Stabilization of Airborne Video Using Sensor Exterior Orientation with Analytical Homography Modeling

Hadi Aliakbarpour, Kannappan Palaniappan, and Guna Seetharaman

[1] Computational Imaging and VisAnalysis (CIVA) Lab, EECS, University of Missouri-Columbia, USA.
[2] Advanced Computing Concepts, U.S. Naval Research Laboratory, Washington, DC, USA.

Abstract. Aerial video captured from an airborne platform has an expanding range of applications including scene understanding, photogrammetry, surveying and mapping, traffic monitoring, bridge and civil infrastructure inspection, architecture and construction, delivery, disaster and emergency response, news and film, precision agriculture and environmental monitoring and conservation. Some of the challenges in analyzing aerial video to track pedestrians, vehicles and objects include small object size, relative motion of the object and platform, sensor jitter and quality of imaging optics. An analytic image stabilization approach is described in this chapter where pixel information from the focal plane of the camera are stabilized and georegistered in a global reference frame. The aerial video is stabilized to maintain a fixed relative displacement between the moving platform and the scene. The proposed algorithm can be used to stabilize aerial imagery even when the available GPS and IMU measurements from the platform and sensor are inaccurate and noisy. Camera 3D poses are optimized using a homography-based robust cost function, but unlike most existing methods, the homography transformations are estimated without using any image-to-image estimation techniques. We derive a direct closed-form analytic expression from 3D camera poses that is robust even in the presence of significant scene parallax (i.e. very tall 3D buildings and manmade or natural structures). A robust nonlinear least squares cost function is used to deal with outliers and speeds up computation by avoiding the use of RANdom SAmple Consensus (RANSAC). The proposed method and its efficiency is validated using several datasets and scenarios including DARPA Video and Image Retrieval and Analysis Tool (VIRAT) and high resolution Wide Area Motion Imagery (WAMI). scenarios.

Keywords: Video stabilization, image registration, georegistration, camera parameter optimization, homography, parallax, 3D

1 Introduction

Wide Area Motion Imagery (WAMI), also known as, wide area aerial surveillance (WAAS), wide-area persistent surveillance (WAPS) or wide field-of-view (WFOV) imaging is an evolving imaging capability that enables persistent coverage of large geographical regions on the order of a few to tens of square miles [1] at tens of centimeter resolution, or very small areas such as bridges and construction projects at very high

resolution from closer range using the same sensor package. It has become even more popular due to performance advances in sensor technologies, computing hardware, battery performance and reduction in size, weight and cost of these components. WAMI sensors can be placed on many types of airborne platforms including fixed wing or multi-rotor Unmanned Aerial Vehicles (UAVs) - both fixed wing and multi-rotor drones, small (manned) aircraft and helicopters [2]. Depending on the imaging sensor characteristics and aircraft altitude, these systems can cover a small city-sized area with an approximate Ground Sampling Distance (GSD) of 10 cm to 30 cm per pixel, tens to hundreds of megapixels at the focal plane using single or multiple optical systems (e.g. 6600×4400 RGB color) with a frame rate of 1 to 10 Hz.

Detection of small and distant moving objects, e.g. cars or pedestrians, in a scene which is observed by a camera that by itself undergoes motions and jitters is extremely challenging. This can be even more challenging considering that small objects like cars may appear as 10 to 25 pixels in their length. To improve detection and tracking in aerial imagery [3–5] in which videos are captured on a moving platform, the images are stabilized (registered) to maintain the relative movement between the moving platform and the scene fixed. An accurate image stabilization in such scenarios can be important for both higher level video analytics and visualization. Traditionally, aerial image registration methods are performed through applying 2D homography transformations in the *image space* [6–8]. Aerial image registration is challenging for urban scenes where there are large 3D structures (tall buildings) causing high amount of occlusion and parallax. In such situations, the presence of parallax can lead to significant error when inter-image 2D registration approaches are used [9].

In this paper, a method to register aerial images is proposed which utilizes available (noisy or approximate) GPS and IMU measurements from the airborne platform and robustly stabilize images by optimizing camera 3D poses using a homography-based cost function. Unlike most existing methods, the homography transformations in our approach are not estimated using any image-to-image estimation techniques, but directly derived as a closed-form analytic expression from the 3D camera poses. In our previous work we leveraged our fast Structure-from-Motion (SfM) technique (BA4S[10, 11]) to derive a novel georegistration approach that did not need to estimate local patch-based homographies and used an analytical model that was both accurate and fast [12]. Although that approach was fast and globally accurate, its cost function is defined over the full 3D space in order to optimize the *retinal plane reprojection pixel error* over the full 3D scene as required by most SfM downstream applications (e.g. dense 3D reconstruction [13–16]). However, as an alternative to full SfM-based georegistration, we propose to stabilize an image sequence or remove jitter, with the objective of deriving a smooth motion trajectory over the sequence of images such that the dominant ground plane is stabilized minimizing a 2D metric distance-based error function. Therefore, in this paper we propose an alternative approach for the parameter optimization with an emphasize on stabilizing the geoprojected aerial imagery by defining a cost function over a *single dominant 2D Euclidean world plane*. The points that do not lie on the dominant are automatically marginalized during the optimization process. In the experiments we will show that the method proposed in this paper is more robust in situations where the available camera sensor pose measurements are extremely inaccurate.

1.1 Related Work

The majority of approaches for image stabilization use pairwise and combinatorial matching and warping transformation for stabilizing the ground plane prior to moving object detection [6–8, 17–26]. Aerial image registration is challenging for urban scenes where there are large 3D structure and tall buildings causing high amount of occlusion and parallax [9, 27, 12]. An aerial image registration method was proposed in [6, 28] which uses a multi-layer (coarse to fine) homography estimation approach to deal with parallax and occlusions. Molina and Zhu [17] proposed a method to register nadir aerial images in which a pyramid block-based correlation method was used to estimate inter-frame affine parameters.

Direct georeferencing of high-resolution unmanned aerial vehicles (UAV) imagery was discussed in [29] while performances of different SfM softwares (Photoscan [30], Pix4D [31] and Bundler [32, 33]) were evaluated. Pritt [34] proposed a fast orthorectification method for registration of thousands of aerial images (acquired from small UAVs). In [35], IMU was used to register laser range measurements to the images captured from a stereo camera. Crispell et al. introduced an image registration technique to deal with parallax, assuming to have a dense 3D reconstructed model of the scene [9]. In [36] GPS and IMU were used to perform an initial (coarse) orthorectification and georeferencing of each image in an aerial video. Then a RANSAC-based method was used to find optimal affine transformations in 2D image space. A method for registering and mosaicking multi-camera images was proposed in [7]. In the proposed method, registration is achieved using control points and projective image-to-image transformations (using a variation of RANSAC). Recently, some image-based methods for robust registration (mosaicking) of long aerial video sequences have been introduced in [37–39].

2 Feature Track Building

In persistent aerial imagery, images are sequentially acquired meaning that one knows that what frame is adjacent to which one. By leveraging the temporal consistency of the images and using them as a prior information, the time complexity of matching can be reduced to $O(n)$. Interest points are extracted from each image using a proper feature extraction method. Starting from the first frame, for each two successive image frames, the descriptors of their interest points are compared. While successively matching them along the sequence, a set of feature *tracks* are generated [40]. A track basically indicates that a potentially unique 3D point in the scene has been observed in a set of image frames.

3 Imaging Model

Fig. 1 shows a world coordinate system W and a dominant ground plane π spanning through its X and Y axes. The scene is observed by n airborne cameras $C_1, C_2 \ldots C_n$. To make the notations succinct, we will omit the camera indices from now on unless otherwise stated. The image homogeneous coordinate of a 3D point $\mathbf{X} = [x \;\; y \;\; z]^\intercal$ from

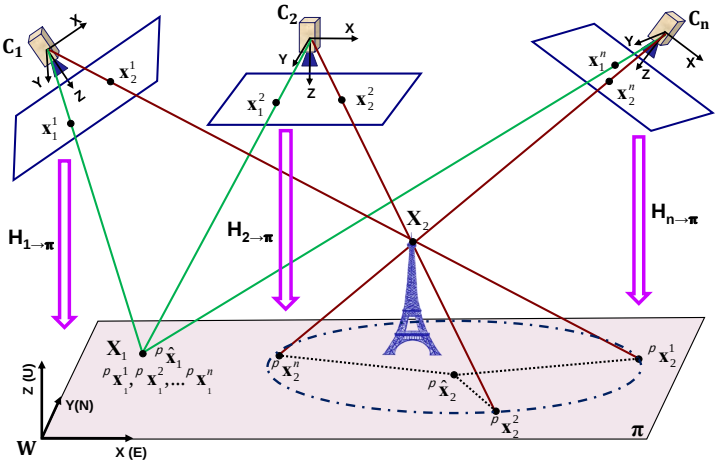


Fig. 1. A scene and its dominant ground plane π is observed by n airborne cameras. For an on-the-plane 3D point such as $\mathbf{X}_1$, it homographic transformation from the image plane of every single camera onto π, all merge together and converge to the same identical 3D point $\mathbf{X}_1$. Whereas, for an off-the-plane 3D point such as $\mathbf{X}_2$, its homographic transformations are spread out (diverged).

the world reference system W projected on the image plane of camera C is obtained as

$$\tilde{\mathbf{x}} = \mathbf{K}(\mathbf{R}\mathbf{X} + \mathbf{t}) \tag{1}$$

where $\mathbf{K}$ is the calibration matrix (intrinsics), $\mathbf{R}$ and $\mathbf{t}$ are respectively the rotation matrix and translation vector from W to C. For a 3D point $\mathbf{X}$ lying on π, its Z component is zero:

$$\tilde{\mathbf{x}} = \mathbf{K}(\begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{r}_3\end{bmatrix} \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} + \mathbf{t}) \tag{2}$$

$\mathbf{r}1$, $\mathbf{r}2$ and $\mathbf{r}3$ being the first, second and third columns of $\mathbf{R}$, respectively. After simplification we have

$$\tilde{\mathbf{x}} = \mathbf{K}\begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix} {}^{\pi}\tilde{\mathbf{x}} \tag{3}$$

where ${}^{\pi}\tilde{\mathbf{x}} = [x \;\; y \;\; 1]^{\mathsf{T}}$ represent the 2D homogeneous coordinates of the 3D point $\mathbf{X}$ on π. One can consider the term $\mathbf{K}[\mathbf{r}_1 \;\; \mathbf{r}_2 \;\; \mathbf{t}]$ as a 3×3 homography transformation matrix which maps any 2D point from π onto the camera image plane as:

$$\tilde{\mathbf{x}} = \mathbf{H}_{\pi \to c} {}^{\pi}\tilde{\mathbf{x}}. \tag{4}$$

Likewise, a homogeneous image point $\tilde{\mathbf{x}}$ can be mapped on π as:

$$^{\pi}\tilde{\mathbf{x}} = \mathbf{H}_{c\to\pi}\,\tilde{\mathbf{x}} \tag{5}$$

where $\mathbf{H}_{c\to\pi}$ is the inverse of $\mathbf{H}_{\pi\to c}$ and is equal to:

$$\mathbf{H}_{c\to\pi} = \begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix}^{-1}\mathbf{K}^{-1}. \tag{6}$$

Assuming $\mathbf{T} = \begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t}\end{bmatrix}$, f as the focal length in pixel, and (u,v) as the camera image principal point, (6) can be expressed as:

$$\mathbf{H}_{c\to\pi} = \mathbf{T}^{-1}\begin{bmatrix} f & 0 & u \\ 0 & f & v \\ 0 & 0 & 1 \end{bmatrix}^{-1} \tag{7}$$

$$\mathbf{H}_{c\to\pi} = \frac{1}{\lambda}\begin{bmatrix} m_{11} & -m_{21} & \begin{bmatrix}-m_{11} & m_{21} & m_{31}\end{bmatrix}\mathbf{v} \\ -m_{12} & m_{22} & \begin{bmatrix}m_{12} & -m_{22} & -m_{32}\end{bmatrix}\mathbf{v} \\ r_{13} & r_{23} & -\mathbf{r}_3^{\intercal}\mathbf{v} \end{bmatrix} \tag{8}$$

where $\mathbf{v} = \begin{bmatrix}u & v & f\end{bmatrix}^{\intercal}$ and λ is a scalar defined as

$$\lambda = f\mathbf{r}_3^{\intercal}\mathbf{t}, \tag{9}$$

and m_{ij} is the $minor(i,j)$ of matrix $\mathbf{T}$. One can omit λ in (8) as a homography matrix is defined up-to-scale, yielding:

$$\mathbf{H}_{c\to\pi} = \begin{bmatrix} m_{11} & -m_{21} & \begin{bmatrix}-m_{11} & m_{21} & m_{31}\end{bmatrix}\mathbf{v} \\ -m_{12} & m_{22} & \begin{bmatrix}m_{12} & -m_{22} & -m_{32}\end{bmatrix}\mathbf{v} \\ r_{13} & r_{23} & -\mathbf{r}_3^{\intercal}\mathbf{v} \end{bmatrix} \tag{10}$$

4 Optimization

Suppose our the global reference system W in Fig. 1 is aligned with NEU (North-East-Up). Reminding that π is the dominant ground plane in the scene and there are n cameras (or one camera in n different poses) observing the scene. The pose of each camera C_i is defined by a rotation matrix $\mathbf{R}_i$ and $\mathbf{t}_i$ which are defined from the global coordinate system to the camera local coordinate system. Also suppose to have m feature tracks in the scene. A feature track is basically a sequence of feature points which are matched across the sequence of image frames. All features within a track are the observations corresponding to a hypothetically identical 3D point in the scene. The homogeneous image coordinates of a 3D point $\mathbf{X}_j$ on the image plane of camera C_i is expressed as $\tilde{\mathbf{x}}_j^i$, and it can be mapped from image plane to the Euclidean plane π as

$$^{\pi}\tilde{\mathbf{x}}_j^i = \mathbf{H}_{i\to\pi}\,\tilde{\mathbf{x}}_j^i. \tag{11}$$

Ideally, if 3D point $\mathbf{X}_j$ lies on the plane π, then mapping of all its corresponding image observations ($\tilde{\mathbf{x}}_j^1, \tilde{\mathbf{x}}_j^2, \ldots \tilde{\mathbf{x}}_j^n$) onto the plane, using (11), have to merge to an identical 2D point on π, which also coincides on the 3D point $\mathbf{X}_j$ itself (see Fig. 1):

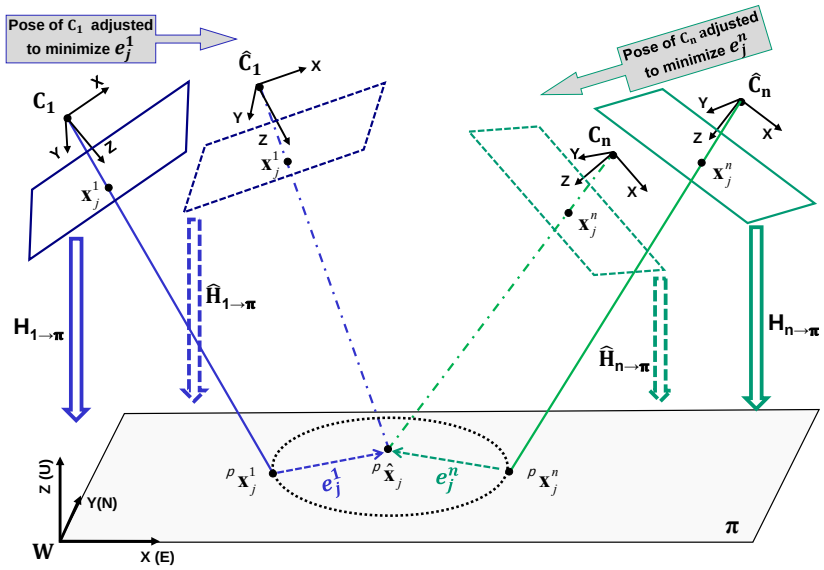


Fig. 2. The optimization scheme: Two matched image points in a features track j, $\mathbf{x}_j^1$ from $\mathbf{C}_1$ and $\mathbf{x}_j^n$ from $\mathbf{C}_n$, are available as the observations corresponding to hypothetically an identical 3D point $\mathbf{X}_j$ in the scene. The features points are projected on π using the analytical homographies defined in (10), where π is the dominant ground plane of the scene. If the 3D point $\mathbf{X}_j$ lies on π, its corresponding mapped homography points should be all close by each other. Indeed, in an ideal case where the camera poses are accurate, all such mapped points on π have to merge and coincide to a single point, however it is not often the case due to different source of noise in IMU and GPS measurements. Here, we use the mean of the homographic transformed points on π as an estimate to initialize the optimization. e_j^1 and e_j^n are the Euclidean distances between each projected point and the mean. The optimization defined in (15) aims to minimize these distance errors by adjusting the camera poses. Notice that if 3D point $\mathbf{X}_j$ does not lie on π, its error values are *automatically marginalized* during the optimization process, thanks to the used robust functions.

$$ {}^{\pi}\tilde{\mathbf{x}}_j^1 = {}^{\pi}\tilde{\mathbf{x}}_j^2 = \cdots = {}^{\pi}\tilde{\mathbf{x}}_j^n \simeq \mathbf{X}_j \tag{12} $$

However, it is not the case in real scenarios due to different source of errors such as inaccuracy in the measured camera poses (e.g. from GPS/IMU). Therefore the set of mapped 2D points, $\left\{ {}^{\pi}\tilde{\mathbf{x}}_j^i \,|\, i = 1 \ldots n \right\}$, corresponding to 3D point $\mathbf{X}_j$, will be dispersed around the actual point ${}^{\pi}\tilde{\mathbf{x}}_j$ on π. One can consider, ${}^{\pi}\hat{\mathbf{x}}_j$, the centroid of the distribution of 2D projected points, as an estimate for the actual point:

$$ {}^{\pi}\hat{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^{n} {}^{\pi}\tilde{\mathbf{x}}_j^i \tag{13} $$

The Euclidean distance between each mapped point ${}^{\pi}\tilde{\mathbf{x}}_j^i$ and the estimated centroid is considered as an error metric:

$$ e_j = \sum_{i=1}^{n} \left\| \mathscr{F}(\mathbf{H}_{i \to \pi}\, \tilde{\mathbf{x}}_j^i) - \mathscr{F}({}^{\pi}\hat{\mathbf{x}}_j) \right\|^2 \tag{14} $$

Overall error for all points and cameras can be used as a cost function to optimize $\mathbf{R}_i$, $\mathbf{t}_i$ and ${}^{\pi}\hat{\mathbf{x}}_j$ (see Fig. 2):

$$ E = \min_{\mathbf{R}_i, \mathbf{t}_i, {}^{\pi}\hat{\mathbf{x}}_j} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \mathscr{F}(\mathbf{H}_{i \to \pi}\, \tilde{\mathbf{x}}_j^i) - \mathscr{F}({}^{\pi}\hat{\mathbf{x}}_j) \right\|^2 \tag{15} $$

where $\mathscr{F}(.)$ designates a function that returns the Euclidean coordinates from 2D homogeneous ones. Such a minimization can be done through using various iterative optimization techniques among which Levenberg-Marquardt methods are well-known and popular in the literature [41]. Here, total number of parameters to be optimized is $6n + 2m$, where n is the number of views and m is number of tracks. Basically, each view i has 6 parameters including 3 for the rotation and 3 for the translation components. Likewise, each track j is represented by the 2D mean position vector, ${}^{\pi}\hat{\mathbf{x}}_j$, as expressed by (13). Total number of parameters in the observation space is $\leqslant 2 \times n \times m$. Note that the length of each track is $\leqslant n$.

The introduced mathematical model for image registration is held only if all 3D points to be imaged lie on the reference ground plane π (assuming to have perfect features correspondences). However, in WAMI and particularly in urban scenarios, the presence of 3D structures/buildings is highly expected. The observed 3D points from such structures once imaged and mapped onto plane π, their corresponding 2D points would not coincide on π and will be dispersed. This phenomenon is known as *parallax* and its magnitude gets stringer as the 3D point get farther from the plane (π) which induces the homography. For example, in Fig. 1, consider $\mathbf{X}_2$ as a 3D point which is off-the-plane. It is imaged as $\mathbf{x}_2^1$, $\mathbf{x}_2^2$ and $\mathbf{x}_2^n$ on the image planes of cameras C_1, C_2 and C_n. Mapping them on π using homography transformations will result ${}^{\pi}\mathbf{x}_2^1$, ${}^{\pi}\mathbf{x}_2^2$ and ${}^{\pi}\mathbf{x}_2^n$. As illustrated in Fig. 1, these mapped points are all spread out on π, and the radius of distribution is proportional to the magnitude of parallax.

There is another type of noise which is likely to exist in the tracks of feature correspondences along the image sequence. The source of such noise can be from the

precision of the feature extraction algorithm or errors in the feature matching algorithm which could lead to many mismatches or outliers. In real scenarios, one can expect to have a considerable percentage of outliers. To deal with outliers mostly RANSAC (or its variations) is used in the literature. In this context, a RANSAC-based approach tries to (jointly) estimate a homography model and at the same time to eliminate the outliers, by looping through a hundreds of iterations. In each iteration of RANSAC, a subset of correspondence candidates is randomly chosen, a homography model is estimated for the chosen population and then the fitness of the whole population of the correspondences is measured using the estimated model. In this randomly exhaustive process, a model that provides the most consensus result would be chosen and at the same time the feature matches which do not obey the estimated model within a threshold will be identified as outliers. Notice that in our work, the homographies are analytically derived and no RANSAC estimation is used and instead the inaccurate sensor measurements from the platform are directly incorporated. Not using RANSAC gives the advantage of avoiding any adverse random behaviour in the model estimation. However, as a consequence of eliminating RANSAC, the existing outliers can not be explicitly identified. In order to address this issue we propose to use a robust error function in an appropriate formulation of the problem.

Robust functions also known as M-estimators are popular in robust statistics and reduce the influence of outliers in estimation problems. We have observed that not every choice of a robust function works well [42] and a proper robust function is critical for achieving a robust minimization of the reprojection error when the initial parameters are too noisy and outliers are not explicitly eliminated beforehand. Two commonly used robust statistics functions are the *Cauchy* (or Lorentzian) and *Huber* [41] measures:

- Cauchy or Lorentzian cost function

$$\rho(s) = b^2 \log(1 + s^2/b^2) \tag{16}$$

- Huber cost function

$$\rho(s) = \begin{cases} s^2 & \text{if } |s| < b \\ 2\,b\,|s| - b^2 & \text{otherwise} \end{cases} \tag{17}$$

where s is the residual (i.e. reprojection error) in (15) and b is usually one or a fixed user defined value. We have chosen Cauchy robust function since it down-weight the residuals more rigidly [43]. This characteristic of Cauchy is appropriate for our purpose specially because there expect to be enormous number of large residuals due to potential parallaxes in the scene. One can consider using other types of robust functions such as a generalization of the *Cauchy/Lorentzian*, *Geman-McClure*, *Welsch*, and *generalized Charbonnier* loss functions [44].

The proposed optimization method is presented in a pseudo code form in Algorithm 1. This method is an alternative approach for the parameter optimization with an emphasize on stabilizing the geoprojected aerial imagery by defining a cost function over a *single dominant 2D Euclidean world plane*. The points that do not lie on the dominant ground plane are *automatically marginalized* during the optimization process, thanks to the used robust functions, instead of using a RANSAC-based outlier elimination approach.

Algorithm 1 Analytical airborne video stabilization.

Input : A set of camera parameters acquired from inaccurate platform sensors, e.g. IMU and GPS: $(\mathbf{R}_i, \mathbf{t}_i, f)$, $i = 1 \dots n$, n being number of cameras/images.
m sets of tracked features along the sequence.

Output : Optimized homography matrices to robustly stabilize the imagery

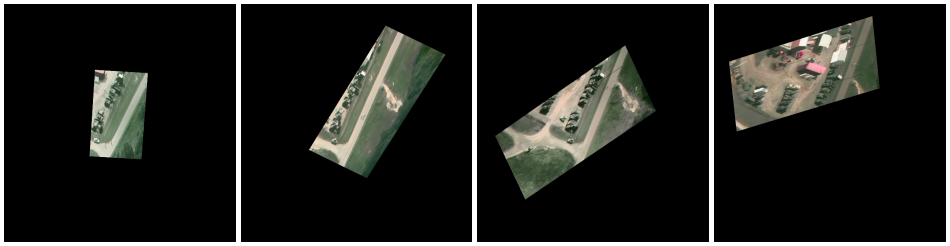
$\mathbf{v} \leftarrow \begin{bmatrix} u & v & f \end{bmatrix}^\mathsf{T}$
for $i = 1$ **to** n **do**
  $\mathbf{T}_i \leftarrow \begin{bmatrix} \mathbf{r}_{1,i} & \mathbf{r}_{2,i} & \mathbf{t}_i \end{bmatrix}$
  Assign $m_{bc,i}$ as the minor(b,c) of matrix $\mathbf{T}_i$
  $\mathbf{H}_{i\to\pi} \leftarrow \begin{bmatrix} m_{11,i} & -m_{21,i} & \begin{bmatrix} -m_{11,i} & m_{21,i} & m_{31,i} \\ m_{12,i} & -m_{22,i} & -m_{32,i} \end{bmatrix} \mathbf{v} \\ -m_{12,i} & m_{22,i} & \\ r_{13,i} & r_{23,i} & -\mathbf{r}_{3,i}^\mathsf{T}\mathbf{v} \end{bmatrix}$
end for
for $j = 1$ **to** m **do**
  for $i = 1$ **to** n **do**
    ${}^{\pi}\tilde{\mathbf{x}}_j^i \leftarrow \mathbf{H}_{i\to\pi}\, \tilde{\mathbf{x}}_j^i$
  end for
  ${}^{\pi}\hat{\mathbf{x}}_j \leftarrow \frac{1}{n}\sum_{i=1}^{n} {}^{\pi}\tilde{\mathbf{x}}_j^i$
end for
$E \leftarrow \sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathscr{F}(\mathbf{H}_{i\to\pi}\, \tilde{\mathbf{x}}_j^i) - \mathscr{F}({}^{\pi}\hat{\mathbf{x}}_j)\|^2$
Optimize $\mathbf{R}_i$, $\mathbf{t}_i$ and ${}^{\pi}\hat{\mathbf{x}}_j$ to minimize the cost function E
for $i = 1$ **to** n **do**
  $\hat{\mathbf{T}}_i \leftarrow \begin{bmatrix} \hat{\mathbf{r}}_{1,i} & \hat{\mathbf{r}}_{2,i} & \hat{\mathbf{t}}_i \end{bmatrix}$
  Assign $\hat{m}_{bc,i}$ as the minor(b,c) of matrix $\hat{\mathbf{T}}_i$
  $\hat{\mathbf{H}}_{i\to\pi} \leftarrow \begin{bmatrix} \hat{m}_{11,i} & -\hat{m}_{21,i} & \begin{bmatrix} -\hat{m}_{11,i} & \hat{m}_{21,i} & \hat{m}_{31,i} \\ \hat{m}_{12,i} & -\hat{m}_{22,i} & -\hat{m}_{32,i} \end{bmatrix} \mathbf{v} \\ -\hat{m}_{12,i} & \hat{m}_{22,i} & \\ \hat{r}_{13,i} & \hat{r}_{23,i} & -\hat{\mathbf{r}}_{3,i}^\mathsf{T}\mathbf{v} \end{bmatrix}$
end for
return optimized homography matrices $\hat{\mathbf{H}}_{i\to\pi}$, $i = 1 \dots n$

5 Experiments

The proposed method was applied to the DARPA Video and Image Retrieval and Analysis Tool (VIRAT) dataset and several WAMI datasets provided by Transparent-Sky[45]. The introduced sequential feature tracking method was used to track the identified SIFT features over each video sequence, followed by applying the proposed optimization method. The top images in Fig. 3 are some sample frames from a shot in the sequence "flight2Tape1_2" of VIRAT dataset which contain 2400 images. The metadata that come with the images are extremely inaccurate. To the best of our knowledge these metadata have not been of use in any SfM, stabilization and geoprojection project. However, our approach managed to seamlessly register the full video shot, smoothly with no jitter or jump. The results corresponding to the frames of the first row are shown in Fig. 3 bottom.

Fig. 4 shows the result of running our method on a WAMI aerial imagery. The metadata and images were provided by Transparent-Sky [45] via flying a fixed wing airplane over the downtown of Berkeley in California. Two exemplary images, with

(a) Original images (frame numbers: 3615, 4610, 5351 and 5901).



(b) Stabilized and geoprojected images (frame numbers: 3615, 4610, 5351 and 5901).

Fig. 3. Stabilized sequence of VIRAT dataset using the proposed method. We used a full shot from "flight2Tape1_2" which contain 2400 frames of 720×480 pixels. The camera metadata in all VIRAT datasets are extremely inaccurate. Our approach managed to perform the georegistration and stabilization on this long sequence without any jump or jitter in the result.

about 45° difference in their viewing angle along (200 frames apart along the sequence), are shown in Fig. 4-top. Their corresponding georegistered frames are plotted in Fig. 4-middle. The bounding boxes of the regions of the interest from the two frames are zoomed and shown in Fig. 4-Bottom. The rectified epipolar line (yellow dotted line) demonstrate the alignments for an exemplary pair of corresponding points (in red) in the two frames after stabilization. A similar evaluation is demonstrated for ABQ (Albuquerque downtown area) WAMI dataset, in Fig. 5. Fig. 6 depicts the original and stabilized images from another WAMI dataset, LA downtown area. As one can see, everything from the dominant ground plane is well aligned between the two registered views and just the building and off the ground objects were wobbled which is due to the existence of parallax. Despites the presence of strong parallax, the method succeeded to seamlessly stabilized the images without any jitter. It is worth to reminding that no RANSAC or any other random-based method has been used in the proposed.

6 Conclusions

We proposed a stabilization and geoprojection method which is able to use available sensor metadata (i.e. GPS and IMU) to register airborne video in a robust and seamless manner. This became possible by deriving a set of analytical homography transformations and defining a *metric* cost function over a dominant 2D *Euclidean* ground plane in the scene. The solution has been formulated such that no RANSAC (any random-based iterative techniques) is used, in contrary to most existing approaches. The robustness in our work is achieved by defining an appropriate (robust) cost function which allows to implicitly marginalize the outliers automatically within the optimization process. Our

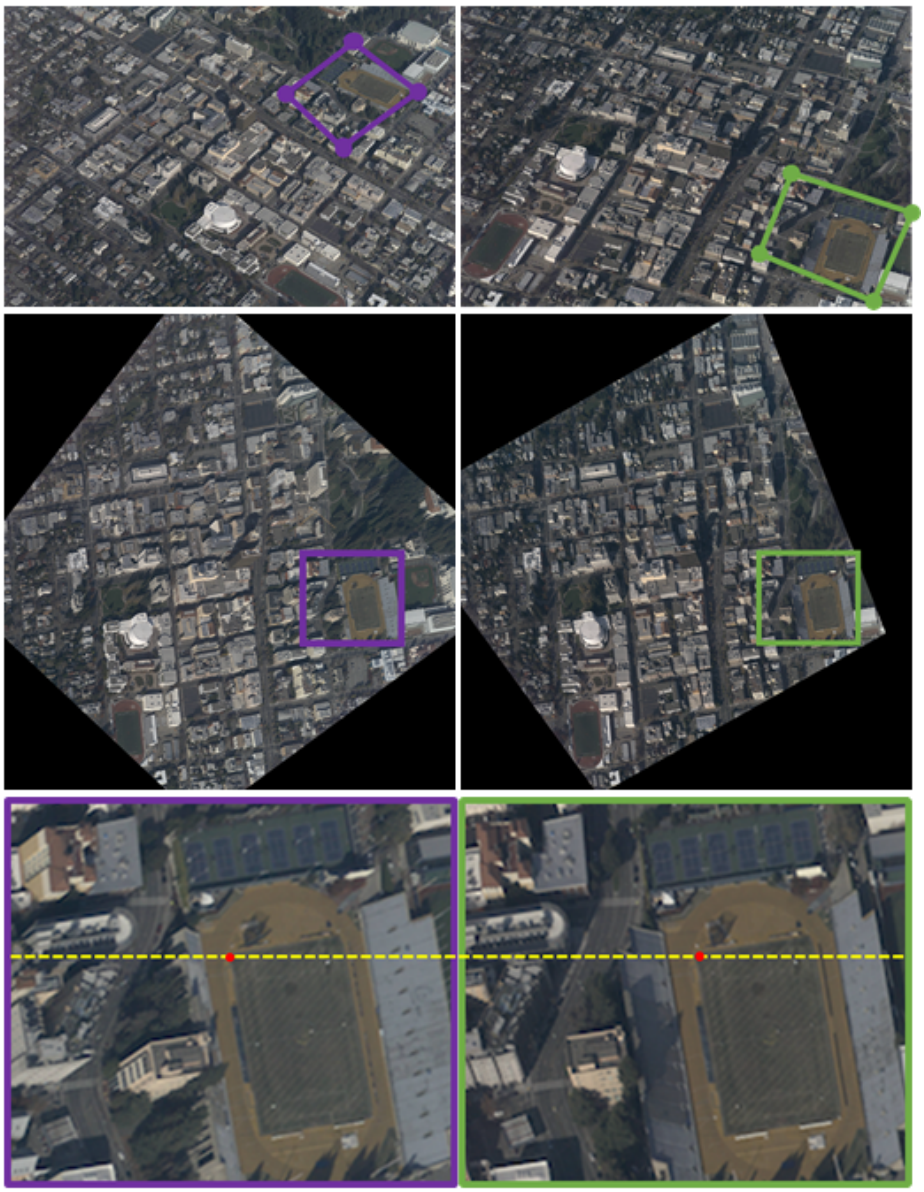


Fig. 4. Stabilization result of Berkeley dataset. **Top**: two raw WAMI images, with size of 6600 × 4400 pixels (frame #0 at left, frame #200 at right). **Middle**: geoprojection of the raw frames after stabilization using the proposed approach. **Bottom**: Zoomed-in versions of the middle row corresponding to the areas which are marked by *purple* and *green* bounding boxes. The rectified epipolar line (yellow dotted line) depicts the alignment for a pair of corresponding points (in red) after stabilization.

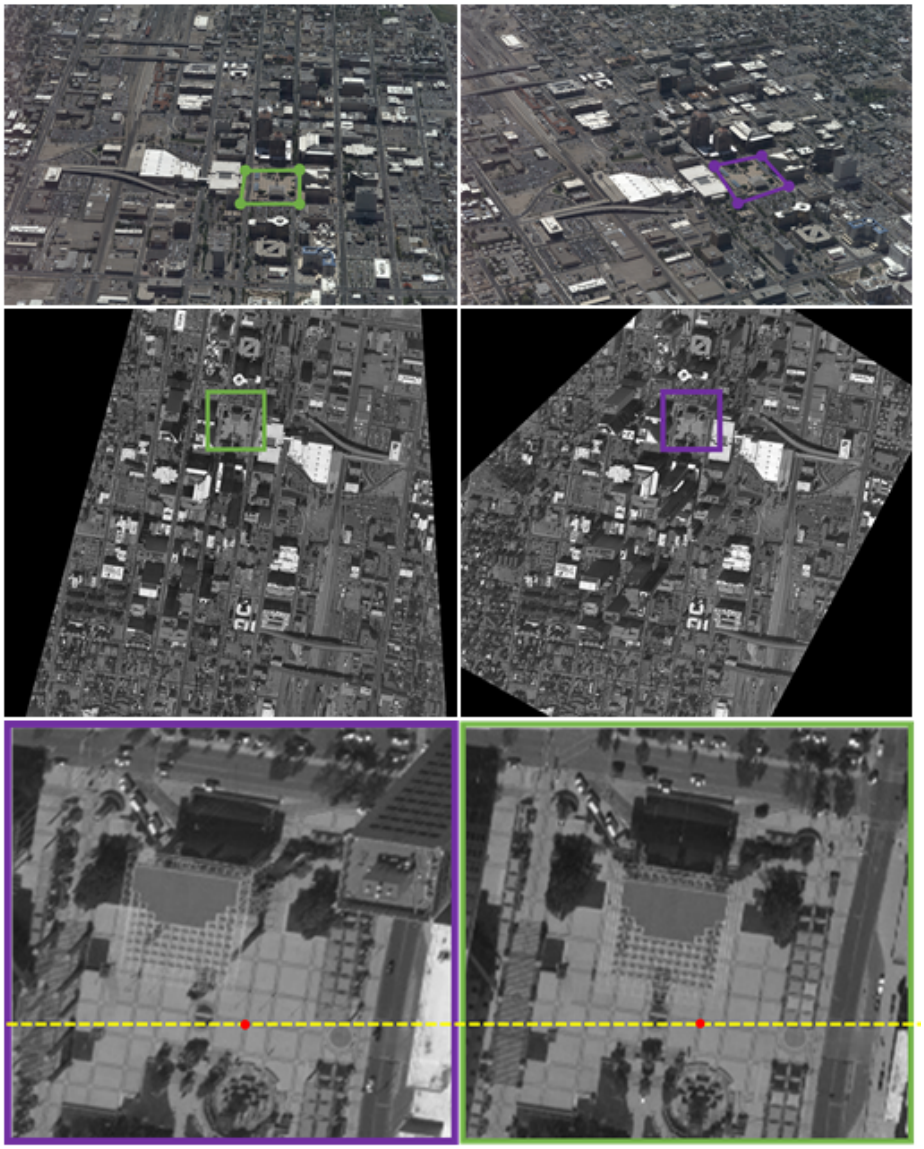
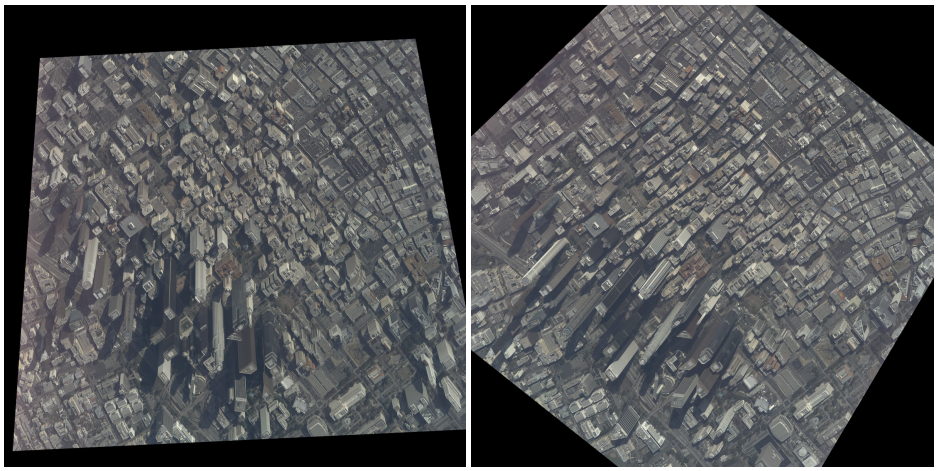


Fig. 5. Stabilization result of Albuquerque dataset. **Top**: two raw WAMI images, with size of 6600×4400 pixels (frame #0 at left, frame #100 at right). **Middle**: geoprojection of the raw frames after stabilization using the proposed approach. **Bottom**: Zoomed-in versions of the middle row corresponding to the areas which are marked by *purple* and *green* bounding boxes. The rectified epipolar line (yellow dotted line) depicts the alignment for a pair of corresponding points (in red) after stabilization.

(a) Original images (frames #0 and #100).



(b) Stabilized and geoprojected images (frames #0 and #100).

Fig. 6. Stabilized sequence of Los Angeles (California) dataset using the proposed method. The high resolution WAMI imagery (with the image size of 6600×4400) along with initial metadata were provided by Transparent-Sky [45]. Despites presence of strong parallax induced by the tall buildings, our method managed to smoothly stabilize the WAMI images.

approach has been tested over a very challenging dataset of DARPA, known as VIRAT. Unlike the imagery component of this dataset is very rich and has been frequently used in different algorithms by several well known research groups, its metadata component is extremely challenging. We know no group or research work which could have relied on the metadata in this dataset and used it in a SfM or stabilization method, as the available sensor measurements are highly inaccurate. Nevertheless, our approach has been tested on this dataset where the challenging metadata was directly utilized to perform a smooth and seamless stabilization on the video sequence. In addition to VIRAT dataset, two high resolution WAMI datasets corresponding to the downtown areas of Berkeley and Los Angeles were successfully tested and stabilized in our experiments.

Acknowledgments

This research was partially sponsored by the Army Research Laboratory and Air Force Research Laboratory under Cooperative Agreements W911NF-18-2-0285 and FA8750-19-2-0001 respectively. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. K. Palaniappan, R. Rao, and G. Seetharaman, "Wide-area persistent airborne video: Architecture and challenges," in *Distributed Video Sensor Networks: Research Challenges and Future Directions*. Springer, 2011, pp. 349–371.
2. R. Porter, A. M. Fraser, and D. Hush, "Wide-area motion imagery," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 56–65, 2010.
3. M. Poostchi, H. Aliakbarpour, R. Viguier, F. Bunyak, K. Palaniappan, and G. Seetharaman, "Semantic Depth Map Fusion for Moving Vehicle Detection in Aerial Video," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1575–1583.
4. K. Palaniappan, M. Poostchi, H. Aliakbarpour, R. Viguier, J. Fraser, F. Bunyak, A. Basharat, S. Suddarth, E. Blash, R. Rao, and G. Seetharaman, "Moving object detection for vehicle tracking in wide area motion imagery using 4D filtering," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2016.
5. D. R. Nilosek, D. J. Walvoord, and C. Salvaggio, "Assessing geoaccuracy of structure from motion point clouds from long-range image collections," *Optical Engineering*, vol. 53, no. 11, pp. 1 – 10, 2014.
6. M. E. Linger and A. Goshtasby, "Aerial image registration for tracking," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2137–2145, apr 2015.
7. D. J. Holtkamp and A. A. Goshtasby, "Precision Registration and Mosaicking of Multicamera Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, pp. 3446–3455 ST – Precision Registration and Mosaick, 2009.
8. J. Lee, X. Cai, C.-B. Schonlieb, and D. A. Coomes, "Nonparametric image registration of airborne LiDAR, hyperspectral and photographic imagery of wooded landscapes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–12, 2015.

9. D. Crispell, J. L. Mundy, and G. Taubin, "Parallax-free registration of aerial video," *British Machine Vision Conference*, pp. 1–4, 2008.
10. H. Aliakbarpour, K. Palaniappan, and G. Seetharaman, "Fast Structure from Motion for Sequential and Wide Area Motion Imagery," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015.
11. ——, "Robust Camera Pose Refinement and Rapid SfM for Multiview Aerial Imagery - Without RANSAC," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2203–2207, 2015.
12. ——, "Parallax-Tolerant Aerial Image Georegistration and Efficient Camera Pose Refinement- Without Piecewise Homographies," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4618–4637, 2017.
13. H. Aliakbarpour, V. B. S. Prasath, K. Palaniappan, G. Seetharaman, and J. Dias, "Heterogeneous Multi-View Information Fusion: Review of 3-D Reconstruction Methods and a New Registration with Uncertainty Modeling," *IEEE Access*, vol. 4, pp. 8264–8285, 2016.
14. H. Aliakbarpour and J. Dias, "Three-dimensional reconstruction based on multiple virtual planes by using fusion-based camera network," *IET Computer Vision*, vol. 6, no. 4, p. 355, 2012.
15. H. Aliakbarpour, L. Almeida, P. Menezes, and J. Dias, "Multi-sensor 3D volumetric reconstruction using CUDA," *3D Research, Springer*, vol. 2, no. 4, pp. 1–14, 2011.
16. H. Aliakbarpour and J. Dias, "PhD forum: Volumetric 3D reconstruction without planar ground assumption," in *Fifth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*. IEEE, 2011, pp. 1–2.
17. E. Molina and Z. Zhu, "Persistent aerial video registration and fast multi-view mosaicing," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2184–2192, 2014.
18. I. Saleemi and M. Shah, "Multiframe Many-Many Point Correspondence for Vehicle Tracking in High Density Wide Area Aerial Videos," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 198–219, 2013.
19. G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 873–889, 2001.
20. Y. Lin, Q. Yu, and G. Medioni, "Efficient detection and tracking of moving objects in geo-coordinates," *Machine Vision and Applications*, pp. 505–520, 2010.
21. V. M. Chellappa, Govindu, and R., "Feature-based image to image registration," in *Image Registration for Remote Sensing*, 2011, pp. 215–239.
22. C. N. Taylor, "Improved evaluation of geo-registration algorithms for airborne EO/IR imagery," in *SPIE, Geospatial Infofusion III*, vol. 8747, 2013, p. 874709.
23. E. Vasquez, Juan and Hytla, Patrick and Asari, Vijayan and Jackovitz, Kevin and Balster, "Registration of Region of Interest for Object Tracking Applications in Wide Area Motion Imagery," in *the IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2012, pp. 1–8.
24. H. S. Stone, M. T. Orchard, E. C. Chang, and S. a. Martucci, "A fast direct Fourier-based algorithm for subpixel registration of images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 10, pp. 2235–2243, 2001.
25. A. Hafiane, K. Palaniappan, and G. Seetharaman, "UAV-Video Registration Using Block-Based Features," in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, no. 1, 2008, pp. II–1104–II–1107.
26. G. S. Palaniappan, G. Gasperas, and K., "A piecewise affine model for image registration in nonrigid motion analysis," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, 2000, pp. 561–564.

27. Z. Zhu, A. R. Hanson, and E. M. Riseman, "Generalized parallel-perspective stereo mosaics from airborne video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 226–237, 2004.
28. B. P. Jackson and A. Goshtasby, "Adaptive registration of very large images," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 351–356, 2014.
29. D. Turner, A. Lucieer, and L. Wallace, "Direct Georeferencing of Ultrahigh-Resolution UAV Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2738–2745, may 2014.
30. "Agisoft, Agisoft Photoscan Professional. http://www.agisoft.com."
31. "Pix4D, http://pix4d.com."
32. "N. Snavely, Bundler: Structure from Motion (SfM) for Unordered Image Collections. http://phototour.cs. washington.edu/bundler."
33. N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from Internet photo collections," *International Journal of Computer Vision*, vol. 80, pp. 189–210, 2008.
34. M. D. Pritt, "Fast Orthorectified Mosaics of Thousands of Aerial Photographs from Small UAVs," *Applied Imagery Pattern Recognition Workshop (AIPR), IEEE*, 2014.
35. H. Aliakbarpour, P. Nuez, J. Prado, K. Khoshhal, and J. Dias, "An efficient algorithm for extrinsic calibration between a 3D laser range finder and a stereo camera for surveillance," *2009 International Conference on Advanced Robotics*, 2009.
36. K. a. Redmill, J. I. Martin, and U. Ozguner, "Aerial image registration incorporating GPS/IMU data," *Proceedings of SPIE*, vol. 7347, no. 1, pp. 73 470H–73 470H–15, 2009.
37. R. Aktar, H. Aliakbarpour, F. Bunyak, T. Kazic, G. Seetharaman, and K. Palaniappan, "Geospatial content summarization of UAV aerial imagery using mosaicking," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 10645, 2018.
38. R. Viguier, C. C. Lin, H. AliAkbarpour, F. Bunyak, S. Pankanti, G. Seetharaman, and K. Palaniappan, "Automatic Video Content Summarization Using Geospatial Mosaics of Aerial Imagery," *2015 IEEE International Symposium on Multimedia (ISM)*, pp. 249–253, 2015.
39. R. Viguier, C. C. Lin, K. Swaminathan, A. Vega, A. Buyuktosunoglu, S. Pankanti, P. Bose, H. Akbarpour, F. Bunyak, K. Palaniappan, and G. Seetharaman, "Resilient mobile cognition: Algorithms, innovations, and architectures," *Proceedings of the 33rd IEEE International Conference on Computer Design, ICCD 2015*, pp. 728–731, 2015.
40. F. Fraundorfer and D. Scaramuzza, "Visual Odometry : Part II: Matching, Robustness, Optimization, and Applications," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 78–90, jun 2012.
41. B. Triggs, P. F. Mclauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle Adjustment A Modern Synthesis," *Vision algorithms: theory and practice. S*, vol. 34099, pp. 298–372, 2000.
42. H. Aliakbarpour, K. Palaniappan, and G. Seetharaman, "Robust camera pose refinement and rapid SfM for multi-view aerial imagery without RANSAC," *IEEE Journal of Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2203–2207, 2015.
43. A. A. Johan A.K. Suykens, Marco Signoretto, *Regularization, Optimization, Kernels, and Support Vector Machines*. Chapman and Hall/CRC, 2014.
44. J. T. Barron, "A More General Robust Loss Function," *Arxiv*, vol. 1, no. 5, pp. 2–5, 2017.
45. "http://www.transparentsky.net."