

# Detection of moving objects from an airborne platform

D. M. Booth, R. Jones and N. J. Redding

Image Analysis and Exploitation Group, Defence Science and Technology Organisation, Australia.

Email: {david.booth, ronald.jones, nick.redding}@dsto.defence.gov.au

## Abstract

The detection of moving targets in airborne surveillance imagery has a fundamental role in ground picture compilation in both the military and anti-terrorism domains. As the number of unmanned air vehicles (UAV's) increases, Imagery Analysts will become a scarce resource and computer aids will become essential, especially in active, urban regions where occlusions are high and UAV's will necessarily operate in groups. This paper outlines some of the novel functionality of the Analysts' Detection Support System, a flexible processing engine developed to assist the Imagery Analyst to detect targets in all-source imagery. It provides the means of structuring a hierarchy of algorithms that, when applied to the data, makes progressively refined decisions on the locations of targets. We present infrastructural developments that facilitate real-time processing of streaming video, together with an outline of some of the algorithms that have been incorporated to support urban surveillance from airborne platforms, principally the detection of moving objects. We examine the potential for transitioning background modelling based techniques normally used in the ground-based surveillance domain into the airborne domain, and discuss the performance requirements this places on image registration algorithms which must align large numbers of frames prior to the construction of a background model. We propose the extension of our feature based registration technique to eliminate these undesirable off-ground-plane matches.

**Keywords:** airborne video, motion detection, UAV surveillance

## 1 Introduction

Airborne video surveillance by unmanned vehicles is becoming increasingly common as a means of improving situational awareness. As aircraft and their associated sensor and datalink technologies become cheaper, they will become deployed more and more in co-operative groups, providing near total ground coverage in difficult terrains, particularly urban ones where ground coverage is highly valued but difficult to achieve due to occlusion from buildings.

In common with other reconnaissance platforms, such as fast jets, the imagery "deluge" will become too great for Imagery Analysts (IA's) to handle without cueing aids. Generally speaking, these aids will take the form of moving target detectors and trackers, but, in the longer term, one would expect to automate the recognition of characteristic behavioural traits. However, the first stage is to develop detectors that are robust to the projective impact of 3-D structures in the scene, and in some sense this implies automating the alignment of corresponding ground features across two or more frames. This alignment process can be disturbed by the comparatively gross relative disparities between elevated features, as well as changes in occlusion.

Partitioning the motion of moving targets from the more coherent motion of the scene is a difficult problem. In our experience, it is normally addressed by detecting anomalies in a frame-to-frame difference image or optic-flow field [1]. These approaches are accommodated within our "Analysts' Detection Support System" (ADSS), however, here we examine the potential for transitioning background modelling based moving target detection techniques normally used in the ground-based surveillance domain into the airborne domain with a moving platform. We will discuss the performance requirements this places on image registration algorithms (a region based and a novel feature based approach) which must align large numbers of frames prior to the construction of a background model. These algorithms have also become components of the ADSS, a flexible processing engine developed to assist the imagery analyst to detect targets in all-source surveillance imagery. It provides the means of structuring a hierarchy of algorithms which, when applied to the data, makes progressively refined decisions on the locations of targets. The ADSS was originally developed to assist in the exploitation of synthetic aperture radar (SAR) imagery [2], but we will briefly discuss the infrastructure developments which facilitate

effective processing of streaming video data, together with a description of the algorithms that have been incorporated to support urban surveillance from airborne platforms. These focus on the detection of moving objects.

## 2 System Infrastructure

The ADSS has been developed to assist Image Analysts to detect targets in large volumes of imagery [2, 3]. It consists of a set of image processing modules that can be selected and configured to solve a particular task. A simple set of protocols allows communication between modules. New image processing algorithms can be developed and/or integrated into ADSS easily, and this facilitates system growth and evolution, and also acts as an enabler for research collaboration.

ADSS classifies its algorithms as: prescreening (detection), discrimination (recognition) and support modules, and these are reflected in their respective interfaces. The detectors extract a set of target positions in the supplied imagery. These are passed to subsequent modules that endeavour to eliminate the false alarms through the use of attribute based filters and classifiers. As well as detection and classification modules there are various support modules, such as image registration, restoration and segmentation algorithms, for pre-processing of the imagery, and a plotter module that allows target positions to be overlaid onto the supplied imagery. ADSS also has various ancillary capabilities including automatic training (parameter setting) by optimising performance based on a set of labelled target examples, and the ability to carry out a detailed statistical analysis of results when tested.

Central to its design is the concept of a growth dimension. In video streams this is the frame dimension of the video, designated by a frame index. In the usual case of a constant frame rate, the frame index and time codes have a simple relationship and are isomorphic. In ADSS the growth dimension can be assigned to be any one of the dimensions of the data volume (the streaming data can have an arbitrary number of dimensions that are named when the algorithm pipeline is instantiated, *e.g.*  $x$ ,  $y$ , level, layer, frame, etc., and the data volume need not be a rectangular prism. This allows it to cater for multi-resolution datasets and multiple sensors on different coordinate systems).

The growth dimension has an index that is used to indicate the available data at the input to ADSS — it denotes the number of frames collected in the video case. Each algorithm in the pipeline reports the amount of the available imagery it has completed at each processing block, and these messages are used by the infrastructure to ensure that

the necessary input to an algorithm or module is met before it commences processing a particular block. This block approach also ensures that the latency through the system is kept to a minimum — downstream modules do not have to wait for upstream modules to finish with all the input data before they commence. This is of course essential for an operational system.

Probably the most important aspect of the ADSS architecture is its algorithm scheduling / synchronisation capability, which is superior to the alternative systems in its relatively unconstrained scheduling of parallel tasks. This is essential if the system is to adopt the multi-sensor/multi-algorithm approach that we intend. In particular, it means that time consuming processing routes with a high pay-off but low likelihood of success could be explored without disturbing the system as a whole.

The design also supports real-time operation on sufficiently capable computer hardware. The system design models the hierarchy of image processing algorithms as a graph of processes connected by pipes, with each algorithm represented by a process having a single input and a single output; each of these is referred to as a module. Messages are defined that are passed to modules and generated by them as part of their processing; the flow of messages denotes the progress of the image processing system. The processes are simply Unix/Linux processes with the messages flowing on the standard input and output streams. The system supports parallel processing and clustering across a computer network.

## 3 Motion Detection

Object motion in a frame is assumed to manifest itself as statistical differences between it and a background model. This approach was chosen because, in ground-based applications, it has shown itself to be robust to sensor noise, it usually yields complete, high quality object segmentations, and it provides persistent target detection should the newly introduced object stop moving momentarily. We consider the major disadvantages of background modelling to be: the constraints on camera motion that it imposes and its heavy computational overhead, particularly during the bootstrap phase. However, we intend to show that the former is becoming less significant given the ability for UAV's to loiter for many hours at a time, as well as recent improvements in the performance of image registration algorithms. Computational load is not as serious a concern to us as it might otherwise be due to the ability of ADSS to invoke parallel processors dynamically during execution.

### 3.1 Selection of a temporal window

Image registration is the process of overlaying two or more images of the same scene taken at different times, from different viewpoints or from different sensors [4]. In order to construct a background model, the sequence of frames is registered to a single frame of reference, typically the middle frame in the sequence, to form a stack. The aim is to remove the apparent motion in the scene caused by camera movement, so that the residual motion in the scene caused by moving targets can be readily identified.

When constructing a background model from a sequence of video frames, the selection of an appropriate temporal window and sampling frequency is crucial. For a static camera, this is not such an issue: sufficient frames are required to build a statistically confident model but not so many that the model lacks the ability to capture gradual variation in the scene, such as might result from sun precession. Therefore, the model should have a bootstrap phase followed by an adaptive phase, or must have a series of bootstrap phases of constrained duration. When the camera is moving, the length of the temporal window is also restricted by the following considerations:

- (i) Frames within the window should overlap by a significant degree (i.e. the background model should be constructed from frames that overlap in order to be robust).
- (ii) The error induced by the change in viewing geometry as the camera moves through the scene should be acceptable. This difference in viewing geometry generally increases with increasing base line between camera positions, and cannot be entirely removed by a registration process based on a parametric registration model (such as a global affine or projective transform). An illustration is shown in Figure 1c which shows residual structure following subtraction of registered frames.
- (iii) The length of the window should be large enough to exclude from the model the undesired contribution made by moving targets in the scene.

Note, these considerations relate directly to the speed of the sensor, its proximity to the scene and the size and speed of the targets in the scene. A suitable choice for temporal window length is therefore dependent on the type of imagery at hand. In our experiments using standard definition video (at 24 frames per second) to detect vehicles of around 10 pixels in length, a window of 100 frames often yields an acceptable background model while minimising the effects of viewing geometry errors.

Given the notion of a temporal window of limited extent, our approach is to break the video sequence

into blocks of  $N$  frames, where  $N$  is the length of the temporal window. MTI is carried out in each block separately, that is, the  $N$  frames in the block are registered to the central frame; a background model is formed from the stack of  $N$  registered frames; and each frame  $f$  is compared in turn with the background model to identify image regions that have changed (that is, moved).

### 3.2 Image Registration

The favoured registration algorithms are the hierarchical, region based (or more precisely image based) correlation technique described by Privett and Kent [5], and more so, the feature-based matching technique described below [6]. In both cases, the image-to-image transformation is modelled with a complexity up to projective.

#### 3.2.1 Feature-based approach

Our feature based method of registration uses a feature tracking algorithm of Kanade, Lucas and Tomasi [5], known as “KLT”, which has shown considerable robustness over a wide range of video imagery. It is a mature feature tracking method that is well established in the vision community for tracking features in video sequences for the purpose of determining structure from motion [6]. In the KLT algorithm, small features such as corner points are extracted and tracked based on a measure of “corneredness”, derived from the eigenvalues of the autocorrelation of the image intensities within a window and the use of a dissimilarity measure to determine the affine transformation. For our purposes, the tracked features are simply used as control points to which a frame-to-frame parametric registration model is fitted (either affine or projective) and used to warp each frame to a common frame of reference.

#### Removal of 3D artefacts

Shown in Fig.1a is a frame from a segment of airborne HDTV video surveillance (a  $445 \times 280$  subset of the full  $1280 \times 720$  frame is shown for clarity). We tracked 1000 feature points through the sequence and the feature results for this frame are shown overlaid in black and white in Fig.1b. The features correspond to areas of the image with a high degree of “corneredness”, including corners and windows on buildings, moving objects such as cars, and textured regions on the ground. For the purposes of image registration however, it is preferable to omit all but the features on the ground level when constructing the registration model.

To begin with, we remove non static features by applying a RANSAC algorithm [6, 7] which is designed to fit models to data in the presence of a significant number of outliers, using a random

sampling and consensus process. In Fig.1b, outliers are shown in black and these generally correspond to moving targets in the scene (though not always; they can also correspond to unreliable features that do not track well). The next step is to remove points that are higher than the ground plane and to this end we are currently exploring techniques used in determining shape from motion [6]. Given a suitable number of tracked features and sufficient knowledge of the camera position and attitude through the sequence (in our collections, highly accurate metadata has been recorded during acquisition), it should then be possible to reconstruct the 3D positions of the tracked points. It is then a simple matter to filter these positions to extract only those features that are lying in the ground plane. We are in the process of implementing this approach.

### 3.3 Registration performance.

Moving target detection in video imagery usually operates on two frames separated by a short time interval, during which the camera motion is relatively uncomplicated; an affine transformation model is usually more than sufficient. Indeed there are strong arguments in favour of keeping the degrees of freedom low to prevent overfitting, particularly if the scene contains unevenly distributed 3-D structure.

Background modelling requires frames to be registered which are separated by longer time intervals during which time the potential exists for more complex platform and camera manoeuvres to have taken place. This presents several problems. Firstly the frame to frame mappings may need to be more complex *i.e.* projective. These additional degrees of freedom offer the potential for overfitting or for misregistration by convergence to a local minima. In a similar vein, the range of durations between pairs of frames means that, in the case of the region-based approach at least, the optimisation schedule for the algorithm must be comparatively loose so as to encompass the more highly separated pairs. At best this will impact on execution time, and at worst it may impact on the precision of the convergence or may cause convergence to implausible minima.

3-D structure presents serious problems. Buildings, although usually being weakly textured, are generally strong exhibitors of the types of features that are often integral to the registration fit metric (*i.e.* lines and corners).

The region-based approach considers the correlation between the two whole images. Given imagery containing a textured ground-plane

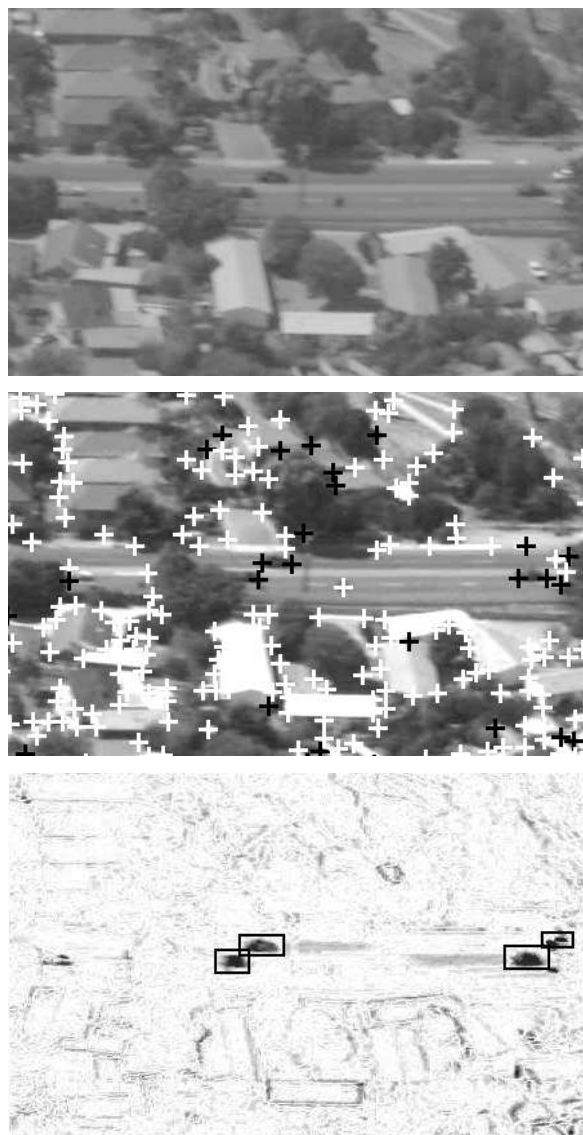


Figure 1: Motion detection in HDTV. Top: Example frame from the sequence. Middle: resulting features after application of RANSAC algorithm; white crosses indicate inliers used for image registration and black crosses indicate outliers. Bottom: Detected targets overlaid on the “difference” image, that is, the number of standard deviations away from the background model.

and some 3D objects it will reach a compromise registration (that could be viewed as an average weighted by image local structure density). Our feature based approach is in some sense more robust. While the proportion of features emanating from the 3-D structure is low, it should have little or no impact. However, because the 3-D structure is high in strong corners, a comparatively low proportion of the imagery being populated by structures such as buildings may cause alignment to building roofs rather than the ground, particularly if the buildings are of uniform height *i.e.* the

relative displacement between corresponding corners is uniform.

These expectations have been borne out by results. We'd expect the region-based approach to recover from translation of 10-20% of the image width and 10-20 degrees of rotation. Performance is reduced when the images exhibit large perspective variation (change in elevation angle). The feature based approach has been applied to a large and varied set of imagery and has performed well, even when the imagery has exhibited relatively weak and sparsely distributed features. In the presence of 3-D structure, the region-based approach has, on occasion, demonstrated a tendency to drift around locally from frame to frame, and there is a gradual increase in registration error as the platform moves further away from the reference image. Under the same conditions, the feature based approach gives precise alignment of building roofs resulting in more pronounced misregistration of the background.

### 3.4 Background Modelling

The main considerations with regard to constructing the background model are: choice of colour space, choice of background model, the methods used for bootstrapping and maintaining the model, and the decision mechanism for identifying statistical outliers. The most appropriate choice depends on the scene content and how it changes over time, imaging conditions, *etc.* As a result, many options are available, and ADSS can be configured accordingly.

All commonly used colour transformations are available in ADSS. In this application RGB was not pursued due to the high degree of correlation between the red, green and blue channels, and the dominance of intensity. Essentially, the discriminatory potential of the colour information is not fully realised. HSV has the desirable characteristic of separating intensity from hue, one benefit of which is the ability to use a simple heuristic to distinguish between object movement and shadow change. However, the continuous, circular nature of the hue feature lead us, in part, to favour a principal component based feature space. We regard perceptual colour as being almost two dimensional, intensity and a colour component, the third feature can be regarded as noise and optionally discarded [8].

The background model of each pixel was represented by the mean and standard deviation of its colour components. These were computed using robust statistics *i.e.* via the median and the median absolute deviation from the median, respectively. This provided a more satisfactory model in the

presence of outliers when dealing with a small number of frames, and in particular, the system can be bootstrapped without a need for the scene to be evacuated of people. Given the background model  $(\mu, \sigma)$ , it is trivial to compute the probability that a particular pixel is consistent with the background model (in terms of standard deviation). Outliers in colour space can be identified using a statistically meaningful multiple of the standard deviation,  $n$ .

In our ground-based work, Gaussian Mixture Models have been used to characterise a scene when, due to motion in a scene, more than one background class may occupy a particular pixel location over time. In common with others, we aim to map one Gaussian to each background class. Each Gaussian is modelled on a cluster generated by the  $k$ -means algorithm, an iterative process based on the distances of samples to current cluster centres, partitions the samples neatly into nonoverlapping labelled clusters.

The sophistication of these techniques varies enormously. The robust statistical technique, though slow, can be re-initialised at any time. Currently, ADSS also uses an update rule, such that the model is updated provided the incoming sample is consistent with the current version. This will accommodate drifts in background characteristics. We are currently implementing an update scheme which offers accommodation of newly introduced objects.

The important feature here is that in an airborne application there is less scope to bootstrap the system on a scene free of targets. Given that the viewing behaviour is quite unpredictable, it means that a background based on mean (*i.e.* Gaussian) or mixture of Gaussian models is probably not desirable. This is particularly true in our use of these models which had been to employ the whole stack of data as bootstrap data, and then look for anomalies within it. This may change when we consider repeated visits.

The impact of registration errors depends on the type of model, and the manner and conditions under which it is computed (in the bootstrap and adaptive phases). We consider the robust, unimodal method in a repeated bootstrap mode (that is, one per stack). Thus, as the sensor moves, the regions of the ground occluded by 3D structures will gradually change. Thus there are two factors impacting on the background model: erroneous ground registration and occlusion inconsistencies. The impact is difficult to predict as it depends on image content. However, relative movement of image structure (whether on the ground or elevated from it) will lead to false alarms, and as the grey level variance estimate close the structure is likely to be raised, changes that take place close to region

boundaries are more likely to remain undetected. This places additional heavy constraints on the size of the temporal window.

### 3.5 Postprocessing

In extreme cases, false alarms have been eliminated by a discrimination module. The false detections are usually caused by changes in occlusion from buildings, and these can be eliminated on the basis of size and aspect-ratio. The aspect ratio is given by the ratio of the eigen-values of the (x,y) pixel locations in the target candidate's mask. Changes (corresponding to potential object motion) are also discarded if they fail to be persistent over a small number of frames.

## 4 Testbeds

Test imagery was collected using the ISR (Intelligence, Surveillance and Reconnaissance) Testbed [9] comprising an airborne platform with an integrated electro-optic & infra-red video sensor (amongst others), line of sight C-band datalink, and a ground element for image exploitation and dissemination. In past trials the video sensor used was the Wescam MX-20, but most recent trials utilised an HDTV sensor based around the Panasonic camera (model AK-HC900) with a 1.2m focal length lens. The video is transported as an MPEG-2 stream from the sensor to the processing/exploitation components. This allows metadata from the sensor (including such things as timecodes, geocoding, and sensor settings) to be encoded with the video stream using SMPTE Standard 336M Key-Length-Value (KLV) triplets [10], and is also an efficient mechanism to compress the video data to reduce datalink and communication requirements. Appropriate application program interfaces and a caching image server architecture within the ADSS's image library provide efficient decompression and reconstruction of the video frames from the MPEG-2 stream.

## 5 Conclusions

We have presented some work on the expansion of ADSS to perform urban surveillance from airborne platforms. It includes adaptations both to the ADSS framework and to the image registration and detection techniques necessary for locating moving objects. The potential to transition background modelling based moving target detection to the airborne domain has been examined, predominantly using a robust, unimodal background model. Future work will attempt to formalise some of the constraints

on this approach and to assess practically the potential of a mixture model to reduce false alarms and weaken the constraints. To alleviate the main problem of disruption by features elevated above the ground plane, we propose the extension of our feature based registration technique to eliminate these undesirable off-ground-plane matches.

## References

- [1] M. Irani and P. Anandan, "A unified approach to moving object detection in 2D and 3D scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, June 1998.
- [2] N. Redding, D. Kettler, G. Blucher, and P. Perry, "The analysts' detection support system for deploying a network of target detection and recognition algorithms in SAR exploitation," *International Conference on RADAR, Adelaide*, pp. 448–453, 2003.
- [3] D. Booth, N. Redding, R. Jones, M. Smith, I. Lucas, and K. Jones, "Federated exploitation of all-source imagery," *Proc. Digital Image Computing: Techniques and Applications, Cairns*, December 2005.
- [4] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.
- [5] J. Shi and C. Tomasi, "Good features to track," *Proc. IEEE Conf. CVPR*, pp. 593–600, June 1994.
- [6] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision, second edition*. Cambridge University Press, 2004.
- [7] R. Jones, B. Ristic, N. Redding, and D. Booth, "Moving target indication and tracking from moving platforms," *Proc. Digital Image Computing: Techniques and Applications, Cairns*, December 2005.
- [8] Y. Ohta, *Knowledge-based interpretation of outdoor natural colour scenes*. Research Notes in Artificial Intelligence 4, Pitman, 1985.
- [9] M. Royce, M. Fiebig, and D. Cannon, "Challenges in the effective exploitation of UAV video — a DSTO ISR testbed instantiation of real-time digital video and metadata delivery," *Proceedings of AUVSI's Unmanned Systems Conference*, pp. 3–5, August 2004.
- [10] SMPTE, "336m-2001:television — data encoding protocol," *Society of Motion Picture and Television Engineers (SMPTE)*.