names, and/or textual propaganda?

- What is the textual content and its context for this posting?

- Does the image/video contain faces, and do the faces exist in the database?

- Does the video contain any events marked for more interest, like object exchange between humans, digging a hole in the ground, throwing an object, etc.?

For each of these in-depth image and video analyses there are special tools and systems working in the background of our architecture. Identifying marks are pulled from the text and metadata with each social media entry. These are fetched with the original stream of postings and used to pull down the necessary extra data related to this posting, like images, videos, geolocation tags, etc.

The rest of the technologies and systems of the Demonstrator listed above will be explained in the following sections.

4.2 Objects marked for special interest

Using machine learning for detecting objects in images is a growing area of research and many machine learning models have been created for this including deep learning techniques [17, 18]. These work out-of-the-box for the predefined classes of objects, but if you need to detect new classes you have to either train your own model or retrain the existing model with new and/or additional data. We want to recognize objects not in the major existing detection models and have used Tensorflow to retrain an existing model with new data from an idea and description made on-line at [19] and [20].

In order to perform the retraining and integrate this into the Demonstrator we downloaded a set of 100+ images of tanks and 100+ set of drones and used Tensorflow to retrain the final layer of an existing deep learning model for object detection. The process of retraining for object detection involves manually tagging each image with items of interest with a corresponding bounding box for the object as shown in Figure 10.

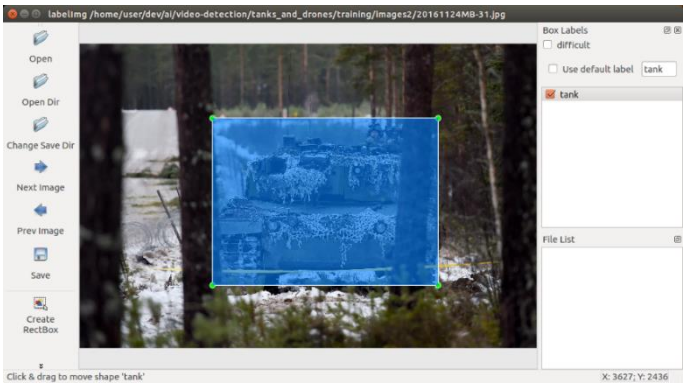

Figure 10: Labeling images for retraining deep learning models to look for new objects

This enables the Demonstrator to return matching objects and include a bounding box for where in the image the object has been detected.

The retraining of the system was performed on the SSD_mobilenet_v1 [20] model which has more focus on speed than accuracy. The main challenge with detection within each image in a video is also in this case the amount of processing power for medium sized videos. In our case a good laptop with a dedicated graphics processor was able to run the detection algorithm on 10-20 frames per second. As described earlier optimizations can be done by looking for significant changes in scenes and/or just using every 10 frames.

4.3 4.3 Pre-existing video clip or image

One problem with social media postings is that they may go viral before any kind of authenticity has happened. If a video posting of a bombing of civilians in a remote area of the world is made, and this goes viral, a lot of opinion can be formed regardless of whether this is true or not. It may be that the posting is from a prior event with scrambled, edited, tiny fractions of clips put together for this purpose only.

Our system will be able to determine whether a video posted on-line does exist in our prior database or not. We only have a system for searching through large amounts of prior videos, but the real challenge is in the startup and collection of the database. The indexing technology being used is called perceptual hashing and is described primarily for use with images, but has also been used for videos [21].

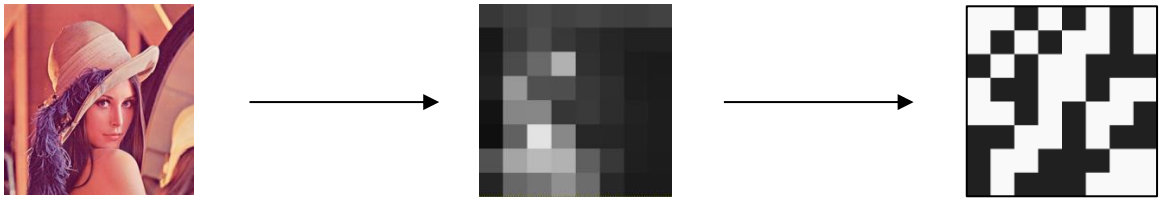

Figure 11: Process of transforming image into a minor representation (middle) and into a 64-bit perceptual hash (right).

There are many ways perceptual hash may be calculated, but the main principle is shown in Figure 11 and we have used the same principle as described in [22]. To convert an image to a perceptual hash you first convert the image into a small representation of your image in "greyscale" values as shown in the middle of Figure 11 as an 8x8 image. This representation can focus on removing minor details and small shifts by just taking an average representation of the underlying area, or it can be a Fourier transformed representation of the image to enable more detailed filtering of the level of details [22]. Usually this representation is made into the same size as the number of bits in the resulting perceptual hash. There are also many methods for converting this representation into the resulting perceptual hash, like average hash, pHash and distance hash. Average hash is made by calculating the average value of the representation pixels and marking each bit with regards to their value being above with "1", and below with "0" and thereby end up with a 64-bit representation of the image.

By calculating this on the stream of images in a video you get identifiers for each period of the video. Our Demonstrator builds a database of these identifiers and compares incoming video clips to this database to see if the video has been published earlier.

4.4 4.4 Textual content of image or video

Optical Character Recognition (OCR) technology is a completely separate area of research outside the scope of our research area, but an open source system [16] has been tested and integrated and works partially with plain English text inside the images on monotone background. As shown in Figure 12 this software does not work very well on angled text nor with text on image backgrounds, but we have still integrated it to be able to make some automatic triggers have information from images and videos.

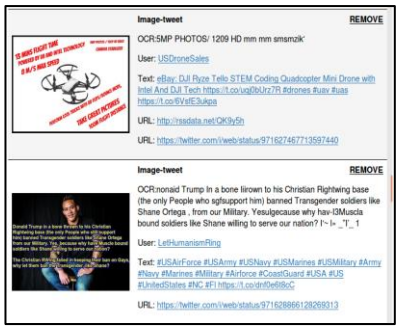

Figure 13: Current integrated OCR tool does not work well with angled text or changing background.

Another important challenge with the open source system is the speed being used for video clips. The best system we have tested uses approximately one tenth of a second for each image, and while this is quite acceptable for images pulled from the feed, a video of a few minutes will now become a challenge. One solution may be to have the video divided up into scenes and only scan each major change in video content for new textual content. The technology for scene splitting does exist [23, 24] and will enable much higher speed for the OCR module, but is not integrated into the current system.

More complex commercial systems can also be integrated into the architecture if they exist. At the time of writing we do not know of any available system with satisfying performance.

5. SOCIAL UNDERSTANDING AND REASONING FROM TEXT

The Social Understanding and Reasoning Framework (SURF) capability has been developed for the US Department of Defense (DOD) for rapid awareness of dynamic threat networks from information obtained from members of adversary groups and their use of social media platforms. While traditional approaches to social network analysis use keyword searches and content analysis of messages, these methods are brittle and incomplete, especially when used with foreign languages. The approach used in the SURF software classifies individuals based on their network motifs. This is a bio-inspired approach using persona-based search and classification based on the structure of a person's network. For this reason, SURF is language agnostic because it does not interpret words contained within messages.

A case study approach is described in this paper to determine the ability of the SURF technology to fully identify adversarial group members contained within a Twitter dataset. We used a scenario to drive the case study that involves a tactical intelligence analyst operating in Afghanistan charged with combatting a growing insurgency by the Islamic State of Iraq and Syria (ISIS). The Priority Intelligence Requirements (PIR) include identifying the most influential individuals associated with ISIS in this region and the network linkages between nodes to understand possible relations to anticipate the nature of future threats. These PIR require the analyst to conduct three tasks: analyse the social network, identify the key influencers, and analyse the message traffic between key influencers (not for message content, but for structure).

The study used Twitter feeds that represented 207 nodes (Twitter users) and 1,534 edges (friend and follower relationships on Twitter). Following a short discussion of the SURF workflow, the three tasks will be shown with screenshots.

5.1 SURF Workflow

In a military setting, 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 and metrics (e.g. tweet frequency).
5. Meaning Determination. Based on the motifs and their feature sets, a machine learning classifier is used to identify person of interest (POI) or forecast an anomalous event 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, cue other sensor collections and ultimately take appropriate action.

The workflow begins with the analyst using the SURF platform to build and extract a social network from a custom watchlist, which is then ingested by a network graphing tool such as Gephi. The network is comprised of Egos (individuals who appear in the analyst's custom watchlist) and their Extended Network (individuals who are friends or followers of those on the custom watchlist). For this case study, the Egos were identified as likely ISIS or ISIS-affiliated. These nodes have an "ISIS" score as an attribute reflecting the degree to which the node is likely ISIS-affiliated, from 0.00-1.00. The Extended Network nodes do not have an ISIS score, but are indicative of who is influencing the network of likely ISIS affiliates. In the SURF platform, an individual is important if the Eigenvector is high (meaning that the person is well connected to other well-connected individuals and is otherwise highly central to the network), In-Degree is high (individual is followed by many other users), Out-Degree is high (person follows many other people), and Degree is high (how many others in the network this person is connected to, regardless of directionality).

The first step in the SURF process is to visualize who in the network are Egos and who are members of their Extended Network. The SURF algorithm performs targeted filtering on large watch lists or incoming data streams to provide usable network sizes and enable graph visualization. Simple network statistics and visualization techniques are then used to further

refine the analysis. The Eigenvector Centrality measure was used to provide an index score of 0-1 of how well connected a user was to other well-connected users. The result was a directed network of all nodes in the network (ISIS ego nodes and ISIS extended network) with green indicating ISIS affiliation. Figure 13 shows the first social network rendering of the entire dataset. The next step is to filter the social network to bring more clarity to the task. Figure 14 shows the filtered social network with only Ego nodes with more than one link appear. These two visualizations of the ISIS social network allow the analyst to understand the configurations of the two main node types in the data (ISIS nodes and friends/followers).

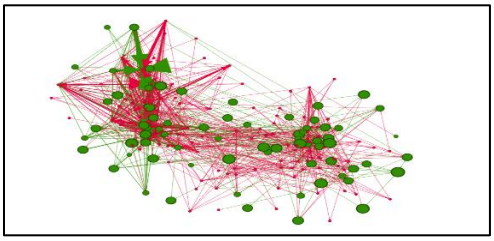

Figure 13: Nodes colored by type by ISIS supporter importance

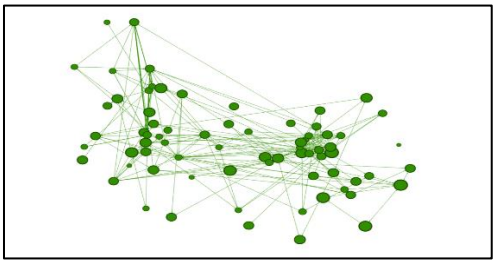

Figure 14: Filtered so only Ego nodes with 1+ link show up

5.2 Step 2: Identify Who in the Extended Network Is Most Influential

It is important for the analyst to understand the friends and follower network of the ISIS nodes to understand to whom the latter are listening. This could help formulate counter-messaging strategies to defeat information campaigns. To calculate the graph in Figure 15, it is important to run the degree statistics on the entire network (including Egos and Extended Network). If the degree statistics were run on only the extended network, In-Degree would be calculated based on how many Extended Network users are following each other. The desired statistic is how many users in the whole network follow Extended Network users giving them high In-Degree. This is achieved by running the statistic first on the entire network before filtering down to the Extended Network. The top nodes in the circle in Figure 15 represent the ones with the highest In-Degree in the Extended Network, meaning users to whom the Egos (likely ISIS affiliates) are listening.

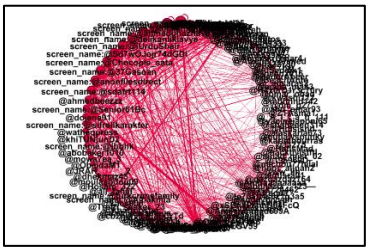

Figure 15: Extended Network with Most Influential Twitter Handle Names Identified

To complete task two the analyst merges the two graphs (ego and extended network) to depict two circles of differing colors comprised of nodes, shown in Figure 16. The Ego nodes (green) are ordered by Eigenvector Centrality to show the top users in this category who should be targeted for further investigation and analysis. The Extended Network nodes (red) are ordered by influence (In-Degree Centrality). The top users in this category are those that could be targeted for further

consideration of action, such as information operations. The final step in this study is for the analyst to extract data for targeting.

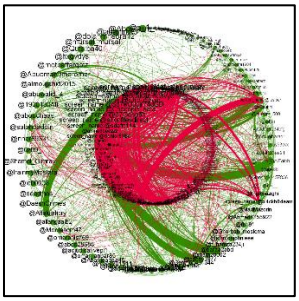

Figure 16: Both circles combined, green = likelihood of Isis affiliation ordered circularly and red = most influential and relevant to operations information campaigns against high Isis affiliations.

5.3 Step 3: Extract Data for Targeting

To extract actionable information from the social network the analyst uses the SURF software to query the network to show only the Egos and sorts the data (decreasing) on Eigenvector Centrality. This yields an ordered list of Ego users who are central to the social network's functionality. This list of Twitter handle names is exported to a .csv file and then repeated for the Extended Network. This produces two outcomes: 1) a list of potential ISIS affiliates and those most important within the network of potential ISIS affiliates and 2) a list of their influencers. Figure 17 depicts a screenshot of such a list.

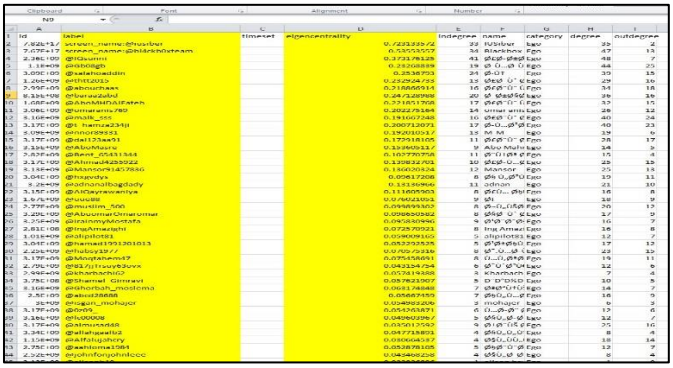

Figure 17: Data exported to .csv file with ranking based on eigenvector centrality (the likeliest ISIS supporter by Twitter name in order of importance)

5.4 Discussion

This case study was undertaken to examine the workflow process of the SURF methodology and the ability of a human analyst to analyse a social network, identify the key influencers, and analyse the message traffic between key influencers (not for message content, but for structure). The Graphical User Interface (GUI) was intuitive and easy to maneuver between the various stages of the analysis. This included accessing the open source Gephi application. Some of the highlights of the analysis capability include the following. First, the technology allows analysts to build custom Watchlists of users and export their social networks. This allows analysts to tailor the tool to their particular needs in a way that fits their desired areas of interest (i.e. ISIS affiliates, hackers, ISIS-affiliated hackers) and size of network. Second, the knowledge that the starting social network has more potential targets than a standard, search term-mined network through the Twitter API provides confidence in the resulting exploitation. Third, the ability to reduce noise in the data achieves a comparable reduction in computational processing, saves analytic time, yields results with higher fidelity, and promotes the use of open source software tools such as Gephi that fail under the computational burden of larger, more noisy data sets.

6. CONCLUSION

This paper demonstrates the level of analytic rigor that is possible with the integration of text and video analytics. We show specific examples of automated cross-cueing between data types and enhanced sensemaking by the analyst who is able to reduce their search for objects.

Tthe presented snippet search tool supports the analyst to find the person of interest in surveillance video data and effectively cross-cue between sources of imagery. The algorithm consists of image selection, object detection, snippet comparison, an embedding, and a graphical user interface. These algorithmic steps enable the analysts to analyze what happened where and when. This is useful to analyze behaviours of interest and to determine intentionality.

The Demonstrator showed the capability to assist the analyst in more rapidly finding relevant social media postings. This will happen through classifying and detecting tagged objects found in videos, locating texts in videos (and images), in addition to common keyword matching. By also classifying and linking video clips to earlier videos found online, the analyst will immediately be able to rule out wrongfully used video clips from prior events. The Demonstrator is currently implemented in an experimental graphical interface, and in need of redesign and reimplementation to be used as a supplemental tool for analysts.

The visualizations produced by the SURF tool were easily understandable. The generation of the watchlist, in working through the three tasks, was quick and easily managed. SURF is unique in that it classifies Twitter users based on network topology and behavior as opposed to the user's content alone. This allows language agnostic analysis, a positive feature when dealing with foreign Twitter accounts.

REFERENCES

[1] Mattis, J., "Information as a joint function," official memorandum, Department of Defense, Washington, DC, USA, 2017 [Online]. Available: https://www.rmda.army.mil/records-management/docs/SECDEF-Endorsement_Information_Joint%20Function_Clean.pdf

[2] William, M., Romanych, M., "The future of IO: Integrated into the fabric of warfighting, "IO Sphere: The Professional Journal of Joint Information Operations," 2014.

[3] Strategy for Operations in the Information Environment, Department of Defense, Washington, DC, USA, 2016 [Online]. Available: https://www.defense.gov/Portals/1/Documents/pubs/DoD-Strategy-for-Operations-in-the-IE-Signed-20160613.pdf

[4] Joint Intelligence Preparation of the Operational Environment, JP2-01.3, Joint Chiefs of Staff, Washington, DC, 2014. [Online]. Available: https://www.hsdl.org/?abstract&did=33212.

[5] Joint Concept for Operating in the Information Environment, Draft Version 0.80. Joint Chiefs of Staff, Washington, DC, September 2017.

[6] Network Engagement: Army Techniques Publication No. 5-0.6. Headquarters, Department of the Army: Washington, DC 19 June 2017.

[7] Burghouts, G. (2017). NATO RTG-144 Experiment TNO: Acquiring intelligence from text and video. TNO: The Hague.

[8] Project Jupyter. On-line interactive programming interface. https://jupyter.org/

[9] Social Understanding and Reasoning Framework (SURF). Available on the Internet: https://www.sbir.gov/sbirsearch/detail/1257717.

[10] Oggero, S. unpublished cartoon. TNO Intelligent Imaging, Oude Waalsdorperweg 63 2597 AK The Hague The Netherlands.

[11] Bowman, E. K., Zimmerman, R. J. (2015) Joint Interagency Field Experimentation 15-2 Final Report (JIFX 15-2) ARL-TR-7562. ARL: Aberdeen Proving Ground, MD.

[12] Dollar, P., Appel, R., Belongie, S., Perona, P., "Fast feature pyramids for object detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1532–1545, 2014.

[13] Xiao, J., Cheng, H., Sawhney, H., Hang, F., "Vehicle Detection and Tracking in Wide Field-of-View Aerial Video", CVPR, 2010.

[14] Bouma, H., Borsboom, S., Hollander, R. den, Landsmeer, S., Worring, M., "Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination", Proc. SPIE, vol. 8359, 2012.

[15] Maaten, L. van der, Hinton, G., “Visualizing data using t-SNE”, Journal of machine learning research, 2579-2605, 2008.
[16] Tesseract Open Source OCR Engine, https://github.com/tesseract-ocr/tesseract/.
[17] Viola, P., Jones, M., "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. I-511-I-518 vol.1.
[18] Krizhevsky, A., Sutskever, I., Hinton, G.E. 2012. "ImageNet classification with deep convolutional neural networks". In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 1. Curran Associates Inc., USA, 1097-1105.
[19] "Tensorflow Object Detection API", On-line resource, Retrieved from https://github.com/tensorflow/models/tree/master/research/object_detection , (Last accessed March 8 2018)
[20] Testing Custom Object Detector - Tensorflow Object Detection API Tutorial, [Blog post], Retrieved from https://pythonprogramming.net/testing-custom-object-detector-tensorflow-object-detection-api-tutorial/, (Last accessed March 8 2018)
[21] Wolfgang, R. B. , Podilchuk, C. I., Delp, E. J., "Perceptual watermarks for digital images and video," in Proceedings of the IEEE, vol. 87, no. 7, pp. 1108-1126, Jul 1999.
[22] Kjelsrud, O., 2014, "Using Perceptual Hash Algorithms to Identify Fragmented and Transformed Video Files", Master Thesis, Gjøvik University College, Gjøvik, Norway.
[23] Lin, T., Zhang, H., "Automatic video scene extraction by shot grouping," *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Barcelona, 2000, pp. 39-42 vol.4.
[24] Maneli Noorkami, Yi Linda Chan, "Video scene detection", 2012, US Patent US8818037B2, https://patents.google.com/patent/US8818037B2/en