Analytical Image Stitching in Non-Rigid Multi-Camera WAMI

Hadi AliAkbarpour[1,2], Tristan Brodeur[2], and Steve Suddarth[2]

[1] *Department of Electrical Engineering and Computer Science, University of Missouri, USA*
[2] *Transparent Sky, NM, USA*

Abstract—Combining the images of multiple airborne cameras is a common way to achieve a higher scene coverage in Wide Area Motion Imagery (WAMI). This paper proposes a novel approach, named Dynamic Homography (DH), to stitch multiple images of an airborne array of 'non-rigid' cameras with 'narrow' overlaps between their FOVs. It estimates the inter-views transformations using their underlying geometry namely the relative angles between their local coordinate systems in 3D. A fast and robust optimization model is used to minimize the errors between the feature correspondences which takes place when a new set of images are acquired from the camera array. The minimum number of required feature correspondences between each adjacent pair of views is only two, which relaxes the need to have a large overlap in their FOVs. Our quantitative and qualitative experiments show the superiority of the proposed method compared to the traditional ones.

Index Terms—WAMI, image registration, homography, image stitching.

I. INTRODUCTION

Wide Area Motion Imaging [1] (WAMI) sensing presents a unique challenge of projecting (georeferencing) many well-resolved, stable pixels into a world map. The best WAMI systems cover areas of tens of square miles while providing the ability to digitally zoom into small areas maintaining high resolution. Large WAMI systems require hundreds of megapixels and even gigapixels of resolution. Achieving this resolution requires the use of multiple cameras mounted in arrays to make a large *virtual camera* covering the area of interest. Stitching the imagery between such cameras requires precise pixel-level alignment, and such alignment generally requires rigid mounts and well-controlled camera geometries. This can require expensive and heavy measures, such as precisely-milled camera mounts that are created out of a single billet of metal, as well as lenses that are specially engineered to ensure constant focal parameters. Most importantly, the rigid mounting requirement greatly increases the cross-sectional area required for large camera arrays, with potentially ruinous effects on the drag properties of an aircraft. By relaxing the requirement for consistently precise camera alignment, new designs for large camera gimbal systems can be contemplated. This includes streamlined designs in which cameras can be placed in long, narrow pods or built into longitudinal aircraft structures. The cameras can have imprecise articulation in which alignment is constantly changing. Furthermore, less expensive, lighter commercial lenses can be used that are not designed for high thermal stability. Camera arrays can be built to arbitrary scales, and they can be placed into structures that are cheaper, lighter, and more well-suited for aerospace uses, such as carbon fiber and many plastics. In order to relax the need for precision camera alignment, however, the stitching parameters must be computed on a frame-by-frame basis in real time using computing hardware that is readily available on the aircraft. The end result of stitching is equivalent to a large virtual image formed by a computed homography, a 3x3 matrix that maps pixel locations between from each camera image into the larger image. In the case of "dynamic" homography, this mapping must be computed for each frame of imagery by aligning the cameras together. In order to precisely estimate such homography transformations, there needs to be a relatively large amount of overlap between the Field-Of-View (FOV) of the cameras. Although this is the most common approach, it is against the idea of maximum utilization of the pixels. Alternatively, one can use one (or more) redundant camera(s) with a wider FOV that is dominant to other cameras' FOV. This approach is not often desired in Unmanned Aerial Vehicles (UAVs) due to its weight, space, and power redundancies.

This paper proposes a novel approach to stitching multiple images (and form a large virtual image) of an airborne array of cameras with narrow overlaps between their FOVs. The proposed approach is called Dynamic Homography (DH) which aims at analytically optimizing the inter-views transformation estimation using their underlying geometry, namely the relative angles between their local coordinate systems in 3D. A fast and robust optimization model is used to minimize the errors between the feature correspondences which takes place when a new set of images are acquired from the camera array. The minimum number of required feature correspondences between each adjacent pair of views is only two, which relaxes the need to have a large overlap in their FOVs.

Related Works: He et. al. [2] proposed a panoramic video stitching method for near ground-mounted cameras (with PTZ) for surveillance. A dynamic estimation of homography for visual-servoing for hand-held/robot is introduced in [3]. A method for addressing parallax artifacts in video stitching of linear camera arrays was discussed in [4]. Zhou et. al in [5] proposed an approach for stitching large area UAV images using Structure-from-Motion technique. Generating a panorama image using a set of multi-camera system is proposed in [6]. An image stitching technique in a multi-camera setup with large FOV is introduced in [7].

II. DYNAMIC HOMOGRAPHY ESTIMATION

A projective Homography is a 3×3 transformation matrix which provides a direct mapping between two (camera) images of a common Euclidean scene plane. The homography transformation mapping the images of a 3D point $\mathbf{X}_j$ lying on an Euclidean plane, π, from the image I_2 (of camera C_2) to the image I_1 (of camera C_1), is expressed by $\tilde{\mathbf{x}}_{1,j} = \mathbf{H}^{\pi}_{2\to1}\tilde{\mathbf{x}}_{2,j}$, where $\tilde{\mathbf{x}}_{1,j}$ and $\tilde{\mathbf{x}}_{2,j}$ are the 2D homogeneous coordinates of the images of $\mathbf{X}_j$ on I_1 and I_2, respectively. The homography $\mathbf{H}_{2\to1}$, induced by plane π, is analytically defined by

$$\mathbf{H}^{\pi} = \mathbf{K}_1(\mathbf{R} + \frac{1}{d}\mathbf{n}.\mathbf{t}^{\intercal})\mathbf{K}_2 \tag{1}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are respectively the intrinsic parameters of the two cameras, defined as 3×3 upper triangular matrices with the focal

length of f (in pixels) and the principal point of (u, v) [8]. In (1), $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector from the coordinate system of C_2 to the one in C_1, $\mathbf{n}$ is the normal of plane π, and d is the euclidean distance between π and the coordinate system of C_2. Notice that this (perspective) homography is valid only for the specific plane π (called induced by the plane). In practice, such a homography matrix is not calculated from its analytical form of Eq. (1), as the involved parameters (i.e. the relative transformation between the two cameras and the plane 3D geometry) are not generally available. Instead, its eight elements (normalized to its 9th element due to being up to a scale, yielding to eight Degree Of Freedom-DOF):

$$\mathbf{H} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & 1 \end{bmatrix} \tag{2}$$

is estimated through the Direct Linear Transform [8] method given a set of point correspondences (minimum of four) between the views (e.g. from an image feature matching process). In some setups a reduced form of (2), known as *affine homography*, is used which ignores the perspective components and has 6-DOF:

$$\mathbf{H}_a = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

In a special case, where the underlying scene is relatively far from the cameras (i.e. $d \longrightarrow \infty$) and/or the translation between the two camera coordinate systems is zero (or negligible), (1) becomes simplified to

$$\mathbf{H}^{\infty}_{2\rightarrow 1} = \mathbf{H}_{2\rightarrow 1} = \mathbf{K}_1\mathbf{R}_{2\rightarrow 1}\mathbf{K}_2^{-1} = \mathbf{K}_1\mathbf{R}_1\mathbf{R}_2^{\intercal}\mathbf{K}_2^{-1} \tag{4}$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ are the rotation matrices of C_1 and C_2, respectively. (4) is known as *infinite homography*. As one can see, this form of homography only has 3-DOF in terms of the extrinsic parameters. For WAMI multi-camera rig, the mentioned criteria (i.e. $d \rightarrow \infty$ and/or $\mathbf{t} \rightarrow \mathbf{0}_{3\times 1}$) are held, allowing to use the infinite homography with a reduced number of parameters instead of its general form in (2). As emphasized in [9], estimating the 3D rotation matrix $\mathbf{R}_{2\rightarrow 1}$ in (4) between two camera coordinates, which only includes 3-DOF, is significantly more stable and robust than estimating a full 8-DOF homography (2), or even its 6-DOF form (3). The reason to this is not only due to having a reduced number of parameters, but also to the ability to enforce the right constraints (only a 3-DOF rotation between two cameras) that matches the geometry of the cameras their physical setup (model based). This prevents estimating other unnecessary constraints (additional 5-DOF in the case of perspective and extra 3-DOF in the case of affine) that do not apply to the actual geometric model. Another advantage of our approach is that the samples (feature correspondences) between the two views do not need to be well distributed over the two image frames, which is a strong requirement in a general homography estimation for achieving accurate results. As a matter of fact, our approach works with only a tiny overlap between the FOVs with as low as only two feature (point) correspondences over the narrow intersected stripe. This helps in maximizing the pixel utilization in an UAV WAMI setup, which is always an important factor.

Knowing the three rotation angles between two cameras:

$$\mathbf{r}_{3\times 1} = \begin{bmatrix} roll & pitch & yaw \end{bmatrix}^{\intercal} \tag{5}$$

one can use (4) to analytically construct the corresponding homography matrix $\mathbf{H} = \mathbf{K}_1\mathbf{R}\mathbf{K}_2^{-1}$, where

$$\mathbf{R}_{3\times 3} = \mathcal{R}(\mathbf{r}_{3\times 1}) \tag{6}$$

and $\mathcal{R}$ is a function which converts a angle-axis rotation vector to a rotation matrix (we use the *Rodrigues* formula for this). A very rough estimate (with a few degrees error tolerance) of the angles in (5), $\hat{\mathbf{r}}$, is assumed to be available (e.g. via the CAD design or using [10]). A non-linear least squares (LS) optimization can be used to refine and optimize the rotation angles. Assume to have m feature correspondences (homogeneous) between the two images, $\{(\tilde{\mathbf{x}}_{1,j}, \tilde{\mathbf{x}}_{2,j}) | \, j = 1 \dots m\}$, the following cost function is defined:

$$e_j = \|\mathbf{x}_{2,j} - \Pi(\mathbf{H}\tilde{\mathbf{x}}_{1,j})\| \tag{7}$$

where, e_j represents the euclidean distance (in pixels) between the jth feature point $(\mathbf{x}_{1,j})$ on the image plane I_1 (of C_1) and its pair $(\mathbf{x}_{2,j})$ on I_2 (of C_2) after mapping to I_1 using the analytical homography

$$\mathbf{H} = \mathbf{K}_1\mathcal{R}(\hat{\mathbf{r}})\mathbf{K}_2^{-1}. \tag{8}$$

$\Pi(\tilde{\mathbf{x}})$ in (7) is a function that normalize a homogeneous vector to its 2D euclidean form. Therefore one can define the LS problem as

$$\mathbf{r} = \arg\min_{\hat{\mathbf{r}}} \sum_{j=1}^{m} \|\mathbf{x}_{1,j} - \Pi(\mathbf{K}_1\mathcal{R}(\hat{\mathbf{r}})\mathbf{K}_2^{-1}\tilde{\mathbf{x}}_{2,j})\|^2 \tag{9}$$

An iterative non-linear solver such as Levenberg Marquardt can be used to find an optimum rotation vector $\mathbf{r}$ in (10), while minimizing the errors between all pairs of feature correspondences.

The DLT homography estimation algorithm is only robust to the type of the noise caused by the measurements of the feature point positions. However, in real scenarios the feature correspondences are often contaminated by false feature matches (known as mismatches). Such spurious feature correspondences (which do not represent an identical 3D point in the scene) are called *outliers*. The presence of outliers often lead to an invalid or erroneous DLT estimation of homography. The most commonly used robust estimator in the traditional homography estimation methods is RANSAC (Random Sample Consensus) [8] which is based on 'randomly' picking the minimal set of the point correspondences (4 pairs, in case of general homography) required for estimating a DLT model, counting the consensus among the remaining samples and repeating this random selection for a predefined number of times. Our proposed method, unlike the traditional ones, do not use a RANSAC-based DLT model estimation. In other words, we do not just algebrically estimate a homography from the point correspondences (samples), but instead we *analytically* construct it from the underlying geometry of the involved cameras (their relative rotations). Thus we wisely avoid RANSAC (or any other similar combinatorial random process) due its drawbacks [11]. However, to battle with the outliers, we use the Cauchy (or Lorentzian) robust function [11], yielding the following non-linear least squares formula:

$$\{\mathbf{r}, f_1, f_2\} = \arg\min_{\hat{\mathbf{r}}, f_1, f_2} \sum_{j=1}^{m} b^2 \log(1 + (\|\mathbf{x}_{1,j} - \Pi(\mathbf{K}_1\mathcal{R}(\hat{\mathbf{r}})\mathbf{K}_2^{-1}\tilde{\mathbf{x}}_{2,j})\|)^2/b^2) \tag{10}$$

where b is a user defined parameter. If the focal lengths are known (e.g. by calibrating each single camera independently using [12]), then (10) will have three variables (the components of the rotation vector $\mathbf{r}$) to solve for. When no focal lengths are available from a calibration process or when they are changed on the fly (either intentionally or by external factors such as temperature/vibration), then a 'rough' estimate of the focal length for each camera can be used for initialization in (10).

The introduced two-camera homography estimation can be extended to more than a pair. With no loss of generality, we extend it

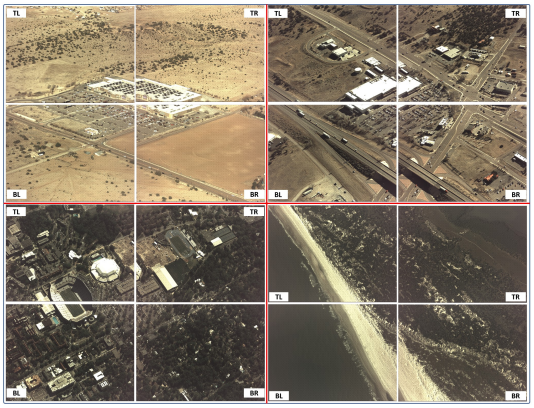

Fig. 1: Raw sample frames used in our experiments. Left to right and top to bottom: Edgewood Walmart (New Mexico), Edgewood Smith's Pharmacy (New Mexico), Gaineville (Florida) and Matanzas Inlet (Florida). Each image in each dataset is labelled by *TL* (Top-Left), *TR* (Top-Right), *BL* (Bottom-Left) and *BR* (Bottom-Right).

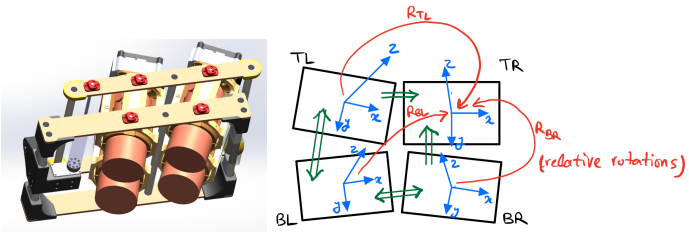

Fig. 2: Left: The four-camera WAMI system with a pinion gimbal. Right: Geometric representation of the cameras' coordinate systems (blue vectors) in a 4-camera setup. One of the camera's coordinate system (*TR*) is assumed as the reference. The other cameras have their coordinate systems defined relative to the reference TR, each by a 3×3 rotation matrix (red arrows).

to a 4-camera setup (Fig. 2-Left), however, the method is scale-able and can be extended to a larger camera array. Fig. 2-Right depicts the geometric relation of the images of this 4-camera system. Each camera/image is labeled as *TL* (top-left), *TR* (top-right), *BR* (bottom-right) and *BL* (bottom-left). They are mounted on a pinion gimbal such a way that in any gimbal lock position, there is a (narrow) strip of overlap on the FOV between these pairs: (TL, TR), (TL, BL), (BL, BR) and (BR, TR). The transformations between the camera references are defined by their relative rotations. Notice that these relative rotations are not constant and are subject to change at every moment due to the mechanical constraint. With no loss of generality, we define TR as the local reference in our solution. We aim to find the following three homography matrices that map the images of TL, BL and BR to the image of TR (see the red arrows in Fig. 2-Right):

$$\{\mathbf{H}_{TL\to TR}, \mathbf{H}_{BL\to TR}, \mathbf{H}_{BR\to TR}\} \tag{11}$$

and compose a single larger image out of the four camera images. As there might be no direct FOV overlap between the BL and the reference (TR), a direct cost function between them can not be defined to optimize their relative angles and consecutively their homography $\mathbf{H}_{BL\to TR}$. We address this by simply defining two transitivity relations to connect the BL image to the TR one (see the green arrows in Fig. 2-Right):

$$\begin{cases} \mathbf{H}_{BL\to TR} \equiv \mathbf{H}_{TL\to TR}\,\mathbf{H}_{BL\to TL} \\ \mathbf{H}_{BL\to TR} \equiv \mathbf{H}_{BR\to TR}\,\mathbf{H}_{BL\to BR} \end{cases} \tag{12}$$

This results to the expansion of the set in (11) to:

$$\{\mathbf{H}_{TL\to TR}, \mathbf{H}_{BL\to TL}, \mathbf{H}_{BL\to BR}, \mathbf{H}_{BR\to TR}\} \tag{13}$$

We define a cost function for each member of the set in (13):

$$e_{TL,TR} = \sum_{i=1}^{m_{TL,TR}} \|\mathbf{x}_{TR,i} - \Pi(\mathbf{K}_{TR}\mathcal{R}(\hat{\mathbf{r}}_{TL,TR})\mathbf{K}_{TL}^{-1}\tilde{\mathbf{x}}_{TL,i})\|^2, \tag{14}$$

$$e_{BL,TL} = \sum_{j=1}^{m_{BL,TL}} \|\mathbf{x}_{TL,j} - \Pi(\mathbf{K}_{TL}\mathcal{R}(\hat{\mathbf{r}}_{BL,TL})\mathbf{K}_{BL}^{-1}\tilde{\mathbf{x}}_{BL,j})\|^2, \tag{15}$$

$$e_{BL,BR} = \sum_{k=1}^{m_{BL,BR}} \|\mathbf{x}_{BR,k} - \Pi(\mathbf{K}_{BR}\mathcal{R}(\hat{\mathbf{r}}_{BL,BR})\mathbf{K}_{BL}^{-1}\tilde{\mathbf{x}}_{BL,k})\|^2, \tag{16}$$

$$e_{BR,TR} = \sum_{l=1}^{m_{BR,TR}} \|\mathbf{x}_{TR,l} - \Pi(\mathbf{K}_{TR}\mathcal{R}(\hat{\mathbf{r}}_{BR,TR})\mathbf{K}_{BR}^{-1}\tilde{\mathbf{x}}_{BR,l})\|^2 \tag{17}$$

The final cost function is made by the aggregation of the four cost functions of (14)- (17), yielding the following optimization formulation:

$$\min_{\hat{\mathbf{r}}_{TL,TR},\hat{\mathbf{r}}_{BL,TL},\hat{\mathbf{r}}_{BL,BR},\hat{\mathbf{r}}_{BR,TR},f_{TL},f_{BL},f_{BR}} \left(e_{TL,TR} + e_{BL,TL} + e_{BL,BR} + e_{BR,TR}\right) \tag{18}$$

Once the four optimized rotation vectors and focal lengths are calculated, the three final homography matrices in (11) can be analytically constructed as:

$$\begin{cases} \mathbf{H}_{TL\to TR} = \mathbf{K}_{TR}\mathcal{R}(\hat{\mathbf{r}}_{TL,TR})\mathbf{K}_{TL}^{-1} \\ \mathbf{H}_{BL\to TR} = \mathbf{K}_{TR}\mathcal{R}(\hat{\mathbf{r}}_{TL,TR})\mathcal{R}(\hat{\mathbf{r}}_{BL,TL})\mathbf{K}_{BL}^{-1} \\ \mathbf{H}_{BR\to TR} = \mathbf{K}_{TR}\mathcal{R}(\hat{\mathbf{r}}_{BR,TR})\mathbf{K}_{BR}^{-1} \end{cases} \tag{19}$$

The number of parameters to optimize in (18) is 12 (four rotation vectors) for the whole optimization problem (plus the focal lengths if desired).

III. EXPERIMENTS

This section presents the experiments carried out on the proposed methods and some comparisons to other methods. The approach has been implemented in C++. The used WAMI datasets consists of four collections: Edgewood Walmart (NM), Edgewood Smith's pharmacy (NM), Gainesville (FL) and Matanzas Inlet (FL)– see some exemplary raw frames in Fig. 1. They were imaged by the four-camera hardware mounted on a pinion gimbal setup shown. As typical in WAMI, the gimbal is locked to the center of area of interest in each data collection, while the airplane flies an orbital pattern around the area. At each gimbal position, each of the four cameras is automatically (mechanically) adjusted to observe the area. The relative rotations between the cameras are unknown and vary at each time instance. The cameras are synced via a hardware trigger. Each single camera provides a color image with the size of 6480×4871 pixels. Once a new set of (four) images are captured by the system, their feature points (we use SIFT [13]) are extracted. The (overlapped are of) two images in each pair of TL-TR, BL-TL, BL-BR and BR-TR, are matched to obtain correspondences among their tiny overlapped FOV. The obtained correspondences are directly fed to the proposed analytical homography estimation model. We quantitatively compared our method against both general (perspective) and affine homography estimation methods. In order to accomplish the quantitative comparisons, we generated a set of ground-truth point correspondences in each pair of images. The geometric errors in the homographies estimated by each of the three methods are computed. Table 1 contains the mean and standard deviation of the errors in each case, shown as (μ, σ). In this table, the 'many-point' label indicates that all available point correspondences (after the feature extraction and matching) are used (including possible

	Perspective			Affine			DH			DH Estimated Relative Angles (deg) to Reference		
	2pt	3pt	many-pt	2pt	3pt	many-pt	2pt	3pt	many-pt	2pt	3pt	many-pt
Edgewood, NM-Walmart	NP	NP	(53.75, 131.01)	NP	(28939.22, 97630.00)	(26.20, 31.08)	(10.13, 13.09)	(4.19, 3.11)	(2.23, 2.63)	TL=(-0.35, -7.99, 0.77) BL=(-6.91,-7.98, 0.66) BR=(-7.11, -0.08, 0.11)	TL=(-0.31, -8.07, -0.19) BL=(-6.89, -7.98, -0.05) BR=(-7.14, -0.06, 0.18)	TL=(-0.32, -8.05, 0.67) BL=(-6.92, -8.02, 0.27) BR= (-7.14, -0.07, 0.13)
Gainesville, FL	NP	NP	(2622.23, 5993.91)	NP	(5312.94, 9259.83)	(66.09, 118.63)	(8.70, 8.75)	(6.42, 5.56)	(1.78, 1.87)	TL=(-0.05,-8.93, 1.47) BL=(-8.16, -9.20, 0.21) BR=(-7.34, 0.14, 0.76)	TL=(0.00, -8.96, 1.39) BL=(-8.15, -9.23, 0.80) BR=(-7.30, 0.08, 0.20)	TL=(-0.01,-9.01, 1.37) BL=(-8.19, -9.28, 0.46) BR=(-7.32, 0.08, -0.09)
Edgewood, NM-Smith's Pharmacy	NP	NP	(97.19, 404.302)	NP	(148.49, 333.59)	(15.54, 34.72)	(12.80, 19.98)	(10.60, 17.31)	(2.42, 1.68)	TL=(-0.32, -8.00, 1.59) BL=(-6.88, -8.03, 1.40) BR= (-7.16, -0.08, 0.29)	TL=(-0.32, -8.01, 1.31) BL=(-6.89, -8.02, 1.12) BR=(-7.14, -0.07, 0.28)	TL = (-0.32, -8.06, 0.71) BL = (-6.91, -8.03, 0.25) BR=(-7.14, -0.07, 0.16)
Matanzas Inlet, FL	NP	NP	(2177.80, 64944.46)	NP	(426.9, 618.7)	(25.44, 37.14)	(5.94, 3.70)	(4.81, 2,95)	(1.70, 2.14)	TL=(-0.02, -8.97, 0.58) BL=(-8.22, -9.17, 0.09) BR=(-7.31, 0.12, -0.58)	TL=(-0.22, -8.96, 0.62) BL=(-8.22, -9.19, 0.09) BR=(-7.32, 0.11, -0.57)	TL=(-0.03, -8.98, 0.75) BL=(-8.23, -9.20, -0.01) BR=(-7.32, 0.12, -0.55)

Table 1: Quantitative comparison between three methods of estimating homographies for composing the images of the four cameras to one image. In each method, three cases are used for each dataset: 2-point only, 3-point only and many points. NP notation indicates that the corresponding method cannot be applied. Notice that only our method can be applied in all three cases. Moreover, our method recovers the relative angles between each of the three TL, BL, BT cameras and the reference camera TR.

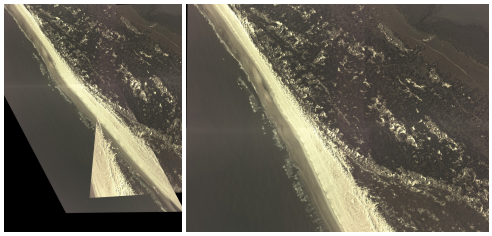

Fig. 3: Results of 4-camera virtual image composition of Matanzas Inlet (Florida) dataset (see the raw frames in Fig. 1) . when only 3 point correspondences are used. Left: The affine approach, Right: The proposed DH method.

outliers) as the inputs. The '3-point' and '2-point' labels signify that only three and two point correspondences (only inliers) are used to estimate a homography matrix, respectively. Our proposed method, DH, has the lowest error values in all combinations in all datasets. In the '3-point' case, apart from the DH, only the affine approach is able to estimate a homography, however, it results in relatively large geometric errors. Among the three compared methods, DH is the only one that is able to estimate a homography matrix using only 2 point correspondences. As described earlier, the proposed method (DH) establishes an analytical homography model directly upon the relative angles between the cameras which corresponds to their physical geometry. The last three columns of Table 1 show the recovered angles (as a direct product) by DH in each dataset. Notice that the presented angles (in degrees) in each row correspond to a specific time instance as they vary time to time. Notice that in WAMI, the cameras are far from the scene and slanted, therefore even very small relative-angle changes can result in significant mapping (projection to the terrain) error. As the error values in Table 1 show, a perspective 8-element homography in most cases is unable to estimate a valid image stitching (due to narrow overlap between the FOVs). The affine homography mostly produce better results than the general perspective one, however the DH results are observed to be always superior in all of our experiments. As mentioned earlier, the minimum number of required point correspondences between each pair of images in affine and DH homography estimations are 3 and 2, respectively. However, the solutions obtained from the affine method have high errors (see Table 1) and are good only locally near their image overlap area. Fig. 3-Left demonstrates an example of using 3-point to stitch the four-camera images in the affine approach, and one can see how invalid the composed virtual image is. However, the proposed DH method is able to provide a superior and decent result even when only 2-point correspondences are used (Fig. 3-Right).

We included more experimental results and additional supporting materials on https://twbot.github.io/DHPaperExtras.

IV. CONCLUSION

We proposed a novel method to produce a large/combined image out of an array of cameras with very tiny overlaps between their FOVs. It uses a model which optimizes the relative angles between the cameras and derives analytical homographies from the recovered angles. The requirements such as the precision camera alignment or large overlap in the FOV of the cameras have been relaxed using the proposed method. The comparative results indicate the superiority and robustness of the proposed method.

Acknowledgment: This work was supported by the The US Air Force Research Laboratory. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof. Authors would like to thank Nicholas Schafer, Enosh Reeder and Francesco Moneta for their valuable helps in the mecatronics and software design.

REFERENCES

[1] K. Palaniappan, R. M. Rao, and G. Seetharaman, *Wide-Area Persistent Airborne Video: Architecture and Challenges*, 2011, p. 349.

[2] B. He, G. Zhao, and Q. Liu, "Panoramic video stitching in multi-camera surveillance system," in *2010 25th International Conference of Image and Vision Computing New Zealand*, 2010, pp. 1–6.

[3] E. Malis, T. Hamel, R. Mahony, and P. Morin, "Dynamic estimation of homography transformations on the special linear group for visual servo control," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 1498–1503.

[4] W. S. Lai, O. Gallo, J. Gu, D. Sun, M. H. Yang, and J. Kautz, "Video stitching for linear camera arrays," in *30th British Machine Vision Conference 2019, BMVC 2019*, 2020.

[5] H. Zhou, D. Zhou, K. Peng, R. Guo, and Y. Liu, "Seamless stitching of large area uav images using modified camera matrix," in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2016, pp. 561–566.

[6] H. Ullah, O. Zia, J. H. Kim, K. Han, and J. W. Lee, "Automatic 360 mono-stereo panorama generation using a cost-effective multi-camera system," *Sensors*, vol. 20, no. 11, 2020.

[7] Y. Lu, K. Wang, and G. Fan, "Photometric calibration and image stitching for a large field of view multi-camera system," *Sensors*, vol. 16, no. 4, p. 516, 2016.

[8] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[9] R. Szeliski, "Image alignment and stitching: A tutorial," Tech. Rep. MSR-TR-2004-92, October 2004.

[10] M. Brown, R. I. Hartley, and D. Nister, "Minimal solutions for panoramic stitching," in *2007 IEEE CVPR*, 2007, pp. 1–8.

[11] H. Aliakbarpour, K. Palaniappan, and G. Seetharaman, "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] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[13] G. Lowe, "Sift-the scale invariant feature transform," *Int. J*, vol. 2, no. 91-110, p. 2, 2004.