Multi-Source Multi-Modal Activity Recognition in Aerial Video Surveillance

Riad I. Hammoud*, Cem S. Sahin*, Erik P. Blasch** and Bradley J. Rhodes*

*BAE Systems, Burlington, MA, USA
**Air Force Research Lab, Rome, NY, USA
riad.hammoud@baesystems.com

Abstract

Recognizing activities in wide aerial/overhead imagery remains a challenging problem due in part to low-resolution video and cluttered scenes with a large number of moving objects. In the context of this research, we deal with two unsynchronized data sources collected in real-world operating scenarios: full-motion videos (FMV) and analyst call-outs (ACO) in the form of chat messages (voice-to-text) made by a human watching the streamed FMV from an aerial platform. We present a multi-source multi-modal activity/event recognition system for surveillance applications, consisting of: (1) detecting and tracking multiple dynamic targets from a moving platform, (2) representing FMV target tracks and chat messages as graphs of attributes, (3) associating FMV tracks and chat messages using a probabilistic graph-based matching approach, and (4) detecting spatial-temporal activity boundaries. We also present an activity pattern learning framework which uses the multi-source associated data as training to index a large archive of FMV videos. Finally, we describe a multi-intelligence user interface for querying an index of activities of interest (AOIs) by movement type and geo-location, and for playing-back a summary of associated text (ACO) and activity video segments of targets-of-interest (TOIs) (in both pixel and geo-coordinates). Such tools help the end-user to quickly search, browse, and prepare mission reports from multi-source data.

1. Introduction

Streaming airborne Wide Area Motion Imagery (WAMI) and Full-Motion Video (FMV) sensor collections afford online analysis for various surveillance applications such as crowded traffic scenes monitoring [18]. In a layered sensing framework, such sensors may be used to simultaneously observe a region of interest to provide complimentary capabilities, including improved resolution for improved target discrimination, identification, and tracking [5]. Typically, forensic analysis, including pattern-of-life detection and activity/event recognition, is conducted off line due to huge volumes of imagery. This big data out-paces users' available time to watch all videos in searching for key activity patterns within the data. To aid users in detecting patterns in aerial imagery, robust and efficient computer vision, pattern analysis and data mining tools are highly desired [9].

1.1. Multi-Source Data and Problem Statement

For data collection and reporting, the aerial video is reviewed by humans (called hereafter *reviewed FMV data*) as the imagery is streamed down from an airborne platform. During a real-time FMV exploitation process, humans could call out significant AOIs, where a voice-to-text tool converts audible ACOs to text (see examples in Figure 5) and a computer then saves the ACOs to storage disks along with the aerial imagery. Additional contextual information besides ACOs include additional reviewers' (internal) chat as well as discussions about the area of coverage of the overhead video from external sources. Together, the ACOs, internal discussions, and external perspectives provide a collective set of "chat messages" [2].

However, these two data sources (chat messages and FMV) are not synchronized in time nor in space. They are not recorded with corresponding time stamps. Furthermore, the called-out targets and activities are not marked in video frames with bounding boxes nor with a start and an end of each activity. It is worth noting that the use of ACOs radically differs from a traditional video annotation paradigm that is typically done manually for training and/or benchmarking of computer vision algorithms. The incorporation of the user's ACO requires advances in automation, human-machine interaction, and multi-modal fusion. In addition, during the overhead imagery review process, there is no advanced equipment such as an eye tracker [11] or touch screen employed to determine screen locations of the TOIs

"Approved for Public Release; Distribution Unlimited. Cleared for Open Publication on April 14, 2014."

corresponding to ACOs.

1.2. Paper Contributions

The ACO messages present a rich source of information allowing for a fast retrieval of activities and providing a summary of events over FMV. They provide a reference ground-truth of the AOIs which occur in the reviewed FMV data. Hence, correlating these two data sources would produce four novel products: (1) a *video summary* of AOIs/TOIs allowing non-linear browsing of video content, (2) *annotated text-over-video* media where only TOIs are highlighted with bounding boxes and synchronized with chat messages, (3) an *activities index* where activities of the same type are grouped together, and (4) *adaptive data playback* allowing for user-selected filtering by geographic location. For instance, the end user may submit a query like this: pull-out all video segments of activity types "turn then stop" near this house on the map (see Figure 7).

In this paper we propose a multi-source probabilistic graph-based association framework to automatically: (1) identify targets-of-interest corresponding to chat messages, (2) detect activity boundaries (i.e., segmenting FMV tracks into semantic sub-tracks/segments), (3) learn activity patterns in low-level feature spaces using the reviewed FMV data, (4) index non-reviewed FMV data (i.e., archived videos), as well as (5) assist FMV analysts with tools for fast querying and non-linear browsing of multi-source data.

Such an automatic linking process of multi-source data enhances data association by eliminating the tedious process of manually collecting and correlating the data. As a side benefit, pattern recognition typically requires training data for activity pattern learning; however, the chat messages provide a notional real-time training template. This problem has been well reported in the literature. For instance, [20] emphasizes the need to collect high-quality activity/event examples with minimal irrelevant pixels for the activity learning modules. Also, during the manual annotation process, Oh *et al.* [20] define very specifically the start and end moments of activities to ensure proper learning on non-noisy data. Here, we demonstrate a paradigm shift in tracking and classification of imagery that does not require training data for real-world deployment of methods.

1.3. Paper Organization

Section 2 details related work. The following sections describe various components of our "Video-Indexed by Voice Annotations" (VIVA) system. Section 3 provides a video processing overview with extensions to our methods. Section 4 describes the mapping of a single FMV target track to multiple graphs of attributes. In Section 4.2 we describe our 2-step algorithm to decompose a single track into semantic segments. Section 5 focuses on parsing of chat messages (or ACO) and their graphical representation. In Section 6 we present the multi-source graph-based association framework and the activity class assignment process. In Section 7 we briefly provide an overview of our approach for learning activity patterns from the reviewed FMV tracks (i.e., training data) and querying the unlabeled FMV data. Sections 8 and 9 outline our Multi-media INdexing and explorER (MINER) interface and evaluates several scenarios to provide performance details of the proposed framework, respectively. We conclude this paper in Section 10.

"Non-Technical Data - Releasable to Foreign Persons."

2. Related Work

Visual activity recognition – the automatic process of recognizing semantic spatio-temporal target patterns such as "person carrying" and "vehicle u-turn" from video data – has been an active research area in the computer vision community for many years [17]. Recently, the focus in the community has shifted toward recognizing activities/actions over large time-scales, wide-area spatial resolutions [10], and multi-source multi-modal frequencies in real-world operating conditions [13]. We assume here that a pattern is bounded by event changes and target movement in between events is an "activity." In such conditions the major challenge arises from the large intra-class variations in activities/events including variations in sensors (e.g., viewpoints, low-resolution, scale), target (e.g., visual appearance, speed of motion), and environment (e.g., lighting condition, occlusion, and clutter). The recognition of activities in overhead imagery poses many more challenges than from a fixed ground-level camera mostly because of imagery's low resolution. Additionally, the need for video stabilization creates noise, tracking, and segmentation difficulties for activity recognition.

The key algorithmic steps in visual activity recognition techniques are (1) extract spatio-temporal interest point detectors and descriptors [7], (2) perform clustering (e.g., K-means) in the feature space (e.g. histogram of gradient (HOG), histogram of flow (HOF), histogram of spatio-temporal gradients (3D-STHOG) and 3D-SIFT) to form codebooks after principal component analysis (PCA)-based dimension reduction, and (3) label tracks using a Bag-Of-Words approach [15, 20]. We follow a similar process when it comes to learning activity patterns from the reviewed FMV tracks. That being said, we first perform multi-source data association to generate training data from the reviewed FMV tracks where FMV tracks are assigned activity labels.

Xiey *et al.* [24] proposed a method for discovering meaningful structures in video through unsupervised learning of temporal clusters and associating the structures with meta data. For a news-domain model, they presented a co-occurrence analysis among structures and observed that temporal models are indeed better at capturing the semantics than non-temporal clusters. Using data from digital TV

news, [8] proposed a framework to determine the correspondences between the video frames and associated text in order to annotate the video frames with more reliable labels and descriptions. The semantic labeling of videos enables a textual query to return more relevant corresponding images, and enables an image-based query response to provide more meaningful descriptors (i.e., content-based image retrieval). Our proposed activity recognition framework discovers meaningful activity structures (e.g., semantically labeled events, activities, patterns) from overhead imagery over challenging scenarios in both reviewed and un-reviewed FMV data.

3. Video Target Tracking

Tracking multiple targets in aerial imagery requires first to stabilize the imagery and then detect automatically any moving target.

Video Stabilization Our *Frame-to-frame stabilization module* aligns successive image frames to compensate for camera motion [23]. There are several steps involved in our 2-frame registration process: (1) extract interest points from the previous image that possess enough texture and contrast to distinguish them from one another, and (2) match the 2D locations of these points between frames using a robust correspondence algorithm. Establishing correspondences consists of two stages: (a) use "guesses", or putative matches, established by correlating regions around pairs of feature points across images, and (b) perform outlier rejection with RANdom SAmple Consensus (RANSAC) to remove bad guesses.

The VIVA stabilization algorithm runs in real-time on commercial off the shelf (COTS) hardware and it was specifically designed to be robust against large motions between frames. The enhanced robustness against large motion changes is essential since analog transmission of electro-optical/infrared (EO/IR) airborne data to the ground can be corrupted, frames can be dropped, time-delays long, and can vary in sample rates. As long as the two frames being registered have greater than 35% overlap, we are usually able to establish enough correspondences for reliable stabilization.

Target Detection and Tracking Our *moving target tracking* algorithm – Cluster Objects Using Recognized Sequence of Estimates (COURSE) – makes few assumptions about the scene content, operates almost exclusively in the focal plane domain, and exploits the spatial and temporal coherence of the video data. It consists of three processing steps. First, the frame-to-frame *registration* is used to find regions of the image where pixel intensities differ – this is done through frame differencing (see Figure 1). Underlying frame differencing is the assumption that pixel intensity differences are due to objects that do not fit the global image motion model. Clearly, other effects – such as parallax – also cause false differences, but these false movers are filtered using subsequent motion analysis. Second, point features with high pixel intensity difference are used to establish *correspondences* between other points in the previous frame, which produces a set of point-velocity pairs. Third, these point-velocity pairs are *clustered* into motion regions that we assume are due to individual targets. Regions that persist over time are reported as multiple target detections. The tracker provides two very important capabilities: (i) it removes false detections generated by the upstream target detection module, and (ii) extends detection associations beyond what can be accomplished by using only the image-based target detection module. COURSE achieves enhanced robustness by (i) removing isolated detections that are inconsistent with the presence of a moving object, and (ii) exploiting large time-event information to deal with brief interruptions caused by minor occlusions such as trees or passing cars. The COURSE tracker generates a mosaic tracking report (see Figures 2 and 3) to be used as input to our multi-source association framework.

Figure 1: VIVA's movement detection module. First registered frames (top left) are differenced to produce a change detection image (lower right). That image is thresholded to detect changing pixels. Point correspondences within those detection pixels are established between the two frames and used to generate motion clusters (right).

4. Multi-Graph Representation of a Single FMV Track

The multi-source association framework is based on a graph representation and matching of target tracks and chat messages. In this section, we describe how to build a graph-based model of a tracked target and how to divide "rich" tracks into semantic track-segments and hence represent a single track with multiple graphs.

"Non-Technical Data - Releasable to Foreign Persons."

Figure 2: Example of track profiles of vehicles generated by COURSE using sample videos from the VIRAT aerial dataset (ApHill) [20].

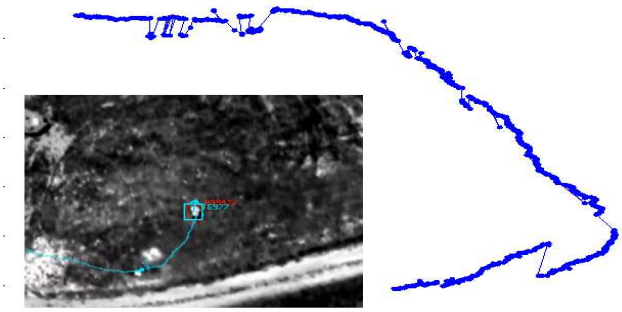
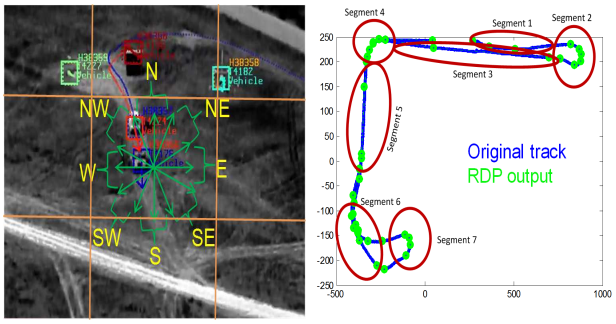


Figure 3: Illustration of a noisy tracking trajectory of a single dismount (from the ApHill VIRAT aerial dataset) generated by COURSE. The track is broken into several segments (i.e., several tracking labels) due to quick changes in motion direction, cluttered background, and multiple stop-and-move scenarios.



(a) (b)

Figure 4: Illustration of our assignment of tracking states into (a) direction and location zones (e.g. south-east direction, top-left screen zone, etc.) and (b) semantic segments based on changes in direction and speed using RDP.

4.1. Mapping Tracks to Graphs

Each target track is cast by a combination of graphs where *nodes* represent targets' attributes and *edges* characterize the relationship between nodes. We divided attributes into *common* and *uncommon* based on their saliency over the life time of a target track. For instance, color and shape of a vehicle remain unchanged, while direction and spatial location vary over time (t). The targets-of-interest are classified into "vehicle" vs. "human" (i.e., *actor* attribute) based on motion, blob size, and shape. The *shape* attribute is divided into "car" vs. "SUV" vs. "truck" for vehicle, and "adult" vs. "child" for human actor/dismount [1]. Each actor is characterized with a unique *color* attribute (e.g., black truck, human with red-shirt, etc.) and a *spatial location* (i.e., xy_s position on the screen and lat/lon on the geographic map). The location is mapped into gross zones (see Figure 4(a)) on the screen to match with gross locations in the chat messages. We divided the video frame into a 3x3 grid (center screen, top left, etc). The *direction* attribute is derived from the velocity vectors $(V_x(t), V_y(t))$ at time t such that $\theta(t) = arctan(\frac{V_y(t)}{V_x(t)})$, which in turn is mapped to a geographical direction using the gross divisions of directions as shown in Figure 4(a). In order to reduce noise in the mapping of θ and xy_s to gross direction and location zones, we applied a sliding window to smooth these values over time. The last attribute is *mobility* which specifies whether the target is moving or stationary (m_t).

4.2. Dividing Tracks Into Semantic Segments

When a track exhibits major changes in uncommon attributes, especially in direction, location and speed; it becomes necessary to break it down into multiple semantic segments, and hence multiple graphs, to match them with multiple chat messages in the association framework. This is the case when multiple chats correspond to a single track generated by our video tracker. Figure 4(b) shows three minutes of a tracked vehicle moving toward the east, making a u-turn then moving toward the west. We apply a 2-step algorithm to break down tracks into semantic segments:

1. Smooth the tracking locations (xy_s) using the Ramer-Douglas-Peucker (RDP) algorithm [21]. This will produce a short list of un-noisy position points (XY_s) (displayed as green points in Figure 4(b)).

2. Detect directional changes computed from $XY_s(t)$ points. The beginning of a new semantic track segment is marked when a peak is detected. The end of a new semantic segment is flagged when the second derivative of $XY_s(t)$ is near zero. Figure 4(b) illustrates the results of this step where 7 segments were detected.

5. Parsing and Graph Representation of Chats

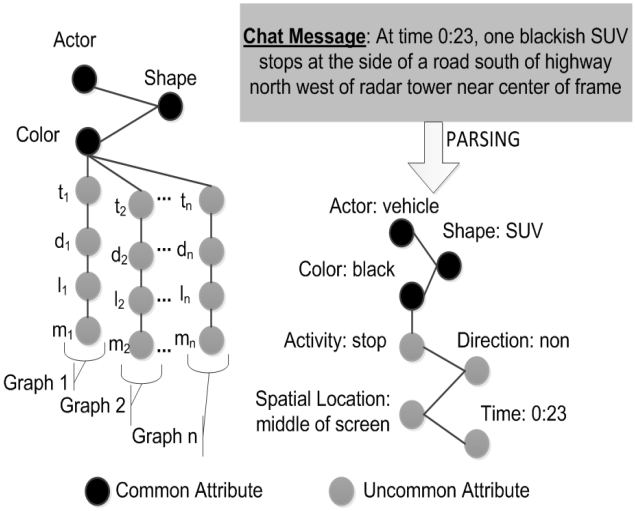
In our data collection setup, the chat messages follow the following format for a target of type vehicle [3]:

"Non-Technical Data - Releasable to Foreign Persons."

```
At <time> <quantity> <color> <vehicle>
<activity> <direction> <location>
where:
<time> = 0000Z - 2359Z
<activity> = (travel | u-Turn ...)
<direction> = (north | south ...)
<location> = screen (middle | left ...)
<color> = (red | black ...)
<shape> = (truck | car ...)
```

Basic search for keywords in a chat message is employed to extract relevant information such as "activity type", "direction", and "location". In our dataset, we have 9 activities (vehicle turn, u-turn, human walking, running, etc.; see Section 9), eight direction zones (north, south, etc.) and nine location zones (middle, top-left screen zone, etc; see Figure 4). These chat messages represent an analyst calling out activities in the FMV, intra-viewer discussions, or other related external discussions. In turn, a chat message is represented as a graph of attributes. However, more elaborated Information Extraction (IE) from a chat message (i.e., micro-text) or a document (e.g., using Sphynx or Apache NLP) as an automated approach [19, 6, 14, 4] could be employed to handle miss-spelled words and larger dictionaries.

Figure 5(b) illustrates a chat message decomposed into multi-modal attributes. An example can come from any modality (e.g., video, text, radar, etc.) so the goal is to decompose the data into these meaningful parts [2].



(a) (b)

Figure 5: Example of representation of a video track (a) and a chat message (b) as graphs.

6. Multi-Source Graph Association And Activity Class Assignment

A mission goal includes allowing the image processing method to answer a user-defined query. The user calling-out significant activities in the image would desire an automated processor to match the target being called out to that of a target-of-interest (TOI). With an image, there could be many movers, targets, and events happening. The system must choose the TOI among several tracked objects in the imagery that corresponds to a meaningful content (attributes) in the chat message by a user. Because users review FMV tracks from streaming airborne video, the call-outs flag AOIs. Association between reviewed FMV tracks and chat messages can be achieved by performing probabilistic matching between graphs from both data sources. It is important to note that, as explained in the introduction, a chat message is the only source to describe the true activity of the TOI. By performing multi-source graph-based association, the true activity of the TOI is mapped to a corresponding track segment from FMV.

The multi-source multi-modal association framework consists of the following stages:

1. In a given time interval, $[t-T, t+T]$ (with t the time stamp from a chat message and T : a pre-defined time window to search for the tracked objects), the chat message and all video tracks are extracted from the data sets.

2. *Graph representations* of video-tracks and chat messages are generated as explained in Sections 4 and 5).

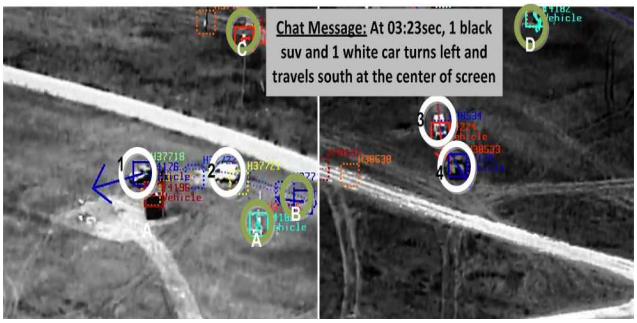
3. *Partial graph matching* uses a probabilistic distance measure (see Equation 1) of ensemble *similarity* between a chat message (j) and track segment (i). There are three main reasons to use a probabilistic distance metric: (i) to associate the graphs even if there are missing attributes, (ii) to reduce the effects of errors coming from the video processor and chat messages (e.g., a user may assign a vehicle color as black while a tracked object from the video processor might be marked as gray), and (iii) to impute the weights of attributes based on the quality of videos. The associated graphs with the highest probabilities are assigned as match.

$$P(T_i|C_j, c_i) = w_a P_a + w_s P_s + w_t P_t + w_{cl} P_{cl} + w_d P_d + w_l P_l + w_{cn} P_{cn} + w_m Pm \quad (1)$$

where $w_a, w_s, w_t, w_{cl}, w_d, w_l, w_{cn}$, and w_m are the user-defined weights of attributes for actor, shape, time, color, direction, spatial location, tracking confidence and target

"Non-Technical Data - Releasable to Foreign Persons."

mobility, respectively. $P_a, P_s, P_t, P_{cl}, P_d, P_l$, and P_m represent the probabilities of corresponding attributes, and P_{cn} is the track confidence value generated by COURSE.



(a) (b)

Figure 6: Successful identifications of AOIs/TOIs in exemplar clips from the ApHill VIRAT aerial dataset using our multi-source association framework. (a) and (b) show multiple vehicle tracks and a single chat message being called-out; the tracks in white circles ($1, 2, 3$, and 4) were highly matched with the chat message graphs while targets in green circles (A, B, C and D) scored low matching probabilities.

An illustrative result of this framework is shown in Figure 6. This framework handles 1-to-1, 1-to-*N*, and *N*-to-*M* association cases. Further this framework not only marks the target of interest but also the rendering of activities. Using labeled track profiles, the boundaries of each activity are determined by using the process described in Section 4.2. For example, after labeling each track segment by associating chat messages, track segments $1, 3$, and 5 are marked as *travel*, segments 2 and 7 are *u-turn* and track segments 4 and 6 are labeled as *turn* in Figure 4(b).

7. Learning Activity Patterns from Multi-Source Associated Data

The chat messages provide the ground truth of the AOIs occurring in the reviewed FMV video (see Section 6). These correlated data serve as training data for activity pattern learning in aerial imagery. Here we employ BAE Systems' Multi-intelligence Activity Pattern Learning and Exploitation (MAPLE) tool which uses the Hyper-Elliptical Learning and Matching (HELM) unsupervised clustering algorithm [22] to learn activity patterns. This is done through extracting features from each labeled track segment, clustering features in each activity space, and finally representing each track by a sequence of clusters (i.e., chain code). In terms of features, we used simple descriptors for vehicles including speed, heading relative to segment start, and position eigenvalue ratio. By measuring the change relative to a fixed starting value, rather than the instantaneous change, the heading feature is robust to variations in how quickly the targets turns from its initial course. The position eigenvalue ratio is a measure of the mobility of the target. It is the ratio of eigenvalues calculated from the target's position within a short time duration. As for people tracking, we compute the histogram of motion flow and neighboring intensity variance which describes the extent to which the target is moving toward or away from potential interaction sites.

The goal of this learning process is to be able to match an unlabeled track (i.e., without chat) to the learned activity patterns. This allows indexing a large amount of unreviewed FMV data. First we use HELM to classify each instance of a new track to one of the clusters of the index. Second we use Temporal Gradient Matching distance to obtain matches between the representation of the new track and the indexed learned patterns. The similarity score between a new track j and an index i is defined as follows:

$$\sigma_{ij} = \frac{t_{ij} + c_{ij}}{2} \tag{2}$$

where t_{ij} represents similarity metric which considers only the common clusters and c_{ij} is the similarity score of temporal gradient for the cluster sequence.

8. Event/Activity Report Visualization and Querying by Activity Type and Geo-Location

The VIVA framework presented in this paper produces three useful products for the end-users to visualize in the same interface (see Figure 7): (1) a *video summary* of AOIs allowing non-linear browsing of video content, (2) *text-over-video media* where only TOIs are highlighted with bounding boxes and synchronized with chat messages which describe their activities, and (3) an *index* of activities. The benefit of the compiled index of videos is that a user (or machine) could find related content over a geographic location and text. For instance, the end-user may submit a query like this: pull-out all video segments of activity types "turn then stop" near this house with specific latitude and longitude coordinates. We converted each track to geo-tracks using both meta-data and reference imagery. If the AOI is detected in archived video using the activity classification framework presented above, the chat panel in our Multimedia INdexing and explorER (MINER) interface shows the automatically generated description (i.e., target category and activity type).

9. Experimental Results

To validate the proposed framework, we used our own dataset consisting of EO/IR airborne videos (about 25 min-

"Non-Technical Data - Releasable to Foreign Persons."

"Non-Technical Data - Releasable to Foreign Persons."

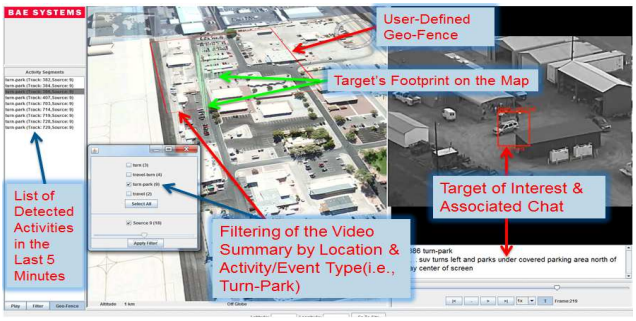


Figure 7: Illustration of the event-report visualization interface (MINER) allowing users to visualize and query correlated chats, Pattern-Of-Life, and activity-labeled track segments.

utes long) and 100 chat messages. The activity list is limited to *vehicle travel, stop, turn, u-turn, maintain-distance, accelerate and decelerate*, and *human walking and running*. In this small dataset we had more vehicles than dismounts. The VIVA video tracker generated about 3200 different tracks. The percentage of false tracking is about 15 percent and could be reduced through fusion of both EO and IR [16]. This is mainly due to camera zoom-ins and zoom-outs. Each object was automatically classified as human vs. vehicle and specific low-level features (see Section 4.1) were computed prior to running the proposed multi-source multi-modal association framework.

Table 1 summarizes the results of the association framework. Correct associations are marked when the tracks or sub-tracks (i.e., semantic segments) in the overhead imagery are associated with their corresponding chat messages. A false association is flagged when the chat message is linked to the wrong target or semantic segment (see Section 4.2). This could occur when multiple targets are moving in the same area at the same time and in the same direction. A miss is defined as a chat message without an associated target in the overhead imagery. On this data set we scored 76.6% correct association, 10.3% misses association, and 12.9% wrong association (i.e., false alarms). During these experiments we set the time window in which to perform the multi-source associations to 15 seconds. Making this window shorter leads to less false alarms but also higher miss rate. Also we only used target's direction, location, and speed as attributes, which do not include other rich content to reduce false alarms.

The association framework for activity recognition handles complex scenarios with multiple tracked objects. Figures 6(a) and (b) show eight different tracks (different track labels) of six moving objects and a single chat message called-out within the same time window. The chat message is parsed automatically into four different graphs which are

Detection	Miss	False
76.6%	10.3%	12.9%

Table 1: Qualitative assessment of the multi-graph association and activity class assignment framework.

matched to all ten graphs representing the video tracks. The additional 2 video graphs (initially we got 8 tracks) came out from the splitting process of a single track into semantic segments (or sub-tracks as described in Section 4.2) due to changes in vehicle direction while traveling. The VIVA framework associated the four chat graphs to the correct four FMV semantic tracks due to strong matches between common attributes. Our approach was also challenged by broken tracks (e.g., case of a dismount/TOI with three different tracking labels in Figure 3). In spite the fact that the same TOI is represented by three consecutive tracks, VIVA provides correct associations with events boundaries (i.e., shorter and semantic track segments). Thus, it is robust to scenario variations.

These preliminary results are very promising. Both the direction and the location attributes play an important role in the association of chat messages to tracks. The list of potential matches is reduced drastically using these attributes. Nevertheless in order to make 1-to-1 association, additional attributes such as shape, color and size, and spatial relationships such as target near an identifiable landmark in the scene, would be very helpful to resolve association ambiguities. Due to the chat description, the extracted target's direction and location are cast to gross zones (i.e., middle screen region, north-east direction, etc.) rather than fine ranges, causing ambiguities in the association. Extracting buildings from available imagery [12] would greatly benefit the association because the chats refer to such attributes when describing activities involving human-object interactions.

Figure 8: Illustration of an exemplar target track (from the ApHill VIRAT aerial dataset) being matched to the proper activity pattern model (a *u-turn* in this example) learned using the training data generated by the proposed multi-source association approach.

We used the multi-source associated data to learn activity patterns and then index un-reviewed data (see Section 7). Preliminary results are illustrated in Figure 8. It shows a

track segment from an unlabeled video correctly matched to a u-turn pattern model with a highest matching score $\sigma_{q,uTurn} \approx 1.0$ compared to other models. This work is still in progress as we are doing more extensive experiments to ensure that we have enough training data to build reliable pattern activity models in challenging conditions with enough intra-class variations using high dimensional activity descriptors over a larger activity list.

10. Conclusion

In this paper, we developed a novel concept of graphical fusion from video and text data to enhance activity analysis from aerial imagery. We detailed the various components including VIVA association framework, COURSE tracker, MAPLE learning tool and the MINER visualization interface. Given the exemplar proof of concept, we highlighted the benefits for a user in reviewing, annotating, and reporting on video content. Future work will explore the metrics and associations used to increase robustness and reduce false alarms. However, it is noted that the end user can check the final results presented in our MINER interface to remove false alarms and generate mission reports effortlessly.

Acknowledgments

This work was supported under contract number FA8750-13-C-0099 from the Air Force Research laboratory. The ideas and opinions expressed here are not official policies of the United States Air Force.

The authors would like to thank Adnan Bubalo (AFRL), Robert Biehl, Brad Galego, Helen Webb and Michael Schneider (BAE Systems) for their support.

References

[1] E. Blasch, H. Ling, Y. Wu, G. Seetharaman, M. Talbert, L. Bai, and G. Chen. Dismount tracking and identification from EO imagery. In *Proceedings of the SPIE*, 2010. 4

[2] E. Blasch, J. Nagy, B. P. A. Aved, M. K. Schneider, R. I. Hammoud, et al. Context aided video-to-text information fusion. In *Int'l Conf. on Information Fusion*, 2014. submitted. 1, 5

[3] E. Blasch, Z. Wang, H. Ling, K. Palaniappan, G. Chen, D. Shen, A. Aved, and G. Seetharaman. Video-based activity analysis using the l1 tracker on VIRAT data. In *IEEE Applied Imagery Pattern Recognition Workshop*, 2013. 4

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3, 2003. 5

[5] A. P. Brown, M. J. Sheffler, and K. E. Dunn. Persistent electro-optical/infrared wide-area sensor exploitation. In *SPIE Vol. 8402 Evol. and Bio-Inspired Comp.*, 2012. 1

"Non-Technical Data - Releasable to Foreign Persons."

[6] E. Cambria and B. White. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9:1–28, 2014. 5

[7] P. Dollr, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VSPETS*, 2005. 2

[8] P. Duygulu and H. D. Wactlar. Associating video frames with text. In *26th ACM SIGIR Conference*, 2003. 3

[9] G. Fan, R. I. Hammoud, F. Sadjadi, and B. Kamgar-Parsi. Special section on advances in machine vis. beyond the visible spectrum. *CVIU Journal*, 117(12):1645–1646, 2013. 1

[10] J. Gao, H. Ling, E. Blasch, K. Pham, Z. Wang, and G. Chen. Pattern of life from wami objects tracking based on visual context-aware tracking and infusion network models. In *Proceedings of the SPIE*, 2013. 2

[11] R. I. Hammoud. *Passive Eye Monitoring: Algorithms, Applications and Experiments*. Springer, 2008. ISBN: 978-3-540-75411-4. 1

[12] R. I. Hammoud, S. A. Kuzdeba, B. Berard, V. Tom, et al. Overhead-based image and video geo-localization framework. In *IEEE CVPR WS*, pages 320–327, 2013. 7

[13] B. Kahler and E. Blasch. Sensor management fusion using operating conditions. In *Proceedings of Aerospace and Electronics Conference, IEEE National*, 2008. 2

[14] M. O. Kulekci and K. Oflazer. An overview of natural language processing techniques in text-to-speech systems. In *IEEE Conf. on Signal Processing and Com. App.*, 2004. 5

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[16] A. Leykin and R. I. Hammoud. Pedestrian tracking by fusion of thermal-visible surveillance videos. *J. of Machine Vision and Applications*, 2010. 7

[17] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *IEEE CVPR*, 2008. 2

[18] B. Maurin, O. Masoud, and N. Papanikolopoulos. Camera surveillance of crowded traffic scenes. In *ITS America 12th Annual Meeting*, page 28, 2002. 1

[19] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26, 2007. 5

[20] S. Oh, A. Hoogs, A. Perera, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *IEEE CVPR*, 2011. 2, 4

[21] D. K. Prasad. Assessing error bound for dominant point detection. *Int'l J. of Image Processing*, 6:326–333, 2012. 4

[22] B. J. Rhodes, N. A. Bomberger, M. Zandipour, et al. Anomaly detection and behavior prediction: Higher-level fusion based on computational neuroscientific principles. *Sensor and Data Fusion*, pages 323–336, 2009. 6

[23] G. Seetharaman, G. Gasperas, and K. Palaniappan. A piecewise affine model for image registration in nonrigid motion analysis. In *IEEE ICIP*, pages 561–564, 2000. 3

[24] L. Xiey, L. Kennedyy, S.-F. Changy, A. Divakaranx, H. Sunx, and C.-Y. Linz. Discovering meaningful multimedia patterns with audio-visual concepts and associated text. In *Image Processing, ICIP*, 2004. 2