

## Visual MTI for UAV Systems

Richard Evans, Esin Turkbeyler  
Roke Manor Research,  
Romsey, Hampshire SO51 0ZN

### Abstract

*This paper addresses the task of detecting objects moving on the ground with a video camera fixed to an unmanned air vehicle. As the camera is on a moving platform all parts of the viewed scene appear to move and the task is one of distinguishing between the real movement of moving objects on the ground and the apparent movement of fixed objects. Here we describe an approach based on extraction of point features, tracking and examination of epipolar geometry using the fundamental matrix. We contrast this approach with the more common approach of image registration and differencing. Finally we outline the work planned to exploit this visual MTI stage in a larger system to map ground movements over a wider area, integrating information from many fields of view.*

Keywords: UAVs, surveillance, vision processing, Moving Target Indication, target detection

### Introduction

Video cameras have considerable advantages over many other sensors for a number of applications because they are compact, cheap and low power. Many of these advantages are the benefit of commercial investment through the development of consumer applications. Video is also a passive sensing technique and this is a clear advantage in military applications. The downside is however that their use generally requires manual operation or manual screening of results to achieve acceptable performance levels.

In recent years the use of cameras in large numbers in civil applications of CCTV has lead to the development of automated intruder detection and other video analysis techniques. Many of these techniques are of potential relevance in military as well as civil applications.

This work builds on earlier research [1] which considered military applications of

an event detection technique originally developed at Roke Manor Research Limited (Roke) for traffic monitoring, security and other civil purposes. The method in question is VMAD (Video Motion Anomaly Detection) [2], a feature-based learning system, in which the original development assumed fixed cameras. The earlier work extended these techniques to apply them to moving cameras.

Two different scenarios of interest have been identified, an emplaced camera (i.e. on pan-tilt mount but set at a fixed location) and a UAV-mounted case where motion in all 6 degrees of freedom was possible. Good progress was made in the emplaced camera case, and the ability to learn normal patterns of movement, at least over short time times, was demonstrated. The UAV case has proved more challenging and in view of the potential military benefits of a UAV solution our current work is concentrating on this application.

## Alternative Processing Strategies

The task of detecting moving ground targets from an airborne platform has been addressed by other researchers [e.g. 3, 4]. The main approach adopted is to register pairs of image frames which have been captured a short time apart and difference the registered images, pixel by pixel. In an idealised case there will only be differences between the images where there are moving targets. Thus moving targets can be detected by thresholding or blob-detection applied to the difference image.

The main problem with this approach is that image registration itself is a non-trivial task and small errors in registration can result in spurious targets. Thus some filtering is required at the detection stage and usually tracking algorithms are applied to the blobs to provide further a false alarm reduction.

Image registration is particularly difficult where there is significant 3D structure in the scene. Accommodating 3D structure in registration is not impossible but complex. In extreme cases of mountainous areas or urban situations with tall buildings, the effect of occlusion/dis-occlusion may be mistaken for target activity and further processing would be required to address these regions.

Rather than “register and detect”, to paraphrase the above method, our approach might be loosely described as “detect and register”, though we do not attempt a formal registration of the extracted features but analyse the apparent motion of the features with respect to a static scene model.

In outline our processing consists of the following elements.

- Detection of point features using the Harris corner operator [5].
- Feature tracking using a 2D tracking algorithm [2].

- Analysis of tracked feature positions, primarily based on use of the fundamental matrix,  $F$ .

In short the fundamental matrix encapsulates information about the apparent motion of features between a pair of frames. Let  $\mathbf{x}_1$  and  $\mathbf{x}_2$  be the position of a stationary feature in images 1 & 2 in pixel coordinates, and  $F$  be the fundamental matrix. Then  $\mathbf{x}_1^T F \mathbf{x}_2 = 0$  (where  $\cdot^T$  indicates transpose). For ease of mathematical expression the 2-dimensional position,  $\mathbf{x}$ , is actually a 3-vector  $(u \ v \ 1)^T$ , where  $(u \ v)$  is the position in 2-dimensional image pixel coordinates and  $F$  is a 3 by 3 matrix.

Our analysis of the feature positions operates in three stages to provide a progressive improvement in target detection and false alarm reduction. These stages will now be described in turn.

### Processing Stage 1: RANSAC

The objective of the first stage is to provide a preliminary classification of tracked features from a pair of image frames as a starting point for further refinement.

Considering that there is no initial classification of features as truly moving or stationary our initial estimation of  $F$  is based on RANSAC (Random Sample Consensus) as this is robust to outliers [6]. RANSAC takes a set of subsets of the tracked features for a pair of frames and for each subset computes  $F$ . There are several methods computation of available. We have adopted a method based on Lagrangian multipliers and eigen-analysis [7]. From the calculated  $F$  matrices the one which is consistent with the largest number of matched features is selected. Each tracked feature is then re-examined in turn and depending on how consistent it is with the selected  $F$  it is given an initial classification as moving or stationary.

A typical output of this stage is shown in figure 1. The blue crosses are the features of the random feature subset used to calculate the preferred  $F$  matrix. These are usually correctly fixed to stationary objects but not always, as shown here by the car at the top of the picture. The colour-coded tracks are those judged to be moving, being inconsistent with the stationary hypothesis. The colour coding indicates the level of inconsistency from red (most consistent, below threshold so classed as stationary), through orange, yellow, green, and blue.

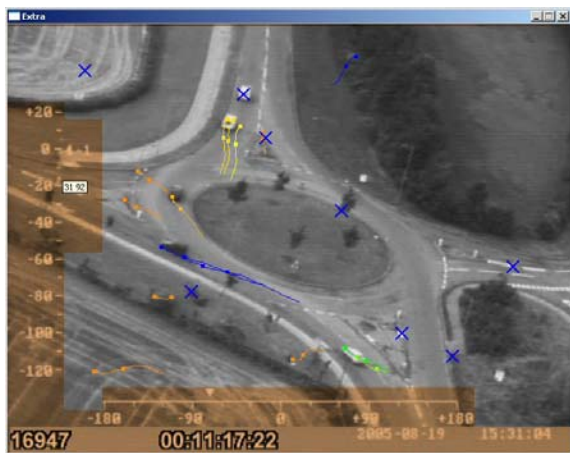


Figure 1 Typical output after stage 1

### Processing Stage 2: 1<sup>st</sup> Temporal Analysis

The objective of the second stage is to exploit temporal information to obtain an improved stationary/moving classification.

Our approach at this stage is to analyse the sequence of (initial) classifications associated with each track for a sequence of processed pairs of frames. Classifications are counted using a cautious voting scheme to identify a set of features which are judged to be stationary with a high confidence.  $F$  is then recomputed using these features and the tracks re-classified.

Figure 2 shows a typical output at this stage. The small blue crosses mark features in the high-confidence stationary set and the coloured tracks mark features classified as

moving, using the same colour coding as before.



Figure 2 Typical output after stage 2

It should be emphasised that although stage two involves a temporal analysis, examining results from many pairs of frames, to identify the high-confidence stationary set, the re-computed  $F$  matrix and the resulting classifications are still based on feature positions in a single pair of frames.

### Processing Stage 3: 2<sup>nd</sup> Temporal Analysis

Though the output of stage two is much improved over stage one, there remain occasional false alarms. There are also occasional missed targets, one reason for this being that the approach taken is blind to some directions of movement – as discussed in more detail below. To improve performance we complete a further temporal analysis based on counting votes at the output from stage two.

A typical output is shown in figure 3. Here we show only tracks classified as moving (indicated in red).



**Figure 3** Typical output after stage 3

### Discussion

The output of the visual MTI chain in three highly contrasting situations is shown in figure 4. In all cases the video has been recorded at a similar platform operating height, i.e. about 1500ft, but with different amounts of camera zoom.

The motorway roundabout has proved the more challenging of our datasets. We have attributed this to that fact that, with the smooth traffic flow on the motorway, there is a large amount of correlated motion in the scene and also the terrain is essentially flat at the image resolution used. Thus the geometrical interpretation of the scene is not very well constrained. It is easy for the RANSAC processing used in stage 1 to be seduced by the moving traffic to generate a preferred  $F$  matrix that corresponds to an interpretation of the scene in which the moving traffic is thought to be stationary and interpreted as objects at some distance above or below the actual ground plan.

The motorway roundabout sequences are also the widest angle viewing geometry of our examples and may be the most prone to lens distortion errors. Calibration issues have not yet been addressed but we plan to do so in the future.



**Figure 4** Further examples of the output of stage 3

The supermarket car park and playing field scene (where two boys are correctly detected running around) are more extreme viewing geometries. In these cases we are moving from perspective to affine geometry. The motorway sequence exhibits a good degree of perspective, the motorway is about 50% wider in pixels in the

foreground than in the distance. With the level of zoom used in these latter sequences, perspective is lacking and as we have used a general perspective form of the  $\mathbf{F}$  matrix, one might be wary of ill conditioning in the calculations. In practice we believe we have not found this a problem but we plan to investigate this in future work.

It should be noted that we have generally found the high-confidence stationary set of features identified in the second processing stage to be very reliable. It is possible that the cautious voting scheme successfully compensates for algorithm behaviour resulting from ill conditioning as well as other sources of error. Though we are pleased with results at the end of the processing chain improvements would obviously be welcome at earlier stages.

The supermarket scene is particularly interesting in view of the complex 3D structure. The apparent motion of the tops of lamp posts and buildings is comparable with that of pedestrians. These significant upstanding features are very largely discounted in the processing while pedestrians, such as those walking between the parked cars at the bottom right of the example image, are detected. In this scene there are some examples of *subjective* features, e. g. where the pole of a lamppost moves in front of a more distant object, create an apparent visual feature. Such features do not fit the stationary world model and can be a source of false alarms.

We have observed that some clearly moving objects are not detected, even at the output of stage two. This is to be expected as the underlying algorithm based on the fundamental matrix is blind to motion in some directions. Recall the definition of the fundamental matrix; if  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are the positions of a stationary feature in images 1 and 2 we have  $\mathbf{x}_1^T \mathbf{F} \mathbf{x}_2 = 0$ . For a given  $\mathbf{x}_1$ ,  $\mathbf{x}_2$  can lie anywhere on a line, so if the

motion is along the line (the *epipolar* line) the feature will be judged to be stationary.

The direction of the epipolar lines will vary with both the position of the features and the motion of the camera. Thus if a moving feature is tracked for a non-trivial period of time we would expect its motion to become exposed at some stage. Thus the temporal analysis in stage three provides this detection opportunity as well as a simple smoothing of errors.

The problem of blindness to the direction of travel could also be addressed by considering the amount of movement along the epipolar lines. This is equivalent to applying 3D constraints and it may be difficult to set suitable threshold values, particularly if slow moving targets are to be detected in an urban situation with tall buildings. It may be possible to determine thresholds adaptively however, analysing the motion of features in the high-confidence stationary subset identified at our second stage of processing.

### Conclusions on Visual MTI

We have described a visual MTI system for use on UAVs and illustrated its results on some sample images. This system processes imagery from essentially overlapping fields of view to distinguish stationary and moving objects on the ground. There remain topics to be addressed in future, namely the effect of lens distortion and limiting cases of the viewing geometry (moving from perspective to affine geometry) and blindness to certain directions of movement. Nevertheless promising results have been obtained in a range of situations.

### Future Work

The main subject of this paper is a visual MTI system, but we aim to incorporate this

is into a larger system able to map activity over a larger area. This will allow the detection of larger movements, such as widely spaced convoys as opposed to independent vehicles, and support higher-level analysis to detect changes in patterns of movement. Thus the concept here is of a UAV system repeatedly sweeping an area of ground or returning regularly to it.

Our feature-based techniques provide a basis for this development as well as for the visual MTI front end. We anticipate a system with two main stages of processing.

- First data from a pair of partially overlapping fields of view is linked together with registration parameters derived from the positions of features tracked through intermediate frames.
- Second a sequence of partially overlapping frames is chained together using the pair-wise links to integrate data from several fields of view onto a single map.

There are a number of issues to be addressed here, such as how to reduce integration drift when chaining sets of information together and also how recognise and exploit return visits to the same location – a topic which is being addressed in projects within the SEAS DTC [e.g. 8, 9]. At the time of writing this part of the work is at a very early stage, but combined together we believe visual MTI and larger area mapping will provide a powerful surveillance tool for UAV systems.

### References

- 1 R J Evans, R G Porges “Video Motion Anomaly Detection for Military Applications”, 3<sup>rd</sup> EMRS DTC Technical Conference, Edinburgh 2006.
- 2 R J Evans, E L Brassington “Video Motion Processing for Event Detection and Other Applications” IEE Annual Conference on Visual Image

Engineering, VIE2003, University of Surrey.

3. I Cohen, G Medioni “Detection and Tracking of Objects in Airborne Video Imagery”.
4. A Mittal, D Huttenlocher “Scene Modelling for Wide Area Surveillance and Image Synthesis.
5. C Harris, M Stephens “A Combined Corner and Edge Detector”, Proc 4<sup>th</sup> Alvey Vision Conference.
6. M A Fischler, R C Bolles “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”, Graphics and Image Processing, June 1981, Volume 24, No 6.
7. Z Zang “Determining the Epipolar Geometry and its Uncertainty: A Review”, International Journal of Computer Vision, 27(2) 161-198 (1998).
- 8 C Harris “Strategies for Visual Exploration of Buildings”, 2<sup>nd</sup> SEAS DTC Technical Conference, Edinburgh 2007.
- 9 D Schroter, I Posner, P Newman, K Ho “Using Vision for Outdoor SLAM”, 1<sup>st</sup> SEAS DTC Conference, Edinburgh 2006.

### Acknowledgements

The work reported in this paper was funded by the Electro-Magnetic Remote Sensing (EMRS) Defence Technology Centre, established by the UK Ministry of Defence and run by a consortium SELEX Sensors and Airborne Systems, Thales Defence, Roke Manor Research and Filtronic. The authors also wish to thank the Police Air Support Unit for assistance in recording the data used in this paper.