# Video georegistration: Algorithm and quantitative evaluation

R. P. Wildes D. J. Hirvonen S. C. Hsu R. Kumar W. B. Lehman B. Matei W.-Y. Zhao

Sarnoff Corporation
Princeton, NJ 08543

## Abstract

*An algorithm is presented for video georegistration, with a particular concern for aerial video, i.e., video captured from an airborne platform. The algorithm's input is a video stream with telemetry (camera model specification sufficient to define an initial estimate of the view) and geodetically calibrated reference imagery (coaligned digital orthoimage and elevation map). The output is a spatial registration of the video to the reference so that it inherits the available geodetic coordinates. The video is processed in a continuous fashion to yield a corresponding stream of georegistered results. Quantitative results of evaluating the developed approach with real world aerial video also are presented. The results suggest that the developed approach may provide valuable input to the analysis and interpretation of aerial video.*

## 1. Introduction

Video is an increasingly common and important source of information about the world. In many situations the usefulness of this information is increased if the imaged objects and events can be precisely localized in the world that has been imaged. The ability to associate 3D world coordinates with video is of particular value for many applications of aerial video. The ability to assign geodetic coordinates to video pixels is an enabling step for a variety of operations, including, targeting, map generation, video annotation from geospatial databases and sensor model refinement. To best serve these applications, the assignment of 3D coordinates to video must be highly robust with accuracy and precision on the order of 10 meters and below. For time critical applications real-time performance also is of importance.

The preceding remarks motivate the research described in this paper. In particular, an algorithm is presented for video georegistration, i.e., the spatial registration of video imagery to geodetically calibrated reference imagery so that the video can inherit the reference coordinates. The application domain of particular concern is aerial video. In this domain, video typically comes with telemetry that can provide an initial estimate of the camera view. Even so, the video georegistration mapping is quite challenging: Spatial correspondence is poorly defined owing to unreliable telemetry, narrow (video) field of view, oblique viewing, rugged terrain and appearance change between the video and reference. Indeed, were telemetry perfect then the desired mapping essentially would be provided as input; however, this is hardly the case in practice, where even precisely calibrated laboratory systems commonly have errors in excess of hundreds of projected ground meters. Nevertheless, results presented in this paper show that the developed video georegistration algorithm is capable of an interesting level of performance in the presence of the noted challenges.

A great deal of research has considered image registration [3]. Most closely related to the work that is presented in the current paper are other efforts that have concentrated on image georegistration. One class of previous approach considers either implicit or explicit recovery of elevation information from video for subsequent matching to a reference elevation map [14, 16, 17]. Appeal to an elevation representation for matching has the potential to be invariant to many sources of video/reference difference; however, it relies on recovery of elevation from video - a difficult task. A second class of approach applies image rendering techniques to account for telemetry supplied information so that the reference and video can be projected to similar views for subsequent appearance based matching [10, 12, 13, 15]. While the current work also falls into this second class, earlier work was more limited in its ability to successfully match the two sets of imagery (even following projection) in the presence of significant remaining appearance differences (e.g., due to poor telemetry or unmodeled seasonal variation).

In the light of previous research, the outstanding contributions of the current work are as follows. First, a novel algorithm for video georegistration is presented. While many of the individual techniques that comprise this algorithm are known in the literature, here they have been selected for principled reasons derived from the challenges at hand and combined to yield an integrated system for continuous video georegistration. Second, the most extensive evaluation published to date of any algorithm for video georegistration is presented. Significantly, the results of this evaluation show that the developed algorithm is capable of dealing with many of the vicissitudes of real-world data.

## 2. Algorithm and system

There are three major algorithmic components to the developed system, see Figure 1. (i) The reference and video are projected to a common coordinate frame based on available telemetry. This projection establishes initial conditions for

0-7695-1143-0/01 $10.00 © 2001 IEEE

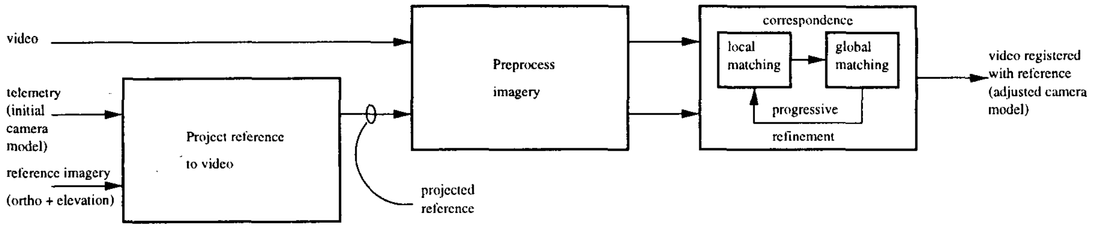
video
telemetry (initial camera model)
reference imagery (ortho + elevation)
Project reference to video
projected reference
Preprocess imagery
correspondence
local matching
global matching
progressive refinement
video registered with reference (adjusted camera model)

Figure 1: Algorithmic Steps to Video Georegistration.

image-based alignment to improve on the telemetry-based estimates of georegistration. (ii) The imagery is preprocessed to bring it under a representation that captures both the geometric and intensity structure of the imagery to support matching of video to reference. Geometrically, video frame-to-frame alignments are calculated to relate successive video frames and extend the spatial context beyond that of any single frame. For image intensity, the imagery is filtered to highlight pattern structure that is invariant between video and reference. (iii) A detailed spatial correspondence is established between the video and reference imagery that effects a precise alignment of the two imagery sources. Integral to the alignment process is an adjustment of the camera model to better reflect the match of the video and reference. A progressive refinement strategy is employed, proceeding coarse-to-fine along several different dimensions, commensurate with uncertainty in the current state of processing.

In this paper mappings between points in various coordinate frames will be presented (e.g., world-to-image, image-to-image). These mappings will be represented in terms of a $4 \times 4$ homogeneous transformation matrix, $\Pi$ operating on $4 \times 1$ column vectors, $\mathbf{m} = (x, y, z, w)^\top$, i.e.,

$$\mathbf{m}_{out} = \Pi \mathbf{m}_{in}. \tag{1}$$

Specific forms for $\Pi$, $\mathbf{m}_{in}$ and $\mathbf{m}_{out}$ will be introduced at appropriates places in the exposition. (Nonlinear effects, e.g., radial lens distortion are neglected here, but could be accounted for via composition of additional transformations.) Video frames and projected reference images will be denoted via $v$ and $r$; particular points will be denoted with $j$. For example, homogeneous coordinates for a corresponding point in video and reference will be symbolized as $\mathbf{m}_{v_j}, \mathbf{m}_{r_j}$, respectively. Two-dimensional image coordinates for the same point will be given as $\mathbf{p}_{v_j} = (\mu_j, \nu_j)_v = (x_j/w_j, y_j/w_j)_v^\top, \mathbf{p}_{r_j} = (\mu_j, \nu_j)_r = (x_j/w_j, y_j/w_j)_r^\top$, as standard. The remainder of this section details the 3 major algorithmic components to video georegistration: projection of reference and video to a common coordinate frame, image preprocessing and correspondence.

## 2.1. Project reference to video frame

The first algorithmic step in the developed approach is to project the reference and video imagery to a common coordinate frame. The goal is to make use of the available telemetry implied camera model to bring the reference and video into an initial alignment that can support further automated registration. Critical to the projection process is to make use of the reference image elevation information to account for as much 3D relief-based variation as possible so that subsequent image matching can be largely 2D in nature. For this reason, the reference orthoimage is projected into the video frame, since the orthoimage already is coregistered to the elevation map. The projection is accomplished via standard texture map-based rendering [7]. The digital elevation map is triangulated to yield a 3D mesh. The orthoimage is regarded as a texture, coregistered to the mesh. The mesh vertices are parametrically mapped to the image plane based on the telemetry implied camera projection matrix. Hidden surfaces are removed via Z-buffering. The output of this first algorithmic step is a view, $r$, of the reference orthoimage according to the rough indication of the camera model provided by telemetry, i.e., as the result of a projective mapping where the general point transformation matrix (1) is specialized to

$$\mathrm{P}^{\mathrm{render}}_{w,r} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ 0 & 0 & 0 & 1 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \tag{2}$$

with the matrix entries derived via composition of the camera interior and exterior orientation parameters [20] that is applied to vertices of the triangulated mesh. Input world points are symbolized as $\mathbf{m}_{w_j}$ and output projected reference points are $\mathbf{m}_{r_j}$. To account for inaccuracies in the telemetry-based projection, the support of the area covered in the projected image is extended beyond the implied field of view. For the sake of efficiency, reference images are not projected for every frame of a video sequence but only on an intermittent basis when the last registered video frame is near the border of the current reference projection.

## 2.2. Preprocess imagery

The second algorithmic step in the developed approach is to bring the video and projected reference imagery under a representation that supports subsequent image-based matching. There are two subcomponents to this stage of processing. (i) A geometric alignment of successive frames in the video is established so that more information than that provided by any one frame can be brought to bear in matching to the reference imagery. (ii) Image intensity structure is processed to facilitate matching between the video and reference imagery by highlighting invariant pattern structure.

**Frame-to-frame alignment** When considered individually, any single frame in a video can lack sufficient distinctive structure to disambiguate matching to a reference image.

This difficulty can be ameliorated by considering collections of frames simultaneously to effectively increase the field of view that is considered in the matching process. To facilitate such matching, the frame-to-frame alignment of adjacent frames in the video sequence is recovered. The recovered parameters subsequently are used as geometric constraints in the matching of collections of frames to reference imagery. Significantly, the alignment parameters are not used to construct a single monolithic mosaic for matching, but rather are reserved for constraints, which allows for greater flexibility in the final match to reference.

In current implementation, the frame-to-frame alignment, $\mathsf{F}_{v,v+1}$, is recovered as an affine mapping, i.e., a specialization of the transformation (1) according to

$$\mathsf{F}^{\text{affine}}_{v,v+1} = \begin{pmatrix} a_{11} & a_{12} & 0 & a_{13} \\ a_{21} & a_{22} & 0 & a_{23} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad (3)$$

with the input and output to the mapping of the forms $\mathbf{m}_{v_j} = (x_j, y_j, 0, 1)^{\mathsf{T}}_v, \mathbf{m}_{v+1_j} = (x_j, y_j, 0, 1)^{\mathsf{T}}_{v+1}$, respectively. Empirically, it has been found that the affine transformation provides an adequate description of frame-to-frame alignment for video rate (30 fps) capture of the aerial imagery of concern. Parameter values for $\mathsf{F}$ are recovered via a gradient-based, coarse-to-fine, Gauss-Newton estimator working over a Laplacian pyramid [2]. For a collection of $n$ video frames, a set of $n-1$ alignment matrices are calculated for subsequent processing.

**Image intensity representation** There can be a great deal of appearance change between a video and its corresponding reference orthoimage, even following projection to a common coordinate frame. Many sources contribute to this change, including, variation in sensor characteristics, diurnal and seasonal cycles and scene structure (e.g., new construction). To ameliorate such difficulties, it is desirable to choose a representation of image intensity that highlights pattern structure that is common to the two image sources that are to be brought into alignment.

Features with the potential to serve in the desired fashion are those that exhibit a local dominant orientation or well localized point-like structures that can be thought of as capturing a range of orientations, e.g., roads, tree lines, edges of buildings, compact isolated structures and the like. Similarly, the local distribution of orientations that are present in an image can be indicative of texture structure. Correspondingly, an image representation is employed that decomposes intensity information according to local spatial orientation. This representation is derived via application of a bank of filters that are tuned for spatial orientation and scale to both the video imagery as well as the (projected) reference image. In particular, the filtering has been implemented in terms of second derivative of Gaussian filters, $G_{2_\theta}$ at orientation $\theta$ and their Hilbert transforms, $H_{2_\theta}$ [11]. The filters are taken in quadrature (i.e., for any given $\theta$, $G_{2_\theta}$ and $H_{2_\theta}$ in tandem) to eliminate phase variation by producing a measure of local energy, $e_\theta(x,y)$ within an orientation band, according to

$$e_\theta(x,y) = (G_{2_\theta}(x,y) * i(x,y))^2 + (H_{2_\theta}(x,y) * i(x,y))^2, \qquad (4)$$

with $*$ symbolizing convolution and $i(x,y)$ images to be filtered, i.e., for current purposes video frames and projected reference images. This filtering is applied at a set of four orientations, vertical, horizontal and two diagonals to yield a corresponding set of "oriented energy images" [8] for both the video and reference imagery. Further, the entire representation is defined over a Gaussian pyramid [11] to support multiresolution analysis during subsequent processing.

Owing to the rectification of the oriented filter outputs, the oriented energy representation is invariant to contrast reversals. Still, the value of any one oriented energy measure is a function of both orientation and contrast. To avoid this confound and get a purer measure of orientation the response of each filter is normalized by the sum of the consort, i.e.,

$$\hat{e}_{\theta_k}(x,y) = \frac{e_{\theta_k}}{\Sigma_k e_{\theta_k(x,y)} + \epsilon} \qquad (5)$$

with $k$ ranging over the four orientations and $\epsilon$ a small bias to prevent instabilities when overall energy is small. (In current implementation, this bias is set to about 1 % of the maximum (expected) energy.) The final set of normalized oriented energies comprise an image representation that captures the local first-order geometric structure of image patterns with robustness to contrast variation. In what follows, matching operations that serve to define the correspondence between the video and projected reference will be performed with respect to normalized oriented energy images, as defined by formulas (4) and (5). For the sake of simplicity, this dependency will not be made explicit in the notation.

## 2.3. Correspondence

The third algorithmic step in the developed approach is to establish a detailed spatial correspondence between the video and projected reference imagery and thereby effect a precise geometric alignment of the two. There are a number of challenges that must be met in establishing such a correspondence. Uncertainty in camera geometry (e.g., as provided by telemetry) necessitates large search ranges. Matching with small spatial support leads to ambiguity; matching with large spatial support can be too sensitive to change between video and reference. In response to these issues, a local to global matching scheme is employed, with progressive refinement. This scheme operates by simultaneously considering a collection of consecutive video frames, that moves across the input stream with a sliding window. For example, in current implementation collections of 3 "key" frames are considered at a time, with key frame selected to have 50% overlap and their frame-to-frame alignments taken as the concatenation of the video rate frame-to-frame estimates.

Initially, correspondences are established on the basis of purely local matching (i.e., matching between single video frames and projected reference). Subsequently, a global alignment is established via a procedure that simultaneously considers all local correspondences for all frames under consideration to estimate a set of alignment parameters optimized for all, i.e., akin to a bundle adjustment [20]. This two stage (local/global) matching scheme iterates in a progressive refinement framework. Early iterations effect coarse alignment via consideration of matching primitives based on low spatial frequency information derived with large spatial support to serve large search ranges but low-order alignment models. Later iterations effect fine alignment via consideration of matching primitives based on higher spatial frequency information to serve smaller search ranges and higher-order alignment models. At each stage, results at the previous stage serve as initial conditions, with telemetry providing the initial estimate for the entire routine. The next several paragraphs detail the local and global matching as well as the progressive refinement strategy.

**Local matching** Local matching concentrates on establishing correspondences between individual video frames and the reference image. Primitives for this stage of processing are spatially overlapping tiles that define a grid over a frame of concern. The size and scale (i.e., pyramid level) of the tiles vary according to the progressive refinement strategy. In current implementation, three refinements are performed, coarse, medium and fine. As examples: For the coarsest stage of refinement, a tile is the entire image taken at level 3 in a Gaussian pyramid; at the finest stage of refinement, a 6 × 8 grid of tiles is formed at pyramid level 1. (Recall that the oriented energy representation is built on top of the (Gaussian) pyramid levels to effect an overall bandpass characteristic in the filtering of the tiles.)

Primitives (i.e., tiles) are matched in terms of correlation search in translation over the (projected) reference image. The search range varies according to the progressive refinement iteration, from hundreds of pixels at the coarsest stage to single pixels at the finest. The resulting match (for a primitive) is represented as a function $\Gamma_{v_j}(\mu, \nu)$ giving "probability" that point $j$ in a given video frame, $v$, has displacement $(\mu, \nu)$ in a reference image, $r$, c.f. [1]. Match functions are computed as normalized correlation scores of a patch about $j$ in $v$ shifted by $(\mu, \nu)$ in $r$ (i.e., a discrete correlation surface). Independent correlation surfaces are computed for each band in the oriented energy representation, which subsequently are multipled together to establish consensus. By representing the local matches in terms of correlation surfaces, it is possible to eschew assigning unique displacement vectors where they are unwarranted (e.g., aperture effects) and make use of more of the available information in subsequent processing.

To serve as a constraint on global matching, the dominant peak, i.e., highest value, in a correlation surface, $\Gamma_{v_j}$ is further characterized in terms of its covariance structure. In particular, let $\gamma(\mu, \nu)$ correspond to a portion of the correlation surface $\Gamma$ that derives from its dominant mode. Support for $\gamma$ is recovered based on a mean-shift procedure that iteratively reassigns points to local maxima in the distribution of $\Gamma$ [4]. The covariance of $\gamma$ is defined as

$$\mathsf{C} = \Sigma^{s}_{\mu,\nu=-s}\gamma(\mu,\nu)\tilde{\mathbf{p}}\tilde{\mathbf{p}}^{\top} / \Sigma^{s}_{\mu,\nu=-s}\gamma(\mu,\nu)$$

where $\tilde{\mathbf{p}} = (\mu - \mu_0, \nu - \nu_0)^{\top}$ with $\mu_0 = \Sigma\mu\gamma(\mu,\nu)/\Sigma\gamma(\mu,\nu)$, similarly for $\nu_0$ and limits on the summation the same as for $\mathsf{C}$, i.e, so as to cover $\gamma$. Finally, the scale of the peak is used as a normalization to yield a measure of covariance shape

$$\tilde{\mathsf{C}} = \mathsf{C}Tr\left(\mathsf{C}^{-1}\right)/2.$$

The final component of local matching is outlier rejection. Due to the difficulty of the matching problem that is under consideration, false local matches are likely and must be removed to avoid corruption of global matching. Indeed, experience has shown that false matches can exceed 50% in the initial matching stage. Since true matches must be consistent with some alignment model, RANSAC [6] is applied on a frame-by-frame basis to estimate that model and remove outliers among the matches. (The specific alignment models considered will be introduced when global matching is considered below. For now it suffices to note that the models can be represented parametrically to map between (projected) reference and video imagery, i.e., according to transformation (1).) For current purposes, the residual, $R^2_j$, for point $j$ used in the RANSAC computation is taken to be the covariance weighted distance

$$R^2_j = (\mathbf{p}_{v_j} - \tilde{\mathbf{p}}_{v_j})^{\top}\tilde{\mathsf{C}}^{-1}(\mathbf{p}_{v_j} - \tilde{\mathbf{p}}_{v_j})$$

with $\tilde{\mathbf{p}}_{v_j} = (\tilde{\mu}_v, \tilde{\nu}_v)_j$ mappings of the reference point $\mathbf{p}_{r_j}$ into the video under the current trial's estimated alignment model. For cases where RANSAC cannot be defined, i.e., coarse matches of entire video frames to projected reference, outliers are rejected by dropping matches that do not derive from unimodal correlation surfaces, $\Gamma$. The overall result is a set of (local) video to reference matches for each frame under current consideration, all of which are to be considered during global matching.

**Global matching** Global matching is accomplished with respect to an operative parametric alignment model, $\mathsf{Q}_{r,v}$, that maps between the (projected) reference(s) $r$ and video frames $v$ to serve in essence as a camera model. Estimation proceeds by simultaneously recovering parameters for a set of mappings for a corresponding set of video frames under consideration (in a sliding temporal window), akin to the photogrammetric notion of bundle adjustment [20]. During this process multiple projected references also can be under consideration as it becomes necessary to project more than one view of the reference orthoimage to accommodate the extent of the current collection of video frames.

Most generally, the form of the mapping $\mathsf{Q}_{r,v}$ is as given in the general transformation (1), with input $\mathbf{m}_{r_j} = (x_j, y_j, 0, 1)^{\top}_{r}$ and output $\mathbf{m}_{v_j} = (x_j, y_j, 0, w_j)^{\top}_{v}$. Depending on the stage of progressive refinement, different alignment models are employed, with models varying from lower to higher order as refinement proceeds coarse to fine. During the coarse iteration, a global shift is employed to establish the center of projection, i.e.,

$$\mathsf{Q}^{\mathsf{shift}}_{r,v} = \begin{pmatrix} 1 & 0 & 0 & a_{14} \\ 0 & 1 & 0 & a_{24} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

During the intermediate iteration, an affine camera model is employed, i.e., $\mathsf{Q}^{\mathsf{affine}}_{r,v}$ has the same form as the frame-to-

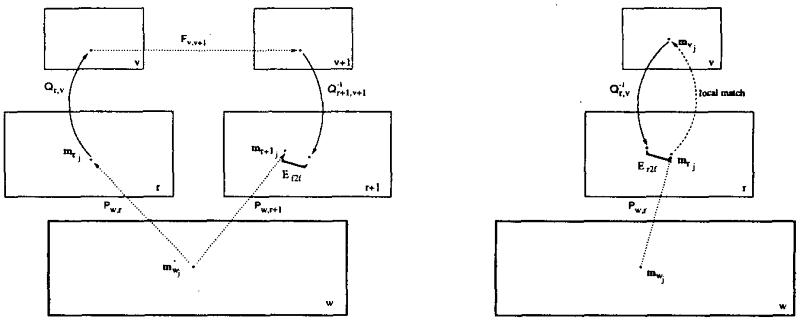

Figure 2: Geometric constraints on global alignment. Frame-to-frame constraints and reference-to-frame constraints are shown in the left and right panels, respectively. Dotted arrows show (i) known relations, $\mathsf{P}_{w,r}, \mathsf{P}_{w,r+1}$ that map the world $w$ to projected references $r$ and $r + 1$ based on telemetry and (ii) constraints, frame-to-frame constraints $\mathsf{F}_{v,v+1}$ that maps video frame $v$ to $v + 1$ and reference-to-frame constraints from local matches between reference and video. Solid arrows show transformations to be recovered, Q's that map between the projected references and video frames. Square brackets show error to be minimized.

frame transformation $\mathsf{F}^{\mathsf{affine}}_{v,v+1}$, (3). Finally, during fine alignment, a 2D projective camera model is employed, i.e.,

$$\mathsf{Q}^{\mathsf{2dProj}}_{r,v} = \begin{pmatrix} a_{11} & a_{12} & 0 & a_{14} \\ a_{21} & a_{22} & 0 & a_{24} \\ 0 & 0 & 1 & 0 \\ a_{31} & a_{32} & 0 & a_{34} \end{pmatrix}.$$

Significantly, this last model is related to the general 3D projective camera model, e.g., as used for projecting the reference to video frame (2), via an assumption that the $z$ terms deviate little from a plane to yield a sparser projection matrix, in particular, an homography mapping between two planes [5]. The local planarity assumption is based on the ability of the projection of the reference to the video coordinate frame to compensate for significant 3D relief effects. In theory, it would be possible to include additional modeling stages, e.g., full 3D projective (2); however, in practice halting with 2D projective has allowed for acceptable alignments without over fitting of the data.

Given an operative alignment model, global matching takes into consideration 2 sets of constraints in order to achieve an overall best alignment between the portion of the video under consideration and the reference imagery, see Figure 2. (i) *Frame-to-frame* constraints are derived from the frame-to-frame alignments that were computed as part of the (video) image representation. (ii) *Reference-to-frame* constraints are derived from the reference-to-(video)-frame matches that were computed by local matching. The next 2 paragraphs detail these constraints.

Frame-to-frame constraints embody the frame-to-frame alignments that were computed as a part of the image preprocessing. Deviations from these constraints are measured in terms of the geometric displacement between the same point $j$ as it appears in a projected reference, and the mapping of that point onto a video frame, then to the next frame and finally back to (projected) reference. In the most general case, two projected references will be involved (i.e., when the two video frames involved are related to separate projected references), leading to an error term of the form

$$E_{f2f} = \delta\left(\mathbf{m}_{r+1_j}, \mathsf{Q}^{-1}_{r+1,v+1}\mathsf{F}_{v,v+1}\mathsf{Q}_{r,v}\mathbf{m}_{r_j}\right),$$

where the reference point pair $\mathbf{m}_{r_j}, \mathbf{m}_{r+1_j}$ are obtained using the (known) world to projected reference mapping (i.e., as provided by transformation (2)), while the composite projection is a chain of mappings from reference $r$ to frame $v$, then to frame $v+1$ and finally to reference $r+1$. Here, frame-to-frame alignment is given by $\mathsf{F}_{v,v+1}$ in accord with transformation (3) as a constraint, while mappings from video frames $v, v + 1$ to (projected) references $r, r + 1$ are described by $\mathsf{Q}^{-1}_{r,v}, \mathsf{Q}^{-1}_{r+1,v+1}$, respectively and $\delta(\mathbf{m}_\alpha, \mathbf{m}_\beta)$ is a distance metric. In current implementation, this metric is instantiated in terms of the covariance weighted Euclidean distance between the equivalent 2D (image) coordinates, i.e., $\mathbf{p}_\alpha = (x_\alpha/w_\alpha, y_\alpha/w_\alpha)^\top$ and similarly for $\beta$ to yield

$$\delta(\mathbf{m}_\alpha, \mathbf{m}_\beta) = (\mathbf{p}_\alpha - \mathbf{p}_\beta)^\top \tilde{\mathsf{C}}^{-1}(\mathbf{p}_\alpha - \mathbf{p}_\beta)$$

As a special case, if the two video frames, $v$ and $v+1$, related via a frame-to-frame constraint, $\mathsf{F}_{v,v+1}$, map to the same projected reference, then the frame-to-frame error term has the same form but with $r$ and $r + 1$ equated.

Reference-to-frame constraints embody the local matches that were computed during the first stage of the correspondence process. Deviations from these constraints are measured in terms of the geometric displacement between a local match and mapping of the same point $j$ onto a common reference from a corresponding video frame, i.e.,

$$E_{r2f} = \delta\left(\mathbf{m}_{r_j}, \mathsf{Q}^{-1}_{r,v}\mathbf{m}_{v_j}\right)$$

with $\mathbf{m}_{r_j}$ the position of point $\mathbf{m}_{v_j}$ in reference $r$ given by local matching and $\mathsf{Q}_{r,v}$ to be estimated in global matching.

Combination of the frame-to-frame and reference-to-frame error terms leads to a total error

$$E = \sum \left(\alpha_1 E^2_{f2f} + \alpha_2 E^2_{r2f}\right)$$

that is to be minimized with respect to the reference-to-frame mappings $\mathsf{Q}_{r,v}$. Here, summation is taken over all local matches computed for a set of video frames under simultaneous consideration and weights $\alpha_{1,2}$ determine the relative contribution of each error terms. In current implementation, $\alpha_1$ and $\alpha_2$ are equal. Minimization of $E$ is accomplished in a

| Test Case | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| GSD (m/pixel) | 0.7 | 2.1 | 0.7 | 2.4 | 2.6 | 1.2 | 0.5 | 0.2 |
| Obliquity (deg) | 45 | 45 | 45 | 60 | 60 | 15 | 70 | 15 |
| 3D relief | flat | hilly | mountain | flat | hilly | flat | flat | flat |
| Ground Cover | urban/forest | farm/forest | bare earth | farm/forest | farm/forest | urban/forest | urban/forest | urban/forest |
| ref/vid similarity | similar | different | similar | different | different | similar | similar | similar |

Figure 3: Characteristics of test data sets.

weighted least squares sense, with weights provided by local match covariance. Estimation of the parameters for the $Q_{r,v}$ is performed via the Levenberg-Marquardt method [9].

From global matching a detailed registration of the two sources of imagery allows the video to inherit the geodetic coordinates of the reference. Also, via a composition of the recovered alignment models with the initial telemetry-based camera model, adjusted model parameters are produced that reflect the available video and reference information.

# 3. Empirical evaluation

This section presents a study that quantitatively evaluates the accuracy of the developed algorithm for video georegistration as a function of several key factors. In particular, variables of concern are (i) video ground sampling distance, GSD, (meters/pixel derived from telemetry), (ii) obliquity of camera angle (degrees from nadir derived from telemetry, i.e., nadir $\equiv$ 0), (iii) terrain 3D relief (e.g., flat, hilly, mountainous, from human inspection of DEM), (iv) terrain surface cover (e.g., urban, farm, forest, bare earth, from human inspection of ortho and video) and (v) reference/video appearance similarity (e.g., same/different due to seasonal variation, from human inspection of ortho and video). These factors have been selected to shed light on the developed approach for both researchers and potential users.

Eight specific test cases are documented in Fig. 3, which cover a wide range of parameter values for the experimental factors of interest. Each test case corresponds to a 2 minute video clip with supporting telemetry captured from an aerial platform flying along (mostly) straight paths at approximately 80 knots. Corresponding reference data consists of USGS digital orthophoto quarter-quads (1 meter GSD) and NIMA Digital Terrain Elevation Data Level 1 ($\approx$100 meter postings) [19]. Registration accuracy is reported in terms of absolute Euclidean distance of registered points from hand mensurated ground truth. Error is calculated in orthoimage coordinates, i.e., 1 pixel implies 1 meter. Hand mensurated ground truth was built by human operators using a GUI to select corresponding points in video and reference. An attempt was made to select both natural and artificial feature correspondences. On average, 150 ground truth points were selected for each video clip.

Fig. 4 presents test case 1 results. Visual inspection of the ortho and video for the featured frame in the top row shows the considerable geometric compensation required to effect registration. Even following projection of the reference via telemetry, gross misalignment remains. Nevertheless, the final results of automated alignment yield a precise registration of the video to reference. Quantitative results compiled across the entire test case are shown in the bottom row. Comparison of the error histograms for telemetry only and final registration shows vast overall improvement from the proposed algorithm. Consideration of mean and maximum error as functions of percentage of total mensurated points further underlines this improvement. The final mean error is under 10 meters (pixels) at all percentiles; the final maximum error is under 10 meters through the 80th percentile and never above 35 meters. Remaining errors in final registration arise mostly when nondistinctive appearance makes fine correspondence ambiguous (e.g., uniform vegetation).

Fig. 5 presents test case 2 results. In this case, there is an extreme appearance change between the reference and video due to their being acquired in summer and winter, respectively. (Test cases 4 and 5 derive from similar conditions.) Here, the algorithm is still capable of providing a precise registration as shown in the final overlay. Note, for example, how the frozen bodies of water in the video have been precisely aligned with their reference image counterparts despite even full contrast reversal, e.g., a bit less than half way down the left side of the final result. The error plots show quantitatively the considerable improvement afforded by the algorithm over telemetry only registration. Finally, note that worst case errors in registration for this test are inflated compared to test 1 due predominantly to the extreme appearance variation between reference and video.

Fig. 6 presents test case 3 results. Challenging aspects of this case include high 3D relief (600m elevation variation across the clip) and the paucity of features to drive registration over bare earth surface cover. In this case the algorithm can still perform well: The projection of the reference orthoimage to the video frame, taking into account reference elevation data compensates for much of the relief effects; the approach to image representation and matching allows for exploitation of the available features (e.g., ridge and gully lines). This allows, e.g., the average error to be under approximately 15m for 90% of the data. Maximum error also is generally improved; although at 90% this is not the case as registration does not adequately compensate over a region of particularly high relief variation. This limitation lies in the coarse resolution of the reference elevation data that has been used as it does not support the projection of the reference to account for the operative 3D effects in the region under consideration. Improved reference data would greatly ameliorate this type of error. Overall, this case represents the algorithm's worst performance for the set under consideration. Still, improvement is demonstrated over telemetry.

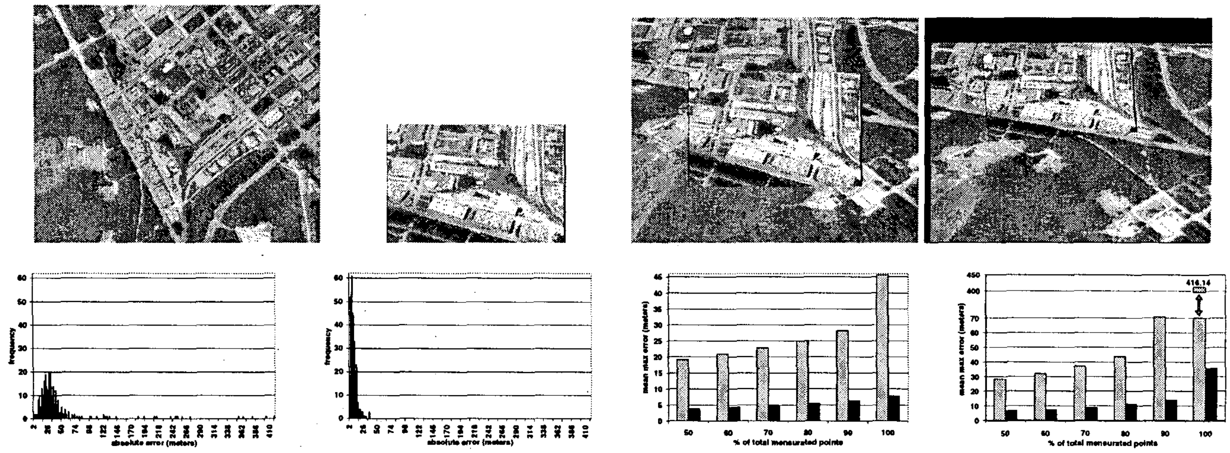

Figure 4: Test Case 1 Results. Top row: Reference orthoimage, source video frame, video overlay on projected reference (telemetry-based), video overlay on projected reference (final result). Same projected reference used in both cases, but windowed differently in display to center approximately the video overlay. Orthoimage and source video not to scale for sake of visualization. Bottom row: Absolute Euclidean error histogram (telemetry-based), Absolute Euclidean error histogram (final result), mean Absolute Euclidean error vs. percentile, maximum error vs. percentile. Light and dark bars show telemetry-based and final results, respectively.

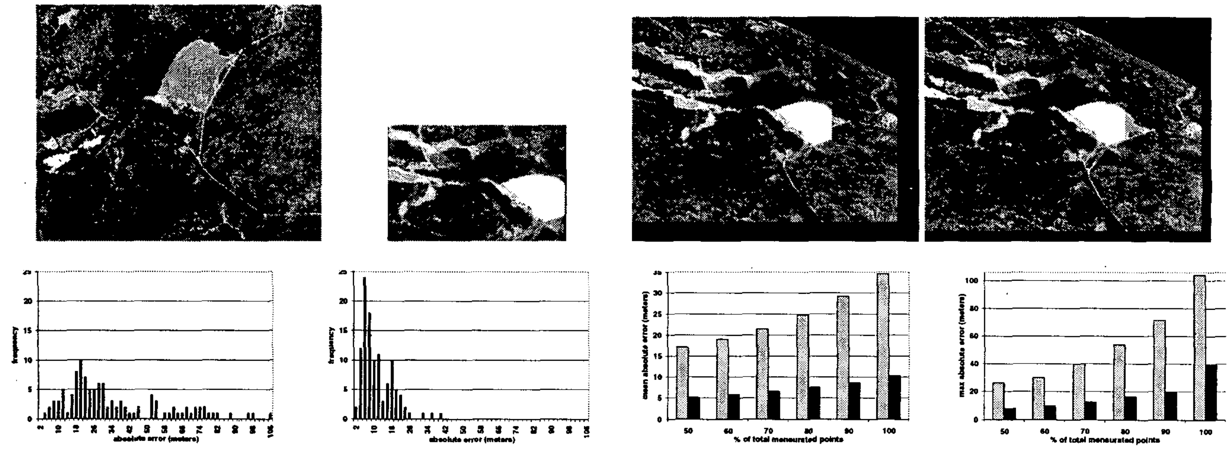

Figure 5: Test Case 2 Results. Format analogous to that of Figure 4.

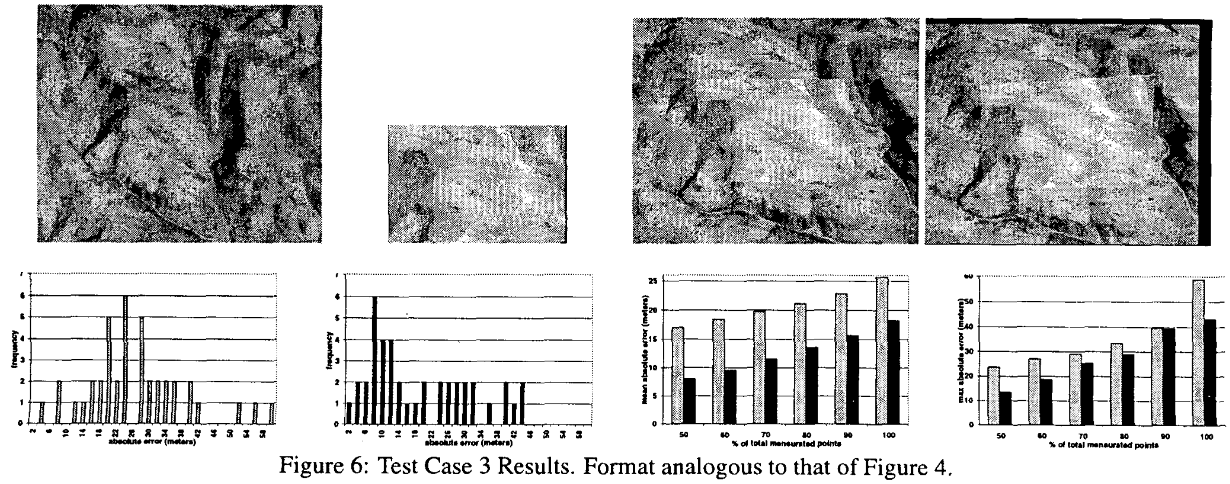

Figure 6: Test Case 3 Results. Format analogous to that of Figure 4.

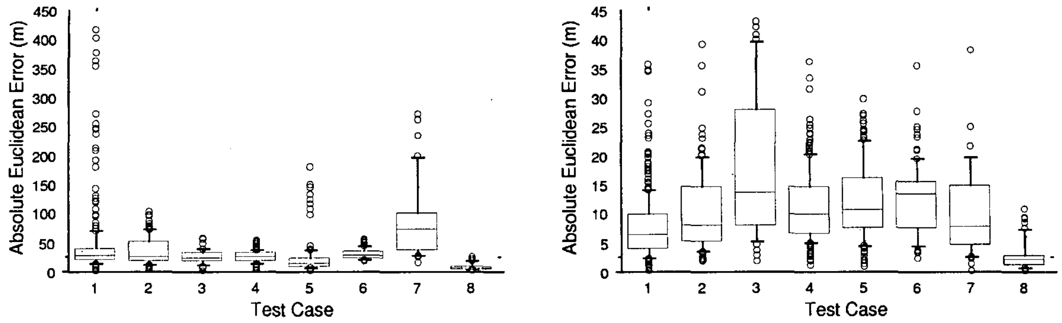

Figure 7: Summary of Telemetry vs. Proposed Algorithm Georegistration Error. Absolute Euclidean error shown as box plots for registration based on telemetry only (left) and proposed algorithm (right) for test cases of Figure 3. *Note that the ordinates differ in scale by a factor of 10.* The extent of a vertical box covers the middle 50% of the data (interquartile range). The line in the interior of a box shows the median score. The lines below and above a box extend to the 10th and 90th percentiles, respectively. Data points that fall beyond the lines are shown as individual circles.

Further, the present limitations lie more in the quality of the reference data (i.e., elevation) than in the algorithm per se.

Summary results for all 8 test clips, both telemetry-based and final algorithm registration, are shown as Box Plots [18] in Fig. 7. (Note that the ordinate scale on the telemetry results is a factor of 10 greater than the final results.) A number of overall observations are possible: (i) The best telemetry-based registration is worse than the worst case algorithm registration, save case 8. Close examination of case 8 shows that the algorithm still improves on the telemetry by a factor of 2. (ii) The central tendencies of the registration results are smaller and less variable than the telemetry results: While the median error scores for the final results span approximately 2-14 meters, the corresponding span for telemetry is approximately 7-75. (iii) The dispersion of the results is smaller following algorithm registration for every case. As noted above, worst case performance is due predominantly to attempts to register in the presence of nondistinctive image appearance (leading to ambiguous correspondence) and low resolution reference elevation data (limiting ability to account for 3D relief). In future work, these limitations could be addressed by making use of larger collections of video frames for global matching (to increase spatial context and disambiguate matches) and incorporating improved resolution reference elevation data. Overall, the results show the algorithm's ability to perform in a robust fashion with accuracy and precision across a wide range of challenging cases.

## 4. Summary

An algorithm for video georegistration has been presented. The input to the algorithm is a video stream with telemetry and geodetically calibrated reference imagery. The output is a spatial registration of the video to the reference so that the video inherits the available geodetic coordinates. The video is processed in a continuous fashion to yield a corresponding stream of georegistered results. The algorithm has been quantitatively evaluated through empirical testing with data derived from the application domain of most interest, aerial video. The results of this evaluation show that the developed approach is robust in the presence of challenging test cases and capable of producing accurate and precise registration of video to reference. On the basis of these results, it is suggested that the developed approach can provide valuable input to the analysis and interpretation of aerial video.

## References

[1] M. Ben-Ezra, S. Peleg, M. Werman, "Robust real-time motion analysis," In *Proc. DARPA IUW*, 207–210, 1998.

[2] J. Bergen, P. Anandan, K. Hanna, R. Hingorani, "Hierarchical model-based motion estimation," In *Proc. ECCV*, 237–252, 1992.

[3] L. Brown, "A survey of image registration techniques," *ACM Computing Surveys 24 (2)*, 325–376, 1992.

[4] D. Comaniciu, P. Meer, "Mean shift analysis and applications," In *Proc. ICCV*, 1197–1203, 1999.

[5] H. Coxeter, *Projective Geometry*, Springer, Berlin, 1994.

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

[7] J. Foley, A. van Dam, S. Feiner, J. Highes, *Computer Graphics*, Addison-Wesley, Reading, MA, 1990.

[8] W. Freeman, E. Adelson, "The design and use of steerable filters," *IEEE PAMI 13 (9)*, 891–906, 1991.

[9] P. Gill, W. Murray, M. Wright, *Practical Optimization*, Academic, NY, NY, 1981.

[10] B. Horn, B. Bachman, "Using synthetic images to register real images with surface models," *CACM 21*, 914-924, 1978.

[11] B. Jähne, *Digital Image Processing*, Springer, Berlin, 1988.

[12] T. Jamison, C. Davis, M. Lucas, "Automated georeferencing of video," In *Proc. ASPRS*. 2000.

[13] R. Kumar, S. Samarasekera, S. Hsu, K. Hanna, "Registration of highly-oblique and zoomed in aerial video to reference imagery," In *Proc. ICPR*, 303-307, 2000.

[14] S. Merhav, Y. Bresler, "On-line vehicle motion estimation from visual terrain information," *TAES 22 (5)*, 588–604, 1986.

[15] P. Pope, F. Scarpace, "Development of a method to geographically register airborne scanner imagery," In *Proc. ASPRS*. 2000.

[16] J. Rodriquez, J. Aggarwal, "Matching aerial images to 3D terrain maps," *IEEE PAMI 12 (12)*, 1138–1149, 1990.

[17] D. Sim, R. Park, "Localization based on the gradient information for DEM matching," In *Proc. MVA*, 266-269, 1998.

[18] J. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

[19] USGS, http://edcwww.cr.usgs.gov/Webglis/glisbin/guide.pl/glis/hyper/guide/usgs_doq.

[20] P. Wolf, *Elements of Photogrammetry*, McGraw, NY, 1983.