Feature-based Georegistration of Aerial Images

Yaser Sheikh[1], Sohaib Khan[2] and Mubarak Shah[1]

[1] Computer Vision Lab,
School of Computer Science,
University of Central Florida,
Orlando, FL 32816-2362,
USA

[2] Department of Computer Science and Computer Engineering,
Lahore University of Management Sciences,
Lahore, Pakistan

1 Introduction

Georegistration is the alignment of an observed image with a geodetically calibrated reference image. Such alignment allows each observed image pixel to inherit the coordinates and elevation of the reference pixel it is aligned to. Accurate georegistration of video has far-reaching implications for the future of automation. An agent (such as a robot or a UAV), equipped with the ability to precisely assign geodetic coordinates to objects or artifacts within its field of view, can be an indispensable tool in applications as diverse as planetary exploration and automated vacuum cleaners. In this chapter, we present an algorithm for the automated registration of aerial video frames to a wide area reference image. The data typically available in this application are the reference imagery, the video imagery and the telemetry information.

The reference imagery is usually a wide area, high-resolution ortho-image. Each pixel in the reference image has a longitude, latitude and elevation associated with it (in the form of a DEM - **D**igital **E**levation **M**ap). Since the reference image is usually dated by the time it is used for georegistration, it contains significant dissimilarities with respect to the aerial video data. The aerial video data is captured from a camera mounted on an aircraft. The orientation and position of the camera are recorded, per-frame, in the telemetry information. Since each frame has this telemetry information associated with it, georegistration would seem to be a trivial task of projecting the image onto the reference image coordinates. Unfortunately, mechanical noise causes fluctuations in the telemetry measurements, which in turn causes significant projection errors, sometimes up to hundreds of pixels. Thus while the telemetry information provides *coarse* alignment of the video frame, georegistration techniques are required to obtain accurate pixel-wise calibration of each aerial image pixel. In this chapter, we use the telemetry information to orthorectify the aerial images, to bring both imageries into a common projection space, and then apply our registration technique to achieve accurate alignment. The challenge in georegistration lies in the stark differences between the video and reference data. While the difference of projection view is accounted for by orthorectification, four types of data distortions are still encountered: (1) Sensor Noise in the form of erroneous Telemetry Data, (2) Lighting and Atmospheric Changes, (3) Blurring, (4) Object Changes in the form of forest growths or

new construction. It should also be noted that remotely sensed terrain imagery has the property of being highly self-correlated both as image data and elevation data. This includes first order correlations (locally similar luminance or elevation values in buildings), second order correlations (edge continuations in roads, forest edges, and ridges), as well as higher order correlations (homogeneous textures in forests and homogenous elevations in plateaus). Therefore, while developing georegistration algorithms the important criterion is the robust handling of outliers caused by this high degree of self-correlation.

1.1 Previous Work

Currently several systems that use geolocation have already been deployed and tested, such as Terrain Contour Matching (TERCOM) [10], SITAN, Inertial Navigation / Guidance Systems (INS/IGS), Global Positioning Systems (GPS) and most recently Digital Scene-Matching and Area Correlation (DSMAC). Due to the limited success of these systems and better understanding of their shortcomings, georegistration has recently received a flurry of research attention. Image-based geolocation (usually in the form of georegistration) has two principal properties that make them of interest: (1) Image capture and alignment is essentially a passive application that does not rely on interceptable emissions (like GPS systems) and (2) Georegistration allows independent per-frame geolocation thus avoiding cumulative errors. Image based techniques can be broadly classified into two approaches: Intensity-based approaches and elevation-based approaches.

The overriding drawback of Elevation-based approaches is that they rely on the accuracy of recovered elevation from two frames, which has been found to be difficult and unreliable. Elevation based algorithms achieve alignment by matching the reference elevation map with an elevation map recovered from video data. Rodrequez and Aggarwal in [24] perform pixel-wise stereo analysis of successive frames to yield a recovered elevation map or REM. A common representation (‘cliff maps’), are used and local extrema in curvature are detected to define critical points. To achieve correspondence, each critical point in the REM is then compared to each critical point in the DEM. From each match, a transformation between REM and DEM contours can be recovered. After transforming the REM cliff map by this transformation, alignment verification is performed by finding the fraction of transformed REM critical points that lie near DEM critical points of similar orientation. While this algorithm is efficient, it runs into similar problems as TERCOM i.e. it is likely to fail in plateaus, ridges and depends highly on the accurate reconstruction of the REM. Finally, no solution was proposed for computing elevation from video data. More recently in ([25]), a relative position estimation algorithm is applied between two successive video frames, and their transformation is recovered using point-matching in stereo. As the error may accumulate while calculating relative position between one frame and the last, an absolute position estimation algorithm is proposed using image based registration in unison with elevation based registration. The image based alignment uses Hausdorff Distance Matching between edges detected in the images. The elevation based approach estimates the absolute position, by calculating the variance of displacements. These algorithms, while having been shown to be highly efficient, restrict degrees of alignment to only two

(translation along x and y), and furthermore do not address the conventional issues associated with elevation recovery from stereo.

Image-based registration, on the other hand, is a well-studied area. A somewhat outdated review of work in this field is available in [4]. Conventional alignment techniques are liable to fail because of the inherent differences between the two imageries we are interested in, since many corresponding pixels are often dissimilar. Mutual Information is another popular similarity measure, [30], and while it provides high levels of robustness it also allows many false positives when matching over a search area of the nature encountered in georegistration. Furthermore, formulating an efficient search strategy is difficult. Work has also been done in developing image-based techniques for the alignment of two sets of reference imageries [32], as well as the registration of two successive video images ([3], [27]). Specific to georegistration, several intensity based approaches to georegistration intensity have been proposed. In [6], Cannata *et al* use the telemetry information to bring a video frame into an orthographic projection view, by associating each pixel with an elevation value from the DEM. As the telemetry information is noisy the association of elevation is erroneous as well. However, for aerial imagery that is taken from high altitude aircrafts the rate of change in elevation may be assumed low enough for the elevation error to be small. By orthorectifying the aerial video frame, the process of alignment is simplified to a strict 2D registration problem. Correspondence is computed by taking 32×32 pixel patches uniformly over the aerial image and correlating them with a larger search patch in the Reference Image, using Normalized Cross Correlation. As the correlation surface is expected to have a significant number of outliers, four of the strongest peaks in each correlation surface are selected and consistency measured to find the best subset of peaks that may be expressed by a four parameter affine transform. Finally, the sensor parameters are updated using a conjugate gradient method, or by a Kalman Filter to stress temporal continuity. An alternate approach is presented by Kumar *et al* in [18] and by Wildes *et al* in [31] following up on that work, where instead of ortho-rectifying the Aerial Video Frame, a perspective projection of the associated area of the Reference Image is performed. In [18], two further data rectification steps are performed. Video frame-to-frame alignment is used to create a mosaic providing greater context for alignment than a single image. For data rectification, a Laplacian filter at multiple scales is then applied to both the video mosaic and reference image. To achieve correspondence, coarse alignment is followed by fine alignment. For coarse alignment feature points are defined as the locations where the response in both scale and space is maximum. Normalized correlation is used as a match measure between salient points and the associated reference patch. One feature point is picked as a reference, and the correlation surfaces for each feature point are then translated to be centered at the reference feature point. In effect, all the correlation surfaces are superimposed, and for each location on the resulting superimposed surface, the top k values (where k is a constant dependant on number of feature points) are multiplied together to establish a consensus surface. The highest resulting point on the correlation surface is then taken to be the true displacement. To achieve fine alignment, a 'direct' method of alignment is employed, minimizing the SSD of user selected areas in the video and reference (filtered) image. The plane-parallax model is employed, expressing the transformation

between images in terms of 11 parameters, and optimization is achieved iteratively using the Levenberg-Marquardt technique.

In the subsequent work, [31], the filter is modified to use the Laplacian of Gaussian filter as well as it's Hilbert Transform, in four directions to yield four oriented energy images for each aerial video frame, and for each perspectively projected reference image. Instead of considering video mosaics for alignment, the authors use a mosaic of 3 'key-frames' from the data stream, each with at least 50 percent overlap. For correspondence, once again a local-global alignment process is used. For local alignment, individual frames are aligned using a three-stage Gaussian pyramid. Tiles centered around feature points from the aerial video frame are correlated with associated patches from the projected reference image. From the correlation surface the dominant peak is expressed by its covariance structure. As outliers are common, RANSAC is applied for each frame on the covariance structures to detect matches consistent to the alignment model. Global alignment is then performed using both the frame to frame correspondence as well as the frame-to-reference correspondence, in three stages of progressive alignment models. A purely translational model is used at the coarsest level, an affine model is then used at the intermediate level, and finally a projective model is used for alignment. To estimate these parameters an error function relating the Euclidean distances of the frame-to-frame and frame-to-reference correspondences is minimized using the Levenberg-Marquardt Optimization.

1.2 Our Work

The focus of this paper is the registration of single frames, which can be extended easily to include multiple frames. Elevation based approaches were avoided in favor of image-based methods due to the unreliability of elevation recovery algorithms, especially in the self-correlated terrains typically encountered. It was observed that the georegistration task is a composite problem, most dependant on a robust correspondence module which in turn requires the effective handling of outliers. While previous works have instituted some outlier handling mechanisms, they typically involve disregarding some correlation information. As outliers are such a common phenomenon, the retention of as much correlation information as possible is required, while maintaining efficiency for real-time implementation. The contribution of this work is the presentation of a feature-based alignment method that searches over the entire set of correlation surface on the basis of a relevant transformation model. As the georegistration is a composite system, greater consistency in correspondence directly translates into greater accuracy in alignment. The algorithm described has three major improvements over previous works: Firstly it selects patches on the basis of their intensity values rather than through uniform grid distributions, thus avoiding outliers in homogenous areas. Secondly, relative strengths of correlation surfaces are considered, so that the degree of correlation is a pivotal factor in the selection of consistent alignment. Finally, complete correlation information retention is achieved, avoiding the loss of data by selection of dominant peaks. By searching over the entire set of correlation surfaces it becomes possible not only to handle outliers, but also to handle the 'aperture effects' effectively. The results demonstrate that the proposed algorithm is capable of handling difficult georegistration problems and is robust to outliers as well.

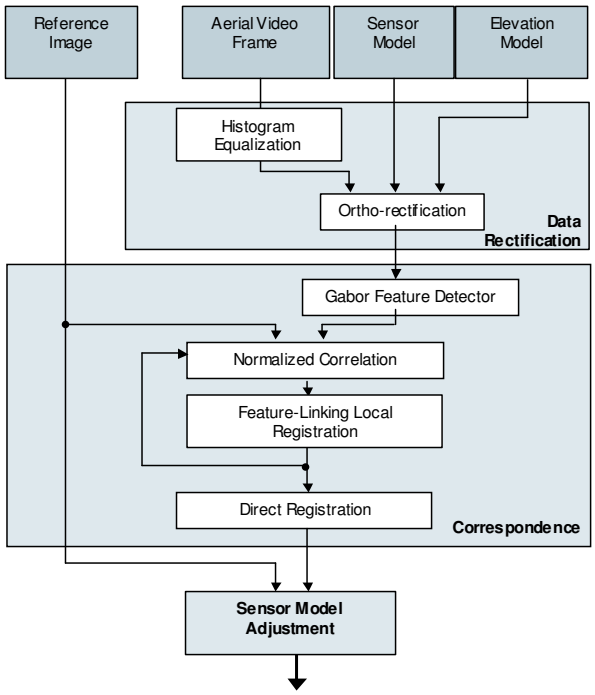


Fig. 1. A diagrammatical representation of the workflow of the proposed alignment algorithm. The four darker gray boxes (Reference Image, Aerial Video Frame, Sensor Model, and Elevation Model) represent the four inputs to the system. The three processes of Data Rectification, Correspondence and Model Update are shown as well.

The structure of the complete system is shown in Figure 1. In the first module Projection View rectification is performed by the orthographic projection of the Aerial Video Image. This approach is chosen over the perspective projection of the reference image to simplify the alignment model, especially since the camera attitude is approximately nadir, and the rate of elevation change is fairly low. Once both images are in a common projection view, feature-based registration is performed by linking correlation surfaces for salient features on the basis of a transformation model followed by direct registration within a single pyramid. Finally, the sensor model parameters are updated on the basis of the alignment achieved, and the next frame is then processed.

The remainder of this chapter is organized as follows. In Section 2 the proposed algorithm for feature-based georegistration is introduced, along with an explanation of feature selection and feature alignment methods. Section 3 discusses the sensor parameter update methods. Results are shown in Section 4 followed by conclusions in Section 5.

2 Image Registration

In this paper, alignment is approached in a hierarchical (coarse-to-fine) manner, using a four level Gaussian pyramid. Feature-based alignment is performed at coarser levels of resolution, followed by direct pixel-based registration at the finest level of resolution. The initial feature-matching is important due to the lack of any distinct global correlation (regular or statistical) between the two imageries. As a result,"direct" alignment techniques, i.e. techniques globally minimizing intensity difference using the brightness constancy constraint, fail on such images since global constraints are often violated in the context of this problem. However, within small patches that contain corresponding image features, statistical correlation is significantly higher. The selection of a similarity measure was normalized cross correlation as it is invariant to localized changes in contrast and mean, and furthermore in a small window it linearly approximates the statistical correlation of the two signals. Feature matching may be approached in two manners. The first approach is to select uniformly distributed pixels (or patches) as matching points as was used in [6]. The advantage of this approach is that pixels, which act as constraints, are spread all over the image, and can therefore be used to calculate global alignment. However, it is argued here that uniformly selected pixels may not necessarily be the most suited to registration, as their selection is not based on actual properties of the pixels intensities themselves (other than their location). For the purposes of this algorithm, selection of points was based on their response to a feature selector. The proposition is that these high response features are more likely to be matched correctly and would therefore lend robustness to the entire process. Furthermore, it is desirable in alignment to have no correspondences at all in a region, rather than have inaccurate ones for it. Because large areas of the image can potentially be textured, blind uniform selection often finds more false matches than genuine ones. To ensure that there is adequate distribution of independent constraints we pick adequately distributed local maximas in the feature space. Figure 2 illustrates the difference between using uniformly distributed points (a) and feature points (b). All selected features lie at buildings, road edges, intersections, points of inflexion etc.

2.1 Feature Selection

As a general rule, features should be independent, computationally inexpensive, robust, insensitive to minor distortions and variations, and rotational invariant. Additionally, one important consideration must be made in particular for the selection of features for remotely sensed land imageries. It has already been mentioned that terrain imagery is highly self-correlated, due to continuous artifacts like roads, forests, water bodies etc. The selection of the basic features should be therefore related to the compactness of signal representation. This means a representation is sought where features are selected that are not locally self-correlated, and it is intuitive that in normalized correlation between the Aerial and Reference Image such features would also have a greater probability of achieving a correct match. In this paper, Gabor Filters are used since they provide such a representation for real signals, [9].

Gabor filters are directional weighted sinusoidals convoluted by a Gaussian window, centered at the origins (in two dimensions) with the Dirac function. They are defined as:

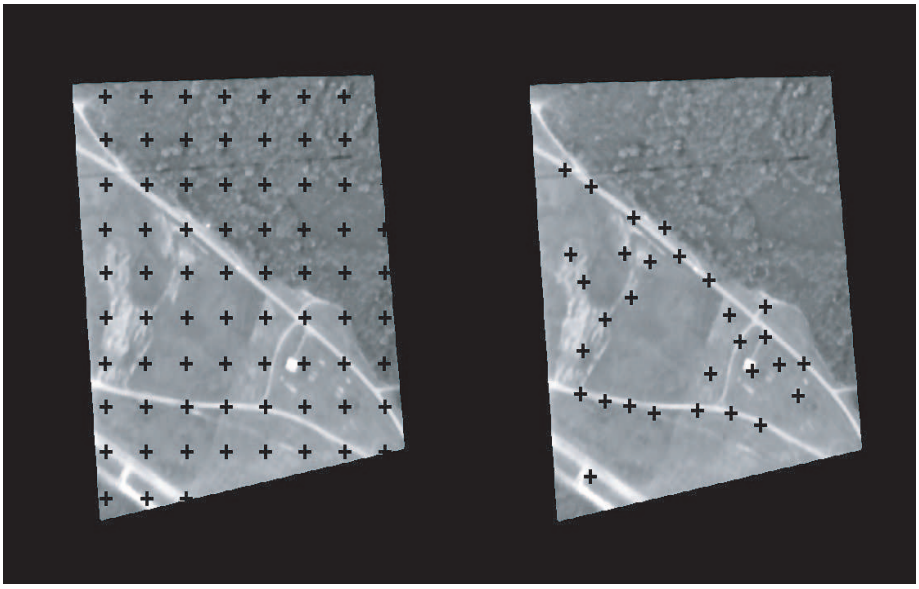


Fig. 2. Perspective Projection of the Reference Image. (a) The Aerial Video Frame displays what the camera *actually* captured during the mission . (b) Orthographic Footprint of the Aerial Video Frame on the Reference Imagery (c) The Perspective projection of Reference Imagery displays what the camera *should* have captured according to the telemetry.

$$G(x, y, \theta, f) = e^{i(f_x x + f_y y)} e^{-(f_x^2 + f_y^2)(x^2+y^2)/2\sigma^2} \tag{1}$$

where x and y are pixel coordinates, $i = \sqrt{-1}$, f is the central frequency, q is the filter orientation, $f_x = f\cos\theta$, $f_y = f\sin\theta$, and s is the variance of the Gaussian window. Figure 3 shows the four orientations of the Gabor filter were used for feature detection on the Aerial Video Frame. The directional filter responses were multiplied to provide a consensus feature surface for selection. To ensure that the features weren't clustered to provide misleading localized constraints, distributed local maximas were picked from the final feature surface. The particular feature points selected are shown in Figure 4. It is worth noting that even in the presence of significant cloud cover, and for occlusion by vehicle parts, in which the uniform selection of feature points would be liable to fail, the algorithm manages to recover points of interest correctly.

2.2 Robust Local Alignment

It is often over-looked that a composite system like georegistration cannot be any better than the weakest of its components. Coherency in correspondence is often the point of failure for many georegistration approaches. To address this issue a new transformation model based correspondence approach is presented in the orthographic projection view, however this approach may easily be extended to more general projection views and transformation models. Transformations in the orthographic viewing space are most closely modelled by affine transforms, as orthography accurately satisfies the

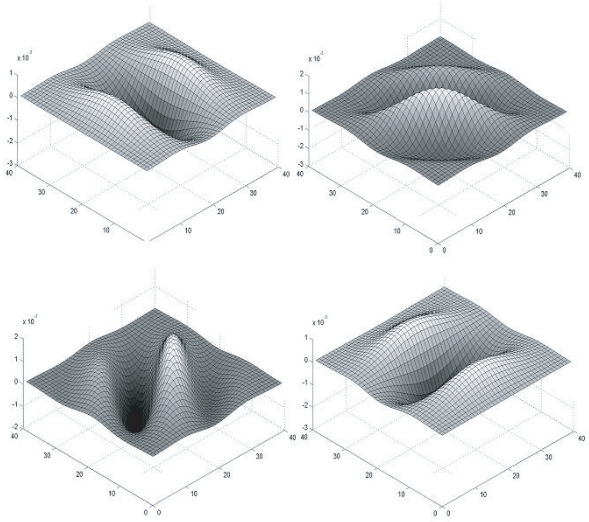


Fig. 3. Gabor filters are directional weighted sinusoidals convoluted by a Gaussian window. Four orientations of the Gabor filter are displayed.

weak-perspective assumption of the affine-model. Furthermore, the weak perspective model may also compensate for some minor errors introduced due to inaccurate elevation mapping. In general, transformation models may be expressed as

$$\mathbf{U}(\mathbf{x}) = T \cdot \mathbf{X}(\mathbf{x}) \tag{2}$$

where $\mathbf{U}$ is the motion vector, $\mathbf{X}$ is the pixel coordinate based matrix, and T is a matrix determined by the transformation model. For the affine case particularly, the transformation model has six parameters:

$$u(x, y) = a_1 x + a_2 y + a_3 \tag{3}$$

$$v(x, y) = a_4 x + a_5 y + a_6 \tag{4}$$

where u and v are the motion vectors in the horizontal and vertical directions.
The six parameters of affine transformation are represented by the vector $\mathbf{a}$,

$$\mathbf{a} = [a_1 \;\; a_2 \;\; a_3 \;\; a_4 \;\; a_5 \;\; a_6]$$

If a planar assumption (the relationship between the two images is planar) is made to simplify calculation, the choice of an orthographic viewing space proves to be superior to the perspective viewing space. All the possible transformations in the orthographic space can be accurately modelled using six parameters of the affine model, and it is easier to compute these parameters robustly compared to a possible twelve-parameter model of planar-perspective transformation (especially, since the displacement can be

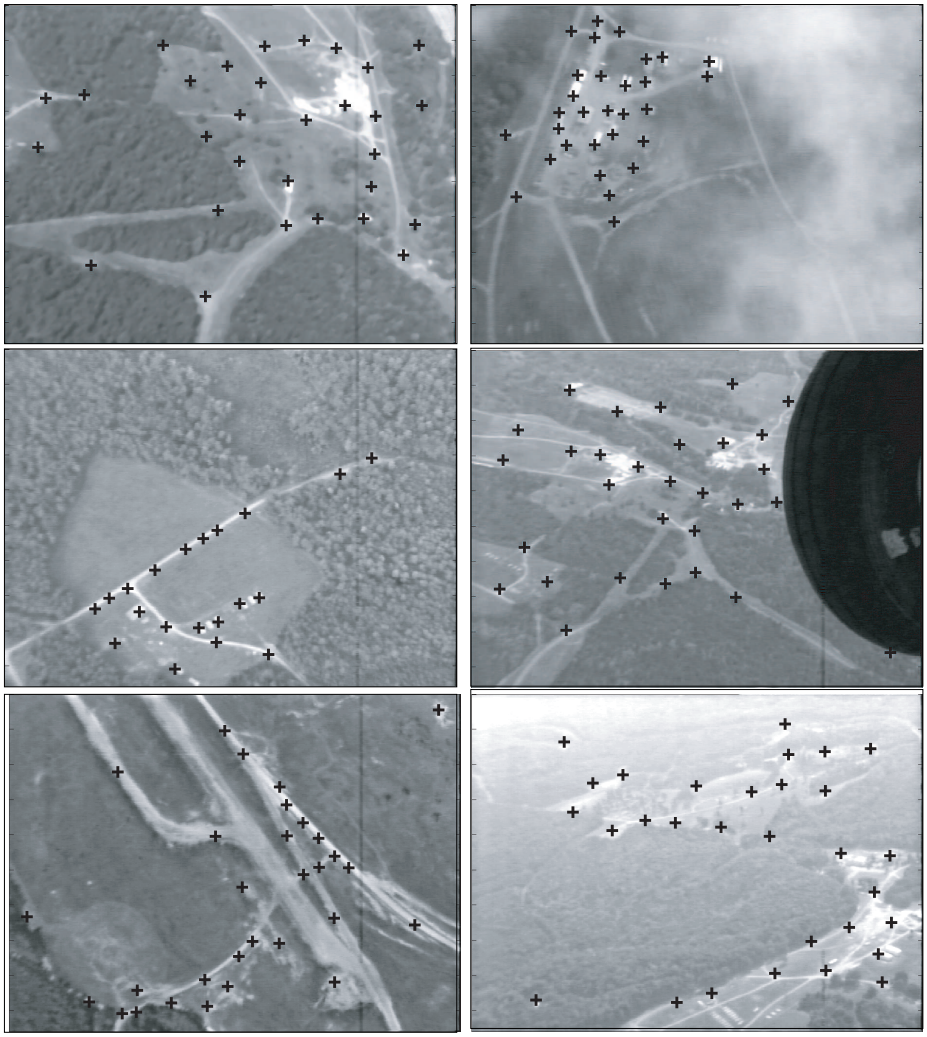


Fig. 4. Examples of features selected in challenging situations. Feature points are indicated by the black '+'s. Points detected as areas of high interest in the Gabor Response Image. Features are used in the correspondence module to ensure that self-correlated areas of the images do not contribute outliers. Despite cloud cover, occlusion by aircraft wheel, and blurring, salient points are selected. These conditions would otherwise cause large outliers and consequently leads to alignment failure.

quite significant). Furthermore, making a planarity assumption for a perspective projection view undermines the benefits of reference projection accuracy. Also, since the displacement between images can be up to hundreds of pixels, the fewer the parameters to estimate the greater the robustness of the algorithm. The affine transformation

is estimated in a hierarchical manner, in a four-level Gaussian pyramid. At the lower resolution levels, the feature-matching algorithm compensates for the large displacements, while a direct method of alignment is used at the finest resolution levels so that information is not lost.

Feature Based Alignment

The Gabor Feature Detector returns n feature points (typically set to find between ten and twenty), to be used in the feature-based registration process. A patch around each feature pixel of the Aerial Video Frame is then correlated with a larger search window from the Cropped Reference Image to yield n correlation surfaces. For T_i, the patch around a feature point, the correlation surface is defined by normalized cross-correlation. For any pair of images $I_2(\mathbf{x})$ and $I_1(\mathbf{x})$, the correlation coefficient r_{ij} between two patches centered at location (x_i, y_j) is defined as

$$r(i,j) = \frac{\sum_{w_x}\sum_{w_y}(\phi_2)(\phi_1)}{\sqrt{\sum_{w_x}\sum_{w_y}(\phi_2)^2 \sum_{w_x}\sum_{w_y}(\phi_1)^2}} \tag{5}$$

where

$$\phi_1 = I_1(\mathbf{x} + [w_x \;\; w_y]^T) - \mu_1 \tag{6}$$

$$\phi_2 = I_2(\mathbf{x} + [w_x \;\; w_y]^T) - \mu_2 \tag{7}$$

and w_x and w_y are the dimensions of the local patch around (x_i, y_j), and μ_1 and μ_2 are the patch sample means.

To formally express the subsequent process of alignment, two coordinate systems are defined for the correlation surface. Each element on a correlation surface has a relative coordinate position (u, v), and an absolute coordinate position $(x_f - u, y_f - v)$, where (x_f, y_f) is the image coordinate of the feature point associated with each surface. The relative coordinate (u, v) of a correlation element is the position relative to the feature point around which the correlation surface was centered and the absolute position of the correlation surface is the position of each element on the image coordinate axes. Each correlation element $\eta_i(u, v)$ can be considered as a magnitude of similarity for the transformation vector from the feature point coordinate (x_f, y_f), to the absolute position of the correlation element $(x_f - u, y_f - v)$. Figure 5 (b) shows the absolute coordinate system and Figure 5 (c) shows the relative positions of each correlation element. Peaks in the correlation surfaces denote points at which there is a high probability of a match, but due to the nature of the Aerial Video Frame and the Reference Image discussed earlier each surface may include multiple peaks or ridges. Now, had the set of possible alignment transformations been only translational, the ideal consensus transformation could have been calculated by observing the peak in the element-wise sum (or product) of the n correlation surfaces. This 'sum-surface' $\eta(u, v)$ is defined over the relative coordinate system as,

$$\eta(u,v) = \sum_{i=1}^{n} \eta_1(u,v) \tag{8}$$

On this 'sum-surface', by picking the translation vector in the relative coordinate system, from the center to the maximum peak the alignment transformation can be recovered. It can also be observed that since translation is a position invariant transform (i.e. translation has the same displacement effect on pixels irrespective of absolute location) the individual correlation surfaces can be treated independent of their horizontal and vertical coordinates. Therefore the search strategy for finding the optimal translational transformation across all the n correlations is simply finding the pixel coordinates (u_{peak}, v_{peak}) of the highest peak on the Sum-Surface. Put another way, a translational vector is selected such that if it were applied simultaneously to all the correlation surfaces, the sum of values of the center position would be maximized. When the vector (u_{peak}, v_{peak}) is applied to the correlation surface in the relative coordinate system, it can be observed that $\eta(0,0)$ would be maximized for

$$\eta(u,v) = \sum_{i=1}^{n} \eta_i(u',v') \tag{9}$$

where

$$u' = u - u_{peak} \tag{10}$$

$$v' = v - v_{peak} \tag{11}$$

However, even though transformations between images are dominantly translational, there usually is significant rotational and scaling as well, and therefore restricting the transformation set to translation is obstructive to precise georegistration. So by extending the concept of correlation surface super-imposition to incorporate a richer motion-model like affine, 'position-dependent' transforms like rotation, scaling and shear are included in the set of possible transformations. Once again the goal is to maximize the sum of the center position on all the correlation surfaces, only this time transformation of the correlation surfaces is not position independent. Each correlation surface, by virtue of the feature point around which it is centered, may have a different transformation associated with it. This transformation would depend on the absolute position of the element on the correlation surface rather than with its relative position as the affine set of transformations is not location invariant. An affine transform may be described by the six parameters specified in Equation 3 and 4. The objective then, is to find such a state of transformation parameters for the correlation surfaces that would maximize the sum of the pixel values at the original feature point locations corresponding to each surface. The affine parameters are estimated by directly applying transformations to the correlation surfaces. Figure 6 shows the correlation surfaces before and after transformation. It can be observed that the positions of the center of correlation surfaces i.e. $\eta(0,0)$ remain fixed in both images. In practice, window sizes are taken to be odd, and the sum of four pixel values around $\eta_i(0,0)$ are considered. The sum of the surfaces is once again expressed as in 9, where η_1 is the set of n affine-transformed correlation surfaces. This time the relationship between (u',v') and (u,v) is defined as,

$$x_f - u' = a_1(x_f - u) + a_3(x_f - u) + a_5 \tag{12}$$

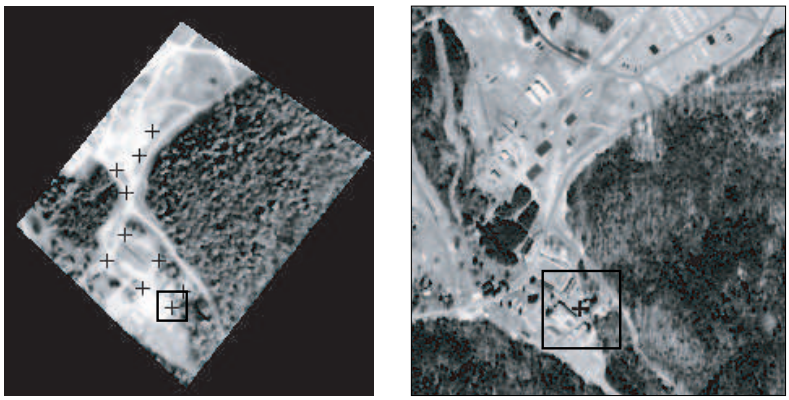
$$y_f - v' = a_2(y_f - v) + a_4(y_f - v) + a_6 \tag{13}$$

and a search is performed over a so as to maximize Thus the function to be maximized is,

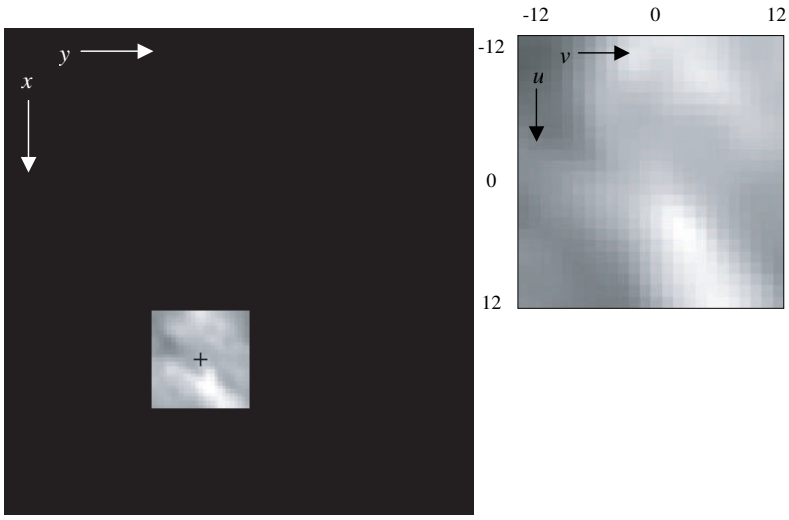
$$F(\mathbf{a}) = \eta(0, 0). \tag{14}$$

In a sense, the correlation surfaces are affine-bound together to recover the most consistent set of peaks. It should be noted that the range of the correlation surface depends on the search window size, which in turn depends on the size of the orthorectified image. This search is performed over a pyramid, and alignment recovered is propagated to the next level. The recovered alignment is also applied to feature points as they are propagated to a higher resolution level, so that correlation may be performed at each level of the pyramid. The benefit of using this hierarchical approach is that it improves computational efficiency and avoids the aliasing of high spatial frequency components that require large displacements. To visualize the entire process, consider a feature point $I_{video}(x_f, y_f)$. A patch of nine by nine pixels around $I_{video}(x_f, y_f)$ is correlated with a fifteen by fifteen pixel search window around $I_{ref}(x_f, y_f)$ to yield the correlation surface η_f. Each element $\eta_f(u, v)$ on the correlation surface is treated as a similarity measure for the vector from $I_{video}(x_f, y_f)$, to the point $I_{video}(x_f + u, y_f + v)$. When the search is performed over the affine parameters, the affine transformation is applied there are n correlation surfaces and each surface is transformed according to the absolute position of the feature point around which it was centered. The task is to find the six affine parameters such that the sum of the values at the center block in each correlation surface (or F) is maximized. Once alignment is recovered it is propagated to a higher resolution level and correlation surfaces are computed around the feature points again and the process is repeated. Maximization of F is achieved by a Quasi-Newton optimization procedure, using a finite-difference computation of the relevant derivatives. Because the positional information is maintained, every iteration places a set of points of the correlation surface onto the feature point around which each surface was initially centered. As the optimization progresses further the method moves towards a consistent set of peaks. Transformations were propagated through the three bottom levels of a Gaussian Pyramid to ensure that large displacements are smoothly captured. It is worth noting that as the set of consistent correlation peaks are being transformed to the feature point locations of the orthorectified image, it is the actually the inverse affine transformation that is computed.

The advantage of maximizing in the process detailed is three-fold. Firstly, by maintaining a 'continuous' correlation surface (rather than thresholding for peaks and performing consistency measurement on them) the most consistent set of peaks in the correlation surface is naturally retrieved. This avoids thresholding and loss of image correlation details. Secondly, by considering surfaces, relative strengths of peaks are maintained: a stronger peak holds greater weight on the overall maximization process. Thirdly, the algorithm returns the optimal affine fit, without the need for an extra consistency step. In effect, the consistency and local alignment process are seamlessly merged into one coherent module.



(a)

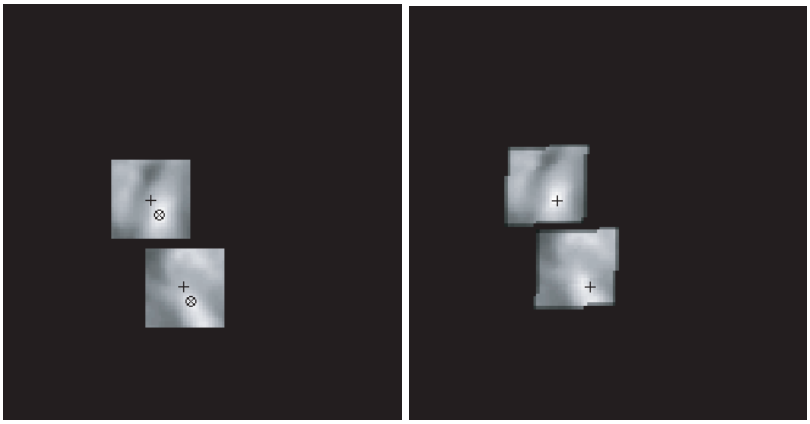


(b)

Fig. 5. The coordinate systems of Correlation Surfaces (a) The Orthorectified Image and the Cropped Reference Image. The smaller window in the Orthorectified Image is the feature patch and the larger window in the Cropped Reference Image is the corresponding search area. (b) The Absolute coordinate system for the resulting correlation. The coordinate system is (x, y) of the original image. The black '+' indicates the position of (x_f, y_f). (c) The relative coordinate system, (u, v) defining distance from the feature point (x_f, y_f) shown as the black '+' in (b). The '$\otimes$' shows the position of the peak in the correlation surface. The lack of any distinct peak should be noted, a typical phenomenon due to the differences between reference and video data.

Direction Registration

Once the Reference Image and the Aerial Video Frame have been aligned through feature-based registration, a direct hierarchical registration method is employed to



(a) (b)

Fig. 6. Absolute Position of Correlation Surfaces before and after transformation. (a) The '+'s mark the positions of the feature points. Two correlation surfaces are shown for illustrative purposes as the other overlap. The '$\otimes$' indicates the position of the dominant peak. (b) The correlation surfaces are transformed according to their absolute positions such the values at the '+'s is maximized. The position of the '+' remains the same in both (a) and (b).

provide a final adjustment. Feature based methods characteristically have a 'window' alignment, thereby losing information in the process of registration. To ensure that the whole image information is used, an affine direct registration is applied as proposed in [3] and [20]. The final transformation between the Aerial Image and the Cropped Reference Image is then the product of the affine transforms recovered from the Local Feature Match and this direct registration. As a general rule of minimization, the closer the initial estimate is to the true solution the more reliable the minimization process will be. The solution obtained after the feature-based alignment provides a close approximate to the answer that is then adjusted using this direct method. To ensure that only a fine adjustment of the feature based method is performed the direct method is implemented for a single level.

3 Sensor Update

So far two-dimensional registration of the ortho-rectified Aerial Image and the Cropped Reference Image has been achieved. The registration is performed in the orthographic viewing space, providing six affine parameters. Using this 2D alignment, it is possible to assign 3D geodetic coordinates to every pixel by simple pixel-to-pixel correspondence from the Reference Image. The final objective of this paper was to recover the adjustment to the sensor model parameters to affect alignment. However, in order to recover the sensor model's nine parameters further processing is required. It is observed that there exists no unique solution (state of sensor parameters) corresponding to any given affine transformation. The following three parameter pairs, in particular, create an infinite space of solutions: (1) The camera focal length and the

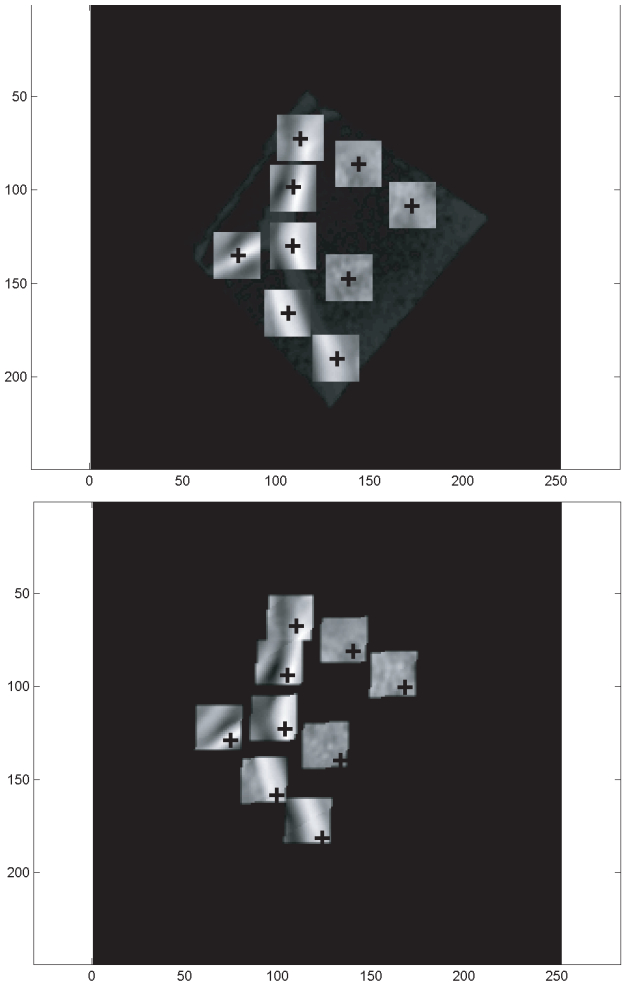


Fig. 7. The top figure show the positions of nine of the twenty feature points marked by '+'s and their correlation surfaces centered at each feature point for the second level of the matching pyramid. The bottom figure shows the results after alignment. It should be noted that the black '+'s do not change their position. After the iterations the correlation surfaces are all positioned so as to maximize the values at the feature points. It is worthwhile to note that the three surfaces in the tree textured area were 'ignored'. For illustration purposes this diagram displays only those surface that do not overlap, at the second level of a four level pyramid.

vehicle height, (2) The camera scan angle and the vehicle heading, and (3) The camera elevation and the vehicle pitch. Each one can have an approximate canceling effect, as the other in the pair, on the image captured. For instance, take the mutual effect of the camera focal length and the height of the camera. Increasing the vehicle height or decreasing the camera focal length achieves an equivalent effect of the captured image. To recover a plausible update of the sensor information two constraints are applied. Firstly, covariance information for each parameter is used while estimating the accurate updates of the sensor parameters and secondly the constraint of least change is applied in the form of a distance measure from the original sensor parameters state.

To recover the sensor adjustments, point correspondences are established between the Aerial Image and the Reference Image using the recovered 2D transformation. The Euclidean distance between those points are then minimized by searching over the nine parameters of the sensor model applying the constraints mentioned. As mentioned earlier the error function is critical to obtaining the fundamentally meaningful adjustments in the sensor geometry. The error function employed here was

$$E = \kappa_1 \Lambda(\mathbf{s}, I_{ref}, I_{video}) + \kappa_2 \Psi(\mathbf{s}, \mathbf{s}') \tag{15}$$

where

$$\mathbf{s} = [\ s_1 \quad s_2 \quad s_3 \quad s_4 \quad s_5 \quad s_6 \quad s_7 \quad s_8 \quad s_9\ ] \tag{16}$$

are the nine sensor parameters, vehicle longitude, vehicle latitude, vehicle height, vehicle roll, vehicle pitch, vehicle heading, camera scan angle, camera elevation and camera focal length, $\mathbf{s}'$ is the initial telemetry state. $\mathbf{\Lambda}$ gives the Euclidean distance between the point correspondences of the two images using the current estimate of sensor parameters. The original set of points is back-projected onto the image plane, and a search is conducted to find a state of $\mathbf{s}$ that maps the projections of the points to their matches on the ground. $\mathbf{\Psi}$ calculates the weighted Euclidean distance of each adjusted sensor parameter from the initial telemetry data (weighted on the basis of the covariance found in the telemetry). $\kappa_\mathbf{1}$ and $\kappa_\mathbf{2}$ are constants whose sum equal one, used to assign a relative importance to the two constraints.

To ensure that the solution obtained from minimization is accurate two safe-checks are employed. First, a least change constraint is placed to ensure that the solution is realistically close to the original values. Second, the covariances provided in the telemetry are used to weight the minimization process to provide unique solutions. To manually calculate the analytical expression for the Jacobian required by the optimization would probably take the better part of a week, so symbolic toolboxes of any commercial mathematics software package can be used to generate the expressions. The expressions would be the expanded form of

$$\overrightarrow{X}_{camera} = \mathit{\Pi}_t \overrightarrow{X}_{world}, \tag{17}$$

where the coordinate transformation matrix $\mathit{\Pi}_t$ is

$$
\begin{aligned}
\Pi_t = & \begin{bmatrix} \cos\omega & 0 & -\sin\omega & 0 \\ 0 & 1 & 0 & 0 \\ \sin\omega & 0 & \cos\omega & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\tau & -\sin\tau & 0 & 0 \\ \sin\tau & \cos\tau & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
& \cdots \begin{bmatrix} \cos\phi & 0 & -\sin\phi & 0 \\ 0 & 1 & 0 & 0 \\ \sin\phi & 0 & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\beta & \sin\beta & 0 \\ 0 & -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
& \cdots \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \Delta T_x \\ 0 & 1 & 0 & \Delta T_y \\ 0 & 0 & 1 & \Delta T_z \\ 0 & 0 & 0 & 1 \end{bmatrix},
\end{aligned} \tag{18}
$$

or more concisely,

$$
\overrightarrow{X}_{camera} = G_y G_z R_y R_x R_z T \overrightarrow{X}_{world}, \tag{19}
$$

where G_y is a rotation matrix in terms of the camera elevation angle ω, G_z is a rotation matrix in terms of the camera scan angle τ, R_y is a rotation matrix in terms of the vehicle pitch angle ϕ, R_x is a rotation matrix in terms of the vehicle roll angle β, R_z is a rotation matrix in terms of the vehicle heading angle α, T is the translation matrix derived from the vehicle latitude, longitude and height. Details of converting vehicle longitude and latitude to meter distances from a reference point can be found using many cartographic texts. Here it is assumed that the vehicle displacements ΔT_x, ΔT_y and ΔT_z have been computed.

4 Results

To demonstrate the algorithm presented in this paper alignment for examples are presented in this section. Despite the substantial illumination change to the extent of contrast reversal (for watery areas), examination of the results shows a significant improvement on the initial estimate. Figure 9, 10 and 11 show the initial Video Frame and Reference Imagery before and after registration. It should be noted that the image sizes are upto 1500x1500 pixels, and figures are not to scale. The misalignments are therefore appear to be scaled down as well. Visual inspection reveals a misalignment after ortho-rectification of the Video Frame using the telemetry and sensor model. Attempts at minimizing this misalignment using brightness consistency constraints fails, but with the proposed Correlation Surface Binding Algorithm proposed in this paper, accurate alignment is achieved. Figure 12 provides further examples of correct registration. White circles are marked on the top two images to highlight the corrected positions of features in the Aerial Video Frame.

The portion of the image set on which the algorithm presented did not perform accurately, were of three types. The first type was images without any features at all, like images of textured areas of trees. As there were no real features to use as constraints, the performance on these images was sub-par. The second problem faced was the aperture problem where features present were linear, and thus only a single dimensional constraint could be retrieved from them. The most convincing method of

addressing both these issue is using some form of bundle adjustment as was used in [6] and [31]. These methods were not used in this work since only video key-frames with little or no overlap were available. The last problem faced was that of occlusion by vehicle parts like tires and wings. This was addressed by calculating the fixed positions of the vehicle parts with respect to the camera in terms of the camera parameters (camera elevation angle, camera scan angle, and camera focal length). The portion of the image is then ignored or if it happened to cover too much of the image space, it is summarily rejected.

The results yielded a pre-registration average error of 39.56 pixels and a post-registration average of error 8.02 pixels per frame. As ground truth was not available to assess the error automatically, manual measurement was performed per frame. The results on a 30 key-frame clip is shown in Figure 8. The key-frames in the clip contained adequate visual context to allow single frame registration. Linear features were encountered causing some of the higher average errors reported.

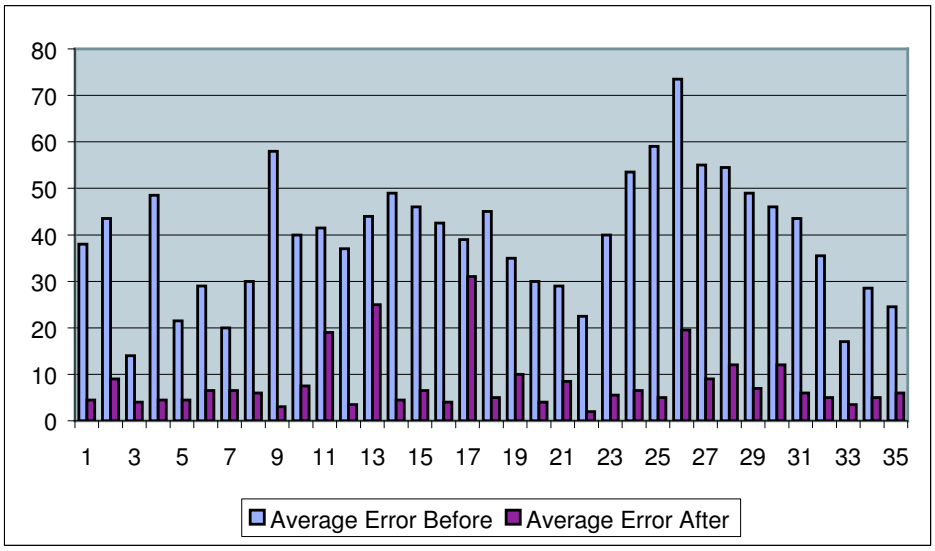


Fig. 8. Average error improvements over a 30 key-frame clip. Frame numbers are numbered along the horizontal axis, while errors in terms of number of pixels are specified along the vertical axis.

5 Conclusion

The objective of this paper was to present an algorithm that robustly aligns an Aerial Video Image to an Area Reference Image and plausibly updates the sensor model parameters, given noisy telemetry information along with elevation values for the Area reference image. The major problems tackled here were rectifying the images

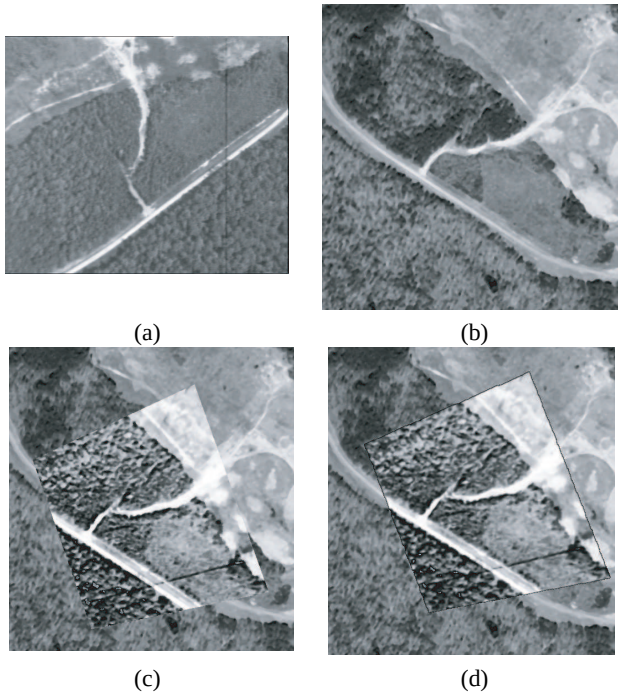


Fig. 9. Geo-Registration Results. (a) Aerial Video Frame. (b) Cropped Area of Reference Image. (c) Orthorectified Video Frame placed upon Cropped Reference Image. (d) Gross Misalignment by parametric direct Registration over a four level pyramid (using an affine transform). (e) Registration by feature linking.

by bringing them into a common viewing space, geodetic assignment for aerial video pixels, and sensor model parameter adjustment. Various forms of distortions were tackled, adjusting for illumination, compensating for texture variation, handling clouds and occlusion by vehicle parts. To achieve registration, the images are equalized and rectified into an orthographic viewing space, after which Gabor features are extracted and used to generate a normalized correlation surface per feature point. The hierarchical affine-based feature alignment provides a robust coarse registration process with outlier rejection, followed by fine alignment using a direct method. The sensor parameters are then adjusted using the affine transformation recovered and the distance of the solution from the original telemetry information. It is to be expected that the sensor data will improve with the forward march of technology, bringing with it the possibilities of more sophisticated models for the georegistration problem. Any improvement in the elevation data in particular would allow more confident use of three-dimensional information and matching. Future directions of the work include solving the initial alignment robustly in the perspective viewing space using more realistic rendering, and performing registration without continuous telemetry information.

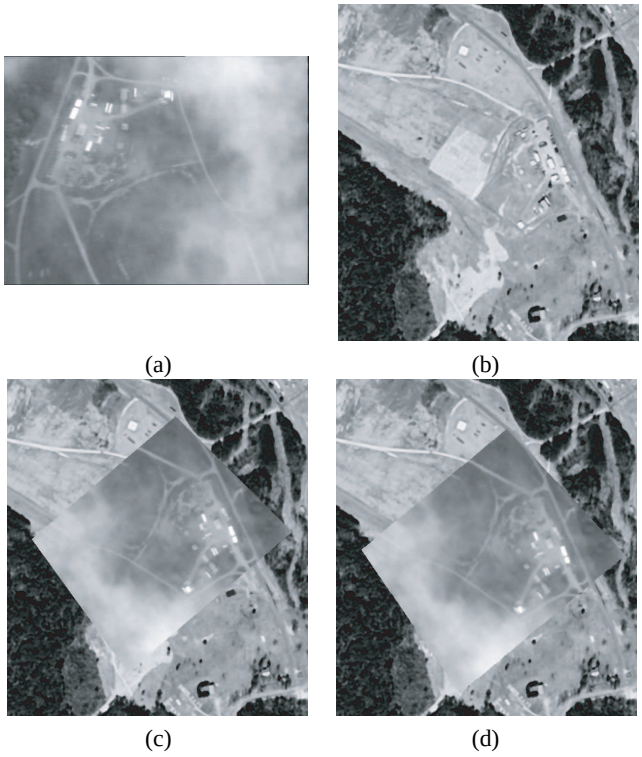


Fig. 10. Geo-Registration Results. (a) Aerial Video Frame. (b) Cropped Area of Reference Image. (c) Orthorectified Video Frame placed upon Cropped Reference Image. (d) Gross Misalignment by parametric direct Registration over a four level pyramid (using an affine transform). (e) Registration by feature linking.

References

1. P. Anandan, *"A computational framework and an algorithm for the measurement of visual motion "*, International Journal of Computer Vision, vol.2, pp. 283-310, 1989.
2. C. Baird and M. Abrarnson, *"A comparison of several digital map-aided navigation techniques"*, Proc. IEEE Position Location and Navigation Symposium, pp. 294-300, 1984.
3. J.Bergen, P. Anandan, K. Hanna, R. Hingorani, *"Hierarchical model-based motion estimation"*, Proc. European Conference on Computer Vision, pp. 237-252, 1992.
4. L. Brown, *"A Survey of Image Registration Techniques"*, ACM Computing Surveys, 24(4), pp. 325-376, 1992.
5. Y. Bresler, S. J. Merhav, *"On-line Vehicle Motion Estimation from Visual Terrain Information Part II: Ground Velocity and Position Estimation"*, IEEE Trans. Aerospace and Electronic System, 22(5), pp. 588-603, 1986.
6. R. Cannata, M. Shah, S. Blask, J. Van Workum, *"Autonomous Video Registration Using Sensor Model Parameter Adjustments"*, Applied Imagery Pattern Recognition Workshop, 2000.
7. P. Curran, *"Principles of Remote Sensing"*, Longman Group Limited, 1985.

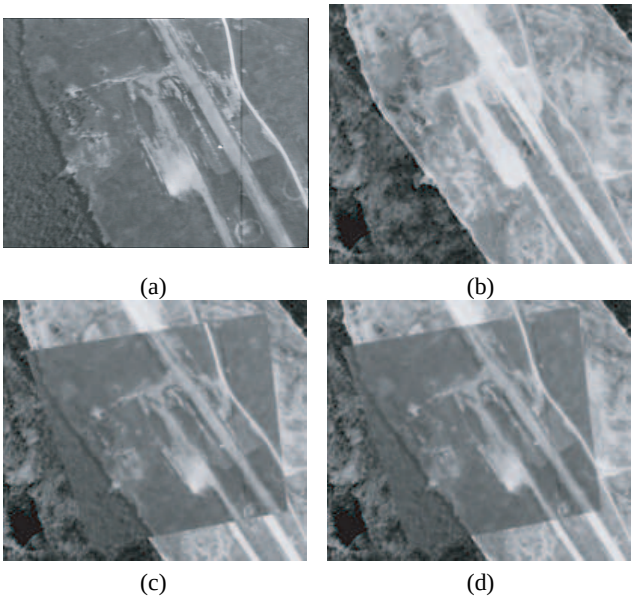


Fig. 11. Geo-Registration Results. (a) Aerial Video Frame. (b) Cropped Area of Reference Image. (c) Orthorectified Video Frame placed upon Cropped Reference Image. (d) Gross Misalignment by parametric direct Registration over a four level pyramid (using an affine transform). (e) Registration by feature linking.

8. J. Foley, A. van Dam, S. Feiner, J. Highes, *"Computer Graphics, Principles and Practices"*, Addison-Wesley, 1990.
9. D. Gabor, *"Theory of Communications"*, IEEE Communications, No. 26, 1946, pp. 429-459.
10. J. P. Golden, *"Terrain Contour Matching (TERCOM): A cruise missile guidance aid"*, Proc. Image Processing Missile Guidance, vol. 238, pp. 10-18, 1980.
11. V. Govindu and C. Shekar, *"Alignment Using Distributions of Local Geometric Properties"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(10), pp. 1031-1043, 1999.
12. K. Hanna, H. Sawhney, R. Kumar, Y. Guo, S. Samarasekara, *"Annotation of video by alignment to reference imagery"*, IEEE International Conference on Multimedia Computing and Systems, vol.1, pp. 38 - 43, 1999.
13. B. Horn, B. Schunk, *"Determining Optical Flow"*, Artificial Intelligence, vol. 17, pp. 185-203, 1981.
14. S. Hsu, *"Geocoded terrestrial mosaics using pose sensors and video registration"*, Computer Vision and Pattern Recognition, 2001. vol. 1, pp. 834 -841, 2001.
15. http://ams.egeo.sai.jrc.it/eurostat/Lot16-SUPCOM95/node1.html
16. M. Irani, P. Anandan, *"Robust Multi-Sensor Image Alignment"*, International Conference on Computer Vision, 1998.
17. B. Kamgar-Parsi, J. Jones, A. Rosenfeld, *"Registration of multiple overlapping range images: scenes without distinctive features"*, Computer Vision and Pattern Recognition, pp. 282-290, 1989.

18. R. Kumar, H. Sawhney, J. Asmuth, A. Pope, S. Hsu, *"Registration of video to geo-referenced imagery"*, Fourteenth International Conference on Pattern Recognition, vol. 2. pp.1393-1400, 1998.
19. B. Lucas and T. Kanade. *"An Iterative image registration technique with an application to stereo vision"*, Proceedings of the 7th International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.
20. S. Manna nd R.W. Picard, *"Video orbits of the projective group a simple approach to featureless estimation of parameters"*, IEEE Transactions on Image Processing, 6(9) , pp. 1281 -1295, 1997.
21. S. J. Merhav, Y. Bresler, *"On-line Vehicle Motion Estimation from Visual Terrain Information Part I: Recursive Image Registration"*, IEEE Trans. Aerospace and Electronic System, 22(5), pp. 583-587, 1986.
22. J. Le Moigne, N. Netanyahu, J. Masek, D. Mount, S. Goward, M. Honzak, *"Geo-registration of Landsat Data by robust matching of wavelet features"*, Proc. Geoscience and Remote Sensing Symposium, IGARSS, vol.4, pp. 1610-1612, 2000.
23. J. Nocedal, S. Wright, *"Numerical Optimization"*, Springer-Verlag, 1999.
24. J. Rodriquez, J. Aggarwal, *"Matching Aerial Images to 3D terrain maps"*, IEEE PAMI, 12(12), pp. 1138-1149, 1990.
25. D.-G. Sim, S.-Y. Jeong, R.-H. Park, R.-C. Kim, S. Lee, I. Kim, *"Navigation parameter estimation from sequential aerial images"*. Proc. International Conference on Image Processing, vol.2, pp. 629-632, 1996.
26. D-G. Sim, R-H Park, R-C. Kim, S. U. Lee, I-C. Kim, *"Integrated Position Estimation Using Aerial Image Sequences"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(1), pp. 1-18, 2002.
27. R. Szeliski, *"Image mosaicing for tele-reality applications"*, IEEE Workshop on Applications of Computer Vision, pp. 44-53, 1994.
28. Y. Sheikh, S. Khan, M. Shah, R. Cannata, *"Geodetic Alignment of Aerial Video Frames"*, In Video Registration, Video Computing Series, KLUWER Academic Publisher, 2003.
29. R. Szeliski, H. Shum, *"Creating Full View Panoramic Image Mosaics and Environment Maps"*, Computer Graphics Proceedings, SIGGRAPH, pp. 252-258, 1997.
30. P. Viola and W. M. Wells, *"Alignment by maximization of mutual information."*, International Journal of Computer Vision, 24(2) pp. 134-154, 1997.
31. R. Wildes, D. Hirvonen, S. Hsu, R. Kumar, W. Lehman, B. Matei, W.-Y. Zhao *"Video Registration: Algorithm and quantitative evaluation"*, Proc. International Conference on Computer Vision, Vol. 2, pp. 343 -350, 2001.
32. Q. Zheng., R. Chellappa. *"A computational vision approach to image registration"*, IEEE Transactions on Image Processing, 2(3), pp. 311 326, 1993.

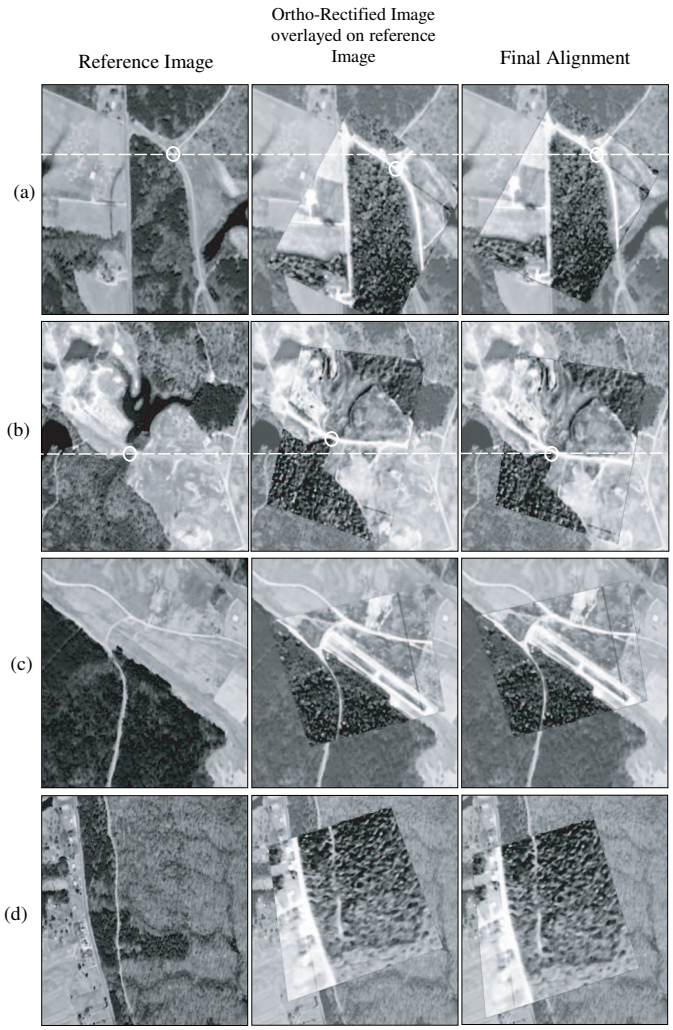


Fig. 12. (a)-(d) The leftmost image is the Cropped Reference Image, the middle image is the Orthorectified Image overlayed onto the Reference Image, and the rightmost image is the Final Registered Image. The white circles highlight initially misaligned features.