AUTOMATED VIDEO GEOREGISTRATION AT REAL-TIME RATE

Charles R. Taylor, Sr. Principal Engineer
John T. Dolloff, Technical Director
Matt Bower, Sr. Engineer
Scott B. Miller, Technical Director
BAE Systems
10920 Technology Pl.
MZ62ENG
San Diego, CA, 92127
charles.taylor@baesystems.com
john.dolloff@baesystems.com
matt.bower@baesystems.com
scott.miller@baesystems.com

ABSTRACT

Video metadata are often of lower absolute accuracy than desired for exploitation and tracking, and are often recorded at a lower rate than the video frame rate. In these circumstances it is beneficial to perform georegistration of the video to reference data in order to improve geopositioning accuracy or to assess the geopositioning accuracy. Manual georegistration is much slower than real-time, and previously existing automated techniques require specialized hardware. The authors have developed a real-time georegistration technique that is demonstrated to run in real-time on a general-purpose workstation with no site model preprocessing. The technique uses a Kalman filter to compute corrections to the interpolated metadata based on automatically measured tie points in every successive pair of video frames, and tie points to reference imagery and a digital elevation model (DEM) at the highest rate permitted by the workstation, currently about one frame in 45. When used with stereo reference imagery the system is also capable of estimating and correcting for DEM vertical bias. The georegistration approach is outlined and results are presented.

INTRODUCTION

The capability of precise geopositioning from video cameras on aerial platforms has long been limited by the accuracy of the metadata needed to support this capability: the camera position and orientation along with any interior orientation parameters (such as zoom level) not available from advanced calibration. Even in the best of circumstances, the jitter in the attitude due to air turbulence along the flight path makes it difficult to know the attitude to sufficient accuracy. On top of this, metadata typically are reported asynchronously from the frames of the video camera, such that the metadata need to be aligned in time and interpolated.

Because of these limitations, video image registration to reference data is an attractive means to achieve the goal of precise geopositioning where it would not otherwise be possible in the presence of low metadata accuracy. (Wang *et. al*. 2008). BAE Systems has long been active in automated registration of still and video imagery, and has recently addressed the challenge of automated video registration in a real-time system.

Automated registration systems that operate in real- or near-real time have been described (Cannata *et. al.*, 2000). Typically these approaches require specialized hardware to achieve a real-time rate, and may also require advanced preparation such as reference image orthorectification and scene model generation. Our approach is to use a general-purpose workstation without special hardware, and to use any available reference data with no special advanced preparation.

Automated video registration is expected to be of continuing interest even as high-accuracy metadata become available for video cameras on aerial platforms. One reason is that a system with the capability of performing automated video registration is also capable of automated quality monitoring of the subsystems that gather and transmit the metadata. Another is that the automated registration capability can still be used in the event of the failure of any part of the metadata collection subsystem.

In this work we will describe our approach to real-time automated video registration. Following this introduction, the next section will touch on our modular design approach, based on frame-to-reference matchers and frame-to-frame matchers supplying measurements to a Kalman filter which in turn supplies adjustments to a geopositioning sensor model. The following section will examine a test case in detail, including ground truth test results for the adjusted sensor model, complete with error modeling diagnostics.

DESIGN OF REAL-TIME PROTOTYPE

The core of a system for real-time video registration can be built as shown in Figure 1. The figure shows a registration manager module that combines a video Kalman filter (KF) with a frame-to-frame matcher and a frame-to-reference matcher. The registration manager is designed to work within a video screening and exploitation application.

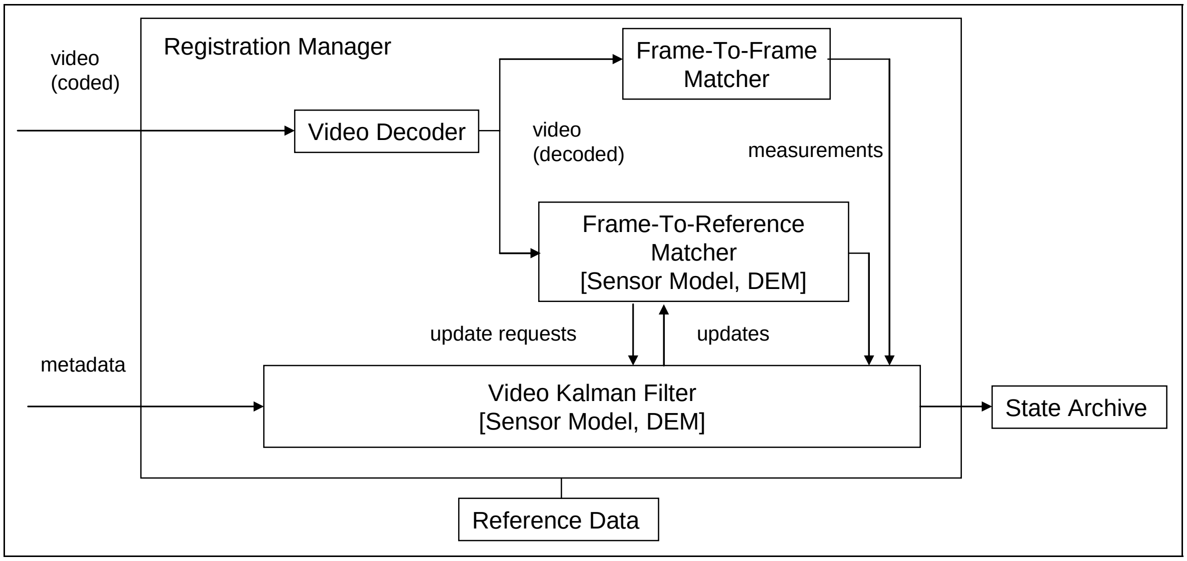


Figure 1. Video Registration Manager.

The frame-to-frame matcher is a module that identifies matching image points between pairs of frames. It is intended to be used for closely spaced video frames, such that perspective changes between frames are not important. As such it requires access to the images only, and not the metadata for each frame, and it does not use a sensor model. We have implemented our prototype in a modular way so that any available frame-to-frame matcher can send its matches to the video KF though a messaging interface. All that is required from the frame-to-frame matcher is identification of the two video frames used, the matching image coordinates, and an estimate of the image-space uncertainty of the matches. The frame-to-frame matcher used for the tests to be described in the following section uses a corner detection algorithm to perform the matching (Förstner and Gülch, 1986), but any fast and reliable frame-to-frame matcher should work equally well.

The frame-to-reference matcher module identifies image points in the video frame that match reference data. As with the frame-to-frame matcher, this module interacts with the video KF through a messaging interface, so that any fast and reliable frame-to-reference matcher can be used. The interface we developed for frame-to-reference matches allows the reference information to be expressed in two different ways. The first is in the form of a ground control point position and error estimate. The second is in the form of an image measurement (and error estimate) in a reference image. In both cases the corresponding image measurement (and error estimate) in a specified video frame accompany the reference information. These two mechanisms can be used in combination. We have also provided the capability in the interface to specify the relationship of the video image measurement to a terrain surface, so that a single reference image and digital elevation model (DEM) are sufficient reference data, and so that DEM corrections can be estimated when there are multiple reference images or 3D control points. The frame-to-reference matcher used for the tests to be described in the following section uses an edge matching algorithm that accounts for changes in perspective and modality (Lofy, 2000).

Since the frame-to-reference matcher must account for varying perspective and imaging modality, depending on the type of reference data, it is more complex than the frame-to-frame matcher. The sensor model for the unregistered video frame is therefore made available to the matcher in the design. The design shows it coupled with the video KF, so that improvements of the video metadata propagated from prior registrations can be used in the frame-to-reference matcher if needed.

Because of the complexity of the frame-to-reference matching task, it is typically not as fast as the frame-to-frame matcher. Our approach is to accept frame-to-reference matches in the video KF as often as we can get them, and maintain accuracy between frame-to-reference matches through the use of frame-to-frame matches for each successive pair of frames.

A Kalman filter (Gelb, 1974) receives measurements from the frame-to-frame matcher and frame-to-reference matcher. The Kalman filter has a state vector (and corresponding error covariance) that is updated for each video frame. The state vector consists of several parts, to handle the video sensor and each source of reference information.

The video sensor portion of the Kalman filter state consists of corrections to the video sensor operational parameters (position, attitude, and field of view) obtained from the metadata. In our implementation the video sensor portion of the state is augmented to use rate corrections (as well as offset corrections) to the sensor position and attitude data. While the metadata are usually asynchronous with the video frames, the measurements input to Kalman filter are synchronous with the frames, and thus the KF state is updated for the time of each frame. An archive is maintained of the state vector and its error covariance so that it can be used by the screener to obtain the updated sensor model for any frame.

The Kalman filter state vector also includes corrections to the reference data: sensor model adjustments for reference images, position adjustments for control points, and DEM height corrections. As the video registration takes place, the video sensor model errors become correlated with the reference data errors. These correlations are captured in the Kalman filter state error covariance, so that they are rigorously handled when the same reference data are used repeatedly in the video sequence, and so that the full covariance is available for rigorous exploitation via the state archive. Equally important, the Kalman filter implementation rigorously accounts for the correlation of video sensor model parameter errors between sequential frames when it applies the frame-to-frame measurements.

The video KF has been implemented to allow out-of-sequence frame-to-reference measurements. This is so that the updated state archive can be kept as accurate as possible by frame-to-frame registration while the frame-to-reference manager is working on its task. It maintains an internal archive of frame-to-frame matches. When a frame-to-reference match is received by the video KF, it rolls back the state to the appropriate time before updating, and then reapplies any previously received frame-to-frame matches. The video KF is a very fast module, so that with a short latency the video screener can always use the best updated state possible for any given frame: updates that have been propagated to the current frame by frame-to-frame matches from a very recent frame that was registered to reference data.

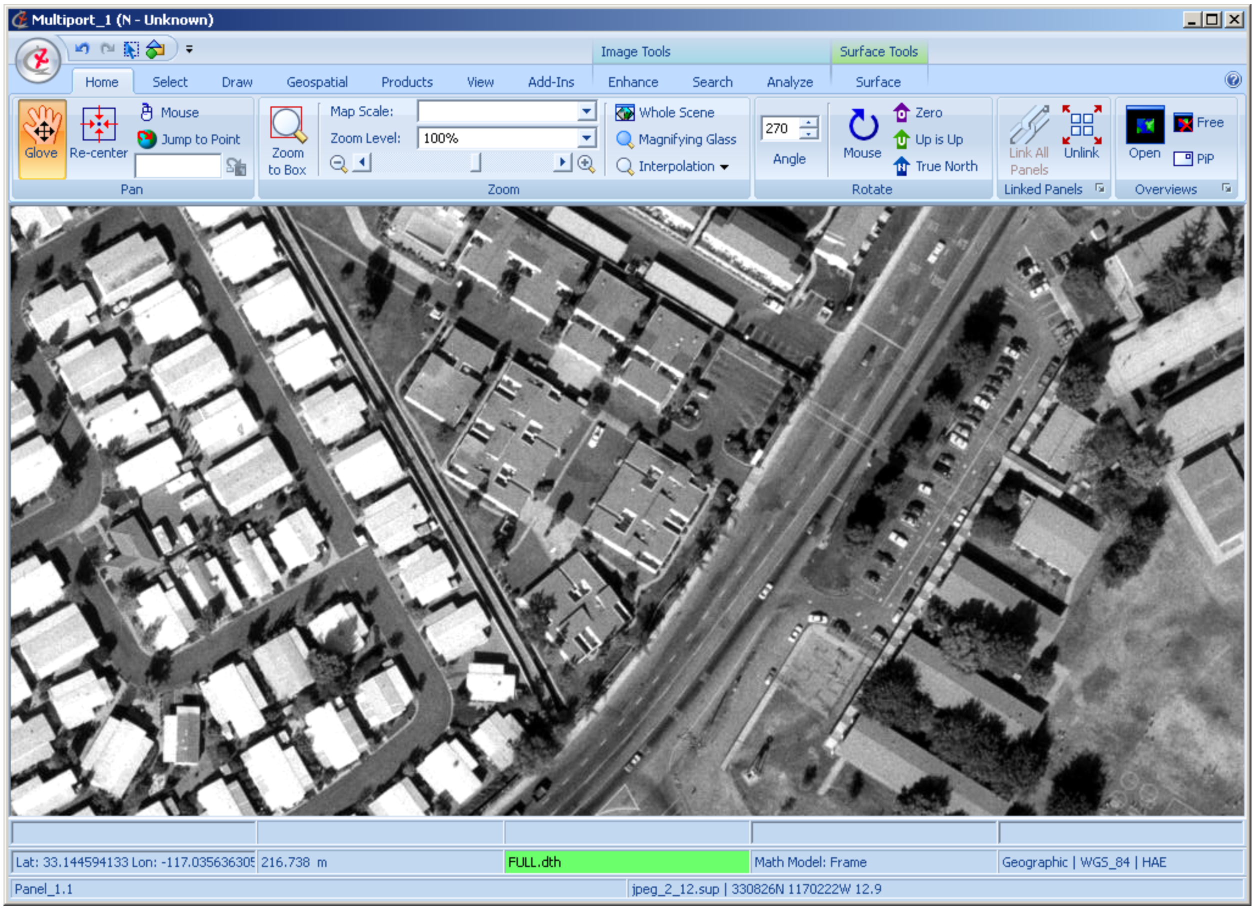
The video registration manager interacts with the video screener application by introducing a minimal latency from the live video stream, but still processes the video at a real-time rate. For each frame, the screener constructs a sensor model from the metadata as corrected by the corresponding entry in the state archive. Once the registration manager has begun its work there is always an updated state available, and even if there were not, the screener could use the uncorrected metadata for a sensor model. When the latency is sufficient, the updated state accuracy can approach the limit imposed by the quality of the reference data.

TEST RESULTS

The design presented above for real-time video registration depends on an initial estimate of the geopositioning metadata, which is to be corrected by the video Kalman filter using matches to frames and to reference imagery. Current practice for video is to place it into an MPEG-2 transport stream, in which it is multiplexed with a metadata stream using one of the Motion Imagery Standards Board (MISB) key-length-value (KLV) metadata standards (MISB 2006, MISB 2009). To date, our implementation of the design has been tested with unmanned aerial vehicle (UAV) systems that use the EG 0104.5 KLV standard. Newer standards, such as EG 0601.3 and EG 0801.2, will only help to improve the effectiveness of the system. In this section, we will describe one set of test data that has been registered with our system, and give the registration accuracy and timing results, and in particular we will describe a ground truth test of the registration results.

Test Data

The test results in this section were obtained with a video recorded with a video camera on a helicopter flight over Escondido, California. The reference data consists of a set of four triangulated aerial frame images, and the DEM extracted from them. The reference image ground sample distance (GSD) is about 0.33 m. The absolute horizontal accuracy is estimated to be 1.32 m CE90. The DEM post spacing is 10 m, and the absolute vertical accuracy is estimated to be 1.64 m LE90. Figure 2 is a screen capture of a portion of one of the reference images. The area on the right-hand side of the figure is covered in a part of the video sequence, and the features will be seen in the video frames to be shown.



Lat: 33.144594133 Lon: -117.03563630 | 216.738 m | FULL.dth | Math Model: Frame | Geographic | WGS_84 | HAE

Panel_1.1 | jpeg_2_12.sup | 330826N 1170222W 12.9

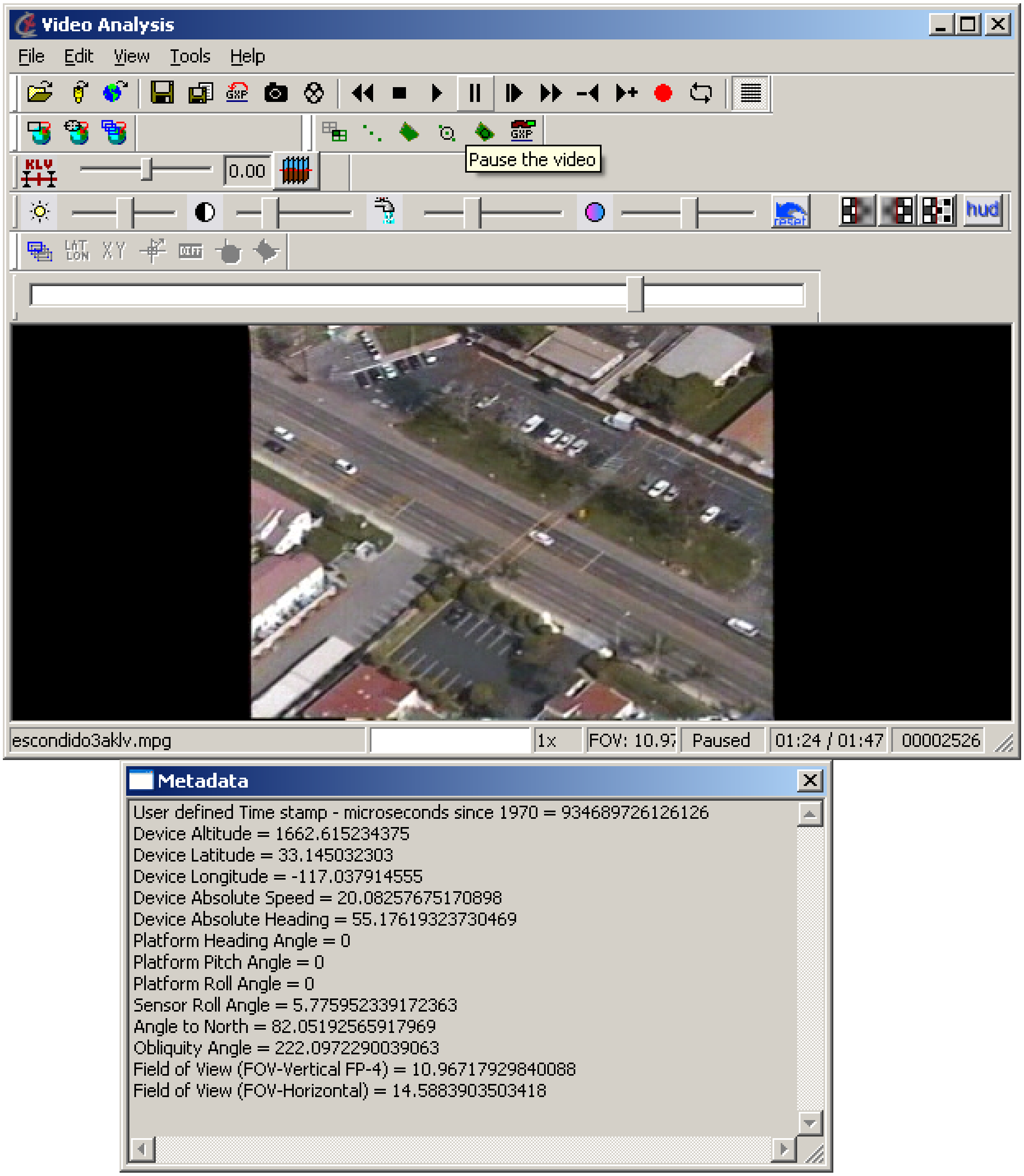
Figure 2. Screen capture of reference image.

The test results in this section were obtained with an MPEG-2 transport stream read from a file on disk. The video was deinterlaced to a 320x240 pixel progressive scan. The GSD ranges from 0.26 m to 0.45 m. The KLV metadata in EG 104.5 format were generated offboard, beginning with a manual resection of one frame to the reference data, followed by a coarse registration of each frame, to which smoothing was applied before multiplexing with the video stream. Note that for many UAV systems the metadata are recorded onboard the platform, so that the previous step is unnecessary. The a priori sensor position error was estimated to be about 20 m (1-sigma) horizontally and 10 m vertically. The a priori attitude angle uncertainties were set at 0.075 rad (1-sigma, 4.3 degrees) on each of three axes. These a priori error estimates are commensurate with what can readily be attained with UAV systems, and they give a predicted horizontal accuracy of about 125 m CE90 at the operating height for the video sensor in the test we present here. A screen capture of a video frame and the corresponding metadata is provided in **Figure 3**.

Accuracy Performance

One part of the video Kalman Filter state consists of adjustments to the geopositioning sensor model parameters for the video sensor itself. When the sensor model is used in a DEM intersection operation to compute the ground coordinates of a measured image point, the adjustments are expected to lead to more accurate coordinates than those obtained with the unadjusted sensor model. Whether or not this has been achieved can be determined by a ground truth test.

A ground truth test was designed as follows. First, a series of seven photoidentifiable check points was identified along the swath of the video sequence, such that a new point is acquired each time another point leaves the video frame. The "true" coordinates of these check points were determined by DEM intersection with their measured image coordinates in reference imagery. An error propagation calculation was also performed for each DEM intersection, taking into account the image measurement uncertainty, reference image sensor model parameter uncertainties, DEM vertical uncertainty, and an estimate of the uncertainty of each point's height relative to the terrain height. The error propagation yields a 3x3 error covariance matrix for each check point.



```
User defined Time stamp - microseconds since 1970 = 934689726126126
Device Altitude = 1662.615234375
Device Latitude = 33.145032303
Device Longitude = -117.037914555
Device Absolute Speed = 20.08257675170898
Device Absolute Heading = 55.17619323730469
Platform Heading Angle = 0
Platform Pitch Angle = 0
Platform Roll Angle = 0
Sensor Roll Angle = 5.775952339172363
Angle to North = 82.05192565917969
Obliquity Angle = 222.0972290039063
Field of View (FOV-Vertical FP-4) = 10.96717929840088
Field of View (FOV-Horizontal) = 14.5883903503418
```

Figure 3. Screen capture of video frame and metadata.

Next, every five frames, for a total of 162 measurements in 811 frames, a manual measurement is made of the image coordinates of the check point that is currently in view. To the extent that the video sensor state is known, it should be possible to match these measurements with a prediction based on the check point ground coordinates and the sensor model for the video frame.

The final step in the ground truth test is to project the check point ground coordinates into the video frame with either the unadjusted or the adjusted sensor model parameters, and compare with the image measurements. Error propagation is also done in this step, taking into account the check point 3x3 ground space error covariance, the sensor model parameter uncertainties, and the uncertainties in the image measurements in the video frame. When this is done with the adjusted sensor model it is particularly important to use the full error covariance from the video Kalman filter.

For the ground truth test, the frame-to-reference match was configured to occur exactly every 30 frames. In real-time operation, the interval between registrations to reference will be somewhat irregular depending on the time required for each registration. We believe the rate of registrations to reference used in this study to be the likely rate that will attained in practice with the real-time configuration in the near future. The frame-to-frame matcher was configured to find five matches between consecutive frames. The frame-to-reference matcher was configured to find up to five matches to reference for each frame attempted, and the actual number was distributed fairly uniformly between zero and five, depending on the scene content of the frame in question.

The differences of the predicted check point image positions from their measured positions are shown in Figure 4 for this ground truth study. The figure shows the results for both the unadjusted and the adjusted sensor model for the video frame. In this figure and those that follow, the abscissa is divided by brown lines to show the portions of the clip in which each of the seven check points is used. It is to be expected that there will be a common bias component in the errors for any one check point, so that only by looking at multiple check points will a more complete picture of the error statistics emerge. This effect is clearly seen in the figure, where (for example) the errors after registration for the sixth check point all tend to be larger and the errors for the seventh check point tend to be smaller.

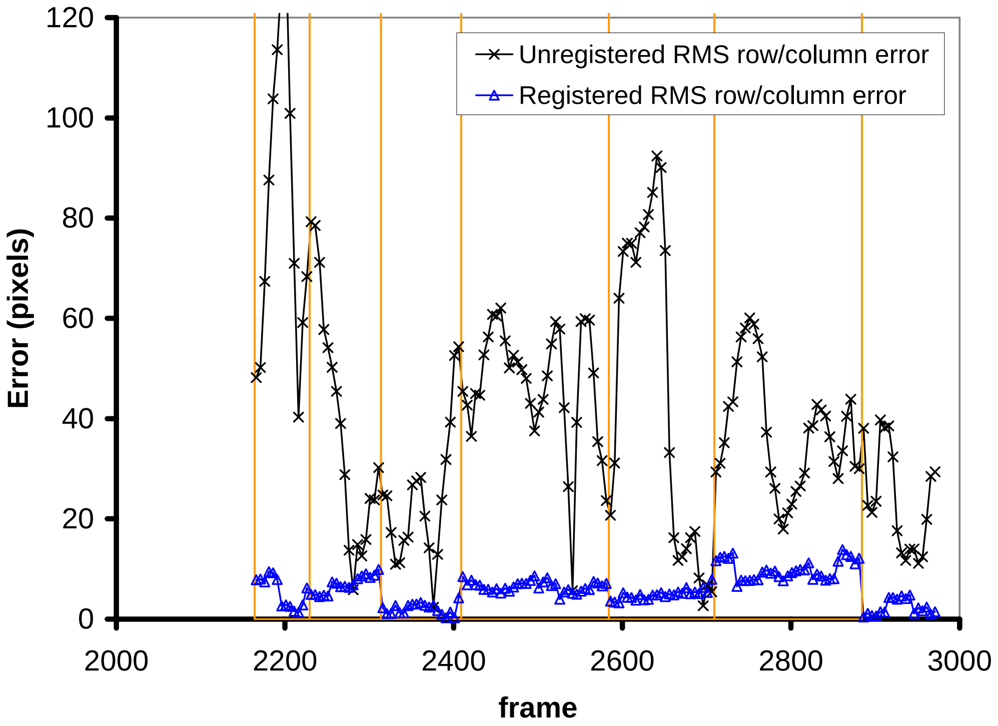


Figure 4. RMS differences in image space between measured and predicted check points for unregistered and registered video frames.

As Figure 4 indicates, the RMS differences in image space have been reduced from a mean of about 41 pixels to about 6 pixels by the registration process. This is a critically important result for an automated process. Even though the reference imagery used to measure the check points manually is the same as was used for the automated registration, it shows that the automated process has not identified false correspondences. This point is important enough that check points should be measured whenever possible when an automated registration process is invoked.

Note that the image space residuals in Figure 4 are not the errors in the measurements made by the matchers, they are the differences between the check point image measurements and the predicted measurements, with errors from all sources. In our testing, the image measurement error estimates for the frame-to-frame matcher are about 1 pixel (1-sigma). The image measurement error estimates for the frame-to reference matcher are about 3 pixels (1-sigma).

It is also possible to express the results of the ground truth test in ground space. This is done by projecting the check point image measurements to the horizontal surface at the known height of the check point. These results are shown in Figure 5, and they show that the RMS differences in horizontal ground space have been reduced from a mean of about 14 m to about 2 m.

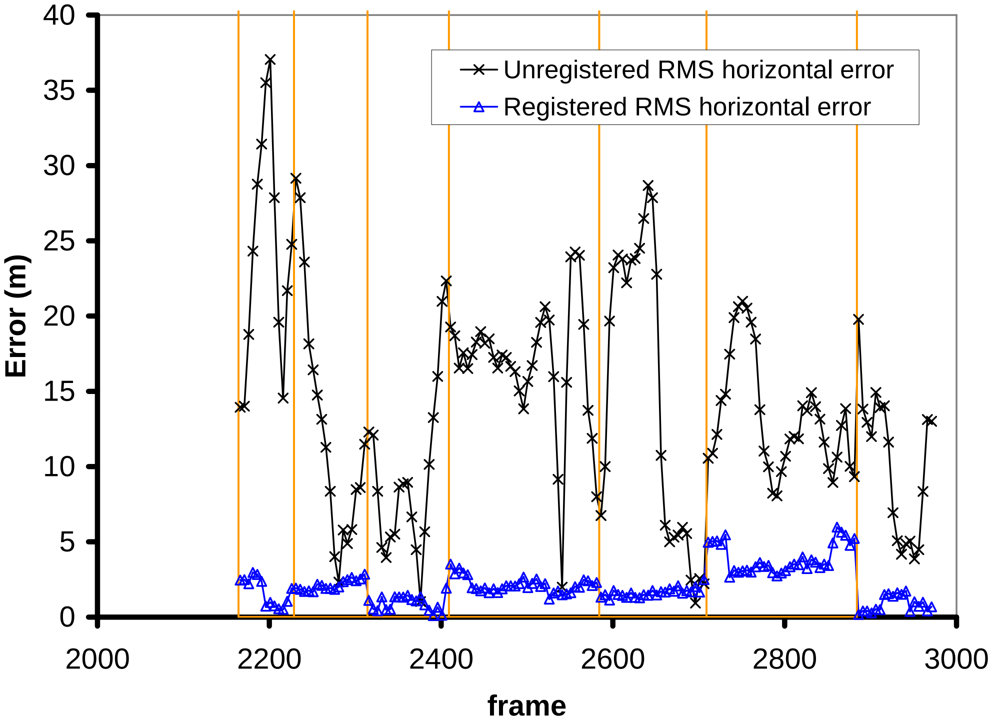


Figure 5. RMS differences in horizontal ground space between measured and predicted check points for unregistered and registered video frames.

Besides showing that the registration process increases the geopositioning accuracy for the video sensor, the ground truth test has another equally important role. That role is to provide confidence in the a posteriori error estimates from the video Kalman filter. In Figure 6 the image space residuals for the check points are shown again for the case of the registered image. In addition, the circular error metric (CE90) is shown in red on the same plot. The CE metric contains all of the sources of error in the difference between the measurements and the predicted measurements as described previously, including the ground space uncertainty of the check point, the sensor parameter error for the video sensor after registration, and the image measurement error for the manual measurements in the video frame. This graph shows that the image space residuals are within the bounds of the CE90 metric for the great majority of the trials, as is to be expected when the registration process is working reliably.

One other bit of information is displayed in Figure 6. There is a fundamental limit to how much accuracy improvement is possible from the registration process, imposed by the accuracy of the reference data as seen at the imaging geometry of the video sensor. This limit can be computed by projecting the accuracy of a point extracted from the reference data into the video frame. In this study, and in all practical applications, the check points are of equal accuracy to any other control point extracted from the reference data (otherwise, if more accurate data were available it would be used as reference data in registration). Therefore, in this study we repeated the projection of the check points into the video frame, but omitted the video sensor model parameter uncertainty from the error propagation. In the figure, it is seen (of course) that the limiting accuracy is always better than the actual predicted accuracy. This is due to the contribution of errors (hopefully well-characterized) in the matchers used in the automatic registration, and because matches will not always be found in the number and spatial distribution required to obtain the ultimate accuracy. We observe that the matches to reference bring the accuracy rather close to the ultimate achievable accuracy, and that the frame-to-frame matches keep the accuracy within about a factor of two of the limit when there is a match to reference about every thirty frames.

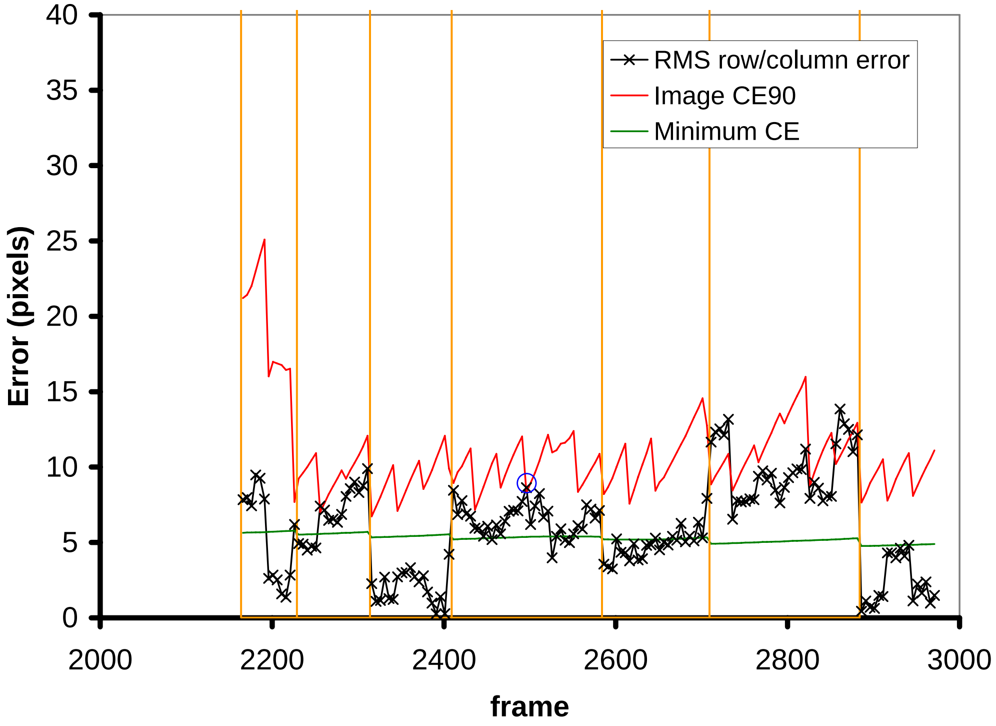
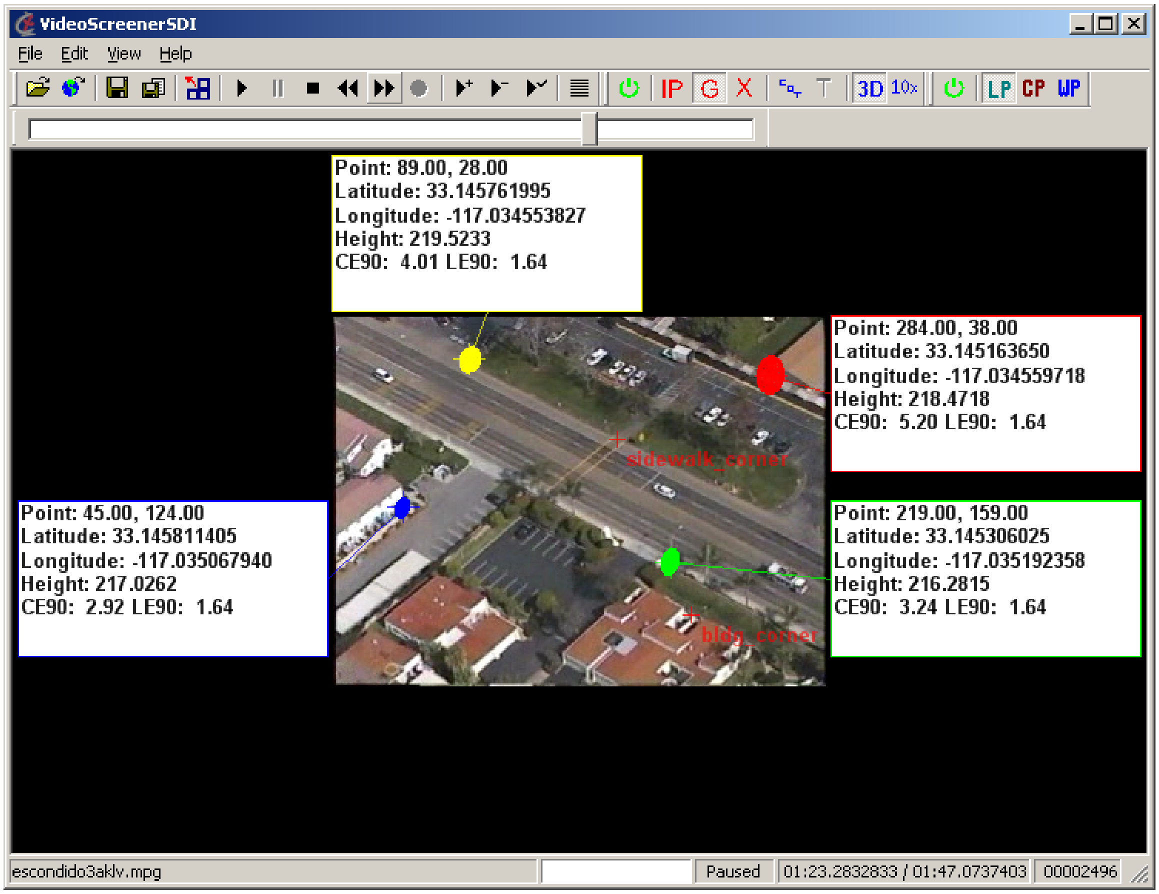


Figure 6. RMS differences in image space between measured and predicted check points for registered video frames, with error estimate from all sources (red) and error estimate excluding video sensor model parameters (green). Frame 2496 is indicated with a blue circle.

The fine-scale "sawtooth" pattern in Figure 6 corresponds to the 30-frame cycle of matches to reference. After each reference match the predicted CE90 falls abruptly, and then it rises slowly since the frame-to-frame matches keep its growth in check. (The predicted CE90 would rise much faster without the frame-to-frame matches.) The sawtooth pattern is interrupted in places where the frame-to-reference matcher fails due to ambiguities or sparseness of features, in which case the predicted CE90 continues to grow until the next successful match to reference.

The blue circle in Figure 6 corresponds to video frame number 2496, which has been registered to the reference image. That is why its CE90 estimate is lower than the preceding frames. A screen capture of frame 2496 is shown in Figure 7. On this screen, the 3D error ellipsoids are shown for geolocation using the registered video frame and reference DEM for 4 arbitrarily chosen points on the terrain. A screen capture of the immediately preceding frame (2495) is shown in Figure 8. This figure shows the 3D error ellipsoids for the same 4 ground points, which are larger because frame 2495 was not registered to reference (the most recent registration to reference having taken place in frame 2466). When the video is viewed in the screener as the registration is being done, the error ellipsoids can be viewed in real time. Their positions track with the real features in the image, and their sizes gradually grow and then snap back to a smaller size as Figure 6 indicates.



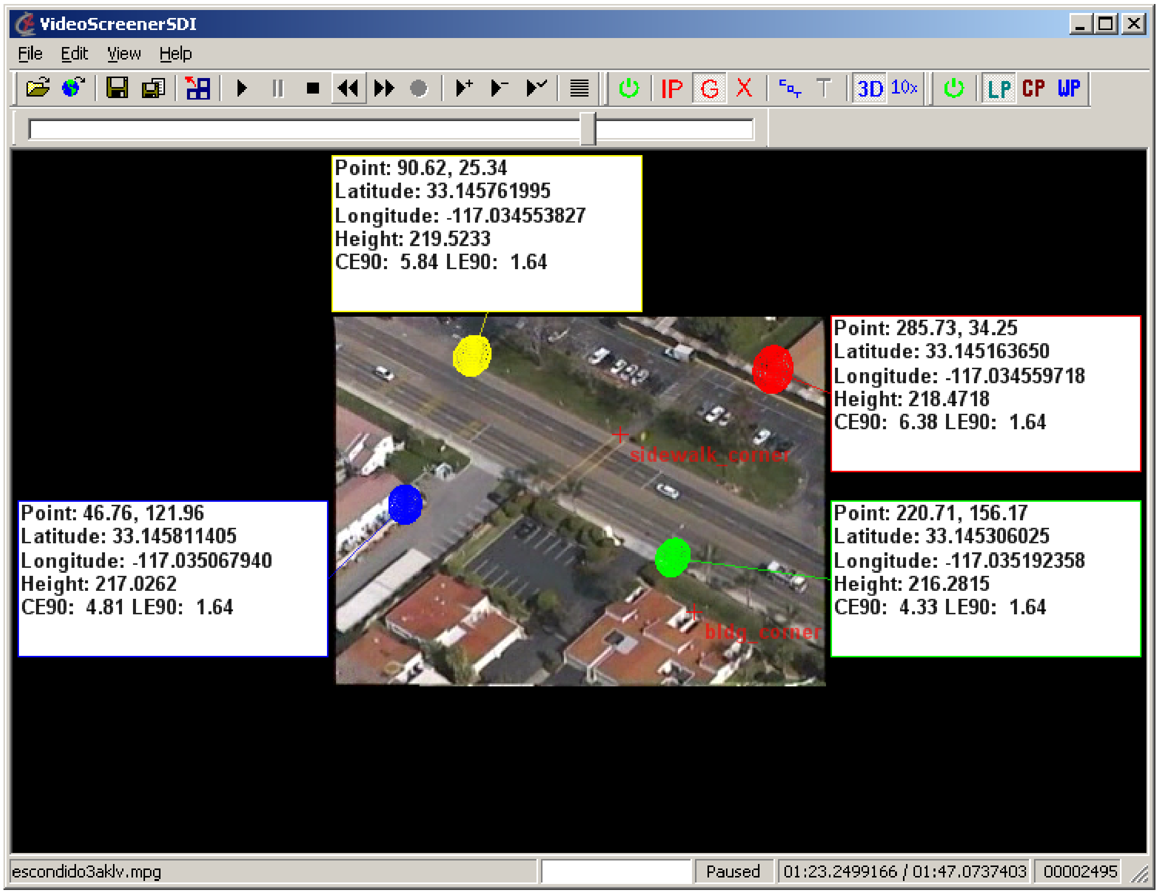
Point: 89.00, 28.00
Latitude: 33.145761995
Longitude: -117.034553827
Height: 219.5233
CE90: 4.01 LE90: 1.64

Point: 284.00, 38.00
Latitude: 33.145163650
Longitude: -117.034559718
Height: 218.4718
CE90: 5.20 LE90: 1.64

Point: 45.00, 124.00
Latitude: 33.145811405
Longitude: -117.035067940
Height: 217.0262
CE90: 2.92 LE90: 1.64

Point: 219.00, 159.00
Latitude: 33.145306025
Longitude: -117.035192358
Height: 216.2815
CE90: 3.24 LE90: 1.64

Figure 7. Screen capture of frame 2496. The screen shows the sizes and shapes of the 3D error ellipsoids for geolocation with the video frame and reference DEM for four arbitrary ground points on the terrain surface. The screen also shows the check point "sidewalk_corner" that was used for this frame, at its predicted location in the frame (which corresponds very closely to its observed image location).



Point: 90.62, 25.34
Latitude: 33.145761995
Longitude: -117.034553827
Height: 219.5233
CE90: 5.84 LE90: 1.64

Point: 285.73, 34.25
Latitude: 33.145163650
Longitude: -117.034559718
Height: 218.4718
CE90: 6.38 LE90: 1.64

Point: 46.76, 121.96
Latitude: 33.145811405
Longitude: -117.035067940
Height: 217.0262
CE90: 4.81 LE90: 1.64

Point: 220.71, 156.17
Latitude: 33.145306025
Longitude: -117.035192358
Height: 216.2815
CE90: 4.33 LE90: 1.64

escondido3aklv.mpg | Paused | 01:23.2499166 / 01:47.0737403 | 00002495

Figure 8. Screen capture of frame 2495. The screen shows the sizes and shapes of the 3D error ellipsoids for geolocation with the video frame and reference DEM for the same four arbitrary ground points as the previous figure. Note that the estimated errors are somewhat larger for this frame, since the most recent registration to reference was 29 frames previously.

A crucial test to build confidence in the error estimates is the chi squared test, which weights the differences between the measurements and predictions by the inverse of the combined error covariance for the difference. With one check point measurement in a given frame, the error vector has two elements and the error covariance is a 2x2 matrix, and the chi squared test has two degrees of freedom. In a random sample, the value of chi squared above which 95 percent of the tests are expected to lie is 0.10 while the value above which 5 percent are expected to lie is 5.99. Ninety percent of trials are expected to lie within this range. The experimental chi squared values are graphed in Figure 9.

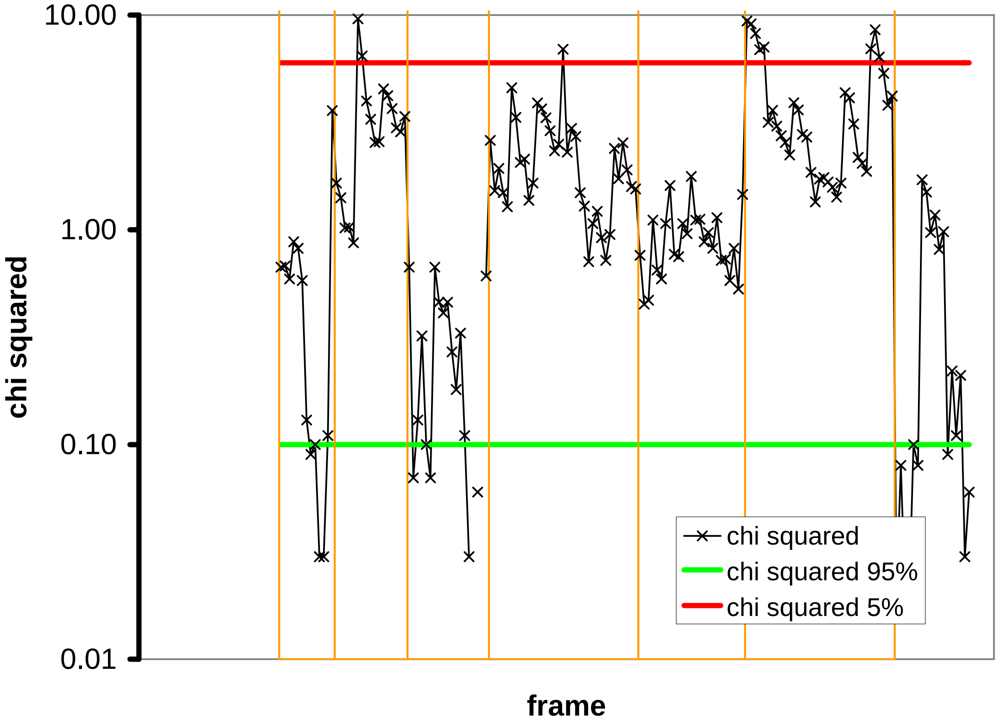


Figure 9. Chi squared metric for image-space residuals (2 degrees of freedom per frame) for 162 check points in 811 video frames, with lower and upper bound of symmetric 90% confidence interval.

Note that a semilog graph is used for Figure 9, because on a linear graph it is difficult to see which points are above and which are below the lower (green) threshold. The results show a healthy registration, with most of the points in the expected range.

Timing Performance

The real-time registration implementation does not require specialized hardware, only a high-powered general purpose workstation. The workstation used for this study is a Hewlett-Packard Z800 with dual Intel® Xeon®* W5580 Quad core processors running at 3.20 GHz. It has 8MB L3 cache, a 1333 MHz DDR3 clock, and 12 Gbyte RAM.

The timing performance is given in the context of all components running simultaneously on the system described above. The time for frame-to-reference matching is about 1.5 s (one in every 45 frames). This process uses up to 12 processing threads. The frame-to-frame matching time is 0.055 s (one every 1.7 frames). This is a bit slower than real time, but a real time rate was achieved by using 3 frame-to-frame matching threads, and it continues to work in real time when every pair of successive frames is processed by the frame-to-frame matcher. The video KF rate is about twice real time, which enables the frame-to-frame matches already found to be reprocessed after every frame-to-reference match within the latency time of about 3 s. The video screener is able to run in real time even with all of the other processes running, and with many wire frame error ellipsoids being computed and rendered for each frame.

CONCLUSION

An approach to real-time, rigorous, automated video registration has been demonstrated on a general purpose workstation with no specialized hardware and no advanced model generation. This was done by integrating a video Kalman filter, frame-to-frame matcher, and frame-to-reference matcher in a modular system that can interact with a

*Intel and Xeon are registered trademarks of Intel Corporation in the U.S. and other countries.

video screener with a short latency from the live stream. Ground truth testing indicates the automated process gives valid sensor model updates and error estimates. Currently available workstations allow the process to work with a latency as small as three seconds.

ACKNOWLEDGEMENTS

We would like to thank software engineers Janelle Beam, Chris Tang, Seth Merickel, Brian Duston, and especially Shannon McDonald, for their contributions to this work.

REFERENCES

Cannata, Richard W., Mubarak Shah, Steven G. Blask, and John A. Van Workum, 2000. Autonomous video registration using sensor model parameter adjustments, in *IEEE Proceedings 29th Applied Imagery Pattern Recognition Workshop*, pp. 215-222.

Förstner, W., and E. Gülch, 1986. A fast operator for detection and precise location of distinct points, corners and centers of circular features, in *Proceedings of the ISPRS Workshop on Fast Processing of Photogrammetric Data*, pp. 281–305.

Gelb, Arthur, 1974. *Applied Optimal Estimation*, MIT Press.

Lofy, Brian, 2000. High Accuracy Registration and Targeting, in *IEEE Proceedings 29th Applied Imagery Pattern Recognition Workshop*, pp. 235-242.

MISB, 2006. *Motion Imagery Standards Board Engineering Guideline: Predator UAV Basic Universal Metadata Set*. MISB EG 0104.5. http://www.gwg.nga.mil/misb/docs/eg/EG010405.pdf
(see also http://www.gwg.nga.mil/misb/docs/standards/Standard060103.pdf
and http://www.gwg.nga.mil/misb/docs/eg/EG080102.pdf.)

MISB, 2009. *Standard: Motion Imagery Standards Board KLV Metadata Dictionary,* MISB Standard 0807.4. http://www.gwg.nga.mil/misb/docs/standards/Standard080704.xls.

Wang, Caixia, Anthony Stefanidis, Arie Croitoru, and Peggy Agouris, 2008. Map registration of image sequences using linear features, *Photogrammetric Engineering and Remote Sensing*, Vol. 74. No. 1. pp. 25 - 38.