

Novel View Synthesis for Change Detection

Aeron Buchanan

2d3 Advanced Imaging Group
14 Minns Business Park, WestWay, Oxford, OX2 0JB, UK

Abstract

The threat of improvised explosive devices (IEDs) is considerable and they are the cause of significant casualties in the current UK military operations. In this project, we demonstrate that the detection of IEDs can be performed using the automated comparison of image sequences taken over roads and terrain susceptible to this kind of attack. Image sets acquired at different times are able to show, either directly or indirectly, the installation of IEDs made in the intervening period. Novel view synthesis techniques provide a way for the detection of any changes present across such image sets to be flagged automatically.

Introduction

Novel view synthesis (NVS), in the context of this work, is the generation of images of a static scene from viewpoints that have not been previously seen, using multiple views of the scene. We suggest that NVS can be used for the task of change detection, and specifically in aid of the identification of improvised explosive devices. IEDs are a major cause of Allied casualties in Iraq and Afghanistan, and, as such, improving the process of their identification has been labelled as a priority by the MOD. Over the course of this project, two NVS solutions for the automation of change detection for IED identification have been developed. One generates an explicit 3D reconstruction of the scene, while the second works implicitly in a probabilistic framework.

Scenario

When military convoys and patrols move along roads in theatre, knowledge of the presence of road-side IEDs is vital. Explosive threats installed on the side of roads vary in size, shape and concealment. Because it is common for IEDs to be hidden, their detection must often be made through indirect observations, such as moved stones or disturbed earth.

Fortunately, advances in the technology and availability of unmanned aerial vehicles (UAVs) provide the deployed forces with the possibility of repeated aerial reconnaissance of routes where IEDs are likely. High viewpoint images along these routes can therefore be obtained and reviewed prior to a vehicle s or convoy s passage. While direct inspection of such footage may give the trained analyst some indication of the presence of suspicious objects, it is their comparison to previously captured images that can provide the best way to identify potentially hostile items. This is the basis of change detection. However, due to the length of the routes in question and the time required to obtain full surveys, operators are expected to review many hours of footage in order to make such change detection comparisons. Reviewing and analysing video for such extended periods of time is undesirable, possibly leading to unnecessary fatigue and the errors can follow from it. The benefits of efficient automation are clear.

We envisage a system where the most recent UAV surveillance footage is processed with and compared to a reference video of the same terrain in an automated pipeline. Significant changes are then flagged for review by the image analyst.



Figure 1. Frames from two video sequences acquired for this project using a micro UAV from Bluebear Systems Research. Top row: first sequence. Bottom row: second sequence.

Test Footage

Multiple flight trials were conducted to obtain test footage for this project using a micro UAV built by Bluebear Systems Research. Some example frames are presented in Figure 1. Considerable motion blur present throughout the video made the sequences particularly challenging (see Figure 2), creating difficulties for both the feature-based tracking (for the structure from motion process; see below for details) and image comparison stages. The footage was taken over a grassy industrial estate and focussed on a section of tarmac road where changes in ground object arrangements were introduced between each fly-over. The differences between the passes shown in Figure 1 are the movement of a white bin-like object along the left edge of the road (third column) and the movement of the two people, who appear as dark blobs, around and away from the light blue metallic car (fourth column).



Figure 2. Motion blur significantly changes the appearance of objects; for example, this white object should be circular.

Processing Pipeline

The full IED detection pipeline comprises two main processing steps:

Input: Two or more image sequences

Stage 1: Image Registration (NVS)

Stage 2: Change Detection

Output: Significant changes highlighted

The majority of the work in this project applies to Stage 1, although as will be explained below, the second of the two approaches we have developed has the advantage of being integrated with Stage 2.

NVS for Change Detection

Change detection relies upon the ability to compare pixel data from two images of the same scene (captured at different times) with, crucially, each comparison being between data for exactly the same point in that scene. Some systems achieve this by guaranteeing that the camera does not move between the times of capture so that raster pixel correspondences directly represent scene correspondences. However, there can be no such guarantee for UAV footage. Therefore, complex scene dependent pixel correspondences are required to ensure meaningful comparisons. We realize this goal by performing image registration through the generation of new images.

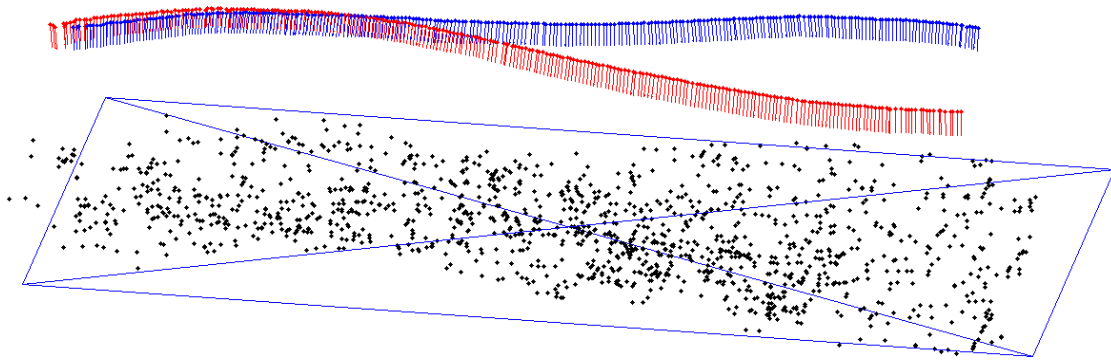


Figure 3. The camera and scene reconstruction of the micro UAV sequence. The two camera paths (flights) have been automatically registered in three dimensions and are shown in blue and red. The black dots are 3D scene points and the blue rectangle is the ground plane.

Novel View Synthesis

We developed two different methods for generating new (registered) images of a scene. The first creates a texture-mapped 3D model of the terrain and creates new views of it by direct re-rendering and is thus termed the *explicit* approach. The second, or *implicit* approach avoids the explicit construction of a scene model and can even be integrated into the change detection pipeline without the need to explicitly generate the novel view image itself.

Both approaches comprise three steps:

1. Solve for the camera parameters
2. Sample the input images
3. Create the new image

1. Camera Solve

The first step is common to both the implicit and explicit approach. We use an advanced version of the structure from motion algorithm [1] to simultaneously solve for the positions, orientations and internal specifications of the cameras, as well as for the 3D structure of the scene. Having done this for each flight path separately, the resulting solves must then be registered so that they all exist in the same coordinate system. In theory affine transformations would suffice, but in

practice a full resolve is required to iron out relative drifts and warps that can accumulate for long sequences. Combining the data from multiple passes has the additional effect of increasing the fidelity of the solution.

2. Image Sampling

For the explicit method we use the solve from the first step to generate a 3D reconstruction of the scene. We use the irregular grid of recovered scene points to generate a continuous surface by interpolating the structure with a thin plate spline. However, the reconstruction of scene geometry alone is not enough; scene appearance must also be reconstructed. We generate a texture map for the reconstructed scene model by simulating orthorectified projection of real image fragments onto the 3D mesh. The fragments are selected from the image sequence using an angular proximity measure.

The implicit method employs an epipolar geometry based sampling method [2] to create a sample colour matrix for each pixel of the new image. There will exist a pattern of consistency within these matrices that implicitly encodes information about the scene behind that virtual pixel [3].

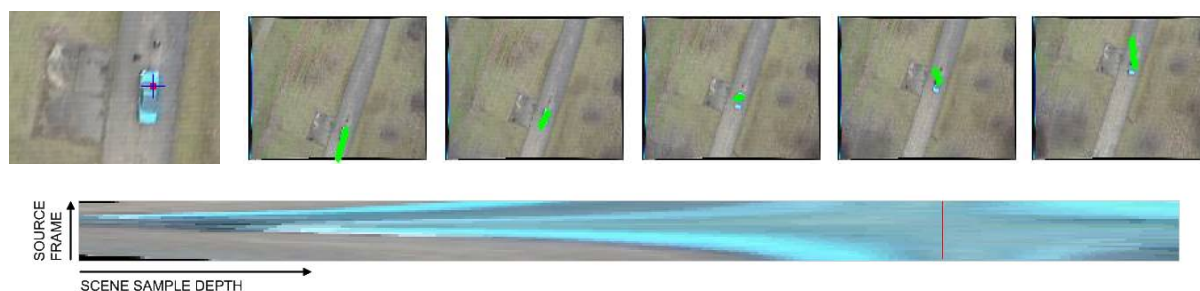


Figure 4. Epipolar sampling for the implicit approach, where image registration is achieved by recreating images from a subsequent (blue) pass as novel views using only images from the first (red) flight path. Top left: a detail of the target image from the blue flight path. The cross denotes the pixel under consideration. Top right: five frames from the red flight path with the reprojected sample points plotted in green. Bottom: the pixel values under the green sample points arranged in a matrix. The correct colour-consistent scene depth is indicated as a red line.

Populating the sample colour matrix for a given pixel is achieved by generating a set of scene depth sample points on the scene ray from the novel camera's centre through the coordinates of the pixel on the camera's image plane and then reprojecting these points into the cameras of the source images. An example of this process is given in Figure 4.

3. Novel View Synthesis.

Once all the source images have been sampled, the gathered pixel data is used to generate the new view itself.

For the explicit method, the 3D camera parameters of the novel view are used to render an image of the texture-mapped model. Any hardware or software renderer can be used. For this project, we employed our own OpenGL based rendering engine.

The implicit method necessarily employs a more involved process and for this project a range of implementations were developed and tested. All are fundamentally methods to determine the scene depth for which the colour entries in the sample colour matrix are most consistent (a single depth corresponding to a column in the matrix) and were investigated in the search for a reasonable speed/quality trade-off. The slowest method optimizes a pseudo

probability density function [4] to enumerate and rank the modal colours of high probability. This method provides the best potential for full pipeline integration and output accuracy, but takes up to 1.5 seconds per pixel in our matlab proof-of-concept implementation. The other methods look directly at the statistics of the colour distribution at each depth sample to determine the most likely scene colour. These methods can process around 8 pixels every second in our comparable implementations: a greater than tenfold increase in speed.

Change Detection

The synthesized views are generated using only images from a single flight over the scene and are created using the specifics of the cameras from the other flight(s). In this way two images for any area of the scene become available: one from a real camera and one generated as described above. Much work already exists regarding change detection and this project's remit did not include any further work in this field. However, as already mentioned, the implicit approach allows for integrated NVS and change detection algorithms that bypass the need to explicitly generate a new image.

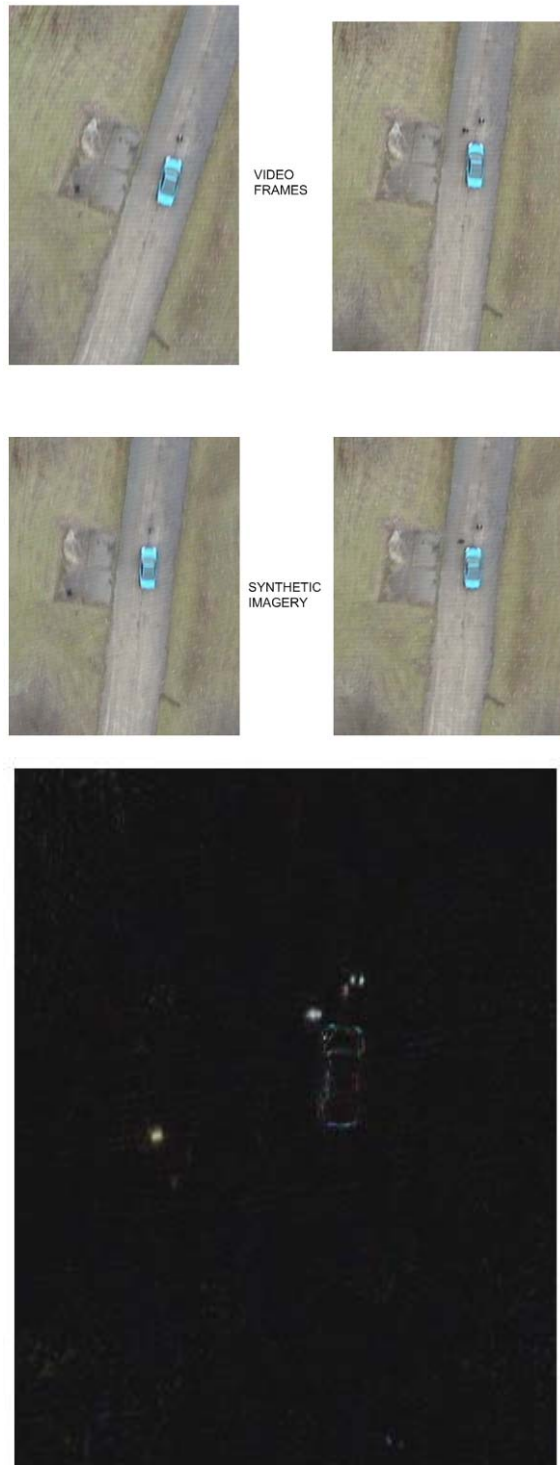


Figure 5. A change detection example from the explicit approach. Top: similar frames from the two flights. Middle: the image sequence represented on the left has been registered to a specific viewpoint from the other sequence, allowing direct comparison. Bottom: the difference image shows the success of the process in detecting the movement of the two people in the images.

Conclusions

The explicit technique has proved to work very well on our test sequences, with no false negatives and very few false positives. It is relatively fast, taking of the order of several hours to process a kilometre of terrain footage. This approach is limited by the maximum fidelity obtainable by the 3D reconstruction process. More sophisticated techniques are possible and should be investigated, but such development was sadly beyond the scope of this project. Nevertheless, we are strongly encouraged by its robust performance (e.g. Figure 5) and can see that there is great potential that further development could realize.

The implicit technique has the potential for greater accuracy, but proof-of-concept processing times are very large (of the order of 1 second per pixel, which corresponds roughly to times of the order of days to fully scan a kilometre of terrain), although the major bottleneck is simply the sampling of the input images: a process that can be greatly enhanced by dedicated hardware. We see the advantages of this implicit approach coming from the ability to allow for the integration of the two processing steps into a single efficient analysis stage, and so enabling a greater amount of information to be available for change detection and thus making for more informed decisions on change occurrence and significance. However, our initial tests suggest increasingly poor performance as the flight path separation distance is increased (see Figure 6).

It is worth noting that both solutions are based on precise knowledge of the 3D attributes and specifications of the cameras that captured the images. As part of our solution, we have developed the ability to automatically determine these parameters. If additional data were available, e.g. GPS location data, it could be used to supplement our automatically recovered data and increase reconstruction fidelity.

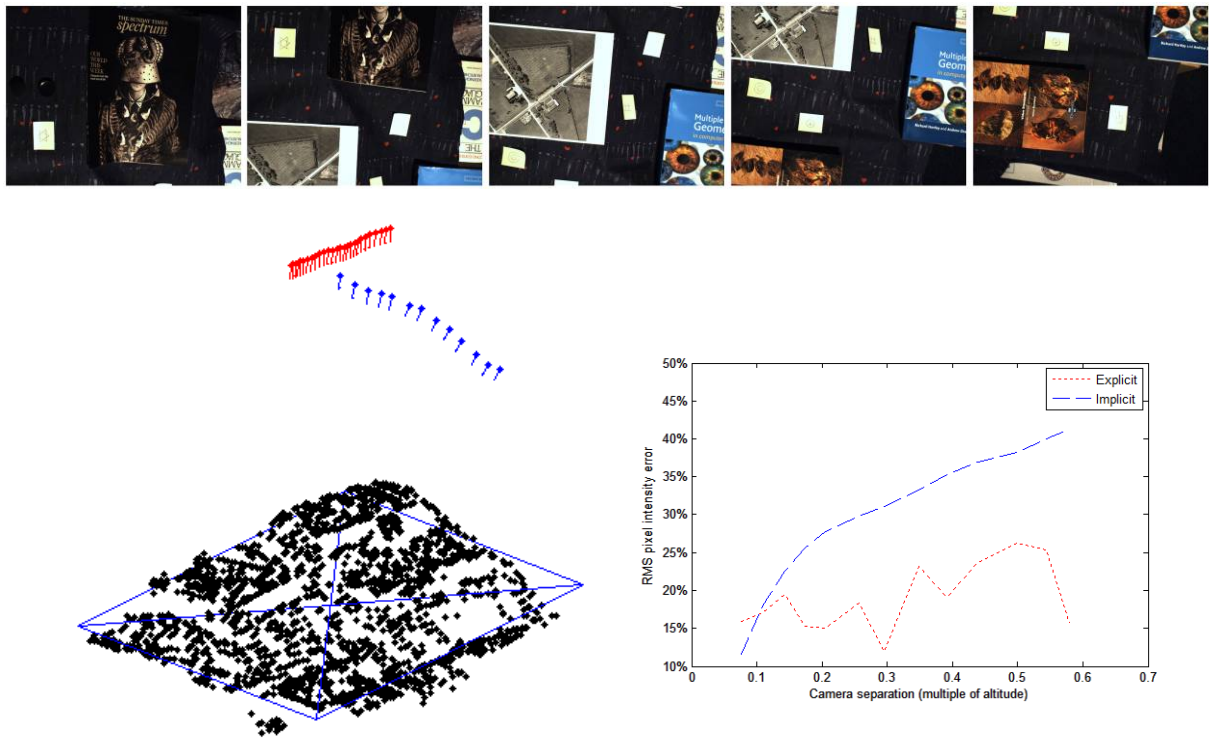


Figure 6. Camera separation test results. Top: example frames from a pass over a simulated scene. Left: the automatically recovered scene structure and camera positions. Right: graph of view synthesis ground truth error using the images from the red cameras to generate views from the positions of the blue cameras for both the implicit and explicit NVS techniques.

It is also advisable to note that in the current form, the techniques described are off-line processes. All the image data must be collected in full before the processing can begin. We envisage that these techniques can be upgraded to work in a more real-time fashion, although that work must be left to future projects.

We see several further advantages in our approach. For example, the scene structure used in the explicit approach to generate the 3D reconstruction gives us the unique advantage of being able to develop structure aware algorithmic enhancements, such as shadow prediction for false positive mitigation. Our framework also provides the ability to extend the process to work with more or fewer channels and with imagery of different wavelengths and modalities. This flexibility makes future

collaborations with other EMRS DTC work possible.

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 of SELEX Galileo, Thales UK and Roke Manor Research.

References

1. A. W. Fitzgibbon and A. Zisserman, "Automatic camera recovery for closed or open image sequences", *Proceedings of the European Conference on Computer Vision*, 1998.

2. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, second edition, Cambridge University Press, 2003
3. M. Irani and T. Hassner and P. Anandan, “What Does the Scene Look Like from a scene point? ”, *Proceedings of the European Conference on Computer Vision*, 2002
4. A. W. Fitzgibbon and Y. Wexler and A. Zisserman, “Image-based rendering using image-based priors”, *Proceedings of the European Conference on Computer Vision*, 2003