Three dimensional scene reconstruction from passive imagery

Malcolm Rollason
QinetiQ Limited, 1105/A7, Cody Technology Park,
Ively Road, Farnborough, GU14 0LX

Abstract

This paper describes a new approach, developed under the auspices of the Electro Magnetic Remote Sensing Defence Technology Centre, for reconstruction of three dimensional scene content using passive video. This gradient based scheme uses the parallax in an image sequence to estimate the distance from sensor to scene. Intensity measurements from a passive camera are combined with a rigid body model of scene content to infer both the motion of the camera relative to the scene, and the 3D scene structure. The latter is stored as a 'depth map', describing the distance from sensor to scene for every pixel. Inference of the camera motion is via an efficient Bayesian scheme, HINTS, which estimates the incremental change in the camera position and pose, between successive images. For the inference of both the camera motion and the scene depth map, a recursive framework is adopted which seeks to minimise an error defined using the intensity in the measured images. The resulting scene depth estimate is illustrated using airborne imagery.

Introduction

An important enabler, for techniques that exploit a sequence of images in which there is relative motion between sensor and scene, is the ability to predict the frame-to-frame motion of the image via a geometric model.

Techniques that use frame-to-frame motion include temporal resolution enhancement (which provides improvements in spatial resolution c.f. the raw sensor imagery); mosaicing; change detection; and track before detect.

For scene content that is approximately flat and viewed in the far field, a valid geometric model for frame-to-frame motion is the affine transform. Frame-to-frame motion is then a warp that preserves parallel lines in the image, and where the same six affine parameters define the transform between frames for all pixel locations.

The flat scene approximation of the affine geometric model is not valid for scene content comprising substantial depth variation, which causes parallax within the imagery. The apparent relative motion of objects in the scene is dependent upon their distance from a moving sensor. A geometric model is required that represents parallax, where objects move in the video according to their distance from the sensor, and the latter must be inferred from the imagery.

Core components of a geometric model for 3D scene content are:

(i) a description of the distance from sensor to scene, - i.e. a depth map;

(ii) a description of the camera motion between frames;

(iii) a method to predict the depth map (and associated pixel intensities) such that they are projected, according to defined sensor motion, from one frame to the next.

The task of inferring scene depth from camera motion has been tackled by a number of researchers. One approach to 3D reconstruction [1] [2] is to use feature tracking (for example, corner detection and frame-to-frame feature association), followed by solving for the 3D co-ordinates of tracked features. Thus a 3D point cloud representing scene depth is computed.

Overlay of Delaunay triangles provides a depth surface by interpolating between detected features.

A new approach, described herein, provides an alternative to such methods. For inference of the scene depth map we have adopted a recursive framework that seeks to minimise an error defined using the intensity in the measured images. The resulting gradient based algorithm provides an estimate of depth to every pixel.

The limitation in extracting scene depth information is in the ability to formulate and efficiently solve the inference problem. Significant progress has been made in this direction through the development of high dimensional Bayesian inference approaches and research in the image processing related disciplines of optical flow, structure from motion and scene reconstruction (see the references contained in [3]). In particular, a Bayesian scheme, Hierarchical Inference via Nested Training Samples (HINTS), described in [4] provides for efficient inference of camera motion.

The remainder of this section outlines the military application of methods for 3D reconstruction. The following section describes the technical approach taken. The results section provides an illustration of the technique, using airborne imagery, to construct a 3D surface for a scene with substantial depth variation.

Military relevance

The technology is applicable to a range of airborne imaging systems (surveillance Electro Optic (EO) turrets, missiles, UAVs or fast jets).

The depth map could provide 3D scene mapping, via construction of a 3D mosaic, and hence supplement Digital Terrain Map data. Modelling of 3D scene content would support detection and tracking in urban environments, and also aid scene matching – e.g. where an urban scene is viewed from differing viewpoints. Further applications could be in collision avoidance, particularly for low flying airborne platforms that provide good triangulation of scene content (including lower cost UAVs), or for helicopter landing.

Technical approach

We have formulated 3D scene depth estimation as an inference problem where the unknown state variables are :

- the scene, comprising two aligned grids
    - depth (i.e. distance of scene points from sensor),
    - intensity;
- the sensor motion (incremental translation and pose) between successive frames.

The states detailed above provide two of the three core components of the geometric model identified in the introduction. The third core component of the algorithm is *dynamic prediction* – i.e. a method to predict the depth map (and associated scene intensities) so that they are projected according to the sensor's motion from one frame to the next. For this we have used a rigid body model that combines the depth map in one frame with the incremental sensor motion between frames so as to predict the depth map at the next frame.

We have developed an efficient framework for solving the inference problem by iterating between an efficient Bayesian alignment process that estimates the geometric parameters (incremental sensor translation and pose), and a scene update stage that computes the scene depth and intensity (**figure 1**).

The choice to decompose the inference in this fashion allows application of methods appropriate to the dimensionality of the problem:

- Scene points ~10^5 (a grid of intensities, and a corresponding grid of distance from sensor to scene);

- Geometric (rigid body motion) and photometric (intensity gain and offset) transformation parameters between successive frames, ~10 times the number of frames. .

The iterative inference scheme is depicted in **figure 1**. On receipt of a new image, the first stage is identification of the 6 parameters in the geometric transform (incremental sensor translation and pose) to bring the incoming image into **alignment** with the scene estimate.

The second stage is an **information update** where the new sensor image is used to update the estimate of scene depth and scene intensity at every pixel.

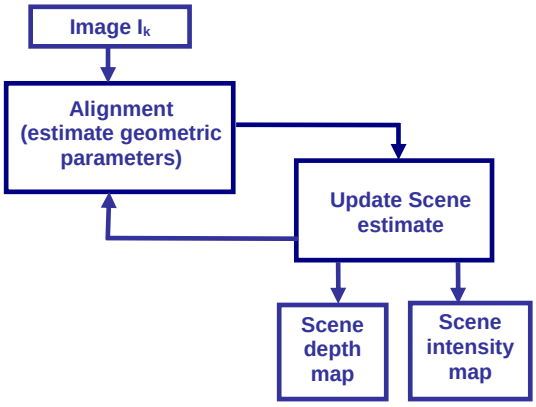

Figure 1 : *Algorithm structure*

Alignment

There are two main components of the alignment problem:

- the *inference scheme* used to identify the parameters in the geometric model;
- the *geometric model* used model depth dependent motion of scene, ie to map scene points from frame *k* to frame *k+1*.

Inference scheme: An efficient Bayesian scheme, Hierarchical Inference via Nested Training Samples (HINTS), [4] is used to infer the six parameters of the geometric transform. The HINTS scheme is ideally suited to the alignment problem because it can exploit the additive structure of the error function to obtain a very rapid search of the multi-dimensional parameter space. (It performs repeated evaluations of single pixel frame-to-frame mapping, under proposed rigid body parameters). The scheme is detailed in [4] and uses a hierarchical structure (**figure 2**) where changes in state at the top (root) level are the result of sequences of Metropolis-Hastings steps taken at the lower levels.

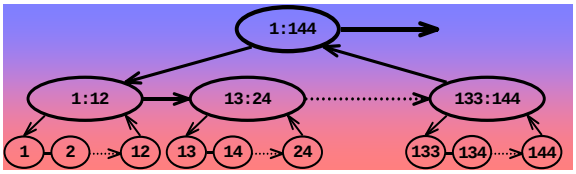

Figure 2 : *HINTS sampling architecture*

Geometric model: An important part of the framework is the geometric model that accounts for the motion between frames in incoming imagery. Its function is to map scene points from frame *k* to frame *k*+1. The single 'rigid body' model that we have adopted assumes that the whole scene moves as one rigid body. It is applicable to a three dimensional scene where substantial depth variation causes parallax within the imagery.

The rigid body model uses the 3D representation of the scene, stored as a scene 'depth map' in co-ordinates aligned with the sensor, and a six dimensional geometric transform describing a 3D translation and 3D rotation of the sensor (which is inferred via HINTS).

Alignment procedure: Suppose that a scene estimate (intensity and depth map) is available at step *k-1*, and is defined in co-ordinates aligned with the sensor. On receipt of an image, *k*, the six rigid body parameters describing the frame to frame motion must be identified. Inference of the rigid body parameters proceeds via HINTS. The scene estimate (of both intensity and depth to every pixel) is then projected into alignment with the inbound image at step *k* using the identified transform and the rigid body model. Hence the predicted scene estimate is aligned with the most recent image, and the scene information update (of

both depth and intensity) can be performed as described in the next section.

Information update

The second processing stage in the recursive processing framework (**figure 1**) updates the scene estimate to take account of the new image.

A generative model, *H*, expresses the optical effects of the sensor (i.e. its point spread function) and is used to produce a predicted sensor image, *g_est*, from the estimate of scene intensity, *X*.

$$g\_est = HX$$

Both the intensity and depth inference are achieved via minimisation of an intensity based reconstruction error, *E*.

$$E = \sum_i [\ g\_est(i,j)\ - g(i,j)\ ]^{\,2}$$

The error is a function of the measured image, *g(i,j)*, and the predicted sensor image, *g_est(i,j)*. The latter is derived from the estimated states – scene intensity, scene depth, and sensor motion. Hence errors in the predicted sensor image derive from errors in the depth map and previous scene intensity estimate. These errors, together with their gradients (dE/dX and dE/dQ) are used to compute refinements to the estimates of intensity and depth as an iterative solution for *X* and *Q*.

The differentiability of the generative model allows an error gradient to be calculated for individual scene points. An iterative solution for the scene intensity estimate, *X*, includes a regularisation term for local smoothness, ψ_X.

$$X_{k+1} = X_k + \tau\ dE/dX + \alpha\, d\,\Psi_X\,/dX$$

The update of reciprocal depth estimate, *Q*, proceeds in a similar fashion :

$$Q_{k+1} = Q_k + \tau_q\ dE/dQ + \beta\, d\,\Psi_q/dQ$$

The *whole scene* gradient update scheme has been implemented using whole image operations (ie convolutions) so that the update scheme could be carried out quickly by parallel architectures such as a FPGA.

Algorithm summary

Key elements of the processing can be summarised as follows. The algorithm:

- estimates camera motion (incremental translation and pose) between frames;
- estimates the scene content – both its intensity and the 3D structure (distance from sensor to scene in every pixel);
- uses a rigid body model to perform depth dependent mapping of scene content from frame-to-frame, according to camera motion;
- combines a generative model of the image formation process with the predicted scene content to define a reconstruction error, *E*, against each received frame;
- uses gradient error minimisation to estimate distance to every pixel in the scene.

Results

Airborne imagery sequences from an MX series turret have been collected in trials funded by the Weapons Technology Centre and Dstl. This section presents results showing inference of depth information for a scene comprising substantial depth variation, and estimated sensor motion.

Depth map for a 3D scene

A region of interest on the boresight of the airborne sensor's field of view (FOV) has been processed. For inference of the six rigid body motion parameters, a reasonably large angular FOV is required (several tens of degrees). Also, since the depth information is derived from passive imagery, and because the images depict the angular position of the scene relative to the sensor, there must be sufficient camera translation for triangulation to make the scene depth information observable.

Results showing the depth map are presented in **figure 3** for a sequence of imagery where Lulworth cove has been

over flown. This imagery has been collected in trials funded by the Weapons Technology Centre and Dstl.

Figure 3 (a), (b) and (c) shows three example sensor images, separated by approx 120 frames between each. **Figure 3 (d)** shows the reciprocal depth map output after 240 frames. It should be noted however that a reasonable depth map has been computed after ~60 frames. This result shows successful inference of scene depth information.

Figure 4 shows the estimated camera motion. Inference of sensor motion is via the HINTS scheme and provides, for each received image, a six dimensional geometric transform describing an incremental 3D translation and a 3D rotation of the sensor. The incremental translations and rotations have been integrated (in sensor axes) to provide the plots in **Figure 4** (a) and (b) respectively.

Discussion

The rate at which the depth map is updated affects the speed and stability of convergence. For a gradient error step size, τ_q, of unity, no prior depth map, or sensor motion, is required. The resulting depth map is however less smooth than for smaller value of τ_q used in the illustration. Regularisation of the depth map is via a local spatial smoothing derived from bilateral total variation [5], with the scale of the smoothing image chosen to maximize the probability of the depth map.

For the obliquely viewed scene presented above, rather than initialise the depth map at iso-range, a crude ramp was applied over the vertical extent of the imagery. Without such a ramp the estimate of the rigid body parameters can become degraded since there can be a tendency to interpret, as image roll, the progression to the left of the foreground scene content (at the bottom of the image) at a greater rate than the scene closer to the horizon (at the top of the image). The cause of this motion is in fact the sensor's translation in a direction tangential to the image, combined with the closer foreground. As an alternative to initialising scene depth with a ramp, prior information on the sensor's translation can be used.

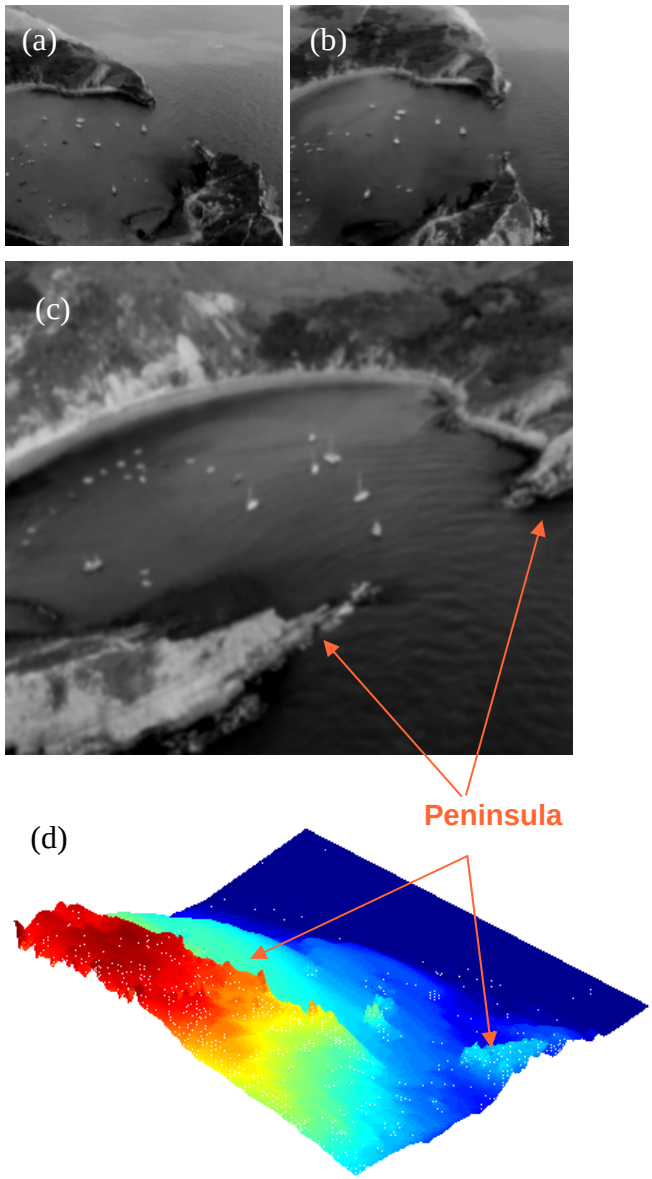

Figure 3 : *(a, b, c) visible band sensor images (309 by 269 pixels covering 27 by 23 degrees); (d) mesh plot of estimated depth map.*

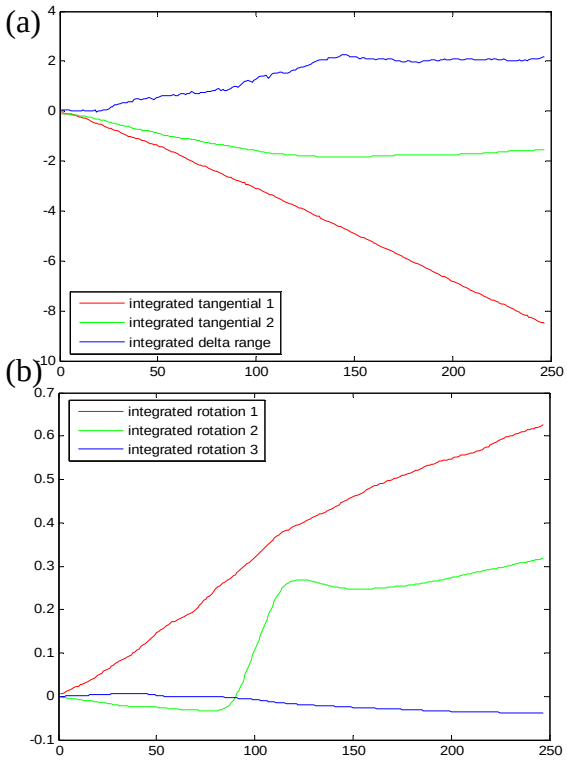

Figure 4 : Sensor motion since start of sequence; *(a) integrated translation; (b) integrated rotation.*

Conclusions and Future work

We have developed a new algorithm for scene depth estimation that has broad applicability to airborne imagery from passive sensor systems for a variety of tasks including:

- 3D scene mapping and mosaicing – as a supplement to Digital Terrain Map data;
- detection and tracking, particularly of targets viewed within urban environments comprising 3D structure, and which may be viewed from differing viewpoints;
- collision avoidance, particularly for low flying airborne platforms that provide good triangulation of scene content (lower cost UAVs), or helicopter landing during brownout.

Central to the above applications would be the projection of the depth map, which is currently maintained in sensor axes, into a common co-ordinate frame on the ground, to enable integration into a 3D mosaic.

The algorithm provides a framework that could be adapted to integrate and mosaic the 3D data from sensors that measure range, such as LIDAR.

Algorithm development work should consider refinements to the scheme for depth estimation by experimenting with the speed of convergence of the depth map, its resulting scale, and interdependence upon the expected magnitude of the camera motion. Estimated integrated camera motion should be compared with truth, where this is available, for example in data collected under the HydraVision 2 trial conducted in 2010.

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.

Imagery of Lulworth cove was collected in trials funded by the Weapons Technology Centre and Dstl.

References

1. Harris C “Stable Scene Surfaces from Computer Vision”, EMRS DTC conference, 2007.
2. Hartley R and Zisserman A, “Multiple View Geometry in Computer Vision”, ISBN 0-5215-4051-8, March 2004.
3. Rollason, M P and Gardner A P, “Temporal Resolution Enhancement from Motion”, *EMRS DTC Annual Technical Conference*, 2008.
4. Strens M J A et al.. “Markov Chain Monte Carlo Sampling using Direct Search Optimization”, Proc. 19th Int. Conf. on Machine Learning, 2002.
5. Farsiu S. et al “Fast and robust multiframe super resolution”, IEEE Trans. Image Proc, Vol.13, 2004.