International Global Navigation Satellite Systems Society
IGNSS Symposium 2007

The University of New South Wales, Sydney, Australia
4 – 6 December, 2007

'Point and Click' Target Location from Georeferenced Video Streams

Steven Mills
Geospatial Research Centre, New Zealand
Tel: +64 3 3643835, Fax: +64 3 3643880, steven.mills@grcnz.com

ABSTRACT

This paper presents research conducted at the Geospatial Research Centre, New Zealand, into developing the ability to locate ground targets through a 'Point and Click' interface from UAV-based imagery. The motivation for this research is to provide a low-cost method for easily and accurately locating ground features of interest from a real-time aerial survey. Potential applications of such a system include search and rescue, fire fighting, and environmental monitoring. The approach described here uses a minimal set of sensors consisting of a GPS unit and video camera. Features are tracked through an image sequence, and the ability of the tracker to operate successfully and to locate and follow features in difficult image sequences is demonstrated. The tracked features are then used to establish a geometric relationship between a pair of images taken some distance apart. This determines the relative orientation of the camera at the two time instants at which the images were captured. Absolute orientation is made from assumptions based on the UAV application domain. This orientation information, along with the GPS locations of the cameras determines the 3D locations relating to tracked image features. Interpolation between these features provides the basis for the 'Point and Click' target localisation.

KEYWORDS: Image processing, UAVs, interactive systems, navigation sensors, sensor integration.

1. INTRODUCTION

At the Geospatial Research Centre (GRC), New Zealand, we are exploring the problem of providing 'Point and Click' target localisation from a UAV. Given an image sequence from a UAV-mounted video camera the user should be able to click on any point in the image and get an estimate of its geographic location (GPS co-ordinates, grid co-ordinates in a local frame, etc). A solution to this problem would allow a ground-based operator to direct resources to geographic locations on the basis of real-time aerial survey or reconnaissance. Potential applications include directing search and rescue teams to people in need of assistance, directing fire crews to hot spots detected in thermal imagery, and locating points or areas of interest for agricultural and environmental applications.

While a number of sensors could prove useful for this problem here we consider a minimal set. A video camera is required to provide imagery, and a GPS or other positioning system is needed to provide a link to a geographic reference frame. The input to this process, then, is a sequence of images and an estimate of the camera position in some absolute reference frame. We assume for now that no other information is available, and we show that under a set of reasonable assumptions no other information is required.

The general approach taken is to estimate the 3D locations of features detected in the image. This provides a sparse set of known image to world location correspondences. Interpolation between these values can provide an estimate of the 3D location that corresponds to any given image point, and thus point and click target localisation. We are interested in solutions using low-cost sensors, and the UAV platform imposes weight and power issues. These constraints make the combination of GPS with a video camera an attractive option, since the required hardware (sensors, transmitters, and base-station receivers) costs less than $1000.

The remainder of this paper is organised as follows: Section 2 gives an outline of the proposed techniques and processes that are applied to the input data in order to estimate the location of features in the scene; Section 3 analyses the sensitivity of these techniques to violations of various assumptions that have been made; and finally Section 4 outlines our future plans and provides summary and concluding remarks.

2. DETERMINING CAMERA LOCATION AND ORIENTATION

In order to determine the location of features on the ground from the images, we need to determine the location and orientation of the camera at two distinct times, along with the image locations of corresponding features. Given this information the ground locations of corresponded features may be computed through triangulation. Correspondences are established through feature tracking, as described in Section 2.1, the camera locations are measured directly from GPS, and a procedure for determining the camera orientations is described in Section 2.2,

2.1 Feature Tracking

In order to establish a correspondence between images at different locations, features are tracked through a video sequence. Tracking is a somewhat simpler problem than direct correspondence because the distance that features move between frames is typically small. This reduces the search for matching features to a small local neighbourhood around the

previous feature location. UAV images, however, can offer particularly challenging tracking problems. Often images are affected by vibration, the low altitude means that motion blur can be an issue, and many of the environments in which we currently test our vehicles are not visually feature rich. Frames from a sample image sequence are shown in Figure 1. These images show farmland near one of our UAV test areas at Swannanoa, North Canterbury.

Figure 1: Sample frames (10 frames apart) from a UAV video sequence.

For tracking features on the ground we use the Kande-Lucas-Tomasi (KLT) feature tracker (Lucas and Kanade, 1981; Tomasi and Kanade, 1991; Shi and Tomasi, 1993) as implemented in OpenCV (Bouguet; OpenCV). At each frame existing features are tracked and new features found to keep the total number of features under consideration constant (100 in our examples). In feature detection and tracking one of the problems is determining what a 'feature' is. The KLT tracker defines a feature as any point that is likely to be tracked well, making it well suited to general purpose tracking.

Features in the KLT tracker are defined as 'good features to track' (Shi and Tomasi, 1993). These features are identified via an eigenvector analysis of the image gradient. Each point has two eigenvalues associated with it, which describe the degree to which the feature is localised. Features with two low eigenvalues are not well localised at all; those with one large and one small eigenvalue are localised in only one direction (for example, edge or linear features); and those with two high eigenvalues are well located in both directions (for example, corner or point features). On this basis, 'features' are defined in terms of the smallest eigenvalue, and local maxima of the second eigenvalue are taken as feature locations. Features detected in the images from Figure 1 are shown in Figure 2. Those shown in red are features that were visible in the last frame, while those in green are new features that have been added to replace those that have left the field of view.

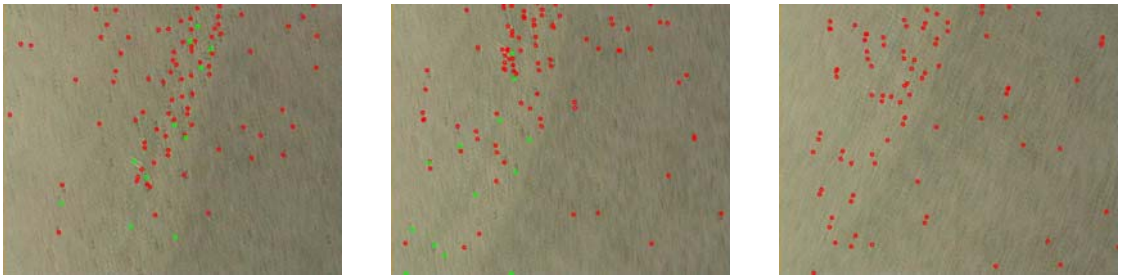

Figure 2: Features detected in the images from Figure 1.

Given a set of features in one image, the KLT tracker looks for corresponding features in the next image from the sequence. This is done using a Newton-Raphson method to minimise the difference between image windows around the two feature locations (Lucas and Kanade, 1981; Tomasi and Kanade, 1991). Image pyramids are used to speed computation and to allow for larger motion of features between frames (Bouguet). The results of tracking features

through the images shown in Figures 1 and 2 are shown in Figure 3. In this figure the red circles are the current image locations of features, and the red lines show the path that the feature has travelled through the image since first being detected.

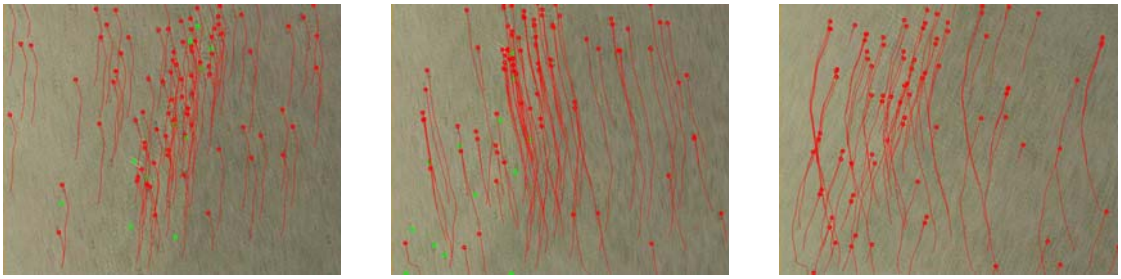

Figure 3: Paths formed from tracked features in the images from Figures 1 and 2.

As Figure 3 illustrates, the KLT feature tracker provides a mechanism to establish sets of corresponding features between images across a video sequence, even in difficult image sequences.

2.2 Visual Determination of Relative Orientation

The next step is to determine the relative orientation of the cameras. We assume that as input we have corresponded image feature locations at two times, and an estimate of the camera location at each time from GPS. At this stage we assume that the first camera is at the origin, and is aligned with the optical axis of the camera pointing down and the *x*-axis aligned with the normal direction of flight. This defines a camera-centred co-ordinate system. C_C, within the global (or world) co-ordinate frame, C_W that we are using, as illustrated in Figure 4.

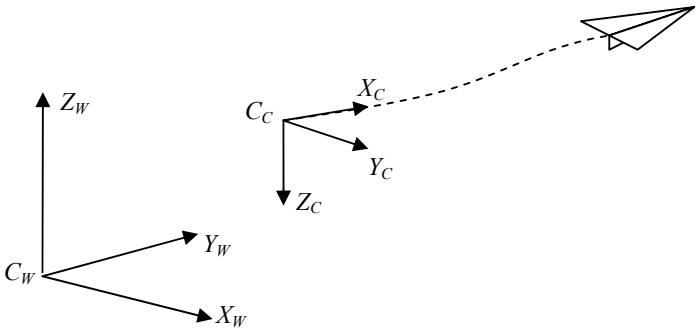

Figure 4: Camera (C_C) and world (C_W) co-ordinate frames.

The relationship between the image features in the two views can now be expressed in terms of the relationship between the two cameras and the three dimensional locations of the feature points:

$$\lambda_1 \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = KRX, \quad \lambda_2 \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = KRQX - RQt,$$

where R is the rotation from C_W to C_C, t is the translation (in world co-ordinates) between the cameras, Q is the rotation between the two cameras, K is the camera calibration matrix, X is the 3D location of a feature in the world, λ_1 and λ_2 are scalars, and (x_1, y_1) and (x_2, y_2) are the locations of the feature in the two images.

The vector t is known from the GPS positions of the two cameras, and the next step is to estimate R. This is also estimated from the GPS path. As illustrated in Figure 4, two of the axes are determined from the fact that the camera x-axis is aligned with the direction of travel, and the z-axis is oriented as near as possible to be vertically down. This means that R may be estimated by assuming that the initial camera roll is 0, and its pitch and heading are chosen to align the x-axis with the vector t.

Given an estimate of the overall orientation of the cameras, the next step is to estimate the relative orientation between the two cameras. The essential matrix (Longuet-Higgins, 1981) relates two views from calibrated cameras by the equation

$$u_2^T E u_1 = u_2^T Q [R^{-1} t]_\times u_1 = 0,$$

where E is the essential matrix, $u_i = R^{-1} K^{-1} x_i$ where x_i is a feature's image co-ordinates in the i^{th} image and K is the camera calibration matrix, Q and t are the rotation and translation between the two cameras, and $[v]_\times$ is a matrix such that $[v]_\times w = v \times w$. All of these terms are known except for Q. Since the UAV is unlikely to have changed orientation significantly over short periods of stabilised flight, we make a small angle approximation to Q:

$$Q \approx \begin{bmatrix} 1 & -\theta_z & \theta_y \\ \theta_z & 1 & -\theta_x \\ -\theta_y & \theta_x & 1 \end{bmatrix}.$$

Substituting this into the essential matrix constraint gives a linear constraint in (θ_x, θ_y, θ_z) for each matched pair of image features. Given a set of image correspondences, these values can be estimated through standard linear least squares or with a RANSAC process (Fischler and Bolles, 1981) in order to eliminate gross outliers.

We now have estimates of all of the camera parameters, although several assumptions have been made to simplify the computation. These initial estimates may now be refined through bundle adjustment (Triggs *et al*., 2000). Although bundle adjustment has traditionally be considered expensive, careful exploitation of sparse matrix structure and increasing computing power means that limited bundle adjustment can be applied even in real time systems on commodity hardware (Engels *et al.*, 2006). The bundle adjustment is based around the Jacobian, J, of the relationship between the camera parameters and the measurements (image feature and GPS locations). J has the sparse structure shown in Figure 5.

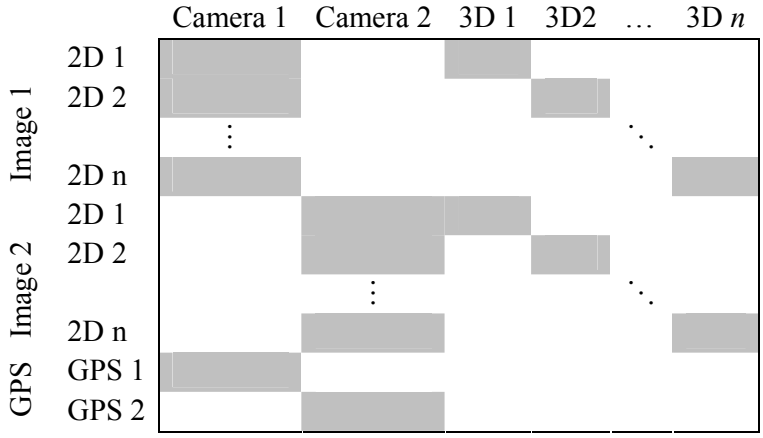

Figure 5: Sparse structure of the Jacobian for parameter refinement bundle adjustment.

2.3 Absolute Orientation Estimation

The methods described in Section 2.2 determine the orientation of the cameras relative to one another. The assumption that one of the camera's axes is aligned with the direction of travel also constrains the absolute heading and pitch of the platform. The absolute roll, however, is not estimated. Estimating this value without additional sensors (such as an inertial sensor to determine the direction of gravity) requires an assumption about the scene. We make the assumption that the ground is roughly level, so that it is, on average, neither higher nor lower to the left or right of the UAV platform's line of flight. While this assumption is not valid in general, it is true for the environments in which initial tests will be conducted. In the longer term we will be investigating the use of other sensors to provide absolute orientation information, or the use of prior terrain models to relax this assumption.

From the earlier analyses we have a set of reconstructed 3D features in the camera-centred co-ordinate frame. We fit a plane to these points through standard least squares fitting and then determine the slope of this plane in the direction perpendicular to the UAV's line of flight. This angle is then our estimate of the UAV's absolute roll.

3. ERROR ANALYSIS

In this Section we analyse the assumptions that were made in the derivation of the techniques in the previous section, and the effects of reasonable violations of these assumptions. The same test scene is used for all of the experiments, so that comparisons can be drawn between them. The world consists of a regular grid of features at 5m intervals on the x-y plane in world co-ordinates. The simulated UAV flies at an altitude of 250m at a heading of 30^{o} relative to the world x-axis for a distance of 40m. The image is taken from a virtual camera having a 40^{o} field of view and, for purposes of representing localisation error, a nominal resolution of 640×480 pixels. Error is measured as the mean difference between the predicted and true world co-ordinates of the target features.

3.1 Errors in Input Data

The analysis presented in Section 2 assumes that the input data is precise and accurate. In reality the image feature co-ordinates and GPS observations will be noisy and therefore uncertain. The use of least squares techniques can mitigate this, and Figure 6 shows the effects of varying levels of Gaussian random noise in the input values on the computed locations of feature points.

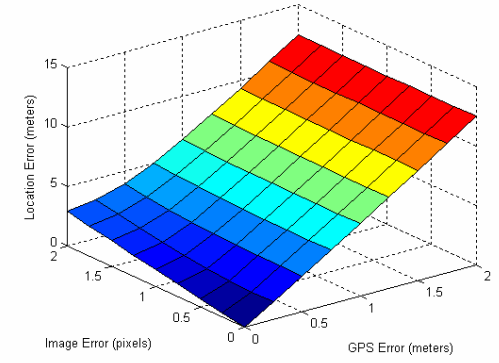

Figure 6: Effects of errors in input data

Errors on the order of 1-2 pixels have a small effect on the computed image locations. The main effect, however, is the error in the GPS estimates of the locations of the cameras. This effectively changes the baseline distance between the two the cameras, which is used to determine the scale of the scene. If this distance (and therefore the scale) is overestimated then the features' reconstructed locations will be too distant from the cameras and vice versa.

3.2 Errors in Orientation Assumptions

The main assumptions used in the derivation of the techniques of Section 2 are based around the orientation of the camera. It is assumed that the camera is initially aligned with the direction of travel and is downward pointing. Furthermore it is assumed that the UAV's orientation is stable over time, so that a small angle approximation to the change in orientation between the two views is valid. We consider variations in the heading, pitch, and roll of the starting camera location; in the change in these values between the two views; and the effect they have on target localisation. Errors are shown for the initial estimate, after 25 and 50 iterations of bundle adjustment, and after iteration to convergence.

The results in errors in the initial camera orientation are shown in Figure 7. In this case 'error' in the angles means misalignment with the direction of motion of the UAV. Note that even quite small changes in the camera orientation can lead to large errors in the initial feature reconstruction estimates. Bundle adjustment, however, improves these estimates, and for errors in the range of $\pm 5^{o}$ brings the errors to within a few meters.

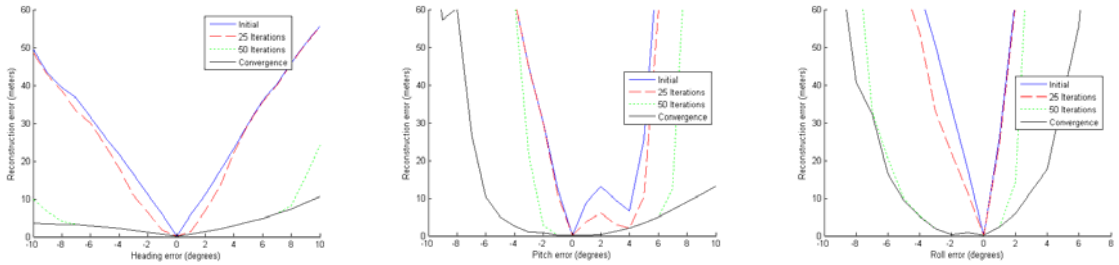

Figure 7: The effects of errors in initial orientation.

The effects of errors in the assumption that there is little change in orientation between the two views are shown in Figure 8. Again, quite small changes can lead to significant errors in the estimated locations of features in the world, but bundle adjustment reduces these to an acceptable level.

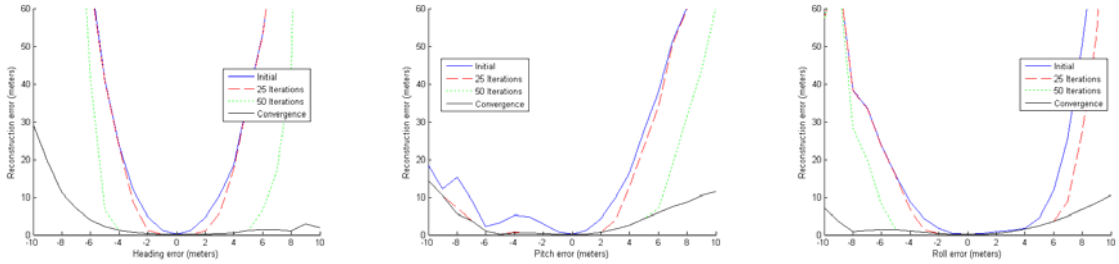

Figure 8: The effects of errors in relative orientation.

4. CONCLUSIONS AND FUTURE PLANS

The methods described in this paper provide a set of image feature locations and their locations in the world co-ordinate frame. Interpolating between these known points provides one potential solution to the 'Point and Click' target location problem that is being researched at the Geospatial Research Centre. Over the coming months we have several flights of our UAV platforms planned to validate these techniques in practical application. A range of GRC UAV platforms will be used for testing, including Kuruwhengi (a modular airframe for initial testing), Korimako (a very low cost platform), and Kakariki (which can lift greater payloads). These tests will include the use of low-cost inertial sensors (such as Crossbow, Microstrain, and Analog Devices IMUs) which will be used to provide an independent estimate of the platform orientation. This will be used to evaluate the visual estimation of orientation determined using the processes outlined in Section 2.

Inertial sensors will also be explored as an additional constraint to the system. Although a solution without inertial sensors may be desirable from a cost perspective, the additional information may be required for situations where greater accuracy is required. Another avenue for future research is the integration of information across more than two images. As shown in Figure 6 there is a direct relationship between the error in the estimated baseline (from GPS) and the accuracy of the scene reconstruction. A sequence of n images, where each image has an associated location estimate potentially gives $n(n\text{-}1)/2$ baseline estimates, one from each pair of images. While these are not independent, there is the possibility of reducing the overall error through least squares or similar approaches.

REFERENCES

Bouguet, J-Y, Pyramidal implementation of the Lucas Kanade feature tracker: Description of the algorithm, http://robots.stanford.edu/cs223b04/algo_tracking.pdf, last accessed 10/09/2007

Engels C, Stewénius H, Nistér D (2006) Bundle adjustment rules, *Photometric Computer Vision*

Fischler MA, Bolles RC (1981) Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM* 24: 381-395

FMA Direct, FMA Direct – High Quality Equipment fpr the R/C Enthusiast, http://www.fmadirect.com/, last accessed 28/09/2007

Longuet-Higgins, HC (1981) A compute algorithm for reconstructing a scene from two projections, *Natur*, 293, 133-135

Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision, *International Conference on Artificial Intelligence*, 674-679

OpenCV Open Source Computer Vision Library (website), http://www.intel.com/technology/computing/opencv/index.htm, last accessed 10/09/07

Shi J, Tomasi C (1994) Good features to track, *IEEE Conference on Computer Vision and Pattern Recognition*, 593-600

Tomasi C, Kanade T (1991) Detection and tracking of point features, *Technical Report CMU-CS-91-132*, Carnegie Mellon University

Triggs B, McLauchlan P, Hartley R, Fitzgibbon A (2000) Bundle adjustment – A modern synthesis, in: Triggs, W, Zisserman A, Szeliski R (eds.), *Vision Algorithms: Theory and Practice*, Springer Verlag, Berlin, 298-375