Integrating GNSS, IMU, and Imagery for Automatic Orthomosaic Generation

S. Mills, D. Park, C. Hide, K. Barnsdale and J. Pinchin *Geospatial Research Centre (NZ) Ltd*

BIOGRAPHY

Dr Steven Mills is a Senior Research Scientist with the Geospatial Research Centre (NZ) Ltd. He holds an honours degree and PhD in Computer Science from the University of Otago, and has 12 years or industrial and academic research experience in the fields of image processing and computer vision. Dr Mills' primary research is in the reconstruction of 3D structure and motion from multiple images.

Dr David Park is CEO of New Zealand based Geospatial Research Centre (NZ) Ltd. He completed his PhD at The University of Nottingham and worked in both industry and academia in the UK before moving to New Zealand in 2006.

Dr Chris Hide is a Senior Research Scientist at the Geospatial Research Centre based in Christchurch, New Zealand. He has a degree in Mathematics and Topographic Science from the University of Wales, Swansea and a PhD in engineering surveying from the IESSG, University of Nottingham, UK. He worked as a Research Fellow for 3 years at the IESSG before starting a secondment at the Geospatial Research Centre in 2006.

Kelvin Barnsdale is a Senior Research Engineer with the Geospatial Research Centre (NZ) Ltd., and is engineering team leader for the Airborne Mapping Package. His previous electronics design experience includes University College London Space Science Lab and Hughes flight simulation division, and has been involved in the GPS industry for the last 12 years, including Navman as their senior RF design engineer. His primary focus is the design of airborne navigation systems and remote sensing thermography.

James Pinchin is a PhD student in the Mechatronics Research Group at the University of Canterbury, New Zealand. His research topic is low cost attitude determination using GNSS. He has a BSc in Physics from the University of Bath, UK and an MSc in Satellite Positioning Technology from the University of Nottingham, UK.

ABSTRACT

The use of orthomosaic images from aerial or satellite data are increasingly common. While current acquisition methods are cost-effective on a national or regional scale, local scale imagery is prohibitively expensive for many target applications. In this paper we present a combined hardware and software solution, developed at the Geospatial Research Centre, which aims to reduce the cost of acquiring and processing imagery and related data in order to produce orthomosaics in a cost-effective manner on a small, local scale.

The hardware component consists of a combined GNSS and inertial solution for determining the position and orientation of a sensor, typically a consumer-grade camera such as a digital SLR. The combination of imagery and navigation metadata allows images to be directly geo-referenced by projecting them on to readily available surface models. Refinements to this initial processing are also presented, which account for boresight and lens calibration error; automatically establishing a correspondence between image features for bundle adjustment; and reducing the visual appearance of any residual misalignments in the final mosaic. The use of commodity sensors and automated processing is an important step in reducing the cost of image acquisition and orthomosaic generation.

The methods described are illustrated using two sample sequences. The first is a set of visible images captured from a digital SLR, and the second a set of frames extracted from a thermal video sequence. These two sequences demonstrate the range of imagery that can be processed, which can support applications ranging from environmental monitoring and precision agriculture to urban planning and infrastructure maintenance.

1 INTRODUCTION

The production and use of orthomosaiced imagery has become pervasive in recent years. From world-wide satellite coverage, through local area aerial imagery, to consumer services such as those offered by Google and Microsoft, a wide variety of people are becoming used to the ready availability of such imagery. The cost of image acquisition however, is still relatively high. Satellites are cost-effective on the large (nation-wide) scale, and traditional aerial photogrammetry on the medium (regional or city-wide) scale. The cost of acquisition on a local (property, farm, or site) scale is too high, however, to justify its use for many emerging application areas. Applications such as precision agriculture, infrastructure maintenance, and environmental monitoring, could benefit from timely capture of

imagery on the local scale in a cost-effective manner.

With this situation in mind, the Geospatial Research Centre (GRC) has developed hardware and software technologies for aerial image capture and processing with the following aims:

- The system should be cost-effective on a small scale.
- The system should support high spatial and temporal capture resolution.
- The hardware should be configurable to meet a range of accuracy and budget requirements.
- Hardware and operating costs should be minimised.
- The processing chain should be automated to reduce costs and turn-around time.

This paper focusses on the last of these objectives, and details an automated process for orthomosaic generation from aerial imagery. Section 2 provides an overview of the data capture process and hardware. Section 3 describes the steps required for processing the data from the navigation sensors. Section 4 describes the basic process of geo-referencing images, with more advanced processing presented in Sections 5, 6, and 7 which cover boresight and lens calibration, feature detection and correspondence, and bundle adjustment respectively. Section 8 describes various approaches to improving the appearance of the final mosaic, and is followed by closing remarks in Section 9.

2 SYSTEM HARDWARE AND DATA CAPTURE

Before presenting details of the image processing chain, a brief overview of the hardware and data capture process is given. The GRC's Aerial Mapping Package (AMP) hardware consists of three main components, which are illustrated in Figure 1. The first of these, the sensor module, contains an inertial measurement unit (IMU) and camera mechanically coupled together and is mounted on the aircraft with a view of the ground below. The second component, which may be mounted internally or externally, contains a GNSS receiver and data logger, along with supporting control and power electronics. The internal lithium battery can support operations for up to 6 hours. The final component is a pilot navigation aid, which shows the target area and flight lines for image capture, and indicates the position of each picture taken.

The system is designed to be used with a variety of IMUs, GNSS receivers, and cameras. It can be attached to a wide range of aircraft, and has been successfully deployed on microlights and several light aeroplanes. This flexibility means that the hardware can be adjusted to meet accuracy, weight, and cost trade-offs, a wide variety of imagery can be captured, and it can quickly and easily be deployed in a range of circumstances. The system can operate using any of the NovAtel range of OEMV series GNSS receivers including the lower cost, single frequency OEMV-1 receiver.

Figure 1. Aerial mapping hardware including camera, navigation equipment (GNSS + IMU) and pilot guidance.

The system can also interface with a range of IMUs including, but not limited to, the iMAR iIMU-FSAS, Honeywell HG1700 and Crossbow IMU family. Again, lower cost IMUs such as the Crossbow IMU440CA have been tested with the system which is of significant interest for developing a cost-effective system where, typically, the IMU is the highest value component. Digital SLR cameras provide an excellent low cost option for acquiring high resolution images in the visible spectrum. Time synchronisation is achieved using the hot shoe on the digital SLR, hence the majority of consumer to professional grade digital SLR cameras can directly interface with the developed system. Figure 1 illustrates one such configuration comprising of Canon 400D digital SLR, Honeywell HG1700-AG62 and NovAtel OEMV-2 GPS receiver.

As well as the images and navigation data from the AMP, the system requires a digital surface (or elevation) model (DSM) as input. This does not need to be of very high resolution or accuracy, but is used as a guide for directly geo-referencing the images. Ground control points (GCPs), or other land survey data is not used because it requires a ground crew to acquire and user-interaction to match GCPs to the imagery. These factors mean that the use of GCPs increases cost and decreases automation. We note also that the requirement for a DSM is no longer a restriction. Such data sets are increasingly common, including the 30m resolution ASTER data set [8] which has near global coverage and is freely available.

Throughout the remainder of this paper, two examples are used to illustrate the techniques described. The first is a set of 90 high-resolution visible images captured from a consumer digital SLR (a Canon 400D with a 28mm lens). These images are 10 megapixels (3888×2592 pixels) and an exposure was made every 2 seconds, and show a subur-

ban area in northern Christchurch, New Zealand. The second is a set of 347 low-resolution frames captured from a thermal video camera over Rangiora, in North Canterbury, New Zealand. This camera produces 320×240 pixel video frames at 10Hz, although not all of the frame contains image data due to a border and super-imposed logo. Sample frames from these two sequences are shown in Figures 2 and 3. Both datasets were collected using a dual frequency NovAtel OEMV series GPS receiver and iMAR Navigation iIMU-FSAS IMU. A Trimble NetR5 reference station was located within 30 kilometres of the survey area for both examples. The IMU and cameras were mounted to the aircraft on a rigid plate.

These two sequences illustrate the range of imagery that can be processed — high and low spatial resolution; high (video) and low (still) frame rate; visible to thermal spectra; and with a high (SLR) to low (video with auto-gain) level of control over the image capture settings. This range of imagery means that the equipment can be easily modified to suit a variety of applications. High resolution visible imagery is suitable for tasks ranging from infrastructure maintenance to disaster evaluation, while thermal imagery can be used for environmental monitoring or search and rescue. Other sensors can also be used such as multi-spectral sensors for land cover analysis or precision agriculture.

Figure 2. Sample frame from the visible image sequence (top) with detail of a small area (bottom) at full resolution.

3 NAVIGATION DATA PROCESSING

The first step in the processing chain after collecting the data is to process the navigation data. The navigation data consists of raw GNSS measurements from the receiver located on the aircraft; high-rate (typically 100-200Hz) time-stamped rotations and accelerations from the IMU; and optionally, raw GNSS measurements from a reference receiver located near to the survey area. Navigation data is processed using the GRC's POINT (Position Orientation INTegration) integration software. POINT provides algorithms for processing both the GNSS and IMU data using loose or tight integration. In both instances, forward and reverse Kalman filters are used to integrate the GNSS and IMU using an 18 state Kalman filter estimating position, attitude, velocity, gyro bias, accelerometer bias and gyro scale factor.

Figure 3. Sample frame from the thermal image sequence. Note that the border area and logo are masked from the final result.

Figure 4 shows the typical processing flow for tight coupled integration of the navigation data in POINT. The system hardware records individual files containing the IMU and GNSS data, and a file containing the time at which the images were taken. It necessary for the user to input the lever arm and installation angles of the IMU with respect to the platform in which it is installed (however this will be recorded by the system when it is permanently installed in an aircraft). The forward and reverse filters are run sequentially, with the alignment of the reverse filter being initialised using the position and attitude from the forward filter. A Kalman filter smoother is used to combine the forward and reverse navigation solutions. As the smoother is run, a navigation solution is generated for each image. So far, all navigation data is processed in the WGS84 coordinate system, however, most final orthomosaic products are required using a projection such as New Zealand Transverse Mercator (NZTM). The coordinate transformation step is calculated after the final navigation solution in WGS84 is generated by POINT. At this stage, both coordinates and attitudes (referenced to WGS84 local level) are recomputed in the mapping coordinate system referred to in the next section as the world-based co-ordinate frame.

This combination of forward and reverse processing is particularly beneficial for integrating lower cost IMUs which,

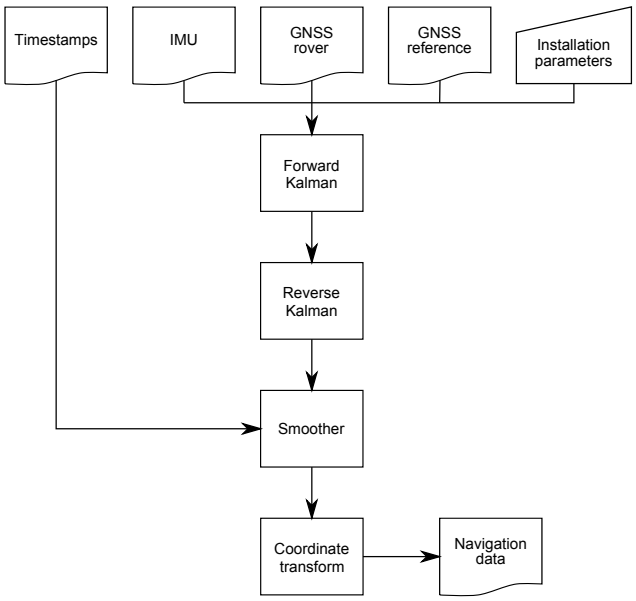

Figure 4. Flow diagram of POINT software for tight coupled integration

as previously mentioned, typically form the most expensive single item of the data collection system. Lower cost sensors such as the Crossbow IMU440CA typically result in much lower attitude accuracy than can be achieved using tactical grade sensors. The combination of forward and reverse filters mean that issues such as initial alignment can be reduced by using the navigation solution from the forward filter to initialise the reverse filter. Also, the drift of weakly observable parameters such as heading can be minimised since the combination of forward and reverse filters effectively minimises the period between which platform dynamics make such parameters observable.

4 DIRECT GEO-REFERENCING OF IMAGERY

The essential process in the automated orthomosaic generation process presented here is direct geo-referencing of individual frames. This is the process by which a transformation is determined from the 2D image co-ordinates to 3D co-ordinates in some geographic co-ordinate frame. This may be visualised as using the navigation data to determine the position and orientation of the camera. The image is then projected from the camera location on to the DSM, and in this manner images are transformed from their 2D co-ordinates to 3D world points.

Mathematically this may be represented as a series of transformations between co-ordinate frames. There are a number of co-ordinate frames of interest to us, which are:

- P, the 2D image- or picture-based co-ordinate frame.
- C, a camera-based co-ordinate frame, aligned to the camera's sensor and optical axis.
- N, a navigation-based co-ordinate frame, centred on and aligned to the IMU's axes.
- W, a world-based co-ordinate frame, fixed with reference to the earth's surface.

These are illustrated in Figure 5.

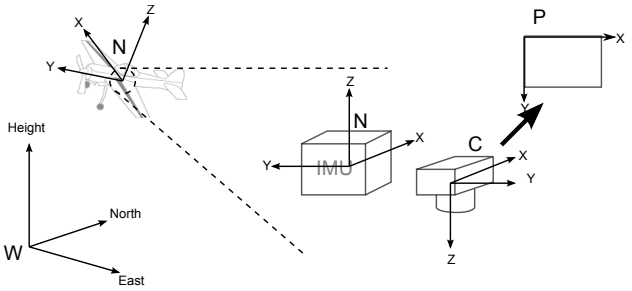

Figure 5. Co-ordinate frames used in direct geo-referencing. We seek a transform that takes us from the image frame, P, through the camera and navigation frames, C and N, to a geographically referenced world frame, W.

We now define a set of transformations between these co-ordinate frames. Denoting the transform from co-ordinate frame A to co-ordinate frame B as $T_{A\to B}$, we can express the transformation from a 3D world point to a 2D image point as

$$p^{(P)} = T_{W\to P}(p^{(W)}) = T_{C\to P}(T_{N\to C}(T_{W\to N}(p^{(W)}))), \quad (1)$$

where $p^{(A)}$ is a point in co-ordinate frame A. In most cases this is a 3D point, but the image point, $p^{(P)}$, is 2D and the final transformation, $T_{C\to P}$ is a projection from 3D to 2D.

The transformation from image to the world is somewhat complicated by the fact that the transform $T_{P\to C}$ takes a 2D point in the image and produces a 3D line in space. This reflects the case that the projective transform $T_{C\to P}$ loses information about the depth of a point in the camera frame. This gives us

$$l^{(W)} = T_{P\to W}(p^{(P)}) = T_{N\to W}(T_{C\to N}(T_{P\to C}(p^{(P)}))), \quad (2)$$

where $l^{(A)}$ is a line in co-ordinate frame A.

To directly geo-reference an image we may take its four corners, and find corresponding lines in W from Equation 2. These lines may be intersected with our DSM to give the corners of a quadrilateral in W. This new quadrilateral may be orthographically projected onto the ground plane by removing the Z-axis. We now have two corresponding quadrilaterals — the original (rectangular) image, and the corresponding region in W. The image may now be resampled into W, which effectively carries out the process of orthorectification.

Applying this process to a sequence of images brings them all into a common co-ordinate frame, W. The orthorectified images may then be overlaid on one another to form an orthomosaic.

In order to carry out this process we need to know the co-ordinate transforms $T_{N\to W}$, $T_{C\to N}$, and $T_{P\to C}$. An estimate of $T_{N\to W}$ comes from the processing of the GNSS

and IMU data; $T_{C\to N}$ is often well approximated by an identity transform if the camera and IMU are aligned; and $T_{P\to C}$ can be initialised using a pinhole camera model if the camera's focal length and sensor dimensions are known.

Results of basic orthomosaicing of the two sample data sets are shown in Figures 6 and 7. Note that while the images roughly align, there are many significant misalignments — most clearly visible in the north-south roads of the thermal data set — and clear boundaries between overlapping frames. These are due to inaccurate estimation of the various transforms and changes in the imaging parameters between frames. The following sections describe how these effects may be reduced. Section 5 describes how $T_{C\to N}$, and $T_{P\to C}$ may be refined from the available data; Section 6 and 7 present an approach to refining $T_{N\to W}$, and potentially the DSM values; and finally, Section 8 puts forward some methods by which the visible effects of any remaining joins and mismatches can be reduced.

Figure 6. Direct orthomosaicing of the visible data set, rendered at a resolution of 1m/pixel.

5 BORESIGHT AND LENS CALIBRATION

The first transforms that we consider are between the image and camera frames and between the camera and navigation frames. Since we use a single camera which is in a fixed, rigid configuration with the IMU, these elements are common across all of the images. The parameters that determine the transform between camera and image co-ordinates are referred to as internal or lens calibration, while those that relate the camera and navigation frames are called boresight calibration parameters.

The basic model for internal calibration is the pinhole or perspective camera model [7, Chapter 6]. This relates camera and image co-ordinates by a calibration matrix, K, defined by the equation

Figure 7. Direct orthomosaicing of the thermal data set, rendered at a resolution of 1m/pixel.

$$p^{(P)} = Kp^{(C)} \tag{3}$$

$$k\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f/p_x & 0 & x_0 \\ 0 & f/p_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{4}$$

where $[x, y]^T$ is the 2D image point, $[X, Y, Z]^T$ is the corresponding point in the camera frame, f is the camera focal length, p_x and p_y are the size of each sensor element, and $[x_0, y_0]^T$ is the principal point, where the optical axis of the camera intersects the image plane. This basic model may be augmented to account for non-linear distortions caused by the lens. Most commonly this is radial (barrel or pincushion) distortion, which is modelled as a polynomial function of distance from the principal point, but other distortions can also be considered [11].

The boresight calibration is represented as a rotation and translation, and so has six degrees of freedom (parameters). For low-resolution sensors, the translational component may be neglected since the camera and IMU are generally in close proximity. If the distance between the camera and IMU is much less than the ground resolution of the images (the physical distance represented by each pixel), then this translational offset will have a negligible effect on the image. Even small rotations, however, must be accounted for as the effects of camera pitch or roll are magnified by the height above ground level.

In order to estimate these parameters, we use a set of n features, and assume that we have determined their image locations in each of k images. Let $p_i^{(P_j)}$ be the measured location of the ith point in the jth image. Each of these corresponds to an estimate, $p_i^{(W_j)}$, of the position of the feature in the world. This estimate may be found by intersecting the corresponding line, $l_i^{(W_j)}$ found from Equation 2 with the DSM. If the transform estimates in Equation 2 were perfectly accurate then all of the estimates of the world location of each point would coincide.

From this observation we form an error metric as the sum of the squared distances between estimates of the points'

world location estimates, given by

$$E = \sum_{i=1}^{n} \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} |p_i^{(W_j)} - p_i^{(W_{j'})}|^2. \qquad (5)$$

For the moment we neglect errors in the navigation estimate (although this will be revisited in the following Sections), and view E as a function of the boresight and lens calibration parameters, x_b and x_l. We now seek a set of parameters to minimise E.

We note that the relationships between co-ordinate frames are almost always non-linear. W, N, and C are typically related by a combination of a rotation and a translation, while C and P are related by a projective transform, possibly with a further non-linear lens distortion model included. We do, however, typically have an initial estimate of the parameters x_b and x_l. Given these characteristics, standard non-linear optimisation algorithms such as Levenberg-Marquardt [12, Chapter 15.5] can be applied to determine updated parameter values.

The results of applying this optimisation to the two sample data sets are shown in Figures 8 and 9. In the case of the visible images 10 points in each of 3 images were identified, and in the thermal imagery 12 points in 3 images were used. These were used to estimate 6 boresight (rotation and translation) and 5 internal (focal length, pixel size, and principal point) calibration parameters. In the case of the visible data set, there was not much misalignment, but the error term was reduced from 30.20 to 2.28 after 100 iterations, with reasonable convergence (an error of 2.59) after just 5 iterations. For the thermal data set the error metric was reduced from 112.17 to 6.64 after 16 iterations when convergence was detected., with an error of 6.84 after 5 iterations. Note that the higher final residual in the thermal data set is due to the lower image resolution — the matching points are only located to the nearest pixel at best, which corresponds to about 1m in the thermal imagery, but about 12cm in the visible.

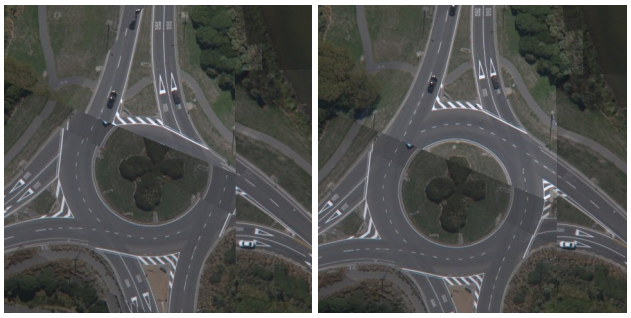

Figure 8. Orthomosaicing of the visible data set before (left) and after (right) refinement of the boresight and internal calibration values. The region shown is the roundabout in the bottom centre of Figure 6 at a resolution of 20cm/pixel

In the examples presented here, the points were manually identified in each image. This is practical because only a small number of parameters are being estimated, and so only a few constraints are required. For fully automated processing, automated feature detection and matching could be used. This is discussed in the following Sections in the context of bundle adjustment, where it is not practical to manually identify all of the points. Automatic feature matching, however, is not an entirely solved problem. As a result, processing on the basis of automatically detected and matched features must be made robust to outliers, as we shall see.

Figure 9. Orthomosaicing of the thermal data set with refined boresight and internal calibration values.

6 FEATURE DETECTION AND MATCHING

The parameter optimisation discussed in the previous section and bundle adjustment presented in the next section require a set of corresponding points between pairs of images. The automated extraction of interest or feature points from images, and the identification of corresponding features between images has received significant attention over the years. Most recently 'scale invariant' feature detectors and descriptors, such as SIFT [9] and SURF [4], have been widely used. These operators are constructed so that the same features can be identified in an image regardless of scale or orientation changes. They also aim to provide descriptions of the image region surrounding the features that are likewise invariant.

In the current context, however, there is significant additional information that can be used to aid the search for corresponding features. Scale, orientation, and translation changes can be estimated from the navigation data and surface model, meaning that such invariance is of much less significance.

Suppose we are given a feature in one image and want to find the corresponding point in another frame. In general this is a challenging problem, as there may be a fairly arbitrary transform between the two images. The navigation data, however, tells us how each of the images relates to the ground surface, and we can use this information to make the problem considerably easier. Given two image co-ordinate frames, P_1 and P_2, and a point, $p^{(P_1)}$, in the first image, we

wish to estimate the corresponding location, $p^{(P_2)}$ in the second frame. We also wish to determine a local region to compute a description of the feature from. We assume that this region is a square box in P_1, centred around $p^{(P_1)}$.

Using the methods described in Section 4 we can find a point on the DSM, $p^{(W)}$, corresponding to $p^{(P_1)}$. We can then project that point into P_2, which yields an estimate of $p^{(P_2)}$. This procedure can then be repeated for the four corners of the region describing the feature, yielding the corresponding region in P_2. This region can then be warped to align with the description extracted from P_1, allowing for easy comparison. This approach is illustrated in Figure 10

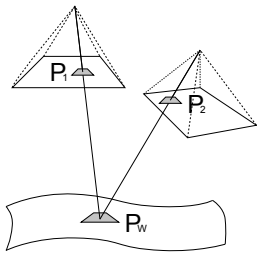

Figure 10. A feature, P_1, in one image is described by a rectangular patch. This relates to a corresponding patch, P_W, on the surface model, which can be projected into a new image to give a location estimate, P_2, and descriptor that is robust to scale and orientation changes.

In most cases, particularly those where the transform parameters need to be refined, the correspondence established by the geometry will not be accurate, but should be close. If the estimates of the camera parameters (position, orientation, and internal calibration) and surface model are well estimated, then the predicted feature appearance and location may be good enough that a simple correlation-based search can be used to refine the correspondence. Alternatively, the gradient-based tracking method of Lucas and Kanade [10] can be used to refine the correspondence. Lucas and Kanade's original method considered only translational motion, which may be sufficient in some cases. If required, however, more complex transforms can be used to correct for errors in orientation or scale [3, 2, 1], and pyramidal techniques [5] allow for correction of large offsets between the predicted and true feature location.

Figure 11 shows the result of matching features between images in the visible image set. In this example, features were detected using the Shi-Tomasi corner detector [13] and were tracked using simple correlation with a sum of squared differences metric. This metric was computed over an area of size 21×21 pixels in the image that the feature was first observed.

7 BUNDLE ADJUSTMENT

Once a set of feature correspondences has been established between the frames, the transform parameters can be refined. This may be extended from the boresight and lens calibration parameters discussed in Section 5, to also include the navigation to world frame transforms for each frame. Furthermore, the 3D locations of the features that have been identified in the images can be estimated. This leads to the process of bundle adjustment, where the structure of the scene, the motion of the cameras, and the camera calibration parameters are all adjusted to best fit measurements made from the world [14]. This sets up a system of equations, $y = f(x)$, where y are the measurements, and x the parameters.

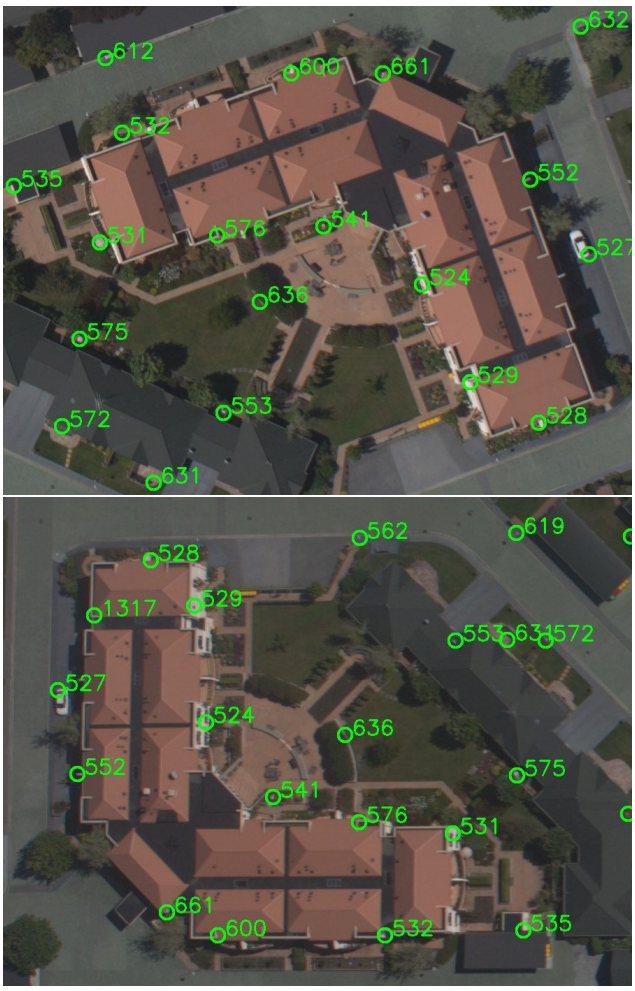

Figure 11. Example of feature matching between two images. Most features are successfully matched between the images with simple correlation, despite a large change in orientation.

In the context of automated aerial image processing with n points detected in each of m camera positions, x may be decomposed into

$$x = [x_b x_l x_{c_1} x_{c_2} \dots x_{c_m} x_{f_1} x_{f_2} \dots x_{f_n}], \qquad (6)$$

where x_b are the boresight parameters (typically a rotation and translation), x_l are the lens calibration parameters (focal length, sensor size, etc.), x_{c_i} are the parameters relating to the ith camera position (typically a rotation + translation), and x_{f_j} are the parameters describing the jth feature in the world (typically a 3D position vector).

Likewise, the measurement vector can be divided up into

$$y = [y_{f_{11}} y_{f_{12}} \dots y_{f_{1n}} y_{f_{21}} \dots y_{f_{2n}} \dots y_{f_{mn}}], \qquad (7)$$

where $y_{f_{ij}}$ is the position of the jth feature in the ith image.

Given an initial estimate of x, non-linear optimisation methods such as Levenberg-Marquardt may be applied to find an optimal parameter set. However, the system of equations becomes very large, and therefore very expensive to solve. The computation required may, however, be reduced significantly by exploiting the structure of the system. Each of the measurements depends on only a few of the parameters, and this leads to a sparse matrix structure. In the course of Levenberg-Marquardt (and other similar methods) the Jacobian matrix is computed, which gives the derivative of each measurement with respect to each parameter. Since each measurement is dependent on only a few parameters, many of the Jacobian entries are zero, as illustrated in Figure 12. The system may be further reduced by the fact that not all features are visible in all images.

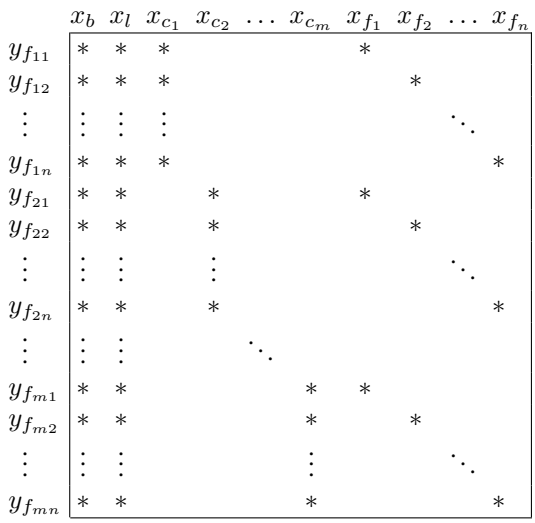

Figure 12. The sparse matrix structure arising in the bundle adjustment process. The Jacobian matrix relates parameters (columns) to measurements (rows), and only those regions of the matrix marked with '$*$' can be non-zero.

As with all automated feature correspondence methods, not all of the matches in Figure 11 are correct. Mismatches can occur to changes in the scene (such as moving cars), appearance changes due to changing view point, two or more similar features in a local neighbourhood, or a variety of other reasons. In order to reduce the effects of poor image matches, a RANSAC process is applied to remove outliers from the image feature measurements [6].

Approximately 200 features are detected in each frame, and then the Fundamental matrix is estimated between all pairs of frames which share at least 50 features [7, Chapter 50]. This estimation is made using RANSAC, and an image feature is removed from consideration unless it is considered an inlier in at least half of all pairs of images in which it is visible. A final check is made to remove any features which are considered inliers in fewer than three frames. A feature visible in only one frame provides no useful constraints on the bundle adjustment process, while a feature visible in just two frames provides 4 constraint equations (two 2D point locations), but introduces 3 extra unknowns (the 3D location of the feature in the world). This is only a marginal benefit for the increased computation required.

The results of RANSAC processing applied to the region shown in Figure 11 are shown in Figure 13. Note that there are significantly fewer features in each image, but that the remaining common features are correct matches. Note also that features which appear to be unmatched in the Figure will be visible in at least two other images from the full data set. The RANSAC approach taken removes features fairly aggressively, but this is desirable because outliers can cause significant errors in further processing, and fewer features mean that less computation is required in the bundle computation.

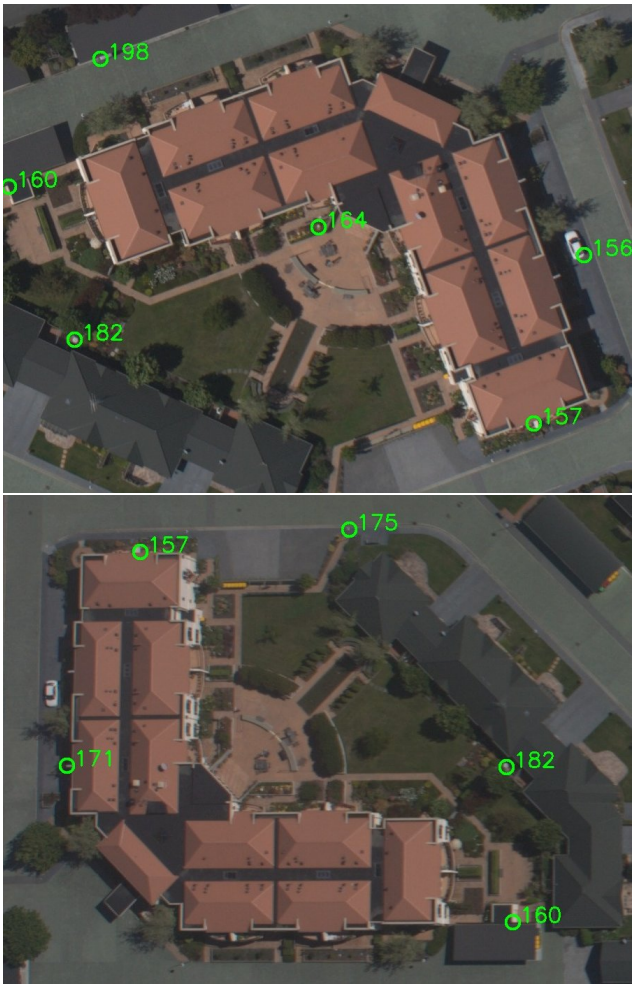

Figure 13. Example of features retained after RANSAC outlier removal.

Once a set of inlier feature correspondences has been determined, the bundle adjustment process can be applied. Bundle adjustment is not applied to the thermal data set, as the alignment is already good compared to the resolution of the images. For higher resolution images, however, even

small changes in the parameters can lead to large misalignments. For the Canon 400D camera and 28mm lens used in the visible example, a change of orientation of 1 degree can lead to a misalignment of 80-90 pixels. For the iMAR IMU used to generate the attitude data, roll and pitch accuracies are typically less than 0.01 degrees however the bosesight calibration parameters can be significantly larger than this. Furthermore, the aim is to develop a processing method that is able to deal with lower cost sensors. For example, typical roll and pitch accuracies of better than 0.1 degrees have been experienced when using the Crossbow IMU440CA with POINT software, although heading errors are typically much larger at approximately 0.5 degrees. Improving the performance of lower cost navigation sensors is an area of continued development and, in particular, heading errors can be substantially reduced using GNSS heading aiding with 2 or more GNSS antennas and receivers. However, for many applications, navigation data of from sensors such as the Crossbow should be sufficient in order to be used with the image processing algorithms described in this paper. Figure 14 shows details of the image mosaic before and after bundle adjustment is used to refine the camera calibration, position, and orientation parameters.

Figure 14. Visible mosaic details at 20cm resolution before (left) and after (right) bundle adjustment.

In most cases the image alignment is visibly improved, but in a few places (such as the last image shown) it is worse after the bundle than before. The reason for this is that the image features are not uniformly distributed across the image. In this particular example, the built up areas are a much more reliable source of feature correspondences than the green space and waterway that lines the main road running through the centre of the image. While the feature detector does find numerous features in all parts of the image, those in the green spaces are not reliably matched, and so are largely eliminated in the RANSAC process. This means that the bundle adjustment has much stronger constraints in the built up areas, leading to better performance in those regions.

The bundle adjustment process presented here may be further extended using other data already available to the system. The navigation solution from the GNSS and IMU processing is one such data source, and provides a constraint on the position and orientation of the camera for each frame. The surface model can also be used as a constraint on the 3D location estimates for each feature point. The points may be either forced to lie on the surface (by estimating their 2D location and deriving the height from the surface model), or near the prior surface estimate (by adding the distance from the surface to the error to be minimised).

Adding these additional constraints raises the issue that the measurements are uncertain, and should not all have equal weight. In the case of the navigation estimates of camera position and orientation, these are typically estimated by a Kalman filtering process which provides a covariance estimate. Covariances may also be available for the surface model, or the estimates may be assigned an estimated uncertainty (typically a few metres for readily available data sources). Representing uncertainty in feature correspondences is somewhat more difficult. A covariance estimate (typically a few pixels) can be given to the locations, but automated feature correspondence typically includes false correspondences or outliers, even after RANSAC processing. Ideally, the bundle adjustment should be made robust to a few gross errors in the feature correspondences. This may be achieved by using robust cost functions, such as the Huber cost function [7, Appendix A6]. The optimisation process aims to minimise the sum of squared errors between the measurements made and their expected values given the parameters. Robust cost function reduce the penalty associated with large deviations — the Huber cost function, for example, changes from a squared penalty for small errors to a linear penalty for larger errors. This reduces the influence that a small number of large errors can have on the system, making it more robust to outliers.

8 IMAGE MOSAICING

The methods discussed in the previous scenes aim to reduce the misalignment between the images that make up an orthomosaic. There will, however, always be some visible seams in the mosaic. These may be due to the uncertainties inherent in any measurement process, which mean that the optimisation methods can never remove all of the misalignments. Other seams are visible because of changing image capture parameters (exposure time, lighting conditions, moving objects, etc.) between frames. Whatever the cause, a variety of image processing techniques can be deployed in order to reduce the visibility of these seams in the final mosaic.

In the case of the sample imagery, the most visible effect is the changing brightness of the images across the sequence. In the case of the visible imagery, the shutter speed was fixed and the aperture allowed to vary, while in the thermal imagery this is due to an auto-gain function of the camera, which cannot be disabled. It is, however, possible to adjust the overall intensity of the frames prior to the mosaicing

process to minimise the effects of this variation. The overlapping regions between frames are used to estimate a shift in intensity for each frame to correct for the overall difference in brightness. For the overlapping region, R, between the ith and jth images, we form a cost function

$$\sum_{(x,y)\in R} ((R_i(x,y)+s_i)-(R_j(x,y)+s_j))^2, \quad (8)$$

where R_k is the region of the kth image corresponding to R, and s_k the shift to be applied to the kth frame. More complex functions (such as a scale and shift, or a non-linear function) could be applied, but the simple shifting alone creates a significant improvement in the visual quality of the mosaic, as shown in Figures 15 and 16.

Figure 15. Visible mosaic with intensity variations corrected.

Figure 16. Thermal mosaic with intensity variations corrected.

9 DISCUSSION AND CONCLUSIONS

This paper has shown how a range of techniques from the computer vision and image processing literature can be used to automate the production of orthomosaics from aerial imagery. Navigation data for the images, along with readily available digital surface models, provides sufficient information to create good quality orthomosaics from a variety of sources. While expert photogrammetric processing and high-quality camera equipment are still required to deliver the best quality results, we have demonstrated that useful data sets can be produced without the expensive cameras or time-consuming manual processing.

Images directly projected onto a DSM on the basis of navigation data from GNSS and IMU integration form the basis of the approach. While the results of such naïve processing are not ideal, they can be done very efficiently, allowing even real-time mosaic generation. Further refinement of the parameters governing the transforms between the image and world co-ordinate frames has been demonstrated using standard non-linear optimisation algorithms. This refinement relies on the identification of corresponding points between image frames, and we have demonstrated that this can be done in an automated fashion that is robust to scale and orientation changes. Finally, we have provided examples of methods that can be used to improve the visual appearance of the final mosaic images.

These methods have been demonstrated on two sample data sets, one a sequence of high resolution visible images from a consumer DSLR, and the other a thermal video sequence. Despite the very different characteristics of these data sets, the same basic processing chain can be applied. This ability to use a variety of sensors means that a wide range of application needs can be met by the system presented. High resolution visible imagery finds a wide variety of applications, from agriculture to urban planning, while the ability to use thermal imagery opens up applications such as environmental monitoring and fire-fighting. Cameras in other spectral bands can also be used, such as near-infrared imagery to support the computation of vegetative indices. It is also important to note that the mosaics produced by this approach are, by virtue of the direct geo-referencing stage, aligned to the chosen world co-ordinate frame. This means that they can be easily combined with other data sources in GIS software, as illustrated in Figures 17 and 18.

ACKNOWLEDGEMENTS

The authors would like to thank Phil Bartie for his assistance with the GIS processing.

REFERENCES

[1] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Robitics Institute Technical Report CMU-RI-TR-03-35, Carnegie Mellon University, 2003.

[2] S. Baker, R. Gross, I. Matthews, and T. Ishikawa. Lucas-Kanade 20 years on: A unifying framework: Part 2. Robitics Institute Technical Report CMU-RI-TR-03-01, Carnegie Mellon University, 2003.

Figure 17. The visible mosaic with road centrelines overlaid in GIS software. Note that the centreline data predates some of the new subdivisions in this image.

Figure 18. The thermal mosaic with road centrelines overlaid in GIS software.

[3] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 1. Robitics Institute Technical Report CMU-RI-TR-02-16, Carnegie Mellon University, 2002.

[4] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, Graz, Austria, May 2006.

[5] J. Y. Bouguet. Pyramidal implementation of the affine Lucas Kanade feature tracker: Description of the algorithm. Online at `http://robots.stanford.edu/cs223b04/algo_affine_tracking.pdf`, last accessed 7 September 2009.

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2000.

[8] NASA Jet Propulsion Laboratory. Aster global digital elevation map. Online at `http://asterweb.jpl.nasa.gov/gdem.asp`, last accessed 7 September 2009.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[10] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[11] Photometrix. Image coordinate correction function in Australis. Online at `http://www.photometrix.com.au/downloads/australis/Image%20Correction%20Model.pdf`, last accessed 7 September 2009, 2001.

[12] W. H. Press, S. A. Teukolsky, W. T. Vettering, and B. P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, second edition, 2002.

[13] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[14] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment — a modern synthesis. In *International Workshop on Vision Algorithms: Theory and Practice*, pages 298–372, 1999.