Object and Activity Detection from Aerial Video

Stephen Se[*], Feng Shi, Xin Liu, Mohsen Ghazel

MDA Systems Ltd.
13800 Commerce Parkway,
Richmond, B.C., V6V 2J3, Canada

ABSTRACT

Aerial video surveillance has advanced significantly in recent years, as inexpensive high-quality video cameras and airborne platforms are becoming more readily available. Video has become an indispensable part of military operations and is now becoming increasingly valuable in the civil and paramilitary sectors. Such surveillance capabilities are useful for battlefield intelligence and reconnaissance as well as monitoring major events, border control and critical infrastructure. However, monitoring this growing flood of video data requires significant effort from increasingly large numbers of video analysts.

We have developed a suite of aerial video exploitation tools that can alleviate mundane monitoring from the analysts, by detecting and alerting objects and activities that require analysts' attention. These tools can be used for both tactical applications and post-mission analytics so that the video data can be exploited more efficiently and timely.

A feature-based approach and a pixel-based approach have been developed for Video Moving Target Indicator (VMTI) to detect moving objects at real-time in aerial video. Such moving objects can then be classified by a person detector algorithm which was trained with representative aerial data. We have also developed an activity detection tool that can detect activities of interests in aerial video, such as person-vehicle interaction.

We have implemented a flexible framework so that new processing modules can be added easily. The Graphical User Interface (GUI) allows the user to configure the processing pipeline at run-time to evaluate different algorithms and parameters. Promising experimental results have been obtained using these tools and an evaluation has been carried out to characterize their performance.

Keywords: Video Exploitation, Aerial Surveillance, Moving Target Indicator, Person Detection, Activity Detection

1. INTRODUCTION

Aerial video surveillance has advanced significantly in recent years, as inexpensive high-quality video cameras and airborne platforms are becoming more readily available. Video has become an indispensable part of military operations and is now becoming increasingly valuable in the civil and paramilitary sectors. Such surveillance capabilities are useful for battlefield intelligence and reconnaissance as well as monitoring major events, border control and critical infrastructure. However, monitoring this growing flood of video data requires significant effort from increasingly large numbers of video analysts.

The detection and recognition of objects and activities are still active research areas in the computer vision community. A comprehensive review of the literature on activity understanding can be found in recent survey papers [1][2][3]. Some of the key challenges include viewpoint variation, occlusion, background clutter, large intra-class variation for the same objects/activities and the small inter-class variation between different objects/activities. Some video exploitation tools are available for fixed surveillance cameras to detect simple events such as abandoned/removed objects, people loitering, etc. However, they do not work on aerial video collected by manned aircrafts or UAV (Unmanned Aerial Vehicles) where the camera is not stationary.

[*] Corresponding author: sse@mdacorporation.com

Geospatial Informatics, Fusion, and Motion Video Analytics V, edited by Matthew F. Pellechia, Kannappan Palaniappan,
Peter J. Doucette, Shiloh L. Dockstader, Gunasekaran Seetharaman, Proc. of SPIE Vol. 9473, 947305
© 2015 SPIE · CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2177094

At MDA, we have developed a suite of aerial video exploitation tools that can detect and alert objects and activities that require analysts' attention. The tools include Video Moving Target Indicator (VMTI), person detector and activity detector. These tools can be used for both tactical applications and post-mission analytics, so that the video data can be exploited more efficiently and timely.

This paper presents the work done during this project. Section 2 provides an overview of the architecture and Graphical User Interface (GUI). Section 3 describes the video exploitation tools developed as well as their performance evaluation results. Section 4 provides conclusions and discusses some future work.

2. SYSTEM OVERVIEW

2.1 Architecture

Figure 1 shows an overview of the architecture. When the user selects a video file, the data ingest module will read the video frames into a buffer. During the processing mode, the various detectors can then process the video frames to find objects or activities of interest. The GUI will then overlay the detection results on the video frames. All the results are also saved to a repository file, which can be reviewed later using the playback mode.

We have developed a flexible architecture to allow new detectors to be added to the system easily, using an object factory design. All the detectors run on separate threads to make use of multi-core CPUs. Moreover, the algorithmic parameters can be configured dynamically at run-time. This provides a test-bed for prototyping and evaluating different algorithms. The system was developed in C++ using OpenCV [4] and QT [5]. OpenCV is a software development kit that provides computer vision and image processing functionalities, while QT provides the GUI framework.

In order to support different scenarios, the user can configure which detectors to use at run-time. For example, the user may want to perform motion detection only or person detection only. The user may create more complex configurations as shown in Figure 2. In Figure 2(a), the person detector can be applied only to the moving regions output by motion detector, instead of the full video frame. In Figure 2(b), the interaction detector can check the proximity between any person and vehicle detected by the person detector and vehicle detector respectively.

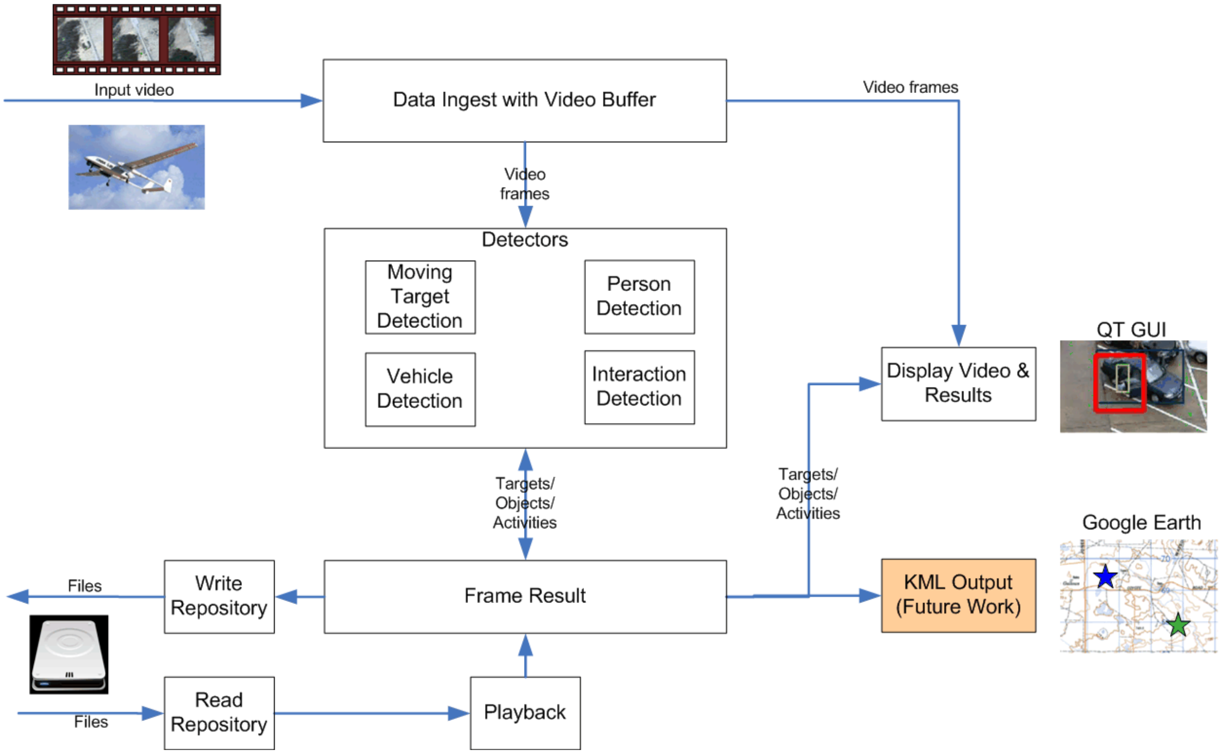


Figure 1 System Architecture (© MacDonald, Dettwiler and Associates Ltd.)

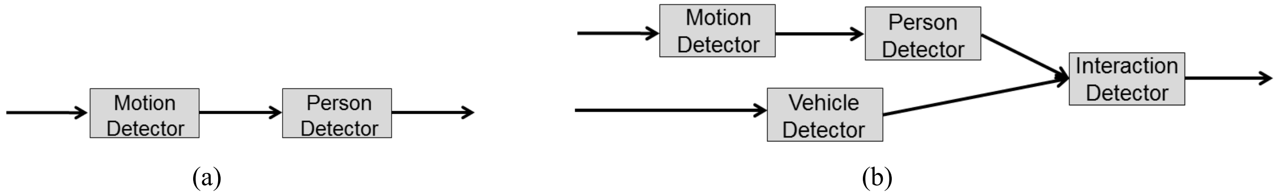


Figure 2 Two Examples of Configurations (© MacDonald, Dettwiler and Associates Ltd.)

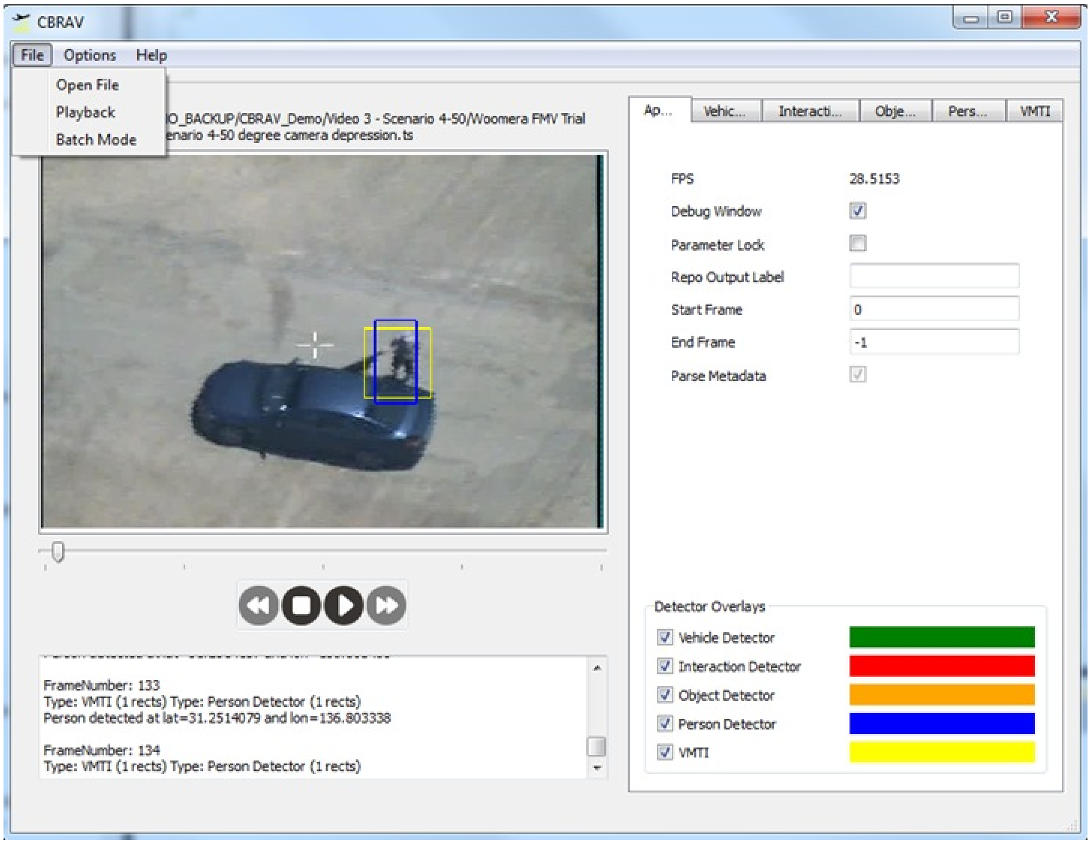
2.2 Graphical User Interface (GUI)

Figure 3 shows the main GUI. The left panel is for displaying the video and results. The detection results are overlaid on the video frames in different colours, with the legend shown in the bottom right window. The bottom left window shows any output messages. The right panel is for the user to adjust the parameters, where the different tabs correspond to the various detectors.

The user can select a video file to process or an output file to play back previous results. The user can also process multiple video files using batch mode and review the results later. The batch mode is also useful to process the same video repeatedly with different parameters for performance evaluation purposes.

Different input video formats are supported including AVI, MPEG and STANAG 4609. As metadata is embedded in the STANAG 4609 file, geographical locations (latitude and longitude) of the detected objects can also be computed. For example, the bottom left window in Figure 3 shows the latitude and longitude of the person detected. The user can optionally specify the start and end frames to process a segment of a video instead of the entire video.

The algorithmic parameters are rendered dynamically based on XML files. This allows the user to change the default values and ranges without re-compiling the software. When a new detector is added to the system, the user can create a new XML file for the new detector's parameters so that a new parameter tab will be rendered automatically.



O_BACKUP/CBRAV_Demo/Video 3 - Scenario 4-50/Woomera FMV Trial
enario 4-50 degree camera depression.ts

FPS 28.5153
Debug Window ☑
Parameter Lock ☐
Repo Output Label
Start Frame 0
End Frame -1
Parse Metadata ☑

```
FrameNumber: 133
Type: VMTI (1 rects) Type: Person Detector (1 rects)
Person detected at lat=31.2514079 and lon=136.803338

FrameNumber: 134
Type: VMTI (1 rects) Type: Person Detector (1 rects)
```

☑ Vehicle Detector 🟩
☑ Interaction Detector 🟥
☑ Object Detector 🟧
☑ Person Detector 🟦
☑ VMTI 🟨

Figure 3 GUI Screenshot (© MacDonald, Dettwiler and Associates Ltd.)

3. VIDEO EXPLOITATION TOOLS

3.1 Video Moving Target Indicator (VMTI)

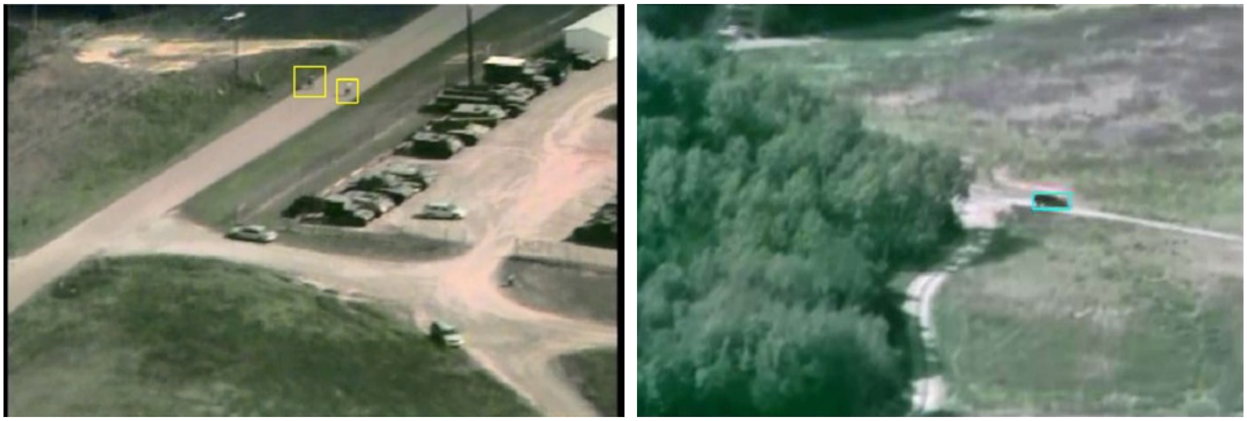
Moving regions typically correspond to objects of interest for the analysts. For aerial video collected by UAV, all the pixels are moving as the camera is moving. Therefore, the objective is to find moving regions that are not due to the camera motion.

We have developed two approaches for VMTI to cater to different scenarios: feature-based and pixel-based. The feature-based approach processes the video at a feature level, i.e. finding image features that do not move consistently with the rest of the image. Similarly, the pixel-based approach processes the video at a pixel level, i.e. finding image pixels that do not move consistently with the rest of the image. Since many potential moving regions are often found, further processing such as filtering and clustering is required to reduce false alarms.

Moreover, an adaptive thresholding technique has been developed that can determine the suitable thresholding parameters automatically based on the input video. Therefore, the user does not need to adjust the parameters for different videos, which is important for operational use.

One of the issues of using operational UAV video is the text and graphics that are sometimes burnt on the video frames. Such text and graphics overlay would be confused as moving objects. In those cases, a pre-processing step is performed to detect the text/graphics or cross-hair so that moving objects found within those regions would be ignored.

Figure 4 shows two examples of VMTI results, in which the correct moving objects are detected in real-time. The test video comes from the aerial dataset released by DARPA's VIRAT project [6]. Figure 4(a) shows two walking people, each of them is less than 10x20 pixels in size. Figure 4(b) shows a moving vehicle in a fairly unstabilized video feed.



(a) (b)

Figure 4 VMTI Detection Examples on VIRAT Aerial Dataset (© DARPA)

3.2 Object Detection

We have implemented several person and vehicle detectors using OpenCV, which are based on supervised machine learning approaches. Publicly available training models are often for ground images only. Therefore, we trained the person detector using positive and negative image chips from representative aerial imagery, some examples of which are shown in Figure 5. Negative examples include randomly chosen chips as well as hard examples that resemble human such as vertical structures. Histogram of Oriented Gradients (HOG) features are extracted from the training images to obtain the Support Vector Machine (SVM) model which is used for detection at run-time [7]. Figure 6 shows examples of person detection on UCF (University of Central Florida) [8] and our own aerial datasets, where it processes the whole video frame to find any stationary or moving human.

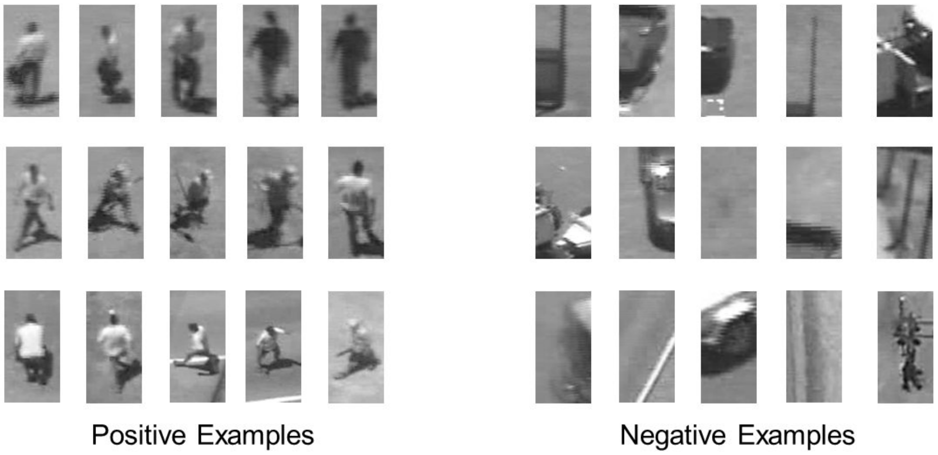


Figure 5 Positive and Negative Aerial Imagery Examples for Person Detector Training (© MacDonald, Dettwiler and Associates Ltd.)

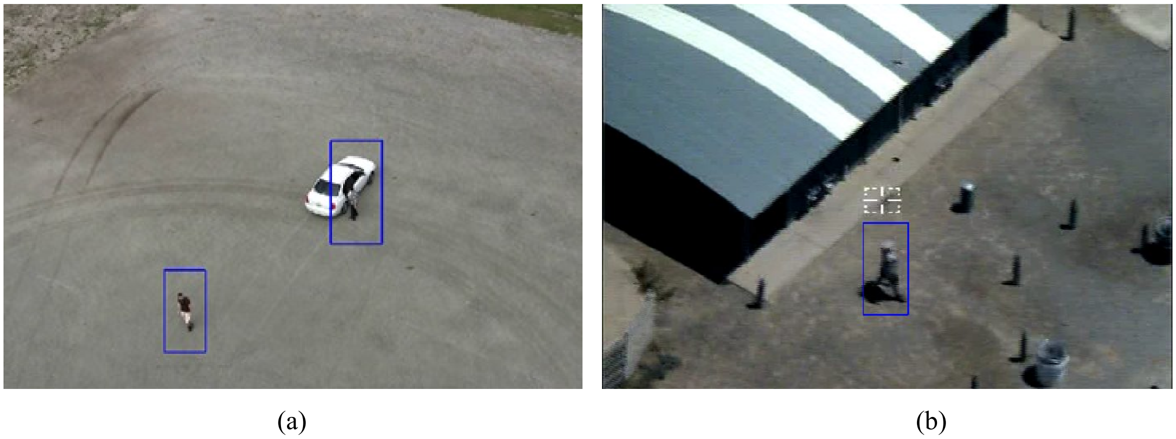


Figure 6 Person Detection Examples (a) UCF aerial dataset (© University of Central Florida) (b) MDA aerial dataset (© MacDonald, Dettwiler and Associates Ltd.)

Figure 7(a) shows the ROC (Receiver Operating Characteristics) curve for person detection on a test video. We used different training datasets to obtain three SVM models to compare with the default OpenCV model:

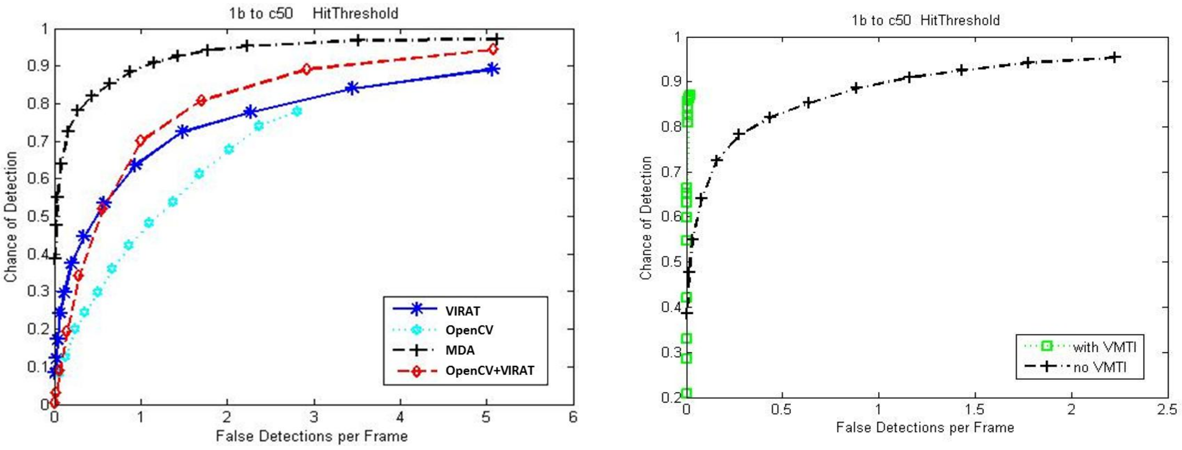
(1) VIRAT is the training model obtained from the VIRAT ground dataset

(2) OpenCV+VIRAT is the training model obtained from combining the VIRAT ground dataset and the dataset used by the OpenCV model

(3) MDA is the training model obtained from our own UAV dataset

The test video is taken from our own UAV dataset that is not used for training. It can be seen that the training models obtained from representative imagery provide better results than the default model in general. In particular, the MDA training model gives the best performance since it resembles the test data the most.

The person detector can be applied only to the moving regions to reduce false alarm using the configuration in Figure 2(a). Although this would miss any stationary persons, they will be detected once they start to move. Figure 7(b) shows the improvement by performing VMTI prior to person detection, where the false alarm rate is reduced to almost zero. The processing speed is also much faster as the person detector only need to process small regions rather than the whole image.



(a) (b)

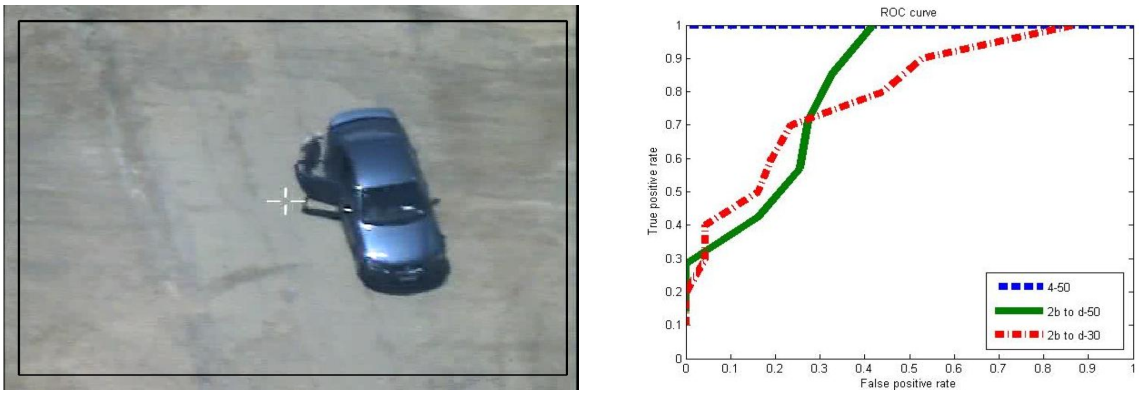
Figure 7 Person Detection ROC Curves Comparison (a) Among different training models (b) With and without VMTI (© MacDonald, Dettwiler and Associates Ltd.)

3.3 Activity Detection

The goal of recognition is to determine the type of activity in the video, implicitly assuming something happened. On the other hand, the goal of detection is to find the temporal and spatial location of an activity, with no prior knowledge on whether or not the video contains an activity. Detection is thus inherently more challenging and computationally demanding as we should both classify activities versus non-activities and specify when and where they occur. For this project, we are interested in detecting person-vehicle interactions to alert the analysts.

While person detection is performed on a frame-by-frame basis, activity detection is performed on a sequence of frames using a sliding window along the video frames. We train the activity detector using positive and negative video clips from the VIRAT ground dataset. Dense spatio-temporal features based on a local part model are extracted from the training videos to obtain SVM model which is used for detection at run-time [9]. For this experiment, we train the system to detect when there is any person getting into or out of a vehicle, but we do not distinguish between these two types of interactions. Figure 8(a) shows an example of such activity being detected. Currently, we only detect when such activity occurs but not where it occurs.

Figure 8(b) shows the ROC curve for activity detection on several test videos. Perfect result was obtained for video 4-50 in which 100% true positive rate is achieved without any false positive, while the other two videos show lower performance. This is expected since video 4-50 has less camera motion and is more similar to the training data such as human motion, view angle and object size. More diverse training data would help improve the results further.



(a) (b)

Figure 8 Activity Detection (a) Person-vehicle interaction example (b) ROC curves for different videos (© MacDonald, Dettwiler and Associates Ltd.)

4. CONCLUSIONS

Aerial video surveillance is useful for military as well as civil and security applications. However, monitoring this growing flood of video data requires significant effort from increasingly large numbers of analysts. We have developed a suite of video exploitation tools to detect objects and activities, including VMTI, person detector and activity detector. They can cue human analysts to inspect video segments that might depict prescribed objects or activities of interests. These tools can be used for both tactical applications and post-mission analytics so that the video data can be exploited more efficiently and timely.

In this paper, we presented the aerial video exploitation tools developed, including the system architecture, the GUI as well as their performance evaluation. The flexible framework allows new detectors to be added easily and algorithmic parameters to be configured dynamically. The VMTI algorithm can achieve real-time performance without user-adjustable parameters. The performance of person detector has been improved by using representative aerial data for training. The activity detector exploits the spatial and temporal information to detect activities of interests after being trained with representative video clips. Promising experimental results have been obtained using these tools.

In the future, the system could also output a KML file to show the detection results in a geographical context. A hybrid VMTI approach that combines the pixel-based and feature-based approaches would be investigated. Moreover, a recent approach using fast feature pyramids could be considered for real-time person detection [10]. We would also extend the activity detector to determine the location of the activity in the video frame and experiment with other types of activities.

ACKNOWLEDGEMENTS

We would like to thank the Canada-Israel Industrial Research and Development Foundation (CIIRDF) for their support. We would also like to thank Eddy Ma and Jared Rewerts for their contribution.

REFERENCES

[1] J.K. Aggarwal and M.S. Ryoo, “Human activity analysis: A review,” ACM Computing Surveys 43(3), 16:1-43 (2011).

[2] R. Poppe, “A survey on vision-based human action recognition,” Image and Vision Computing 28(6), 976-990 (2010).

[3] D. Weinland, R. Ronfard and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” Computer Vision and Image Understanding 115(2), 224-241 (2011).

[4] OpenCV, http://opencv.org/

[5] QT, http://qt-project.org/

[6] VIRAT Video Dataset, http://www.viratdata.org/

[7] N. Dalal and B. Triggs, “Histogram of oriented gradients for human detection,” IEEE Conference on Computer Vision and Pattern Recognition (2005).

[8] UCF Aerial Action Dataset, http://crcv.ucf.edu/data/UCF_Aerial_Action.php

[9] F. Shi, E. Petriu and R. Laganiere, “Sampling strategies for real-time action recognition,” IEEE Conference on Computer Vision and Pattern Recognition (2013).

[10] P. Dollar, R. Appel, S. Belongie and P. Perona, “Fast feature pyramids for object detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence 36(8), 1532-1545 (2014).