Persistent Aerial Video Registration and Fast Multi-view Mosaicing

Edgardo Molina, *Student Member, IEEE,* Zhigang Zhu, *Senior Member, IEEE*

***Abstract*—Capturing aerial imagery at high resolutions often leads to very low frame rate video streams, well under Full Motion Video standards, due to bandwidth, storage, and cost constraints. Low frame rates make registration difficult when an aircraft is moving at high speeds or when GPS contains large errors or it fails. We present a method that takes advantage of persistent cyclic video data collections to perform an online registration with drift correction. We split the persistent aerial imagery collection into individual cycles of the scene, identify and correct the registration errors on the first cycle in a batch operation, and then use the corrected base cycle as a reference pass to register and correct subsequent passes online. A set of multi-view panoramic mosaics is then constructed for each aerial pass for representation, presentation and exploitation of the 3D dynamic scene. These sets of mosaics are all in alignment to the reference cycle allowing their direct use in change detection, tracking, and 3D reconstruction/visualization algorithms. Stereo viewing with adaptive baselines and varying view angles is realized by choosing a pair of mosaics from a set of multi-view mosaics. Further, the mosaics for the second pass and later can be generated and visualized online as their is no further batch error correction.**

***Index Terms*—Image registration, aerial imagery, stereo-viewing, drift correction, wide field-of-view, persistent imaging**

I. Introduction

Persistent aerial imaging at high resolutions is vital for many applications such as search and rescue, surveillance, and mapping applications. The challenges that arise in storing, presenting, and exploiting aerial data are especially critical for time-sensitive operations such as search and rescue, and surveillance. The ability to create wide field-of-view panoramas, 3D reconstructions, 3D visualizations, and automated change detection and tracking would all help operators parse information quicker.

While there are many challenges related to sensor limitations, scene geometry, computation, and storage, we address two problems with a fast solution. First, persistent capture at high resolutions leads to low frame rates, well under Full Motion Video standards. In aerial capture scenarios, this leads to images with a low amount of overlap from frame to frame, 50% or less. Scene features can sometimes only be seen from two consecutive frames making registration problematic. The second problem is the large amount of frame data being generated. At high resolutions it is difficult or prohibitively costly to capture full motion video, a storage device may be running at its bandwidth capacity while capturing 20+ megapixel images at 2-5 frames per second, continuously for up to hours at a time. There is a need to efficiently summarize and present (visualize) all of the data to operators and algorithms exploiting the video data. For the use of this paper, we will use the term "video" to refer to both full motion video and low frame rate video capture.

Aerial vehicles and platforms can image a large area of interest by continuously sweeping an area, either by circling above the area or by forward and back sweeps of the area. In most cases these continuous data collections are interested in detecting changes and tracking movers in unpredictable and cluttered environments. This can generate hours of video collection which may have to be analyzed by human operators. In these cases it is useful to provide operators a mapping of the aerial data as it is received, as well as cues to the changes that are detected over time. The video being captured in these cyclic patterns also exhibit strong motion parallax as the collection is often happening at various altitudes.

Typically the first critical step in making sense of the enormous amounts of persistent aerial video data is registration. Bundle adjustment is a robust registration method that produces accurate results but it is a computationally costly method. There is a need for online methods that produce results that approach the accuracy of bundle adjustment while still generating results immediately. Subsequent exploitation algorithms, such as 3D reconstruction, change detection, tracking, will depend on the correctness of the imageries registration. SLAM (simultaneous localization and mapping) methods are often applied in such scenarios, but SLAM can also be affected by the small amount of video frame overlap.

Multi-view stereo mosaics produce a representation of aerial videos that preserves 3D and moving target information while spatially condensing unchanging static structures in the video. These mosaics work well at representing an area after a single sweep, but in persistent surveillance cases their use needs to be extended to handle the multiple sweeps of the area of interest.

In this paper we present a method that takes advantage of the properties of persistent cyclic video data collections to perform an online registration with drift correction. We assume that there is an area of interest being imaged by an aircraft which continuously circles above or sweeps an area with a continuous flight pattern. We split the persistent aerial imagery collection into individual cycles of the scene. We identify and correct the registration errors on the first cycle in

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

E. Molina and Z. Zhu are with the Department of Computer Science, City College of New York - CUNY, New York, NY, 10031 USA e-mail: {molina,zhu}@cs.ccny.cuny.edu

Manuscript received March 5, 2013; revised October 25, 2013; accepted March 5, 2014.

a batch operation, and then use the corrected base cycle as a reference pass to register and correct subsequent passes online. Multi-view stereo mosaics are then constructed for each aerial pass for representation, presentation and exploitation of the 3D scene that includes moving targets (movers). These sets of mosaics are all in alignment to the reference cycle allowing their direct use in change detection, tracking, and 3D reconstruction/visualization algorithms. Further, the mosaics for the second pass and later can be generated online as there is no further batch error correction.

The main contributions of the paper relate to the tasks of image registration and mosaicing for video data collection patterns that are persistent and cyclic. First, we present a method that produces a base reference for the alignment of video frames, from the video itself (a base cycle from the cyclic video). With this, there isn't a need to use an external reference image for alignment (such as satellite images or DEM's, Digital Elevation Maps) which typically present a problem to alignment as the video frames and external reference image are captured with different sensors at different times. Second, both the base cycle and the subsequent cycles have multiple views (layers) that exhibit motion parallax that can be used for interactive stereo-viewing, 3D reconstruction and mover detection. These multi-view panoramic mosaics that are generated for each cycle allow for interactive stereo viewing with varying baselines and varying view angles. These sets of multi-view mosaics can also be used for more accurate and robust 3D reconstruction [1]. Third, instead of directly registering the upcoming video frames to a layer of the base cycle, which is a multi-perspective image instead of a perspective image, only the motion parameters of the video frames that produce the base cycle are used to find the closest image among them to the current video frame, therefore the registration is between these two video frames. This allows all the video frames from different cycles to be aligned together, and multi-view mosaics can even be generated using frames from multiple cycles.

The remainder of this paper is organized as follows: Section II lists related works, Section III describes the details of our method. Section IV presents results on real data and Section V concludes the paper and presents future research directions.

II. Related Work

Video panoramas and mosaics have become a common way of representing the imagery of a scene captured by a moving camera. Although panoramas have been studied and understood for many years, the research indicates that one of the biggest challenges to creating seamless and drift-free panoramas is still in computing a correct image alignment for long sequences with obvious motion parallax.

The works by Hsu et. al. [2], Shum and Szeliski [3], and recently Lovegrove and Davidson [4] have proposed methods that construct globally aligned panoramas for rotating cameras by optimizing camera poses. In the case of [4] the method is real-time with the use of GPU acceleration. The work of Zhu et. al. [5] proposed a method for fast panorama generation for rotating cameras with a global constraint. Alignment of video images that are augmented with position metadata (from GPS and INS/IMU sensors) has been studied in the work of Heiner and Taylor [6], Oskiper et. al. [7], and Zhu et. al. [8]. The panoramas obtained in these methods provide the benefit of being geo-referenced, but the results will require local alignment improvements due to GPS/INS sensor errors. These also require additional hardware. Other approaches for the global alignment of imagery are to use reference images (e.g. satellite images), a known 3D model (such as a DEM) or known scene landmarks. These methods have recently been studied by Lin et. al. [9], and Zhu et. al. [10] and Oskiper et. al. [11]. Olson et. al. [12] describe techniques for registering images from sequences of aerial images captured of the same terrain on different days. The goal of their work is to make registration robust to changes in weather, by using both robust template matching and SIFT-like feature matching. Motion parallax is not involved in registration and no mosaics are generated.

Much of the work has dealt with rotating cameras, or cameras that translate small distances. For panoramas from aerial imagery, which causes large translations, the research has focused on methods that make use of GPS/INS data [6], [7], [8] and/or a 3D model and landmarks for global imagery alignment [10], [11].

Constructing panoramas also require an algorithm to combine the imagery such that seams between the images are not noticeable. Peleg and Herman, in [13], described panoramas that perform a blending of panorama pixels where multiple aligned images overlap. The works of Agarwala et. al. [14], Burt and Adelson [15], and Jia and Tang [16] have all described general purpose methods for smooth seams in combined imagery. Recently, Gao et. al. [17], have proposed a method that uses a dual-homography warping to combine images with two dominant planes with smooth seams. Constructing panoramas for long sequences has been studied by Peleg and Herman [13], in Zheng's work on Route Panoramas [18], the parallel-perspective stereo mosaics of Zhu et. al. [19], and the multi-viewpoint panorama work of Agarwala et. al. [20].

The work of Sawhney, et.al. [21] considers video from multiple passes over a scene, determines a topology for the video frames and constructs a mosaic by combining all imagery into one sequence. Our work differs in that we consider long video sequences over large areas and spans of time, and we preserve 3D parallax information by constructing multi-view mosaics for each pass of the scene.

The general registration problem is addressed by bundle adjustment [22] and SLAM [23] and Visual Odometry [24] alignment components. Bundle Adjustment and its implementations work robustly in scenes where specific features are viewed from many image frames with different views. In persistent aerial scenarios where we have little scene view overlap, bundle adjustment will achieve its best results only after multiple sweeps of the scene have been taken, providing enough features viewed from multiple frames. SLAM and visual odometry will also face similar problems with a lack of features to match in cases where the view overlap is small. The use of keyframes would increase computation significantly

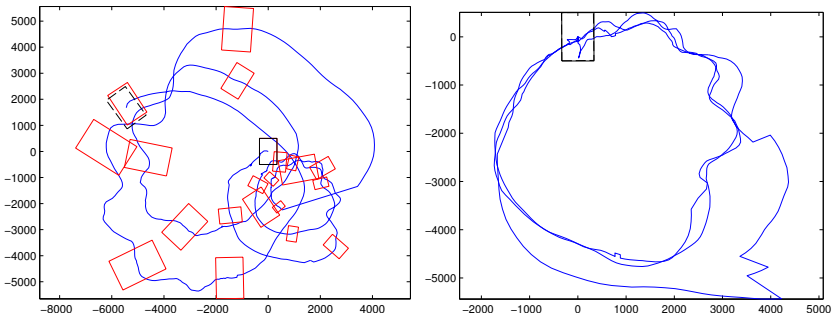


Fig. 1. The CLIF 2007 sequence makes 4 approximately circular passes of an area. The units in both horizontal and vertical axes are in pixels. The dashed black box represents the last frame. Red boxes represent every 50th frame. (a) Image alignment without error correction. (b) Image alignment after error correction.

since every frame or two may become a keyframe.

In this work we have focused on long persistent video sequences with large translational motion with an approximate nadir view. However, the motion model is not a simple linear motion that can be modeled by a linear push broom imaging geometry [19]. The data we consider in this paper has circular motion camera trajectories. We perform global alignment using only the imagery itself since it is possible for GPS/INS to fail and since some areas may not have accurate or existing 3D DEM's. We also propose a method to combine long video sequences into multi-view stereo panoramas using a layering approach.

III. Persistent Aerial Video Registration with Drift Correction

In scenarios where persistent imaging of a large area is required, an aerial vehicle will "circle" above the area of interest. Generally, the aerial vehicle will attempt to follow some ideal cyclic path many times. Due to wind, vehicle speed, and environment conditions each run of the path varies from the ideal path and will likely capture images at different locations and with different camera poses. Image registration is therefore required to align the imagery across runs, and ultimately to a georeferenced space.

Using on-board GPS/INS sensors is required to georeference the imagery, and can also be used directly to align the imagery for panorama purposes. But, the measures made by GPS/INS devices come with errors due to hardware, and used directly will produce panoramas with large apparent errors and discontinuities. Additionally, GPS availability can be compromised or spotty in many environments.

Using a solely image-based registration approach can produce seamless panorama results, as the errors between consecutive frame-to-frame registration are small. But solely image-based registration approaches suffer from error accumulation which often leads to large drifts over long imagery sequences. Figure 1a, shows how drift can accumulate over 4 passes of an area, the correct image path for this example was a continuous circular flight as shown in Figure 1b. Using results with significant drift in sequences that are known to produce a cycle, will produce results like those in Figure 4a, where the path is not closed. In the following subsections we present our image-based approach to align images while eliminating global drift for multi-pass aerial video. The base cycle is first corrected in a batch operation and used as a reference for online registration of subsequent cycles.

A. Motion Estimation

Working with video data benefits image-based registration methods since video frames are captured sequentially providing images with considerable overlap. Additionally illumination changes in naturally lit scenes will occur gradually in a video sequence. For these reasons we use a standard frame-to-frame registration approach to estimate camera motion. We use a pyramidal block-based correlation method [5] to determine the interframe affine parameters (translation (tu, tv), rotation θ, and scale s) between consecutive frames. In our simplified motion model the translation (tu, tv) accounts for 2D translational changes along the X and Y axis and small pan and tilt angular changes of the camera; the rotation θ accounts for heading changes between two frames; and the scale s accounts for focal length and small altitude changes of the camera. We are making an assumption about the planarity of the scene, that while not true, it captures camera motion based on a dominant plane that we exploit to construct multi-view stereo mosaics.

B. Base Cycle Registration and Error Correction

We use the method presented in [25] and based on [26] to estimate and correct the registration errors for a single pass on a cyclic path. In [26], Sharp et. al., present an analytic method to produce globally consistent registrations of multi-view data in cycles. By performing pairwise registrations (and measuring error) of each view of their 3D scan data, they were able to compute the global error obtained, and they redistributed that error in a weighted manner across all pairwise registrations, yielding a globally and locally consistent registration. Unlike bundle adjustment and other optimization methods, this method only requires an error criterion, and does not require that features be matched across all views or for registration to continuously take place. Since the interframe alignment errors are typically small (∼2-5 pixels), the redistribution fairly

accurately reflects the true motion; additionally the weighted process ensures a reasonable redistribution of errors.

This method is applied to the first full pass in a multi-pass video sequence,which we call the base cycle, and apply as follows:

We compute the interframe motion $\mathbf{m}_k$ between consecutive frames $k-1$ to k of the base cycle. The global motion, $\mathbf{M}_k$, for any frame k is computed by taking the first frame of the sequence as the global reference frame, such that:

$$\mathbf{M}_k = \mathbf{m}_k \mathbf{M}_{k-1} \tag{1}$$

where:

$$\mathbf{M}_1 = \mathbf{I}. \tag{2}$$

With a complete base cycle consisting of K frames, the computed $\mathbf{M}_K$ should close the cycle assuming no errors. More specifically, for testing purposes and to check the error, we insert a copy of frame 1 at the end of the sequence as frame $(K+1)$; ideally without errors we expect the global position of frame $(K+1)$ as computed through accumulation, to be identity, $\mathbf{I}$, with regard to the first frame. (In practice we do not use frame $K+1$, and we note that identifying frame K to terminate the first cycle in the video is a separate problem. We have taken the approach to automatically search for frame K by matching each frame with frame 1, and selecting the frame that best matches as in [5].) But due to the accumulation of small errors at each frame that is not the case:

$$\mathbf{I} = \mathbf{M}_1 \neq \mathbf{M}_{K+1} = \mathbf{m}_{K+1}\mathbf{M}_K \tag{3}$$

Instead, $\mathbf{M}_{K+1}$ represents the total accumulated residual error in the cycle for each motion model parameter, $(Tu_{res}, Tv_{res}, \Theta_{res}, S_{res})$. We redistribute these errors across all K base cycle frames such that the corrected global path is constrained to a cycle. We take a weighted error redistribution approach by assigning an error weight to each frame based on a SAD (sum of absolute differences) error measure on the pixel intensities of the overlapped region between frames. Thus, each frame k is assigned:

$$w_k = (SAD_k) / \sum_{i=1}^{K} SAD_i \tag{4}$$

where the higher the weight, the more error is distributed to that frame. We chose SAD empirically, but other error criteria such as SSD, MSE, or RMSE can be used.

The procedure for base cycle registration and error correction is then:

1) Compute global position of all $K+1$ frames ($\mathbf{M}$)
2) Determine the residual rotation and scale (Θ_{res}, S_{res}) from $\mathbf{M}_{K+1}$
3) Compute updated global positions ($\mathbf{M}'$) redistributing (Θ_{res}, S_{res})
4) Determine the residual translation (Tu'_{res}, Tv'_{res}) from $\mathbf{M}'_{K+1}$
5) Compute updated global positions ($\mathbf{M}''$) redistributing $(Tu'_{res}, Tv'_{res}, \Theta_{res}, S_{res})$

The translation parameters are dependent on rotation and scale requiring that they be treated first. Once complete we have:

$$\mathbf{M}''_{K+1} = \mathbf{M}''_1 = \mathbf{I}. \tag{5}$$

This yields a result which is globally consistent by redistributing the error across the entire sequence. Section IV will show examples of its use. The base cycle registration and error correction presented here is a batch operation, that cannot be applied until all of the frames are available. But it is not a costly operation, because interframe computations are done online, and the global batch operation simply operates on the parameters of $\mathbf{M}$, and does not require any additional image matching and registration operations.

C. Online Registration of Subsequent Cycles

In persistent aerial video we may have an aerial platform image the area of interest for hours. By using the method described in the previous section we can isolate a single cycle and correct any errors that accumulate over the sequence. It would be possible then to apply the same method to each subsequent cycle after the base cycle. The disadvantage with such a strategy is that it can only be applied at the end of each cycle. Furthermore, even though such a strategy could correct the end-to-end errors in each cycle, there is no guarantee that all the frames from across two or more cycles can be aligned. It is more desirable to provide immediate results to operators that are also registered to a single common reference. Here we present a spatial-temporal registration method that uses the corrected base cycle as a reference to perform an online registration of the subsequent cycles. Such a registration is desirable because it treats the error online to produce a corrected global motion estimate for each frame, and cycles are identified automatically.

We proceed by assuming that we have the corrected base sequence $\mathbf{M}''$, which we now call $\mathbf{M}^B$ from the previous section for the K base frames. We then treat the remaining sequence of frames as t, from 1 to the end of the sequence n (which consists of an unknown number of cycles). As each frame is being captured we compute a *temporal* interframe parameter estimate from frames $(t-1)$ to t which we denote as $\mathbf{m}_{t \leftarrow t-1}$. We also compute a *spatial* interframe parameter estimate from frame k to t which we denote as $\mathbf{m}_{t \leftarrow k}$, where k is a frame from all previous cycles (initially only the base cycle) located nearest to frame t globally. We now compute the global position estimate, $\mathbf{M}_t$, by integrating the spatial and temporal motion estimates for each frame t.

We take

$$\mathbf{M}_0 = \mathbf{M}^B_K \tag{6}$$

so that we can compute

$$\mathbf{M}^T_t = \mathbf{m}_{t \leftarrow t-1}\mathbf{M}_{t-1} \tag{7}$$

as our temporal estimate. We also compute an error measure for the temporal estimate, SAD^T_t , using the SAD measure. The estimated global location $\mathbf{M}^T_t$ is then used to determine the nearest frame k from the base cycle, which we use to compute

$$\mathbf{M}^S_t = \mathbf{m}_{t \leftarrow k}\mathbf{M}^B_k \tag{8}$$

as our spatial estimate. We once again compute an error measure for the spatial estimate, SAD^S_t, between the overlapped intensities of frames t and k using the SAD measure. We use

Require: All $Panos$ set to $empty$

procedure GENERATEPANORAMAS($Imgs, Panos$)
  $N \leftarrow$ COUNT($Imgs$)
  $L \leftarrow$ COUNT($Panos$)
  for $t \leftarrow 1$ to N **do**
    $img \leftarrow Imgs[t]$
    for $l \leftarrow 1$ to L **do**
      $pano \leftarrow Panos[l]$
      for all $img(x, y) \neq empty$ **do**
        $(x', y') \leftarrow$ PROJECT(img, x, y)
        if $pano(x', y')$ is $empty$ **then**
          $pano(x', y') \leftarrow img(x, y)$
          $img(x, y) \leftarrow empty$

Fig. 2. GENERATEPANORAMAS procedure.

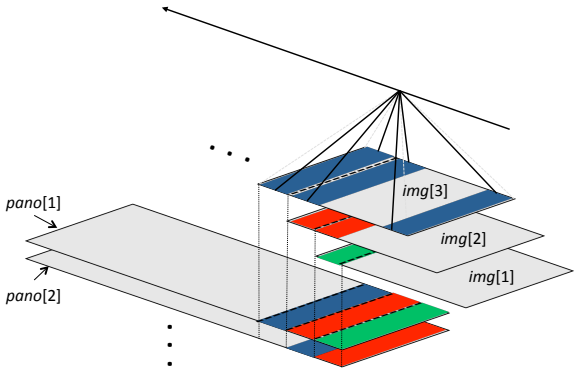


Fig. 3. Illustration of the GENERATEPANORAMAS procedure. Under ideal conditions, each multi-viewpoint panorama $pano[i]$ is constructed from a similar view angle on the scene.

the error measures to weigh both results and integrate them into a final result

$$\mathbf{M}_t = w_t^T \mathbf{M}_t^T + w_t^S \mathbf{M}_t^S \tag{9}$$

where

$$w_t^T = (SAD_t^S)/(SAD_t^T + SAD_t^S) \tag{10}$$

$$w_t^S = (SAD_t^T)/(SAD_t^T + SAD_t^S). \tag{11}$$

We note that on occasion a frame k may be unavailable due to the motion trajectory temporarily leaving the ideal cyclic path, or because errors are too great. To handle these cases, we maintain a standard deviation of the spatial errors, along with a mean motion model. When the errors are outside of 3 standard deviations we rely only on the temporal estimate. As more cycles are taken, there will also be more frames in the history to locate a frame k, we also use this for correction.

This weighing approach adds robustness to the registration of persistent videos. In our implementation we have relied on the SAD error measure as it has worked well in real world scenarios. But other suitable error metrics (MSE, RMSE, and SSD) can be substituted in. Furthermore, we have found the simple weighting approach produces very good results, even though this spatial-temporal registration approach can be generalized to a more general framework such as one using Extended Kalman Filtering; this would incorporate the error dynamics of both spatial and temporal registration.

Additionally, as we have presented the method, no further batch error correction is required. This permits us to begin generating multi-view mosaics immediately on the subsequent cycles. It also permits us to begin other video exploitation such as change detection and moving object detection, since the frames are all in reference to the error corrected base cycle.

D. A Layered Approach for Fast Multi-view Mosaic Generation

After computing the camera motion and correcting its global drift we can construct a set of multi-view panoramas for the video scene. We use a layered approach for fast multi-view mosaic generation. The basic principle is the following. Based on the global motion parameters, buffers for multiple empty layers are created, and are laid out in the order of the 1st layer, the 2nd layer, and so on, each below the previous one, as can be seen in Figure 3. All of the original frames are warped based on their global motion parameters and the warped frames are laid out in the mosaic space. Starting from the first layer at the top, pixels of each warped frame are placed down through all the layers until they hit the first empty layer and it gets drawn there. If it hits a layer (layer n) with existing pixels at its location, it continues onto the next layer (layer $n + 1$), until it falls on an empty layer.

The procedure GENERATEPANORAMAS (Figure 2) outlines our approach to constructing a set of multi-view panoramic mosaics. The procedure takes as input an ordered sequence of the video images (along with its computed parameters), $Imgs$, and a reference to the ordered sequence of multi-view panoramas, $Panos$. The procedure iterates through all of the images in the sequence; then for each panorama, starting with the top one, we draw all non-empty pixels in img onto the current $pano$, and set those drawn pixels to $empty$; this operation is performed for each of the multi-view panoramas. The function PROJECT is not outlined, but simply maps (x, y) in its local img coordinate space to (x', y') in the global $pano$ coordinate space using $\mathbf{M}_{img}$ (which is the corrected global position $\mathbf{M}_k^B$ in the base frame, or $\mathbf{M}_t$ from eq. 9 in subsequent cycles). Figure 3 shows an illustration of drawing three image frames onto the first two multi-view panoramas. Here we see that $img[3]$ paints its leading slit onto $pano[1]$, and the following slit onto $pano[2]$ and so on. Note that the slits contributing to a particular panoramic layer come from very similar perspective directions in the original images, thus minimizing misalignments between slits. In the ideal case, the camera performs a pure 1D translation in the Y-direction, and each image contributes a single X column to each panoramic layer and therefore it forms a perfect parallel-perspective (pushbroom) mosaic [19]. In more general cases, each layer approximates a multi-perspective panorama with similar viewing directions. Thus two layers can form a pair of multi-perspective stereo panoramas.

Our goal with this method is to provide 3D viewable results quickly. Warping operations in particular can get computationally expensive as the imagery resolution keeps increasing. Blending operations can produce ghosting which can make

it distracting for some users when viewing the imagery on a 3D display. Instead we rely on the fact that our drift correction method has produced results with very minor local misalignments that the viewer can cope with, in particular because we performed the layered construction approach, where similar views align into each layer. We understand that this is a subjective issue and will vary from user to user.

The layered approach can be used to generate sets of multi-view mosaics from both the base cycle as well as the subsequent cycles. In fact multi-view mosaics can also be generated from aligned image frames that come from all cycles, providing the best imagery overlap for multi-view mosaics. This will be left as a future work topic.

IV. Results

We ran our implementation of the above algorithms on the CLIF (Columbus Large Image Format) 2007 [27] data. This dataset was captured by a persistent flyover of the Ohio State University campus in Columbus, Ohio. The images are captured by 6 cameras at approximately 2 frames per second. In this work we have only considered a single camera, number 0 specifically. The dataset covers about four and a half cycles over the campus. In the following experiments we consider 890 frames covering 4 complete cycles. We scaled (50%) and cropped the images for runtime considerations, but full resolution results are comparable to these results.

In the CLIF data set we observe that under real world conditions small unexpected changes will pose problems to our registration and mosaicing algorithms. In this dataset we observed that at different locations on different passes, it appears that the vehicle speed increased enough that there was not enough overlap to perform image based registration. One solution to this problem is to have onboard GPS and inertial measurement devices to estimate the motion. With the solution presented here we show that the base cycle provides a reference to correct such cases when they occur after the base cycle.

A. Registration Results

First we ran the entire sequence with our frame-to-frame image based matcher and plotted the results in Figure 1a. The estimated path begins at position $(0, 0)$ illustrated by the solid black frame. We plot every 50th frame in red, and the final frame with a dashed black outline. While the four cycles of the flight should actually cover the same area, it is obvious from the plot that the estimated motion parameters of the camera drift a lot in both location and scale due to the accumulating errors in image registration. Figure 4a is a plot of the original motion estimates for the base cycle and we can see that even in the first cycle the last frame does not meet the first frame of this cycle. Figure 4b shows it after applying drift correction. Figure 4c shows the estimated motion paths for three subsequent passes using the online spatial-temporal registration method outlined above.

From Figure 4 we can see that the online spatial-temporal approach maintains the corrected global path without large drift for 3 subsequent cycles. At the bottom right of Figure 4c we can see that the last cycle significantly drifted from the path, by inspecting the original imagery it can be seen that this actually did happen in the imagery, and was handled appropriately by our spatial-temporal method. For the result in 4c we did not separate any cycles beyond the base, all frames where automatically aligned as they came in for 3 continuous cycles after the base.

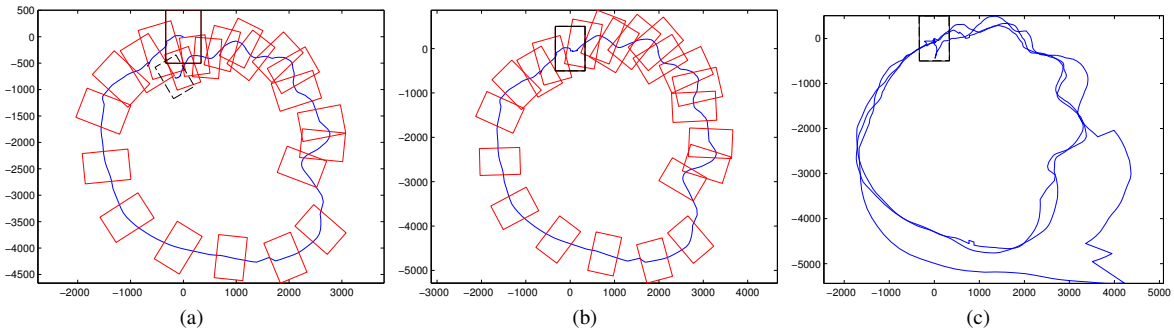
For comparison, Figure 5 shows the results of separating each subsequent cycle and attempting to register them independently. We can see that all of the cycles accumulate significant drift, and while the cycles are over the same area, the drift changes in different passes. For example, in cycles 3 and 4 we can see that the scale got larger, where in cycle 2 and the base cycle, drift made the scale get smaller.

B. Mosaicing Results

Figure 6 shows the mosaicing results on the base cycle and on cycle 2. Cycle 2, like the base cycle, has very minor misalignments in a few places, particularly those with small overlap. Figure 7 shows a close up view of buildings before and after error correction in the mosaics of the base cycle. Figure 7b shows that error correction rotated the scene and changed its scale, and shows that it does not have easily detectable misalignments. Overall the mosaics show that the registration for subsequent cycles remains robust and without significant drift. Additionally the mosaics across different passes are aligned on the same reference coordinates, allowing them to be used for change detection across passes, moving object detection, and 3D viewing and reconstruction. Figure 8 shows a close up stereo view of two multi-view mosaics as a red-cyan anaglyph. Viewing this image in 3D requires red-cyan glasses (for left and right eye respectively) on a color printout of this paper. Using the multi-view mosaics generated from multiple cycles, we will be able to create a virtual 3D fly-through with the freedom of choosing viewing locations, viewing angles, and cycles for 3D perception, motion detection and change detection, leading to a stronger 3D and motion perception than simply viewing the original video sequences.

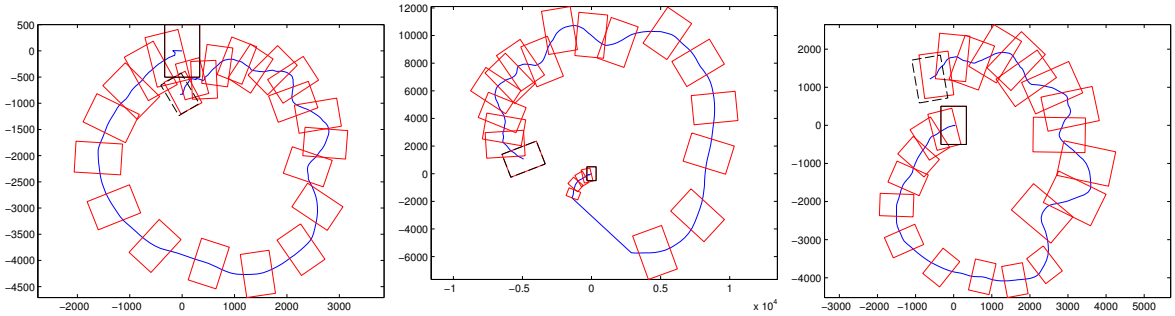
C. Error Analysis

To determine how well our spatial-temporal registration method performed we ran the complete four cycle (890 frame) sequence three times, each time inserting frame 1 at a different location to measure its alignment error. In the first run we inserted the frame only at the end of cycle 2. In the second run, we inserted frame 1 only at the end of cycle 3, and in the third run, only at the end of cycle 4. This was done separately for each cycle, so that inserting the frame in cycle 2, would not affect the computed global position of frame 1 at the ends of cycles 3 and 4, and so on. The purpose of this experiment was to determine after some number of cycles whether the spatial-temporal method continued to maintain a globally consistent alignment, even without batch post-processing of each cycle. Under ideal conditions, the inserted frame should always be $\mathbf{I}$, since the frame should pick itself as frame k in the spatial estimate, but in practice this is not always the case. Table I shows the results of the test for all runs. The table specifies



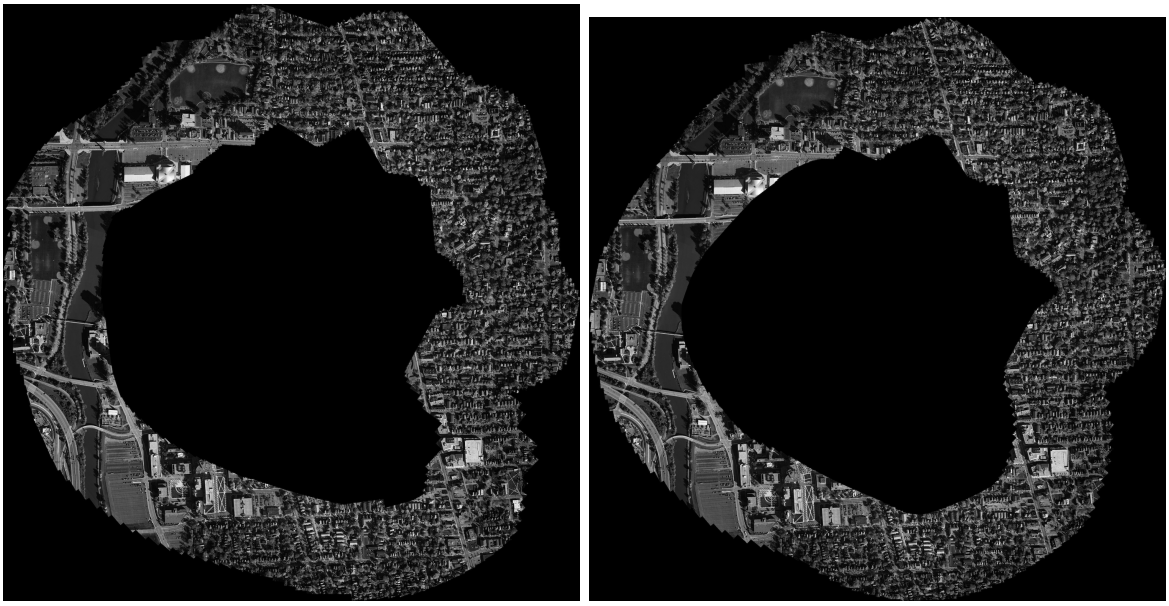
(a) (b) (c)

Fig. 4. (a) Original Base Cycle. (b) Corrected Base Cycle. (c) 3 registered subsequent passes. Red boxes represent every 10th frame. The dashed black box is the last frame.



(a) (b) (c)

Fig. 5. Original registrations for subsequent cycles: (a) cycle 2 (b) cycle 3 (c) cycle 4. Red boxes represent every 10th frame. The dashed black box is the last frame.



(a) (b)

Fig. 6. Mosaics for (a) base cycle and (b) cycle 2.

Inserted in Cycle	Spatial Pick k	Temporal Parameters (Tu, Tv, Θ, S)	Spatial Parameters (Tu, Tv, Θ, S)
2	0	-0.4982 , 1.3306 , 0.0018 , 0.9986	0, 0, 0, 1
3	228	-13.9932, -75.6169, 0.0890, 0.9832	-0.5681 , 1.3147 , 0.0020 , 0.9986
4	0	-3.0000 , 5.4349 , 0.0107 , 0.9937	0, 0, 0, 1

TABLE I
ERROR ANALYSIS: THE TABLE LISTS THE ONLINE SPATIAL-TEMPORAL ESTIMATED PARAMETERS FOR THE REFERENCE FRAME AT THE END OF EACH CYCLE. TRANSLATION IS MEASURED IN PIXELS.

(a) (b)

Fig. 7. Closeup of error correction in base cycle: (a) shows building prior to error correction, (b) shows the same buildings after correction.

which frame k was nearest in the spatial estimate. A spatial pick of 0 means the frame matched itself, and for cycles 2 and 4 we can see that is the case, further we can see that the temporal parameters were within 5 pixels of the correct alignment in a mosaic area of about 5000x5000 pixels. With cycle 3 we observe that the temporal parameter was off by over 13 pixels in the X axis and 75 pixels in the Y axis, and this caused the spatial pick k to be a frame other than itself. Frame 228, is 2 frames from 0 and closing the loop in the base cycle, and with this spatial estimate we see that we are placed within a pixel and half from the correct position. This demonstrates how the online weighing solution manages to maintain the flyovers path without accumulating the large drift as observed in Figure 1a and Figure 5.

To determine the effects of redistributing the error in the base cycle we ran an experiment with each of the four cycles as the base. In this experiment we found that the average redistributed tu and tv shift to any one frame in all 4 cycles ranged from 3.1 pixels to 5.5 pixels. Testing showed that applying this error uniformly created a lot of local artifacts, while the weighted method produced locally and globally consistent results.

V. CONCLUSIONS

Here we have presented a spatial-temporal solution to improve the registration of persistent aerial videos that have low frame rates and low consecutive frame overlap. We first identify a base cycle of the imagery which we use to correct errors and construct a reference frame for the further registration of subsequent cycles over the same area. We have shown that for subsequent cycles the method can be applied online and can be used to construct sets of multi-pass multi-view mosaics, which are a useful representation for video data archival, exploitation and visualization.

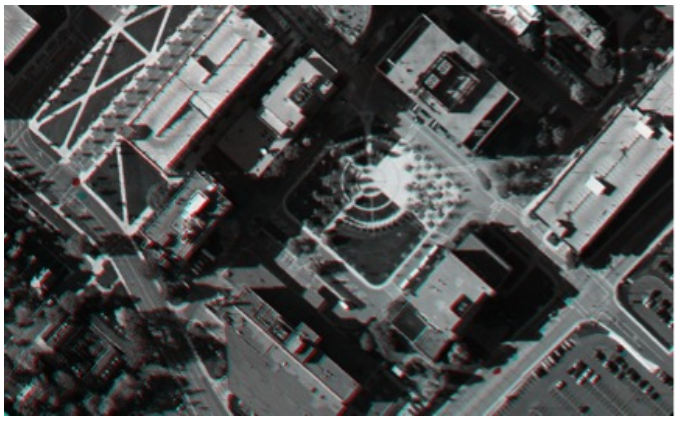


Fig. 8. A close up 3D view of the CLIF scene. Use red-cyan anaglyph glasses to view on a color copy.

Our future work in data exploitation include using the multi-pass multi-view mosaic representation for: mover detection, change detection, 3D reconstruction. In addition, for image registration, we will apply a more principled spatial-temporal integrated registration framework such as the Extended Kalman Filter. For stereo mosaicing, we will investigate methods to further improve alignment quality by utilizing optimal seam selection and view interpolation techniques.

ACKNOWLEDGEMENTS

This work has been supported by AFRL Award #FA8650-05-1-1853, AFOSR Award #FA9550-08-1-0199, the 2011 Air Force Summer Faculty Fellow Program (SFFP), and by a PSC-CUNY Research Award. The work is also partially supported by NSF under Award #EFRI-1137172, Award # CNS-0551598, and ARO Award #W911NF-08-1-0531. We thank Mrs. Olga Mendoza-Schrock, Dr. Todd Rovito, Dr. Clark N.

Taylor and Dr. Kevin L. Priddy for their inspiring discussions and advising during our summer research at the Air Force Research Laboratory, WPAFB.

REFERENCES

[1] H. Tang and Z. Zhu, "Content-based 3d mosaics for representing videos of dynamic urban scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22(2), pp. 295–308, 2012. 2

[2] S. Hsu, H. S. Sawhney, and R. Kumar, "Automated mosaics via topology inference," *IEEE Computer Graphics and Applications*, vol. 22, pp. 44–54, 2002. 2

[3] H.-Y. Shum and R. Szeliski, "Construction and refinement of panoramic mosaics with global and local alignment," *IEEE ICCV*, p. 953, 1998. 2

[4] S. Lovegrove and A. Davison, "Real-time spherical mosaicing using whole image alignment," *ECCV*, pp. 73–86, 2010. 2

[5] Z. Zhu, G. Xu, E. M. Riseman, and A. R. Hanson, "Fast construction of dynamic and multi-resolution 360 degrees panoramas from video sequences," *Image Vision Comput.*, vol. 24, no. 1, pp. 13–26, 2006. 2, 3, 4

[6] B. Heiner and C. N. Taylor, "Creation of geo-referenced mosaics from mav video and telemetry using constrained optimization bundle adjustment," in *IROS*, 2009, pp. 5173–5178. 2

[7] T. Oskiper, Z. Zhu, S. Samarasekera, and R. Kumar, "Visual odometry system using multiple stereo cameras and inertial measurement unit," in *CVPR*, 2007. 2

[8] Z. Zhu, E. M. Riseman, A. R. Hanson, and H. J. Schultz, "An efficient method for geo-referenced video mosaicing for environmental monitoring," *Mach. Vis. Appl.*, vol. 16, no. 4, pp. 203–216, 2005. 2

[9] Y. Lin, Q. Yu, and G. Medioni, "Map-enhanced uav image sequence registration," *WACV*, vol. 0, p. 15, 2007. 2

[10] Z. Zhu, T. Oskiper, S. Samarasekera, R. Kumar, and H. S. Sawhney, "Real-time global localization with a pre-built visual landmark database," in *CVPR*, 2008. 2

[11] T. Oskiper, H.-P. Chiu, Z. Zhu, S. Samarasekera, and R. Kumar, "Stable vision-aided navigation for large-area augmented reality," in *VR*, 2011, pp. 63–70. 2

[12] C. F. Olson, A. I. Ansar, and C. W. Padgett, "Robust registration of aerial image sequences," in *Proc. ISVC(2)*, ser. ISVC '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 325–334. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-10520-3_30 2

[13] S. Peleg and J. Herman, "Panoramic mosaics by manifold projection," *CVPR*, vol. 0, p. 338, 1997. 2

[14] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, "Interactive digital photomontage," *ACM Trans. Graph.*, vol. 23, pp. 294–302, August 2004. 2

[15] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Trans. Graph.*, vol. 2, pp. 217–236, October 1983. 2

[16] J. Jia and C.-K. Tang, "Image stitching using structure deformation," *IEEE Trans. PAMI*, vol. 30, pp. 617–631, 2008. 2

[17] J. Gao, S. Kim, and M. Brown, "Constructing image panoramas using dual-homography warping," *CVPR*, 2011. 2

[18] J. Zheng, "Digital route panoramas," *IEEE Multimedia*, vol. 10, no. 3, pp. 57–67, 2003. 2

[19] Z. Zhu, A. R. Hanson, and E. M. Riseman, "Generalized parallel-perspective stereo mosaics from airborne video," *IEEE Trans. PAMI*, vol. 26, pp. 226–237, 2004. 2, 3, 5

[20] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski, "Photographing long scenes with multi-viewpoint panoramas," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 853–861, 2006. 2

[21] H. S. Sawhney, S. Hsu, and R. Kumar, "Robust Video Mosaicing through Topology Inference and Local to Global Alignment," in *ECCV '98: Proceedings of the 5th European Conference on Computer Vision*, 1998. 2

[22] Y. Jeong, D. Nister, D. Steedly, R. Szeliski, and I.-S. Kweon, "Pushing the Envelope of Modern Methods for Bundle Adjustment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 2

[23] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *AAAI/IAAI*, 2002, pp. 593–598. 2

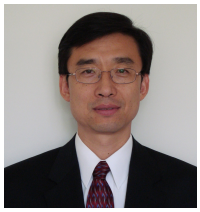
[24] D. Scaramuzza and F. Fraundorfer, "Visual Odometry [Tutorial]," *Robotics & Automation Magazine, IEEE*, 2011. 2

[25] E. Molina, Z. Zhu, and C. N. Taylor, "A layered approach for fast multi-view stereo panorama generation," *IWVP*, 2011. 3

[26] G. Sharp, S. Lee, and D. Wehe, "Multiview registration of 3d scenes by minimizing error between coordinate frames," *IEEE Trans. PAMI*, vol. 26, no. 8, pp. 1037 –1050, aug. 2004. 3

[27] O. Mendoza-Schrock, J. Patrick, and M. Garing, "Exploring image registration techniques for layered sensing," in *Proceedings of SPIE*, vol. 7347, 2009, p. 73470V. 6

Edgardo Molina is a Computer Science doctoral student at the CUNY Graduate Center in New York. He received a BS in Computer Science from the Macaulay Honors College at the City College of New York CUNY in 2005. His research interests are in computer vision, human-computer interaction, 3D visualization, and applications for assistive technologies. He is a student member of the IEEE.



Zhigang Zhu is the Herbert G. Kayser Chair Professor of Computer Science, at the CUNY City College and the Graduate Center. He directs the City College Visual Computing Laboratory (CcvcL), and co-directs the Center for Perceptual Robotics, Intelligent Sensors and Machines (PRISM) at CCNY. His research interests include 3D computer vision, multimodal sensing, virtual/augmented reality, video representation, and various applications in assistive technology, environment, robotics, surveillance and transportation. He has published over 140 technical papers in the related fields. He is an Associate Editor of the Machine Vision Applications Journal, and a Technical Editor of the ASME/IEEE Transactions on Mechatronics.