

Visual Change Detection for Route Monitoring

C. Stennett, R.J. Evans

Roke Manor Research Ltd, Old Salisbury Lane, Romsey, SO51 0ZN, UK

Abstract

This project has developed computer vision algorithms to detect differences observed between different journeys along a route, with the prime objective of finding differences resulting from the placement of improvised explosive devices (IEDs). The processing algorithms developed are based on accommodating viewpoint differences by warping imagery using matched features as tie points. Here we focus on the key processing stage of inter-video feature matching. Both grey level correlation and edge matching algorithms have been implemented for change detection. On sample data sets, both methods successfully detected the changes in the environment, albeit with some outliers. It was found that the grey level correlation method yielded superior results, but the edge-based matching method is expected to perform more reliably after more development. Identified future work includes refinement to the low-level processing algorithms and the development of high-level interpretation functions, which are required if a change detection system is to be operationally viable.

Keywords: Vision processing, tracking, feature matching, IED, ATR.

Introduction

Many aspects of military operations – and civil operations too – rely on, or can be improved by, an individual person's familiarity with their operating environment. A soldier can observe and compare his current observations with previous ones; "This pile of rubbish wasn't here yesterday". This familiarity and ability to detect change enables soldiers to be more wary in the presence of potential threats; they can use this to prioritise search activity or simply behave in a way to mitigate risks.

Unfortunately humans tire and become less observant, and they change shift or rotate deployment so they are replaced by newcomers wholly unfamiliar with the particular environment. Automating visual change detection is one of a number of techniques aiming to reduce reliance on individual's familiarity with the environment.

In this project we are primarily concerned with the application to monitoring convoy routes. In an example scenario, imagery from a video camera mounted on a lead vehicle in a convoy is compared, in real-time, with a previous video recorded during a previous journey along the route. Depending on the sensor resolution and platform scenario, we expect to be able to detect changes in road-side artefacts and furniture, such as might disguise an IED. Change detection could be used simply to provide a warning to the convoy commander, or it could be used to trigger automatic target recognition algorithms or cue the deployment of dedicated sensors and processing.

While our primary interest here is in monitoring convoy routes, the developed techniques may also be applicable to a host of similar linear tasks such as monitoring power lines, pipelines, perimeters and borders.

Current and historic video will not be identical, even where there is no change, because viewpoints and lighting will differ. Thus comparison techniques will require a high degree of intelligence and the aim of the work presented here has been to develop the necessary processing algorithms.

We originally planned to address change detection with either a downward-looking airborne camera or a forward-looking ground-vehicle mounted camera. In view of other work [1] proceeding in parallel with this project which is concentrating on the downward-looking UAV-mounted camera case, we have concentrated on the ground vehicle case. This is considered more readily exploitable from a military point of view, though it is possibly the more challenging from a technical point of view.

Project Overview

The essence of this project is the comparison of images of a scene taken at different times. In very controlled situations, such as prevail in industrial machine inspection, this is readily achieved by image differencing. We have found very little literature on the subject of change detection outside of such controlled environments and a few specialist domains such as retinal and other medical analysis tasks.

In this case, several factors make simple differencing methods ineffective, in particular differences in viewpoint and illumination. Coping with differing viewpoints is highly problematic, not least because of the effects of the 3D nature of the observed scene. In the outdoors, the scene illumination usually varies significantly with the weather and the sun angle. For this reason, simple photometric image comparisons (e.g. pixel differencing) are not reliable even if imagery has been successfully registered.

On the other hand, there are three important characteristics of the data that apply when

attempting to use the change detection concept for monitoring convoy routes:

- The path taken by the camera varies only a moderate amount from one journey to the next.
- The platform speed is likely to be different from one journey to the next, and so the synchronisation between the video sequences does not stay constant.
- The convoy passes the same points along the route in the same order each time.

In light of these characteristics, we have broken down the task of visual change detection into six key stages, addressing in turn the key issues of each task:

- Feature extraction
- Intra-video (i.e. frame to frame) feature matching
- Inter-video feature matching
- Video synchronisation
- Viewpoint accommodation
- Change detection

Feature extraction is the task of extracting localisable, trackable features from an image.

Frame-to-frame **intra-video feature matching** is the task of associating observations of the same feature seen over a number of frames into a single object. Here we call the observations *features* and the list of associations a *tracked feature*. For processing intra-video features we use existing feature-extraction and tracking algorithms.

Inter-video feature matching is the task of matching features seen in both current and historic video. This is described in more detail below.

Video synchronisation is the task of ensuring that a video frame of the current video is compared with the most appropriate frame of the historic video, compensating for changes in platform speed along the journey route. We developed a

simple approach to maintaining synchronisation using the triangulation of inter-video feature matches as used below in viewpoint accommodation.

To address **viewpoint accommodation**, we developed techniques to warp a current image to achieve a pixel-to-pixel correspondence with a historic reference. These algorithms are based on inter-video feature matching. We use the matched features to provide a set of tie-points and interpolate between them using a triangulation of the image.

With the imagery now registered by the process of viewpoint accommodation, to address **change detection**, we developed algorithms to detect areas of visual change. In this project, we considered and implemented two approaches both based on dividing the imagery into cells, and scoring corresponding cells (after viewpoint accommodation) according to their similarity in:

- Grey level correlation
- Edge matching – Spatial (position and orientation) correlation of edge information

The six processing stages above provide the nucleus of a technical/algorithmic solution for a change detection system. Each processing stage is too complex to describe fully here but the most challenging, inter-video feature matching, is outlined below.

Inter-Video Feature Matching

Inter-video feature matching can exploit three kinds of matching constraint:

- Appearance: the most obvious constraint answering “Do the features look like each other?”
- Spatial: on similar views, a feature in, for example, the top left of one video must be near to the top left of the other video.
- Temporal: inter-video feature matching should output matches, which are consistent with intra-video matching.

Those that are inconsistent are unreliable and so can be rejected.

Even with these constraints available, inter-video matching was found to be highly problematic, with the main factors being:

- Differences in the time of image capture mean that the scene is viewed in different lighting and weather conditions. Thus the scene may be generally lighter or darker, perhaps with a different colour balance, or have differing shadow orientations which may affect the relative brightness of different areas if viewed in monochrome.
- Differences in viewpoint, in a 3D structured scene, mean that there is no simple image transformation which brings all the features into alignment. However, by synchronising the video stream, we can assume that the field of view is broadly similar.
- Different sets of features are detected in each of the videos. Thus, there is no one-to-one mapping between the sets of features. This may be caused by image noise but also because changes in illumination and viewpoint may affect the relative prominence of different features.
- Lens distortions mean that the relative positioning of features vary with position in the image. These effects are more pronounced for wider angle lenses as lines which are straight in 3D appear curved on the image. Fortunately, this has not been pronounced in the scenarios we have addressed.

We considered a number of feature types for this application and after experiments we chose Harris corner features [1] in preference to SIFT [5] features. This was because we found Harris corners substantially more plentiful than SIFT and well positioned in relation to scene structure, reflecting their point-like, as opposed to blob-like, nature. There remained two major issues:

- The Harris corner detector, as originally formulated, works at a single scale.
- The Harris corner detector's attributes are effective for matching frames close in time and orientation (i.e. for intra-video feature matching), but are not so good for inter-video feature matching.

The solutions we implemented to solve these two problems are thus:

- Perform Harris corner detection at multiple resolutions, i.e. not only extract features at full-resolution, but also at half and possibly quarter resolution as well. This has proven to work well and incurs ~40% CPU overhead.
- The SIFT local image descriptor (LID), as described in [5], provides a vector of information per feature which can be used for wide-baseline comparison purposes. SIFT uses this descriptor for matching purposes, and there is no reason why it cannot be used with other feature types, including Harris corner features.

We found however that matching based on SIFT LIDS and weak (2D) spatial constraints was only partially successful even when applied to well synchronised videos. Mismatches were typically under 1%, but this is unacceptable given their impact on later processing. Stronger spatial constraints were required and it was not clear how to choose the required seed region to start the algorithm.

Therefore, we opted to employ a stronger spatial constraint based on the fundamental matrix which encapsulates information about the relative viewing geometry for a pair of views of a static scene. The fundamental matrix, F , is a 3×3 matrix. If $\mathbf{x}_1 = (u_1 \ v_1 \ 1)^T$ and $\mathbf{x}_2 = (u_2 \ v_2 \ 1)^T$, $(u_1 \ v_1)$ and $(u_2 \ v_2)$ being the image coordinates of corresponding features as viewed from two camera positions, 1 and 2, then

$$\mathbf{x}_2^T F \mathbf{x}_1 = 0$$

Once F is calculated for a pair of inter-

video frames, we have a powerful constraint on \mathbf{x}_1 and \mathbf{x}_2 and this can in theory be used in place of the weaker spatial constraint used in our initial algorithm. For a given \mathbf{x}_1 , $F\mathbf{x}_1 = 0$ defines an epipolar line in the second image on which the feature \mathbf{x}_2 must lie. This holds for an idealised pin-hole camera, but it does not hold in the presence of radial distortions. Therefore, in practice, we cannot apply this constraint extremely tightly. Even so, its use resulted in a reduction in mismatches by about a factor of 2 of without having to resort to using a seed region.

Calculation of the fundamental matrix itself is not straightforward. It is possible to calculate the fundamental matrix given a set of matched features but this is a "chicken and egg" situation, made worse by the fact that calculation of the fundamental matrix is sensitive to errors in measured feature positions and a putative set of matches may include some mismatched features.

Roke has experience of this in other DTC project work [6] where we applied a RANSAC algorithms estimate F . We were, therefore, able to use existing techniques, albeit with some adaptation.

In this implementation, the putative matches are selected frame by frame using SIFT LIDs in conjunction with very weak spatial constraints. To ensure that we do not exclude many correct matches we allow each feature to match (putatively) to as many as three others. Thus we know there will be a high level of mismatch amongst the putative matches. Initially we used a standard RANSAC algorithm, which we combined with the Longuet-Higgins 8-point algorithm [2] for computing F given a set of matches. With our datasets however, we found that matched features were mostly concentrated in a part of the image and in these cases the "majority rules" RANSAC process frequently produced a fundamental matrix which did not perform well across the whole image. To improve the likelihood of finding an F matrix appropriate to the

whole image, subsets of selected features to be used were chosen to be well-distributed across the field of view rather than purely randomly. Also we adopted the MSAC (M-estimator SAmple Consensus) [4] approach.

Road Data Example

Figure 1 and Figure 2 show frames from a pair of test sequences. This video data consists of a pair of sequences recorded a few weeks apart in March 2008 – one on a sunny day and the other cloudy. Figure 3 and Figure 4 show change detection results. Although these are colour videos only the grey-level data has been used. Techniques might be extended to colour but note some terrains contain very little colour information and so being reliant on colour would reduce performance.



Figure 1: Raw Image from Video 2



Figure 2: Raw Image from Video 1

Example output of the grey-level correlation method is shown in Figure 3 where the red overlay marks cells with low inter-video correlations indicating high levels of change. Areas correctly mark red

are the dustbin placed by the kerb on the left and the moving traffic in the distance, though some areas of change are incorrectly marked among the bare tree branches.

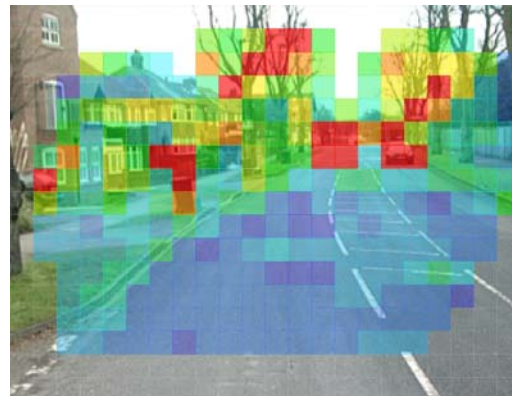


Figure 3: Grey level correlation method



Figure 4: Edge matching method

Performance of the edge-based method (Figure 4) is harder to appreciate, though it follows a similar pattern. In each image cell, the numbers of unmatched and matched edgels are indicated respectively by the areas of magenta and yellow overlay. Thus a high proportion of magenta in a cell indicates a high level of change though in some cases this is not a reliable indicator being based on a small number of edgels.

At this early stage of development, the grey level correlation method performs better than the edge-based method. However, the edge-based approach is expected to improve with further development and eventually provide a more reliable solution in more challenging scenarios.

Discussion of Results to Date

Work has concentrated on algorithm development in the context of four pairs of

video sequences. This experience has highlighted the need for work in two areas

- Continued refinement of low-level processing techniques, i.e. processing at the level addressed here.
- Development of higher-level processing if threat detection based on change detection is to be a viable proposition.

Concerning low-level processing we have found that most errors in the final change detection result can be attributed to the inter-video feature matching stage where a lack of successfully matched features may result in faults in viewpoint accommodation. The inter-video feature matching algorithms are however considered to be robust and so provide a good basis for further processing that would “infill” sufficient additional matches to properly support viewpoint accommodation.

Concerning high level processing, by working with real video examples, it has become clear that not all changes are of military interest and some high-level interpretation and screening will be needed. An example is parked cars, which in the UK may not be significant. In a desert village, an unexpected abandoned vehicle is likely to be very significant. Other issues include shadows, pedestrians and moving vehicles, and 3D structures, such as road signs. At present individual issues are considered soluble, some using off-the-shelf solutions. Addressing these issues will require the integration of several low-level and high-level techniques.

Outline of Future Work

Proposals building on the work to date have been submitted to the EMRS DTC. These cover the first of three future phases of work aimed to lead to a first experimental operational demonstrator.

Phase 1 is the refinement of low-level processing algorithms, concentrating on the infilling of inter-video feature matches obtained by current algorithms. This work

would be coupled with an identification of high-level interpretation issues and the short-listing potential solutions.

Phase 2 covers the development and integration of a selected set of high-level interpretation functions.

Phase 3 addresses system and user interface issues such as system initialisation and user displays.

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.

We would like to thank Omnicom Engineering Limited who provided the video data shown in this paper.

References

1. A. M. Buchanan “Novel View Synthesis for Change Detection”, *6th EMRS DTC Conference*, Edinburgh, 2009.
2. H. C. Longuet-Higgins, “A computer program for reconstructing a scene from two projections”, *Nature*, vol. 293, 1981, pp. 133-135, 1981.
3. C. Harris and M. Stephens, “A combined corner and edge detector”, *Fourth Alvey Vision Conference*, Manchester, UK, pp. 147-151, 1988.
4. P. H. S. Torr and A. Zisserman, “MLE-SAC: A New Robust Estimator with Application to Estimating Image Geometry”, *CVIU*, vol. 78, pp. 138-156, 2000
5. D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, *International Journal of Computer Vision*, 2004.
6. R.J. Evans and E. Turkbeyler, “Visual MTI for UAV Systems”, *4th EMRS DTC Conference*, Edinburgh, 2007.