Autonomous Video Registration Using Sensor Model Parameter Adjustments

Richard W. Cannata[1], Mubarak Shah[2], Steven G. Blask[1], John A. Van Workum[1]

[1] *Harris Corporation, GCSD, P.O. Box 37, Melbourne, FL 32902*
[2] *School of EECS, University of Central Florida Orlando, FL 32816*

Abstract

Recently, airborne video surveillance platforms have gained greater acceptance for use in a variety of DoD missions due to their utility, affordability and autonomy. While a variety of airborne collectors and unmanned aerial vehicles may be equipped for video surveillance to support diverse mission needs, ground processing systems cannot handle this high data rate medium without some degree of autonomous processing to simplify and extend the exploitation process for video imagery.

In this paper, we outline a novel approach for near-real-time video registration based on sensor model parameter adjustments and the application of a Kalman filter. The goal of our Precision Video Registration (PVR) development is to register video with a reference image to provide accurate 3-D geolocations. Our sensor-based 3-D treatment is unique since most registration approaches employ only simple image-to-image mappings, such as affine transformations. In our approach, we explicitly model the projections between the 3-D world and 2-D images and perform registration in 3-D with greater accuracy and fidelity. PVR performance results show significant accuracy improvement over unregistered frame geolocation, and autonomously generated video mosaics appear smooth and seamless.

1. Introduction

The availability of low-cost, lightweight video camera systems, high-bandwidth VHF communications links and a growing inventory of DoD unmanned aerial vehicles (UAVs) has resulted in dramatic new opportunities to conduct remote battlefield surveillance and reconnaissance missions from the relative safety of distant ground stations. As with any new remote sensing capability, advances in capabilities also present new challenges in data processing and exploitation. In the case of video, streaming data is difficult to exploit and situational context is difficult to achieve due to the narrow fields of view and imprecise accuracy of camera pointing.

In this paper, we describe a new video processing capability that overcomes these limitations by autonomously georegistering sequences of video imagery, in near real time, to reference imagery having high geodetic accuracy. The resulting process yields accurate, georeferenced video frames that can be easily displayed as mosaiced products, reprojected onto maps or other frames for situational context, or quickly exploited to derive high-precision 3-D geolocations of objects within each frame. These capabilities will provide analysts and decision-makers with greater situational awareness and high-accuracy geolocation.

Image registration is the process of establishing correspondence between two or more images [1]. It is often the critical prerequisite step for exploitation of information contained within video image sequences. Generally, the goal of any registration algorithm is to find the best transformation, T, which relates two images, M_1 and M_2. Images are said to be registered when $TM_1 = M_2$ for all pixels within the same scene. Extending this principle to video registration, we solve for the global transformation, which relates a mission video frame, M_1 to a reference image, R. The georegistration process transfers the high geodetic accuracy of controlled satellite reference imagery to the more recent mission video imagery of a tactical UAV.

To solve for correspondence in scenes collected at different times or from different perspectives, a video registration process must be robust, agile and capable of accommodating diverse sensor capabilities, limited fields of view, and considerable perspective distortion. When descriptive information about the camera orientation is not specified, the registration problem is most complex. However, when additional information is available, the registration problem may be simplified and more accurate results obtained.

Our video georegistration approach uses *a priori* knowledge of the sensor and a digital elevation model (DEM) approximation in addition to scene (pixel) information to derive a robust 3-D transformation mapping between sequences of frames. This approach is described by Bryan *et al.* [2] for satellite image registration and later extended to a fully autonomous data-driven solution by Hackett *et al.* [3] to support autonomous geopositioning solutions for fixed frame aircraft imagery. Their georegistration treatment solved the pixel correspondence problem through the use of an area-based registration approach, whereby each image to be registered has candidate match regions projected to a common three-dimensional surface using model-based photogrammetric principles and *a priori* knowledge of sensor state variables.

This model-based approach has distinct advantages over the more common transformation treatments such as affine [4] or perspective in determining the transformation function since the overlap areas can be corrected for different relief distortions and will therefore yield more accurate transformation solutions. Approaches using affine transformations, or any other polynomial transformations, produce approximate registration solutions since they employ only 2-D pixel information. These treatments produce acceptable results for airborne video collections over flat terrain however they fail to produce accurate scene alignments in areas of high terrain relief where large and/or complex relief distortions are present in each frame. In the model-based approach, knowledge about the image acquisition physics and collection geometry are used to derive a more accurate transformation between 2-D pixel space and the 3-D surface of the Earth. In the present context, metadata (“telemetry data”), describing the sensor attitude, position and velocity, is recorded concurrently with the video stream and used with a DEM to align each video frame to a reference image.

In this paper, we describe extensions and refinements to the approach by Hackett *et al.* [3] that permit fully autonomous, near-real-time, 3-D georegistration of sequences of video frames. In Section 2, we describe the video registration process in terms of match point (correspondence) selection, sensor parameter adjustment strategies and adjustment strategies using a Kalman filter. In Section 3, we summarize a near-real-time processing architecture for ingest, registration and output of georegistered video frames and sequences. We discuss our preliminary quantitative performance assessment in Section 4 and provide conclusions in Section 5.

2. Core registration process

The registration process for Precision Video Registration (PVR) consists of two distinct steps. The first is the autonomous generation of correspondence points between the video frame and a referenced image. The second is adjustment of the video frame sensor parameters so that it is aligned with the reference image. After registration, the accuracy of the reference image has been transferred to the video image and accurate estimates of ground point locations can be made directly using the adjusted video frame. In addition, *a posteriori* covariance of the parameters is available and may be used to estimate the accuracy of the estimated ground point.

An accurate, rigorous sensor model is essential to the successful registration of the video and reference image. The sensor model developed for the PVR program completely models all parameters of the PVR sensor. It includes all telemetry parameters and all static parameters of the installation. The telemetry parameters include the aircraft position given in latitude, longitude and height, as well as the attitude parameters of heading, roll and pitch. In addition, the camera gimbal readings for sensor azimuth and elevation are included, as well as the camera focal length. The static installation parameters include the orientation of the camera gimbal coordinate system with respect to the aircraft coordinate system, the displacement between the camera and the aircraft position sensor, as well as the size of the camera detector and the focal length of the lens. The model includes a capability to model the distortions of the lens, but that was not included for this study.

The model provides the basic ability to map from image space pixel coordinates of line and sample to ground space coordinates of latitude, longitude and height. The ground space may be a surface at a constant height above the earth ellipsoid, or a detailed digital elevation model (DEM).

An advanced feature of the sensor model is the ability to compute estimates of the error associated with the ground points. The model includes an *a priori* covariance matrix for all the parameters, and methods to propagate those covariances into estimates of the ground point errors. After registration, the covariance matrix may be updated to reflect the *a posteriori* values that may be used for error propagation.

A reliable, robust correspondence point generation process is essential to successfully register video images to a reference image. For PVR, we use a process of normalized cross correlation of the video with the reference image to obtain these correspondence points. In this process, the images are first orthorectified then reduced in resolution so that a grid of 16 (or more) 32x32-pixel patches cover the overlap area between images. Orthorectification is the process of projecting the imagery onto a DEM and synthesizing an overhead view. The effects due to depth variations and viewing aspect are minimized by this process so that correspondence may be more easily determined. Each patch in the video frame is correlated against a corresponding patch from the image in orthorectified space. The size of the reference patch is larger than the video patch and is determined by the uncertainty in the location of the ground points from the video frame. Usually the reference patch is 2 to 3 times larger than the video frame patch.

The correlation process produces a correlation surface that is irregular in shape and usually has multiple peaks. We select up to four of the strongest peaks from each patch and save them for later processing. We need some way of finding the subset of peaks, from among all patches, that represents the correct correspondence points between the video and reference images. Figure 1 illustrates the problem. The "+" signs in the figure identify match points in the video frame. Each "+" sign is labeled with a match point ID. The lines identify locations of the correlation peaks that indicate candidate coordinates for the same features in the reference image. Each match point has from zero to four peak locations.

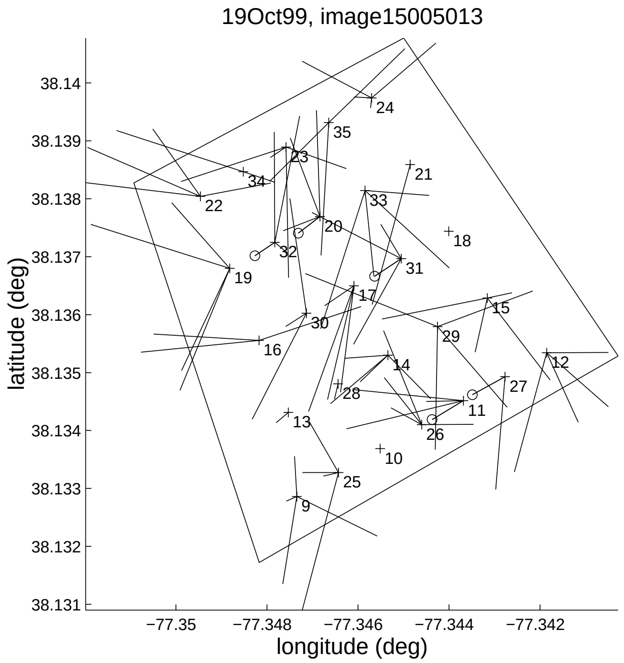

Figure 1. Candidate correspondence points

In PVR, we select the best subset of peaks that can be represented by a four parameter affine transform between the locations of the "+" signs and the end of the lines.

$$\mathbf{y}' = sA\mathbf{y} + \mathbf{b}$$

$$A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

where:

$\mathbf{y}$ is the location of the reference point.
θ is the rotation angle of the transform.
$\mathbf{b}$ is the translation of the transform.
s is the scale change of the transform

In practice, we find several different subsets with this process, but the one with the most points is invariably the best choice. In Figure 1, the selected subset is indicated with circles drawn at the end of the line. In this case, only 5 consistent match points were selected. Comparing the lengths of the lines at 32 and 20 with those at 26 and 27 provides evidence for image rotation.

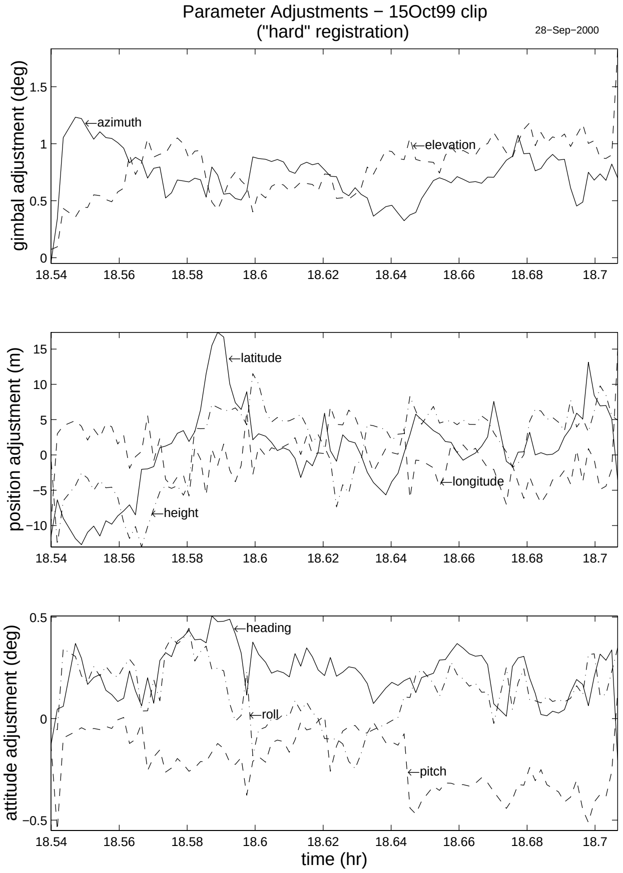

Figure 2. Sensor parameter adjustments

After the match points have been obtained, the sensor parameters may be adjusted to bring the video frames into alignment with the reference image. There are numerous ways to do this. One method would be to perform a least-squares triangulation adjustment of the video frame parameters. For PVR, the image quality difference between the video and the reference often results in very few match points with large errors in them. This would produce unreliable registrations. One way to deal with this problem would be to include several frames in the solution and use both video-to-reference match points and video-to-video tie points. This will produce better adjustments for the images, but the disadvantage is that we have to wait until all frames are available in order to make the adjustments.

For PVR, we are using a Kalman filter to adjust the images. This has the effect of using multiple video frames in a solution, but it allows us to process a frame as soon as the match points are available, thus reducing latency. In effect, we are averaging over time using the state model of the Kalman filter rather than spatial averaging using the triangulation approach. The Kalman filter equations are:

$$\begin{aligned}
\mathbf{x}(t',t) &= \phi(t'-t)\mathbf{x}(t,t) \\
P(t',t) &= \phi(t'-t)P(t,t)\phi(t'-t)^T + Q(t',t) \\
k(t',t) &= P(t',t)H(t',t)^T[H(t',t)P(t',t)H(t',t)^T + R(t',t)]^{-1} \\
\mathbf{x}(t',t') &= \mathbf{x}(t',t) + k(t',t)[\mathbf{z}(t') - H(t',t)\mathbf{x}(t',t)] \\
P(t',t') &= [I - k(t',t)H(t',t)]P(t',t)
\end{aligned}$$

where:

$\mathbf{x}$ is the state vector
P is the covariance of the state vector
k is the Kalman gain
$\mathbf{z}$ is the observation vector
R is the covariance of the observations
ϕ is the state transition matrix
H is the state to observation Jacobian

The state model chosen for PVR represents the sensor parameter adjustments as constant in time. This allows the filter to track slowly varying effects, such as GPS errors, aircraft flexure, bias errors in the installation of the equipment, sensor drift, etc. High frequency effects, such as aircraft vibration and turbulence, would not be modeled. The state vector then is a nine-parameter vector containing adjustments to aircraft position (3 parameters) and attitude (3 parameters), gimbal readings (2 parameters), and camera focal length (1 parameter). The state transition matrix for this case is simply an identity matrix. Initial values for the covariance matrix, P, were obtained from an analysis of the results of triangulation with manually dropped calibration points on 500+ video frames.

The observation vector, **z**, contains the line, sample values of video frame correspondence points obtained during the match point generation process. Ground points for these image points are obtained by propagating the corresponding reference image points to ground. R is the covariance of the video image points computed during match point generation. We refer to the collection of video frame correspondence points, ground points and covariances as a “Match Point Observation.”

Figure 2 shows an example of the image adjustments for a set of video frames collected over the VA site on 15 Oct 99. The individual frames were extracted about 6 seconds apart. The horizontal axis on the plots is time of day in hours. The adjustments show some fluctuation, but the frame-to-frame changes are generally small.

Figure 3 shows plots of the variance of the adjustments as a function of time. The plots start at near the *a priori* values for the adjustments and rapidly converge to smaller values. There are some small oscillations in the values, which were found to correspond to times when the aircraft was turning. Given the nature of the constant state model, it is reasonable to expect the filter to have bigger errors when the aircraft is turning, gradually catching up again when the aircraft is flying straight and level.

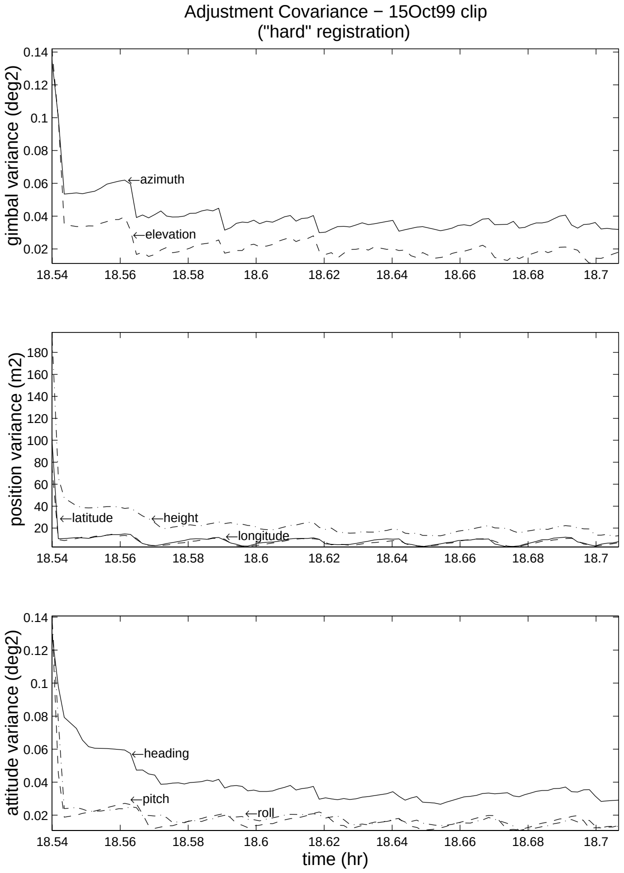

Figure 3. **Sensor parameter adjustment covariance**

3. PVR processing threads

Figure 4 illustrates the processes that comprise the core Real Time Video Registration (RTVR) architecture. At the heart of the system is a dynamically updated Telemetry Database that resides in shared memory so that all processes in the PVR system have real-time, direct access to the latest available data. A raw telemetry ingest process unpacks the streaming telemetry data read from the decoder hardware and logs it in the Telemetry Database tables. Raw telemetry packets for aircraft position, attitude, gimbal pointing, and sensor parameters arrive at different rates, and the “queue” nature of the database tables keeps the information sorted in time-order if packets arrive out of sequence. The adjustment table of the Telemetry Database logs sensor model parameter adjustments produced by RTVR, currently at 0.22 Hz rate for the SGI Onyx2 Infinite Reality with 6 Match Point Observation Generators running simultaneously.

The RTVR Process Control application monitors the Telemetry Database and chooses frames for registration processing based on its current frame select strategy. Recall that our core registration process has two steps: correspon-

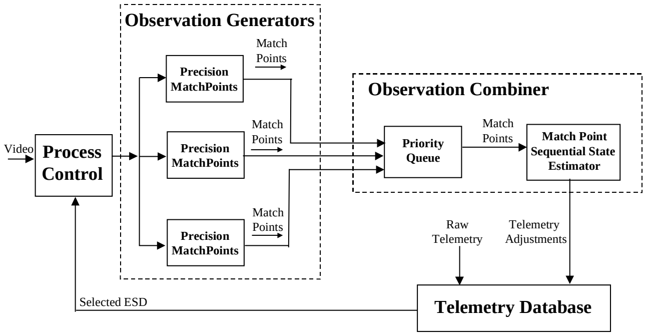

Figure 4. RTVR architecture

dence determination followed by sensor model adjustment. Correspondence determination is the processing bottleneck, so a bank of match point Observation Generators processes N frames in parallel to meet the desired frame-registration rate of 1 Hz. Since the processing time for an individual frame depends somewhat on scene content, a priority queue within the match point Observation Combiner process ensures that the frames are re-sorted into time order before processing by the Kalman filter adjustment algorithm. The Observation Combiner logs its telemetry corrections in the adjustment table of the Telemetry Database.

The Precision Video Registration (PVR) system architecture of Figure 5 is designed to accommodate the real-time streaming of video images and telemetry support data from the Common Air/Ground System (CAGS) services into the PVR processing modules in support of PVR client services. This architecture produces an asynchronous flow of improved telemetry to PVR clients and also supports *ad hoc* requests for georeferenced orthomosaics, precision geolocation results and improved telemetry for specified video frames. All of these user services depend on the adjusted telemetry that is the primary output product of the RTVR subsystem.

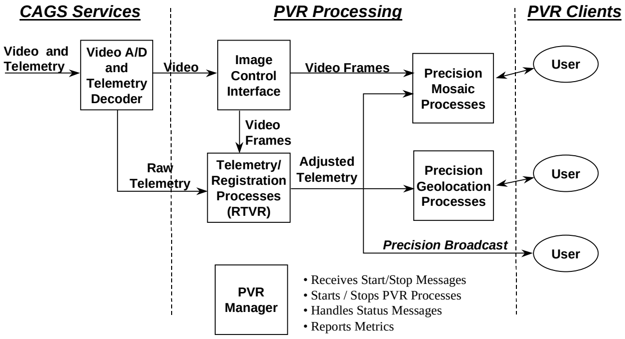
CAGS Services
PVR Processing
PVR Clients
Video and Telemetry
Video A/D and Telemetry Decoder
Video
Image Control Interface
Video Frames
Precision Mosaic Processes
User
Video Frames
Raw Telemetry
Telemetry/ Registration Processes (RTVR)
Adjusted Telemetry
Precision Geolocation Processes
User
Precision Broadcast
User
PVR Manager
• Receives Start/Stop Messages
• Starts / Stops PVR Processes
• Handles Status Messages
• Reports Metrics

Figure 5. PVR system architecture

The processes implementing the RTVR subsystem and the Precision Mosaic, Precision Geolocation, and Precision Broadcast user service subsystems are under the control of the PVR Manager application. The inter-process communication and control mechanism uses the well-known Parallel Virtual Machine (PVM) framework and messaging protocol. This protocol is also used within the RTVR architecture of Figure 4.

Telemetry processing is the essence of our video image registration paradigm. Our registration approach produces adjusted telemetry data as its output, as opposed to new registered image raster files. The telemetry support data allows the sensor model to define a 3-D ray through any pixel in the image, which may be intersected with a DEM to produce a geolocation or orthorectify a video frame. The adjusted telemetry produced by RTVR improves the geodetic accuracy of pixels undergoing this transformation process. Examples of orthophoto mosaics produced by PVR and quantitative characterization of PVR geolocation accuracy are presented in the next section.

4. Performance evaluation

For performance evaluation purposes, we tested our precision video registration processor against a variety of collection conditions, scene content and terrain relief. Our goal is to collect and process the video sequences as they might be obtained under real-world surveillance conditions so that performance characterization may ultimately be used to develop a predictive performance model. The field collection matrix of collection locations, imaging conditions, terrain and ground cover descriptions is shown in Table 1.

Table 1. Collection locations and conditions

Look Angle	Terrain	Low Relief (flat)	Med Relief (hilly)	High Relief (Mountains)
	GSD			
High >55°	> 3 m	VA Site	NY Site	NV Site
	< 3 m	VA Site	NY Site	NV Site
Mod. 55° –35°	> 3 m	NC/VA Site	NY Site	NV Site
	< 3 m	NC Site	NY Site	NV Site
Low <35°	> 3 m	VA Site	NY Site	NV Site
	< 3 m	NC Site	NY Site	NV Site

All data were collected using a DeHavilland DHC-6 Twin Otter aircraft operated by the U.S. Army's Night Vision and Electronic Systems Directorate. The aircraft was equipped with a Wescam 14-inch Skyball gimbal with EO and FLIR video cameras, a Litton LN-100G integrated GPS/INS unit, an IRIG B-compliant VITC time code generator, and an onboard SGI Octane with two R10000 250 MHz processors for collecting and recording video and telemetry data to digital tapes. All registration processing was performed off board the aircraft at our video processing laboratory, although a similar PVR processor has been tested and delivered to the Government for field deployment. The video data were processed on an SGI R10000 processor under the IRIX6.5 operating system. Since all field collections have not yet been processed, we report on initial representative results for the NY Site and the NC Site overflights. In all cases, we autonomously register each video frame to a precision controlled image for geodetic accuracy. These standard products are available through the DoD (satellite imagery) or the USGS (7 arc minute Digital Ortho Quads).

By comparing initial (unregistered) alignment error against post-registration alignments, we obtain a measure of accuracy improvement in terms of Euclidean distance. Alignment errors for initial and post registration results are derived manually by first projecting the video frame and reference image to common screen coordinates based on support data. Next, we identify 3-8 (average 5) salient features in each video frame and the reference image. When fewer than 3 points are present due to sparse scene content, these cases are not included in our statistics. Ground point estimates for the geodetic coordinates of each salient point are computed based on the reference imagery and are treated as truth. Video image points are projected to ground space using raw and adjusted telemetry. Residual misalignment in ground space is averaged for all points and reported as the alignment error for each frame.

The New York (NY) site collection occurred on 23-24 February 2000, and the North Carolina site collection occurred 28-31 March 2000. In all, the aircraft collected approximately 25 hours of EO and IR video data, which were later partitioned by resolution, collection (grazing) angle, scene content, terrain and collection mode. Video clips, consisting of 2 minutes of continuous video frames, were selected after sorting by terrain, GSD and look angles and then processed for analyses. Due to the laborious nature of the manual analysis required for error characterization (described above), only a few clips for each terrain and imaging condition were selected for analysis. Nevertheless, we used over 700 frames of autonomously registered video, manually evaluated and measured for accuracy, to statistically quantify our registration accuracy.

For the NY site video collection, we processed 4 data sets. One representative example is shown in Figure 6. For this event, the aircraft slowly circled while the sensor stared at a road intersection. The grazing angles varied between 35^o and 55^o, and the pixel resolution ranged from 1 to 2 meters depending on varying slant range. The terrain was hilly, with bare trees and shrubs visible. Most notably, the ground was covered with snow to a depth of 0.1-1.0m, except for several cleared roads. For this clip, the reference imagery was collected 9 years earlier during Summer under full foliage and no snow cover conditions.

The initial alignment error (based on manually specified ground truth) shown in Figure 6, reflects random and systematic errors in the sensor state measurements used for the initial projection. For this aircraft, gimbal azimuth and gimbal elevation uncertainties in the support data appear to contribute the most to initial alignment errors. Both uncertainty elements produce large translation error (i.e., 10-100s meters) when used to project the video frame to ground coordinates. Rotational error, due principally to unreported pitch or roll components (for highly squinted cases), is present as well, but to a far smaller degree than translation error. We reduce both error types through the registration process by allocating correlation-derived adjustments to the sensor state parameters (gimbal, aircraft location and orientation) and reprojecting the video frame to a new position. The “after registration” alignment error is derived by measuring the average distance between common salient features as observed in the reference and the reprojected video frame positions as described above.

Despite large differences in scene content between the reference and video imagery, relative error was reduced by a factor of 2.5. Alignment errors were reduced 95% of the time with respect to unregistered video frames with an average residual error, after registration, of approximately 12 meters. In three cases, the registration process failed to make improvements. A closer examination of these failures identified two factors, operating simultaneously, that caused the registration solution to converge on a weak correspondence solution. Low scene content resulted in a minimum number of match points as a snow- covered wooded area filled the sensor field of view. The presence of long linear features, in this case a road, also offers weaker constraints since only displacement in the normal direction can be solved for. Linear alignment of match points (“correlation saddle”) may thus result in ambiguous solutions. This problem, also referred to as “aperture effect”, was responsible for almost half of our registration failures and appears whenever dense forest and single linear (road) features were present.

Figure 7 illustrates the results for a video registration sequence for a different video and reference dataset collected in N.C. This video was collected on 28 March 2000 over a flat urban setting at approximately 4000 ft. AGL. Grass and low shrub vegetation were present throughout the imagery. The aircraft flew in a nearly straight path with active gimbal slewing present as the video camera scanned across a 35-55 degree range of grazing angles. Resolution remained approximately constant at 1-meter ground sample distance (GSD). The reference image used in this registration sequence was approximately 8 years

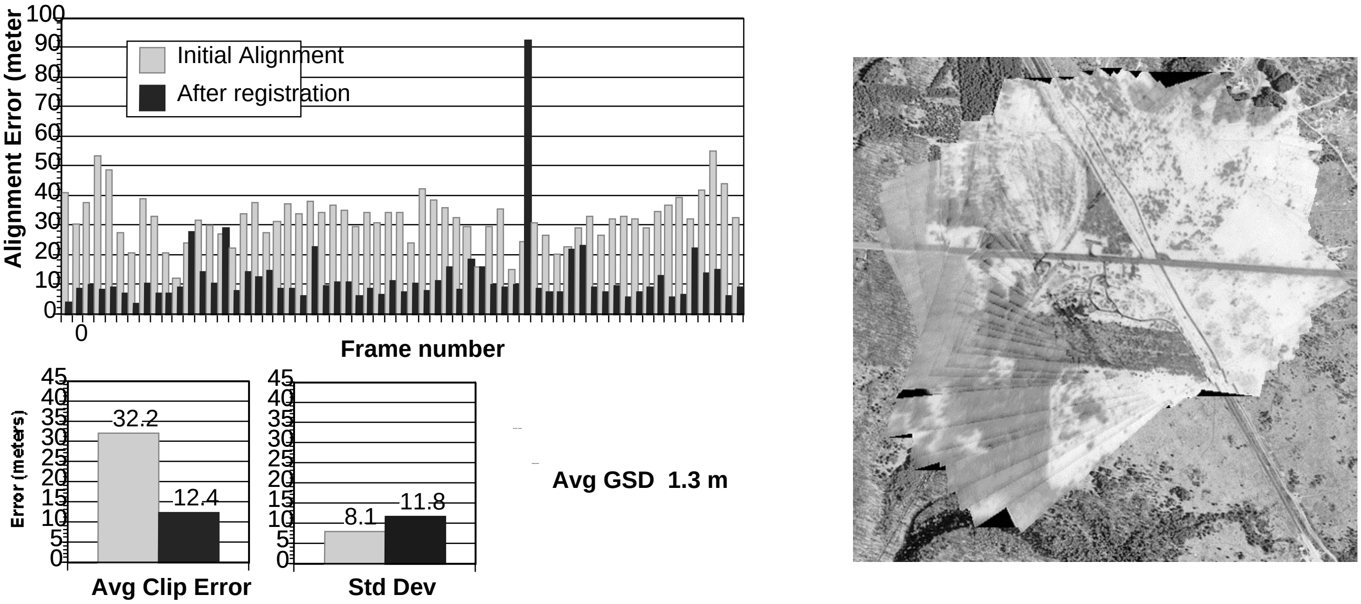

Figure 6. NY site georegistration accuracy and autonomously generated scene mosiac.

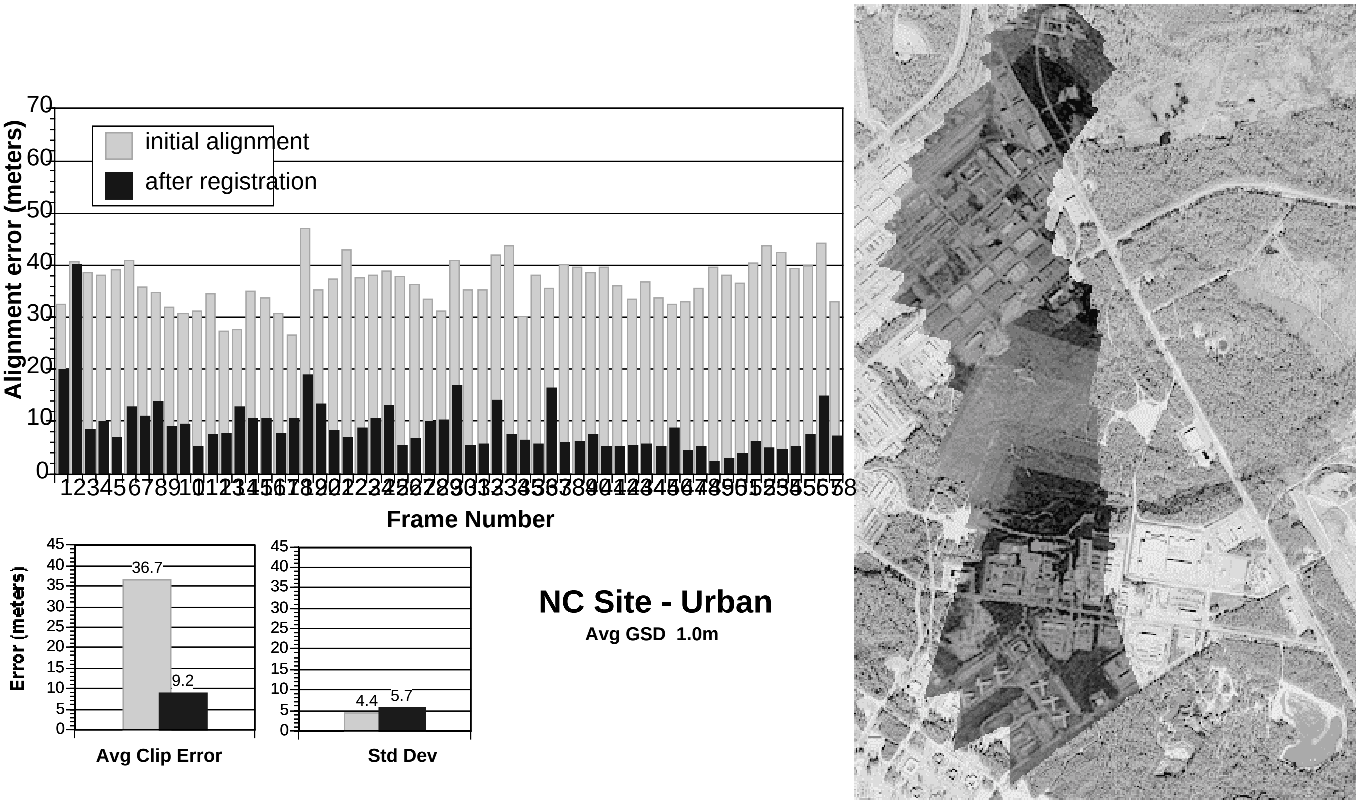

Figure 7. NC Site georegistration accuracy and autonomously generated scene mosiac.

old. For this case, registration performance was somewhat greater than the NY site case, with a factor of 4 reduction in error. With this dataset, almost every registration event resulted in improvement to geolocation accuracy. We attribute these accuracy improvements and the robust performance to increased scene complexity, compared to the NY site data, and greater similarities in scene content between the reference and video imagery (no snow cover). As scene content changes from uniform rural content to complex structured urban settings, the density of useful match points increases, and the accuracy of the correspondence solution increases accordingly. Registration solutions for the first two video frames were degraded due to low scene content (a lack of match points) and the presence of long linear features (aperture effect).

5. Conclusion

Autonomous video georegistration is a valuable tool for a host of analytical and operational needs. Our model-based approach for georegistration offers a unique capability based on a rigorous photogrammetric approach that corrects for relief distortion using projections to a common 3-D correlation surface. An iterative treatment corrects for translational and rotational errors present in pixel space by making constrained adjustments, aided by a Kalman filter, to the modeled video sensor and collection parameters. Initial results confirm a robust formulation with a 5X accuracy improvement over a range of terrain, feature and imaging conditions. In some instances (<5%), registration solutions were aliased by aperture effects. To correct this behavior, we are examining alternative registration approaches to augment our current registration treatment.

6. Acknowledgements

The authors would like to thank Mr. Steve Hennessy, Dr. Robert Hummel and Dr. Thomas Strat at DARPA for their continued support and stimulating discussions on this project. We would like to thank Mr. Elliott Lloyd at the Army NVESD for technical assistance in planning and supporting our field collection efforts. We would also like to recognize and thank Richard Collier and Carl Williams for their technical contributions as well as Nikki Tucker for her assistance in manuscript preparation. Portions of this research were funded by the Defense Advanced Research Projects Agency.

7. References

[1] Brown, L., "A Survey of Image Registration Techniques", ACM Computing Surveys, 24(4): 325-376, Dec. 1992.

[2] J. Bryan, D. Bell, A Lee, N. Carender and F. Baker, "A New Image Registration Paradigm", *Electronic Image Int'l Conf. Proc.*, Boston, MA, Sep 1993.

[3] J. Hackett, D.Trask and R. Cannata, "Automated Near-Real Time Registration And Geopositioning Based Upon Rigourous Photogrametric Modeling", *ASPRS-RTI Conf. Proc.*, Tampa, FL, 1998.

[4] M. Hansen, P. Anandan, K. Dana, G. Van der Wal and P. Burt, "Real Time Scene Stabilization and Mosaic Construction", *Proc. DARPA Image Understanding Workshop*, Nov 1996, pp. 457-465.