Robust Registration of Aerial Image Sequences

Clark F. Olson[1], Adnan I. Ansar[2], and Curtis W. Padgett[2]

[1] University of Washington, Bothell
Computing and Software Systems, Box 358534
18115 Campus Way N.E.
Bothell, WA 98011-8246
cfolson@uw.edu
[2] Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109-8099

Abstract. We describe techniques for registering images from sequences of aerial images captured of the same terrain on different days. The techniques are robust to changes in weather, including variable lighting conditions, shadows, and sparse intervening clouds. The primary underlying technique is robust feature matching between images, which is performed using both robust template matching and SIFT-like feature matching. Outlier rejection is performed in multiple stages to remove incorrect matches. With the remaining matches, we can compute homographies between images or use non-linear optimization to update the external camera parameters. We give results on real aerial image sequences.

1 Introduction

Persistent, high resolution aerial imagery can provide substantial scene detail over wide areas. Sensor platforms such as Angel Fire, Constant Hawk and ARGUS that dwell over an area give wide area coverage at high resolution. Both Angel Fire and ARGUS serve these images at low latency to multiple ground users at update rates sufficient for real time situational awareness. To increase the utility of these persistent surveillance platforms to ground users, there is a need to align the imagery to existing databases (road, city, waterway maps, etc.) and to fuse the data from multiple platforms servicing the same area, possibly of different modalities.

One of the simplest ways to accomplish this is to use onboard Inertial Navigation Sensors (INS) and GPS to geo-register the collected imagery to a terrain map (either a pre-existing Digital Elevation Map, or one generated from the collected imagery). Requested portions of the image product can then be served directly to ground users from the aerial platform thus eliminating the need to send the full image stream (requiring greater than 100 Mbits/second over limited bandwidth) to the ground for further processing. A well registered image stream also improves the efficiency of the image compression routines allowing for higher quality imagery (or more coverage) to be passed directly to the user.

G. Bebis et al. (Eds.): ISVC 2009, Part II, LNCS 5876, pp. 325–334, 2009.

Image analysts need information much more than they need raw data, and utility grows at each level of information that is successfully and reliably extracted. Fusion of data from multiple sensors has long been known as an effective way of highlighting and interpreting significant events while reducing false indications. Often the challenge, however, is ensuring proper association between data items collected from dissimilar sensors. In other words, you need to line images up before you can find out where information coincides. The value of even simple pixel-level image fusion has been shown for well over a decade. Targets pop out and images become easy to interpret [1]. Pixel level fusion, however, depends upon lining up images to a very high level of precision, a task that is difficult enough when the sensors are mounted on the same platform.

Although very precise INS/GPS sensors exist, the challenge of localizing a fraction of square meter (the typical pixel size on the ground) from more than a mile away is extremely high, even frame to frame on the same image. Further, future persistent surveillance platforms will come in all size classes (Shadow, Scan Eagle, Predator, etc.) precluding the use of the best INS sensors due to size/power constraints. Biases in the INS/GPS pointing systems also introduce misalignments that can result in substantial ground errors.

Ideally, the data products collected onboard should be referenced to a standard, conical map to align the locally produced imagery prior to dissemination. This would provide a unified view of the data products collected across time (prior flights of the same sensor), platforms (multiple views of the same ground), and modalities (different sensors, wavelengths, resolution, etc.) providing a robust, easy to manipulate view of a scene for exploitation by image analysts or automated recognition algorithms.

Given these constraints, we are interested in a problem where persistent surveillance of a site is performed on multiple days and it is desirable to align the imagery from the current day with the previously acquired data in real-time. This requires the computation of an offset in the external camera parameters over some initial set of images that can be propagated to future images allowing precise registration of the future images with minimal processing. In this scenario, we have high resolution data (the captured images are 4872×3248, although we perform most processing at 1218×812). In general, our scenario allows us to assume that the images are captured at roughly the same elevation, since this is typical with the surveillance flights (and we can rescale according to the estimated altitude, if necessary). Similarly, since the aircraft circles a particular area of interest, we can extract images from the previous sequence from approximately the same viewpoint.

To accomplish the registration, we use a feature matching strategy with careful outlier rejection. We then optimize the offset in the external camera parameters using multiple images in each sequence in order to determine a precise relative positioning between the sequences and allow real-time alignment. We can also use these techniques to align multiple sequences with a single previously generated map.

2 Previous Work

There has been extensive work on image registration [2,3,4], including work on aerial images [5,6,7,8] and aerial image sequences [9,10,11].

A common strategy has been feature detection and matching, followed by a process to optimize the alignment of features. Zheng and Chellapa [5] described such a technique for finding the homography aligning the ground planes for the registration of oblique aerial images. Tuo et al. [6] perform registration after modifying the images to fit a specified brightness histogram. Features are then detected and aligned. Yasein and Agathoklis [7] solve only for a similarity transformation, but use an iterative optimization where the points are weighted according to the current residual.

Xiong and Quek [8] perform registration up to similarity transformations without explicitly finding correspondences. After detecting features in both images, an orientation is computed for each feature and all possible correspondences are mapped into a histogram according to the orientation differences. The peak in the histogram is chosen as the rotation between the images. Scale is determined through the use of angle histograms computed with multiple image patch sizes and selecting the histogram with the highest peak. Niranjan et al. [9] build upon the work of Xiong and Quek in order to register images in an image sequence up to a homography.

Lin et al. [10] concentrate on registering consecutive aerial images from an image sequence. They use a reference image (such as a map image) in order to eliminate errors that accumulate from local methods. Their two-step process first performs registration between images and then uses this as an initial estimate for registration with the reference image. Wu and Luo [11] also examine registration in an aerial image sequence. In their technique, the movement of the camera is predicted from previous results in the sequence. This information is used to rank possible correspondences. A variation of RANSAC [12] is used to validate correspondences and to refine the motion estimate.

3 Feature Matching

In order to register images from different sequences (and from the same sequence), we match features or landmarks identified in the images. Two methods are used for feature matching that may be used independently or in combination. In both methods, we select a set of discrete features to be matched in one or more of the images.

Given our mission scenario, we work under the assumption that the images are captured from roughly the same altitude and that we can find images captured with roughly the same camera axis from the two image sequences. This allows us to neglect major scale and orientation differences in the feature matching process and gains us robustness to false positives that might occur between features of different scales or orientations. The techniques are able to handle variation of up to 15 degrees or 15% scale change. In cases where these assumptions are not warranted, we can easily replace the techniques with those invariant to scale and rotation changes. This is expected to be rare, since the INS provides us with data that can be used to warp the images into roughly the same scale and rotation.

3.1 Feature Selection

Features are selected in images using a two step process. First, each pixel in the image is assigned a score based on the second moment matrix computed in a neighborhood around each pixel:

$$\sum\sum\begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \tag{1}$$

Following Shi and Tomasi [13] we use the smallest eigenvalue of this matrix as a measure of how easy the feature is to track. (If the larger eigenvalue is low, this indicates a featureless region. If the larger eigenvalue is high, but the smaller eigenvalue is low, this indicates a linear edge.)

Given the scores, we select features from rectangular subimages in order to distribute the features over the entire image. Features are selected greedily starting with the largest scores and discarding those that are too close to previously selected features. The current implementation uses 16 subimages in a 4×4 grid with 16 features selected in each for a total of 256 features.

3.2 Robust Template Matching

Of the two feature matching techniques we use, template matching using gradient or entropy images is the more robust, but also more time consuming, unless multiresolution search techniques are used. The technique is based previous work for aligning entropy images [14]. We first compute a new representation of each image, replacing each pixel with either a local measure of either the image gradient or entropy.

Feature matches are detected using normalized correlation between the templates encompassing the detected features and the search image. We look for matches over the entire search image efficiently using the Fast Fourier Transform (FFT) to implement normalized correlation. This allows each template to be processed in $O(n \log n)$ time, where n is the number of pixels in the search image. In order to improve the speed, a multi-resolution search option is included.

Featureless regions in the search image are undesirable and lead to poor quality matches, so we discount template windows in the search image with below average root-mean-square (RMS) intensity (in the gradient or entropy image) using the following function:

$$S(r,c) = \begin{cases} NC(r,c), & \text{if } w(r,c) \geq \overline{w} \\ \frac{2w(r,c)\cdot NC(r,c)}{w(r,c)+\overline{w}}, & \text{if } w(r,c) < \overline{w} \end{cases} \tag{2}$$

where $NC(r,c)$ is the normalized correlation of the template with the window centered at (r,c), $w(r,c)$ is the RMS intensity of the window, and $\overline{w}$ is the average window intensity over the search image.

3.3 SIFT-Like Feature Matching

We also use a method based on the SIFT technique [15]. However, we do not employ the scale and rotation invariance aspects of SIFT, since the images are already (or can be transformed to be) at roughly the same scale and rotation.

Feature extraction is first performed on both images using the technique described above. Each feature is characterized using the SIFT method as a vector with 128 entries representing a histogram of gradients in the feature neighborhood at various positions and orientations. Features are compared using normalized correlation and the best match is tentatively accepted if the normalized correlation exceeds 0.75.

One disadvantage to this technique is that, even if the same feature is located in both images, it may be localized at slightly different locations. To correct this, we refine each feature match using a brute-force search that considers positive and negative displacements in row and column of up to two pixels.

3.4 Comparison

Over a test set consisting of 108 image pairs, the entropy techniques averaged 119.3 inliers found, while the gradient techniques averaged 116.6 inliers and the SIFT-based techniques averaged 79.7 inliers. The entropy and gradient techniques required approximately 22 seconds working on 1024×812 images[1] using a multi-resolution search, including file I/O, feature detection, matching, and refinement. The SIFT-based techniques required approximately 30 seconds, but were able to work on the full 1218×812 images. When the multi-resolution search is not used, the average number of inliers increases to 143.1 for entropy matching and 140.4 for gradient matching, but the computation time increases to 350 seconds. Figure 1 shows examples of features extracted and matched using aerial images.

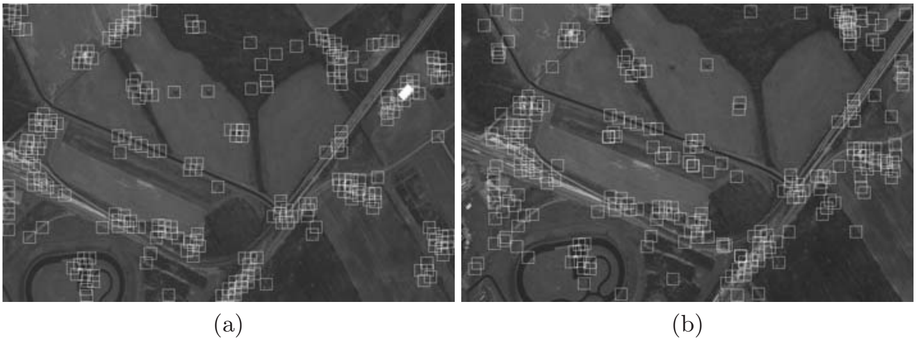
4 Outlier Rejection

We use a multiple step process to reject outliers in the detected feature matches. The first step is designed to reject gross outliers using sampling. The second step (which is iterated) computes a homography between the points and discards those with larger residuals. Both steps are very efficient and require a fraction of the time used for feature matching.

4.1 Gross Outlier Rejection

Gross outliers are detected using a variation of the RUDR strategy for model fitting [16]. In this strategy, trials are used that hypothesize sets of correct matches that are one match less than the number sufficient to calculate the model parameters. The parameters are estimated by combining the hypothesized matches

[1] The images were reduced from 1218 columns to 1024 in order to quarter the time required by the FFT.



(a) (b)

Fig. 1. Feature selection and matching. (a) Features selected in first image. (b) Best matches found in second image. Outlier rejection has not been performed at this stage.

with each possible remaining match and detecting clusters among the estimated parameters. A trial is expected to succeed whenever the hypothesized set of matches contains no outliers.

To reject gross outliers, we use a simple model of the motion that allows only similarity transformations between the images (rotation, translation, and scale). For this model, two matches are sufficient to solve for the model parameters. Therefore, each trial fixes one match and considers it in combination with all of the other matches. Since the number of matches is not large, we perform a trial for each match, rather than using random sampling as previously described [16].

In our scenario, the clustering in each trial is relatively simple, since the images have roughly the same scale and orientation. We can simply eliminate those matches that produce a rotation estimate that varies significantly from zero or a scale estimate that varies significantly from one. (Our experiments allow 15% scale change and 15° rotation.) The trial with the largest set of inliers (that is, the fewest eliminated matches) is selected and the remaining matches are not considered further.

4.2 Careful Outlier Rejection

The matches that survive the previous step are examined more carefully for further outliers. In this step, we solve for the homography that best aligns the matches using a least-squares criterion and compute the residual for each match. Any match with residual greater than twice the median residual is eliminated. This is iterated until one of the following conditions hold:

1. No outliers are found.
2. The median residual falls below a threshold (1.0).
3. The number of matches falls below a threshold (20).

Note that the third condition was never met in our experiments, but is present in the implementation to ensure that sufficient matches remain to perform the optimization. Figure 2 shows the matches from Fig. 1 with outliers removed.

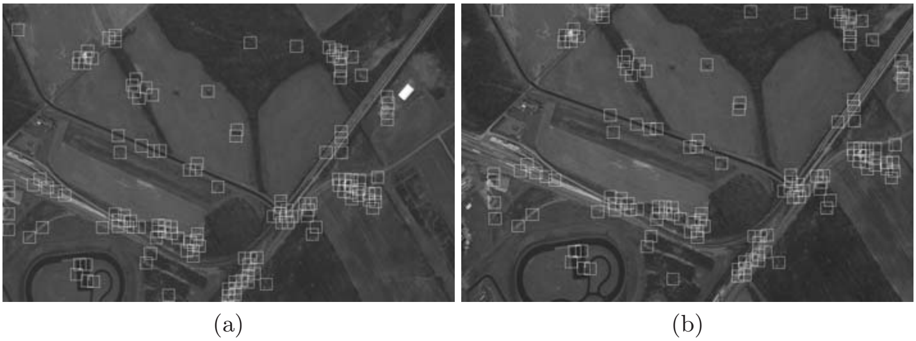

(a)
(b)

Fig. 2. Features after outlier rejection. (a) Features selected with rejected matches excluded. (b) Best matches found in second image with rejected matches excluded.

5 Nonlinear Parameter Optimization

Given the matches computed between the images sequences, we can now perform a nonlinear optimization step to refine the motion between the sequences. Our goal is to compute a single six degree-of-freedom (DOF) transformation between the two sequences, under the assumption that the relative errors within each sequence are small. However, only five parameters can be extracted without additional information, owing to the scale ambiguity. We set the scale by requiring the average depth of the points from the camera to agree with the elevation specified by the INS for the first image. This allows us to optimize the six motion parameters without ambiguity.

As in previous work [17], our optimization uses a state vector that includes not only the motion parameters, but also the elevation of each feature point matched. (Initially, each point is estimated to be on a flat ground plane.) With this formulation, we can use an objective function based on the distance between the predicted feature location (according to the estimated motion) and the matched feature location. The objective function is augmented with a penalty term that enforces the scale constraint. Overall, this yields a state vector with $m + 6$ variables to optimize, where m is the number of distinct features matched, and an objective function with $2n + 1$ constraints, where n is the number of feature matches between the sequences. The values of m and n are not necessarily the same, since a feature may be matched in multiple images of a sequence.

We optimize the objective function using the Levenberg-Marquardt method with lmfit, a public domain software package based on MINPACK routines [18].

6 Results

Figure 3 shows three registration results using these techniques. In each case, an image has been warped into the frame of an image from a previous sequence. The images are merged by taking the average of the pixel values after the warping

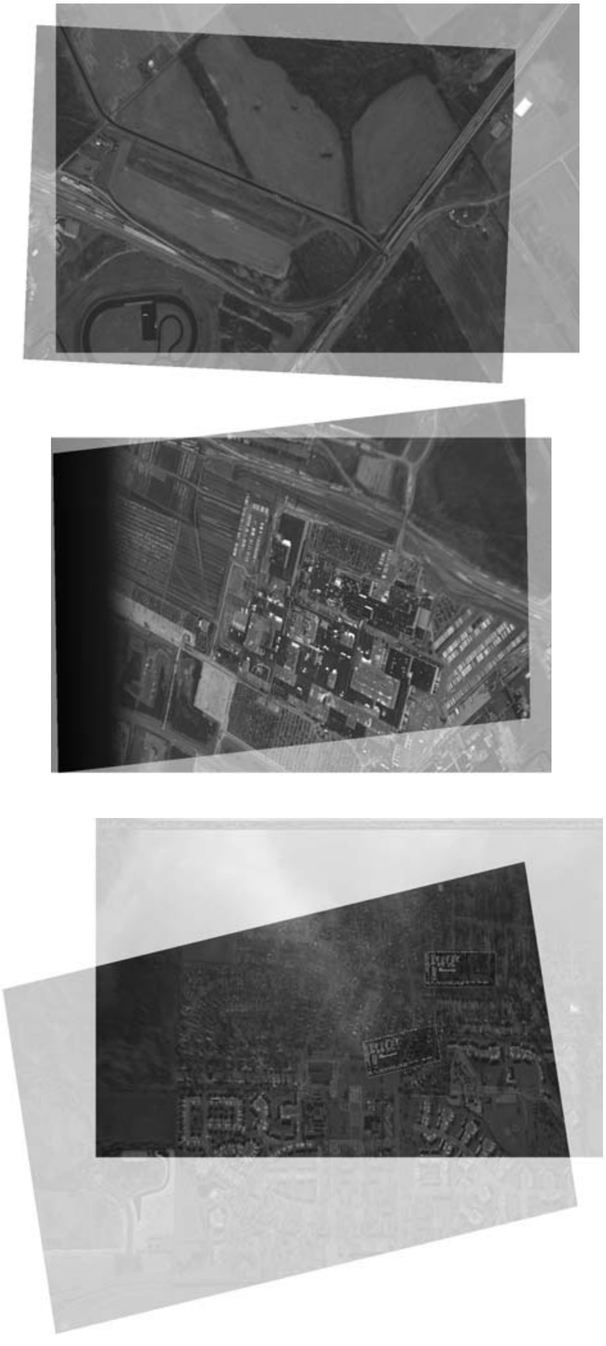


Fig. 3. Registration results. Pairs of images are merged by taking the average of the pixel values after warping the second image into the frame of the first image.

has been performed. Locations outside of the image are left as white. (Areas covered by only one image are lightly illustrated.)

The first example shows the relatively straightforward example that has been used to illustrate the components of the system in Figures 1 and 2. It can be observed that the registration is good, since there is little blurring in the averaged pixels and the landmarks (such as roads) align well. The second example shows a scene with buildings and that has increased warping between the images. The left side of these images is dark owing to occlusion.

The final example shows a more complex case, where clouds occluded the terrain in one of the sequences and a data box produced distractors that did not move with the terrain. Furthermore, the overlap between the images is reduced in this case. Despite these issues, our algorithm was able to correctly register the images. In practice, our operational scenario will yield images with greater overlap and less rotation than in this example.

7 Summary

The registration of aerial image sequences is important in persistent surveillance applications in order to accurately fuse current images with previously collected data. We have described techniques for the robust registration of such image sequences. We first match landmarks between the sequences by selecting interesting image features and using robust matching techniques. Outlier rejection is performed carefully in order to extract a set of high quality matches. Finally, the external camera parameters are refined using nonlinear optimization. Results on real images sequences indicate that the method is effective.

References

1. Waxman, A.M., Aguilar, M., Fay, D.A., Ireland, D.B., Racamato Jr., J.P., Ross, W.D., Carrick, J.E., Gove, A.N., Seibert, M.C., Savoye, E.D., Reich, R.K., Burke, B.E., McGonagle, W.H., Craig, D.M.: Solid-state color night vision: Fusion of low-light visible and thermal infrared imagery. Lincoln Laboratory Journal 11, 41–60 (1998)
2. Brown, L.G.: A survey of image registration techniques. ACM Computing Surveys 24, 325–376 (1992)
3. Maintz, J.B.A., Viergever, M.A.: A survey of medical image registration. Medical Image Analysis 2, 1–16 (1998)
4. Zitova, B., Flusser, J.: Image registration methods: A survey. Image and Vision Computing 21, 977–1000 (2003)
5. Zheng, Q., Chellappa, R.: Automatic registration of oblique aerial images. In: Proceedings of the IEEE International Conference on Image Processing, vol. 1, pp. 218–222 (1994)
6. Tuo, H., Zhang, L., Liu, Y.: Multisensor aerial image registration using direct histogram specification. In: Proceedings of the IEEE International Conference on Networking, Sensing and Control, pp. 807–812 (2004)

7. Yasein, M.S., Agathoklis, P.: A robust, feature-based algorithm for aerial image registration. In: Proceedings of the IEEE International Symposium on Industrial Electronics, pp. 1731–1736 (2007)
8. Xiong, Y., Quek, F.: Automatic aerial image registration without correspondence. In: Proceedings of the 4th International Conference on Computer Vision Systems (2006)
9. Niranjan, S., Gupta, G., Mukerjee, A., Gupta, S.: Efficient registration of aerial image sequences without camera priors. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 394–403. Springer, Heidelberg (2007)
10. Lin, Y., Yu, Q., Medioni, G.: Map-enhanced UAV image sequence registration. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (2007)
11. Wu, Y., Luo, X.: A robust method for airborne video registration using prediction model. In: Proceedings of the International Conference on Computer Science and Information Technology, pp. 518–523 (2008)
12. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–396 (1981)
13. Shi, J., Tomasi, C.: Good features to track. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)
14. Olson, C.F.: Image registration by aligning entropies. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 331–336 (2001)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
16. Olson, C.F.: A general method for geometric feature matching and model extraction. International Journal of Computer Vision 45, 39–54 (2001)
17. Xiong, Y., Olson, C.F., Matthies, L.H.: Computing depth maps from descent images. Machine Vision and Applications 16, 139–147 (2005)
18. Wuttke, J.: lmfit - a C/C++ routine for Levenberg-Marquardt minimization with wrapper for least-squares curve fitting (2008) based on work by B.S. Garbow, K.E. Hillstrom, J.J. More, and S. Moshier.: Version 2.4, http://www.messen-und-deuten.de/lmfit/ (retrieved on June 2, 2009)