Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 16-002

Large Scale Image Mosaic Construction for Agricultural Applications

Zhengqi Li, Volkan Isler

January 28, 2016

Large Scale Image Mosaic Construction for Agricultural Applications

Zhengqi Li and Volkan Isler

Abstract—We present a novel technique for stitching images including those obtained from aerial vehicles flying at low altitudes. Existing image stitching/mosaicking methods rely on inter-image homography computation based on a planar scene assumption. This assumption holds when images are taken from high-altitudes (hence the depth variation is negligible). It is often violated when flying at low altitudes. Further, to avoid scale and resolution changes, existing methods rely on primarily translational motion at fixed altitudes.

Our method removes these limitations and performs well even when aerial images are taken from low altitudes by an aerial vehicle performing complex motions. It starts by extracting the ground geometry from a sparse reconstruction of the scene obtained from a small fraction of the input images. Next, it selects the best image (from the entire sequence) for each location on the ground using a novel camera selection criterion. This image is then *independently* rectified to obtain the corresponding portion of the mosaic. Therefore, the technique avoids performing costly joint-optimization over the entire sequence. It is validated using challenging input sequences motivated by agricultural applications.

I. INTRODUCTION

ONE of the driving forces behind the proliferation of commercial Unmanned Aerial Vehicles (UAVs) is their applications in agriculture. In particular, obtaining frequent aerial images of their farms allow the farmers to make informed decisions related to fertilizer application, irrigation, pesticide application and other farm practices. In a typical application, a UAV equipped with (often multi-spectral) cameras is flown over the field. The images are geo-tagged and a mosaic is built to inform farm decisions.

Existing mosaic construction algorithms rely on the assumption that the world is planar and therefore two images viewing the same plane are related by a homography. To enforce this assumption in practice, UAVs are typically flown at high altitudes. This way, depth variation due to objects such as trees become negligible. Since scale and resolution changes can introduce errors in mosaic computation, the UAV motion is constrained so that it flies at a fixed altitude with the camera plane essentially parallel to the ground. Violation of these assumptions results in misalignment, perspective distortion or even complete failure in mosaic generation.

However, there are many reasons, including regulations, safety and privacy concerns, to avoid flying at high-altitudes. From a technical perspective, flying high can cause loss of

resolution. Smaller UAVs may not have the robustness and payload capacity to fly at high altitudes. In this paper, we present a novel mosaic construction algorithm which alleviates these limitations.

Rather than making an implicit planar world assumption, we start out by reconstructing the scene using a small subset of the images using general Structure from Motion (SfM) techniques. This allows us to explicitly compute the ground plane (or planes if there are multiple dominant planes). At the heart of our method lies a novel view selection technique which allows us to select, for each ground patch, the best view across *the entire sequence*. This view is then *independently* rectified and warped onto the ground patch.

In a sense, our approach falls between a SfM-based full reconstruction and standard mosaic generation. On the one hand, performing SfM over the entire input sequence would be prohibitively costly and would yield only a sparse reconstruction of the scene. On the other, homography-based mosaic construction is prone to accumulation of errors. Joint optimization methods such as bundle adjustment to avoid error accumulation would again make use of the entire sequence infeasible. Discarding images to reduce the input size would yield significant loss of resolution especially if wide-angle lenses (e.g. a fish eye camera) are used to obtain images. Our approach circumvents these difficulties by choosing the best view across the entire sequence and independently warping it to a carefully chosen world frame. As we show in Section IV-B, its performance is superior to state-of-the-art mosaicking software.

Before we present further technical details and experimental results, we start with an overview of related work.

II. RELATED WORK

Stitching images to generate image mosaics and panoramas are fundamental computer vision problems. The reader is referred to the book by Szeliski for technical background [1].

Image panorama generation starts out with the property that if the camera motion between two images of a scene is purely rotational, then the images are related by a homography. The basic method to create a panorama is to warp new images onto the panorama one at a time and to align each new image with the images that have already been warped into the panorama [2]. This method is intuitive but prone to large error accumulation if the number of images that need to be stitched is more than a hundred. Szeliski [3] reviews image alignment and stitching algorithms, and presents a standard method to stitch a large number of images. Brown and Lowe

The authors are with the Computer Science & Engineering Department, University of Minnesota. Emails: {lix2938, isler}@umn.edu. This work is supported in part by the MnDrive initiative, National Science Foundation grants #1317788 and #1111638. We thank Kent Cavender-Bares, Haluk Bayram and Nikolaos Stefas for the images.

Fig. 1: The orchard dataset. The camera intrinsics for this footage were precomputed.

Fig. 2: The corn field dataset. The camera intrinsics for this footage were unknown.

[4], [5] propose a standard pipeline to stitch multiple images based on rotation-only Bundle Adjustment (BA) using SIFT features. These methods are optimized for the case when the camera motion is purely rotational.

If the camera motion includes translation, then the relationship between two images is no longer restricted to be a homography unless the scene is planar. In image mosaicking, views of a roughly planar scene are stitched together by warping the images to a base view. Most of the research on image mosaicking focuses on maintaining high resolution and minimizing misalignment.

A common application domain for image mosaicking is underwater navigation and exploration. Gracias and Santos-Victor [6] use feature alignment to estimate frame-to-frame motion parameters which are then cascaded to form a global motion model. Afterwards, all images are mapped onto a common, arbitrarily chosen, image reference frame. This limits the applicability of the method to small scale environments since in the long run, it will cause perspective distortion whenever an image plane is not parallel to the sea floor. Later, they proposed an improved algorithm that first estimates image motion in a sequential manner to get the topology of image overlap. The algorithm then refines motion estimates by minimizing the matching error across pairwise images and then refining global motion by using all matches [7], [8]. In [9], [10], the authors create a mosaic of the sea-floor by finding image registration parameters that best comply with image matches. Gracias and Negahdaripour [11] proposed a mosaic creation method using image sequences from different altitudes. The algorithm

takes advantage of the motions extracted from higher altitude images to refine the topology of motions for lower altitude images. However, all of these methods are demonstrated only in small scale underwater environments. Ferrer [12] present an approach for photo-mosaicking using 3D position estimates provided by a navigation sensor such as an IMU.

One of the most popular mosaicking technique is a graph-based method. Kang et al. [13] propose a global image registration method based on the frame-graph. The algorithm searches the optimal path in the frame-graph and constructs the final mosaic using the images along this path. Fusiello et al. [14] also use graph-based global registration techniques to create super-resolution mosaic based on nonlinear least square optimization. Pfingsthorn et al. [15] introduce a spectral image registration method based on Phased-Only Matched Filter (POMF). The results of POMF are then used in pose graph optimization to create mosaics.

There is also significant work on creating mosaics of aerial images taken at sufficiently high altitudes (more than 100 meters). Steder et al. [16] present a system that performs SLAM using an attitude sensor and downward camera attached to blimps and helicopters. Botterill [17] describes a scheme for real time aerial image mosaicking which warps every image to a single image chosen under the assumption that the camera is looking straightly downward to the ground. Lin et al. [18] construct mosaics of UAV images based on two methods: registration between image pairs and registration between images and reference map. The recent work of Marburg and Hayes [19] decomposes the homography onto a base image to obtain

a similarity and a fronto-parallel projection which is then used to warp the images. Error accumulation is controlled by a global refinement procedure. A graph based on image overlaps is created whose structure is exploited to make the method efficient. The joint optimization step limits the application of the method to a large number of images (Marburg and Heyes report 144 images taken from a high altitude in [19]). It is not clear how the fronto-parallel decomposition relates to the ground plane. As we show in Section III-B, homography estimation errors can adversely affect ground plane estimation.

Our method differentiates itself from the existing body of work in a number of significant ways. It can accommodate a large number of images because the joint optimization required for estimating the world structure in a robust fashion is performed on a very small fraction of the images (about 2% in some cases). However, all images in the input sequence are utilized. We present a novel method to choose the best image for each ground patch. Experiment results also indicate that view selection reduces misalignment and improves resolution. Explicitly reconstructing the ground geometry allows us to find a canonical rectification for each image. The method can be easily extended for the case when the ground is composed of multiple planes. We demonstrate the results in challenging data sets obtained by off the shelf aerial vehicles flying over orchards and corn-fields at low altitudes and/or performing complex motions.

III. GENERALIZED MOSAIC METHOD FOR AGRICULTURE ENVIRONMENT

Our algorithm, whose input is an image sequence $\mathcal{I}$, consists of six primary steps shown in Figure 3. In the first two steps, a small portion $\mathcal{S} \subseteq \mathcal{I}$ is chosen and input to a standard SfM algorithm to obtain a sparse ground reconstruction and camera poses. In the third step, the ground plane $\mathcal{P}$ is estimated using a robust plane fitting algorithm. At this stage it is possible to determine if the ground structure contains multiple planes. However, for our data sets it sufficed to fit the reconstruction to a single plane. In the next step, the algorithm rectifies each image used in $\mathcal{S}$. This step uses the ground plane $\mathcal{P}$ and the camera poses for each image and computes a homography which corresponds to rotating each camera pose in place so as to make the image plane parallel to the ground. At this stage, the size of the final mosaic is determined. In the last two steps, the algorithm considers all input images. For every ground patch in the final mosaic, it chooses the best view which can see this patch and has the highest resolution. It then warps this image onto the ground patch so as to create a high resolution mosaic. The details are presented in the succeeding sections.

A. Estimating Ground Geometry

Let $\mathcal{I} = I_1, \dots, I_n$ be the input sequence. We start by choosing a small subset $\mathcal{S} \subseteq \mathcal{I}$ with sufficient overlap. For example, suppose the UAV is moving at 1 m/s and recording images at 30 fps. Further suppose that due to the low altitude flight, the image footprint is only 10 meters. When the UAV is flying straight, it would move 2 meters

Fig. 3: Method overview: $\mathcal{I}$: original image sequence, $\mathcal{S}$: a subsequence (Section III-A), $\mathbf{K}$: intrinsics, $\mathcal{P}$: world plane. $\mathbf{H}$: rectification homography

in 2 seconds and therefore $\frac{4}{5}$ of image I_j would overlap with image $I_{j+2 \times 30 - 1} = I_{j+59}$. Therefore, keeping around 2% of $\mathcal{I}$ by picking every 50th frame would safely yield sufficient overlap. The images in $\mathcal{S}$ are input to a robust Structure from Motion (SfM) pipeline based on Bundle Adjustment [20] to get globally optimal camera poses as well as 3D locations of robust features such as SIFT [21]. We used the Visual SfM [22], [23] tool. At this stage, camera intrinsics including distortion parameters can be estimated if not already available. The output of this stage for various input sequences are shown in Figures 1c and 2c.

At this stage, there are two possible extensions: (1) If the scene is truly planar, some SfM methods can fail. To remedy this, automatic model selection methods, such as the one used in [24], can be used. (2) If the speed or pose of the UAV is highly varying, the output of SfM can be investigated to adjust the selection of $\mathcal{S}$ by adaptively choosing key frames in $\mathcal{I}$ based on the amount of overlap for image pairs similar to the keyframe method presented in [25]. We have not explored these extensions further in this paper since the basic (uniform) selection methods turned out to be sufficient for our datasets.

B. Ground Plane Estimation

When flying at low altitudes, the planar scene assumption can fail locally. See Figure 4. Further, objects such as trees which cause height variations are good sources of robust features and therefore they are frequently picked up for homography estimation. As a result, homographies obtained directly from the images can be inconsistent across the sequence. In particular, when the normal vector of the ground plane is directly estimated from a single homography, if the deviation from the true ground plane is large, it can cause significant perspective distortion in the mosaic. This is evident in the orchard data set. The output of step 2 shown in Figure 4 is obtained from 230 images taken by a UAV flying at about 10 meters.

Our method avoids such errors for two reasons: (i) In step 2, the world geometry is estimated without making any planarity assumptions, and (ii) in step 3, the ground plane is directly and *globally* estimated using a robust plane fitting algorithm. Specifically, we first use RANSAC to fit a plane to the 3D feature points and perform Principal Component Analysis (PCA) on the inliers. The first two major eigenvectors

and the mean provide the world reference plane. Features corresponding to trees etc. above the ground are removed as outliers.

Fig. 4: Top: Height variation due to trees. Bottom: The angle between the ground normal and camera optical axis with respect to each camera. Red dashed line with circles indicates ground normal estimated from inter-image homography; blue solid line indicates the ground normal estimated by fitting a plane after SfM.

To further quantify the effect of height variation, we first computed the homography between consecutive images and estimated the poses and the ground normal vector directly from the homography. The derivation can be found in [26]. The angle between camera optical axis and estimated ground normal is plotted as the dashed red line in Figure 4-bottom across the sequence. We also computed the ground plane $\mathcal{P}$ using our method and computed the difference between the camera optical axis (from SfM) and the plane normal. This is plotted as the solid blue line in the figure. The average difference between the two methods is 10.097 degrees with standard deviation of 9.470 degrees. This gap introduces a large variance for the plane normals estimated purely from image homographies.

C. Metric Rectification

At this step we have the world coordinate frame $\mathbf{w}$, given by the ground plane $\mathcal{P}$ estimated in step 3, and camera poses for each image in the subsequence $\mathcal{S}$. We use the notation ${}^{c_i}\mathbf{T}_{\mathbf{w}}$ and ${}^{\mathbf{w}}\mathbf{R}_{c_i}$ to denote the translation and rotation from the camera frame c_i to the world frame $\mathbf{w}$.

Warping an image directly to another image can cause significant perspective distortion, especially if the camera motion includes rotations. The ground plane $\mathcal{P}$ computed in

step 3 provides a canonical frame to warp the images. When images are rectified with respect to $\mathcal{P}$ perspective distortion and misalignment is significantly reduced (Figures 8f and 9f).

We now present the details of the rectification procedure. Let t be the index of the camera to be rectified. Let ${}^{\mathbf{w}}\mathbf{N}$ be the unit normal vector of $\mathcal{P}$. In camera coordinates, this vector is expressed as

$${}^{c_t}\mathbf{N} = {}^{c_t}\mathbf{R} {}^{\mathbf{w}}\mathbf{N} \quad (1)$$

We compute the rotation which aligns the optical axis of the camera with ${}^{c_t}\mathbf{N}$. This is given by a rotation around the axis $\hat{\mathbf{k}} = {}^{c_t}\mathbf{Z} \times {}^{c_t}\mathbf{N}$ by an angle of $\theta = \arccos(\frac{{}^{c_t}\mathbf{Z} \cdot {}^{c_t}\mathbf{N}}{\|{}^{c_t}\mathbf{Z}\| \|{}^{c_t}\mathbf{N}\|})$. The rotation matrix can be obtained by a standard angle-axis to matrix conversion procedure.

D. Homography Based Ground Plane Warping

The homography which warps image i to the mosaic plane $\mathcal{P}'$ that is parallel to the ground plane $\mathcal{P}$ is given by:

$${}^{\mathcal{P}'}\mathbf{p} = \mathbf{K} \underbrace{({}^{\mathcal{P}'}\mathbf{R} + {}^{\mathcal{P}'}\mathbf{T}_{c_i} \frac{{}^{c_i}\mathbf{N}^T}{{}^{c_i}d})}_{\mathbf{H}_{c_i-\mathcal{P}'}} \mathbf{K}^{-1} c_i \mathbf{p} \quad (2)$$

where $\mathbf{K}$ is the camera calibration matrix; ${}^{\mathcal{P}'}\mathbf{R}$ and ${}^{\mathcal{P}'}\mathbf{T}_{c_i}$ are respectively the orientation and translation of camera c_i with respect to $\mathcal{P}'$; ${}^{c_i}\mathbf{N}$ is unit normal vector with respect to camera c_i ; ${}^{c_i}d$ is distance between optical center of camera c_i and ground plane with respect to world coordinate; $\mathbf{H}_{c_i-\mathcal{P}'}$ is the homography from camera c_i to $\mathcal{P}'$. ${}^{\mathcal{P}'}\mathbf{p}$ and $c_i \mathbf{p}$ are position of feature points in image coordinates with respect to mosaic plane and camera c_i . By rectifying each image in $\mathcal{S}$, we can determine the size of final mosaic.

E. Automatic View Selection

So far, we have only used the images in the subsequence $\mathcal{S}$. We now turn our attention to the larger input sequence. Let c_i be the pose of an arbitrary image in $\mathcal{I} \setminus \mathcal{S}$ and let c_j be the image with the nearest index in $\mathcal{S}$ which has been computed in Step 2. We first compute the camera pose of each such c_i by computing the essential matrix between c_i and c_j and then extracting the relative pose of c_i with respect to c_j . Since the pose of c_j has already been computed with respect to the world frame (in a robust fashion), we can easily compute the pose of c_i with respect to the world frame.

When generating image mosaics, usually the images in the data set are warped to mosaic in an arbitrary or sequential order. When a scene point (or patch) is visible from multiple viewpoints, the selection of the correct image to represent this area is crucial to maximize resolution and to minimize distortion. This is especially true for cameras with large lens distortion or UAV motions including tilting of the camera and motion at different altitudes. Such images can induce visible misalignment and poor resolution in the final mosaic.

We now present the view-selection step of our algorithm.

We first partition the mosaic (whose size was determined in step 4) into square blocks of size $n \times n$. In theory, the block size can be chosen arbitrarily small. However this would adversely affect the running time of the algorithm. In practice,

its choice depends on the size of the mosaic and individual camera poses. For the orchard data set, to generate a mosaic of size 2500×3000 pixels, we used a block size of 100×100 . When projected to the original images, the corresponding window area ranged from 6% to 21% of the image area. For each block b , let $T(b)$ be the list of images that can see it entirely.

In order to pick the best image for block b , our algorithm chooses the best view using a novel metric given by the number of pixels dedicated to b in an image normalized by the distortion in these pixels. The quality measure assigned to image $I \in T(b)$ for block b is given by

$$Q_b(I) = \sum_{p \in I} w_p \rho_p \quad (3)$$

where the summation is over the pixel indices $\mathbf{p} = [x \ y]^T$ and ρ_p is an indicator variable which is 1 if and only if $\mathbf{p}$ is the projection of a point in block b . The distortion weight w_p is computed using a distortion model. We use the following standard model:

$$x_{\text{offset}} = f_x(x_n(1 + k_1 r^2 + k_2 r^4 + k_3 r^6)) \quad (4)$$

$$y_{\text{offset}} = f_y(y_n(1 + k_1 r^2 + k_2 r^4 + k_3 r^6)) \quad (5)$$

where x_{offset} and y_{offset} are the displacements in x and y directions measured in pixels. x_n and y_n are the coordinates of the pixel $\mathbf{p}$ in normalized image plane before distortion correction and r is distance between the pixel and principal point in normalized image plane before correction.

Based on the lens distortion model, we generate a weight kernel whose size is the same as original image size and for each pixel $\mathbf{p}$ it is defined as below:

$$w(\mathbf{p}) = \exp\left(-\frac{d(\mathbf{p})}{d_{\text{max}}}\right) \quad (6)$$

where $w(\mathbf{p})$ is weight for each pixel at position $\mathbf{p}$ and $d(\mathbf{p}) = \sqrt{x_{\text{offset}}^2 + y_{\text{offset}}^2}$; d_{max} is the largest corrected displacement over all the pixels in image. See Figure 5.

Fig. 5: The image kernel based on standard lens distortion model. Darker colors indicate higher distortion.

The proposed view selection algorithm is given in Algorithm 1.

IV. RESULTS

In this section, we present two sets of results to evaluate our approach. First, we compare our method with a standard

Input : block $b = 1 \sim N$ in mosaic, image sequence $j=1 \sim M$, homography from mosaic plane to the image $\mathbf{H}_{\mathcal{P}'-c_j}$, image distortion coefficients k_1, k_2, k_3 , block visible vector $T(b)$

Output: Mosaic (with best view selection)

```

for  $b \leftarrow 1$  to  $N$  do
  for  $j \leftarrow 1$  to  $M$  do
    if projection of 4 corners of block  $b$  are all
      within image  $j$  then
      | add image index  $j$  to  $T(b)$ 
    end
  end
end
for  $b \leftarrow 1$  to  $N$  do
   $\text{maxQ} = -1$ ;
  if  $T(b)$  is empty then
  | continue
  end
  for camera  $j \in T(b)$  do
    Warp pixels within block  $b$  to image  $j$  using
       $\mathbf{H}_{\mathcal{P}'-c_j}$ 
    Compute  $Q_b(I) = \sum_{p \in I} w_p \rho_p$  (Equation 3)
    if  $Q_b(I) > \text{maxQ}$  then
      |  $\text{maxQ} = Q_b(I)$ 
      | OptimalIndex =  $j$ 
    end
  end
  Warp the portion of image with index OptimalIndex
  that belongs to block  $i$  to the corresponding block  $i$ 
  in the final mosaic
end

```

Algorithm 1: The view selection algorithm

stitching approach and present quantitative results on reprojection error. Next, we compare our technique with commercially available software.

A. Quantitative Evaluation

We compare our algorithm with a “baseline” approach which, given an input sequence, chooses the middle frame I_m as reference and warps all frames onto I_m using inter homography derived from SIFT correspondences. This method is compared with the output of our approach before and after the view selection steps.

We use two image sequences we refer to as “orchard” and “corn field.” The orchard dataset contains images of UMN Horticulture Field Station taken from a UAV flying at around 13 m altitude and at approximately constant speed of 1m/s (Figure 1c). The area covered is about 2200 m². The camera mounted on the UAV was GoPro Hero 3 whose intrinsics are available on the accompanying website. The video was recorded as an MP4 file at 29 fps at 1920×1080 resolution. We picked every fifth frame from the video to generate an image sequence with 3500 images. In the corn field dataset, the images are collected from a UAV exhibiting a complex motion: First, the UAV flies at low altitude (around 5 m) and then it lifts up to about 25 m (Figure 2c). In this dataset, the high flying speed causes motion blur in the Middle section

of the video. We also picked every fifth frame to generate an image sequence with 540 images. Camera parameters were unknown.

To evaluate the algorithms, we generate the histogram of reprojection error using the same set of SIFT features for 3 different methods. For all methods, we first project each inlier 3D point $\mathbf{P}$ from SfM (i.e. the point above the ground and within final mosaic) onto ground plane $\mathcal{P}$ to obtain the 2D feature $\mathbf{P}'$. For the baseline method, we first transform $\mathbf{P}'$ to the coordinate of reference mosaic I_m with the same scale based on camera poses. Then, we project each feature in I_m back to the original image using homography computed from baseline method and evaluate root mean squared reprojection error (RMSRE) for each $\mathbf{P}$ given by: $\sqrt{\frac{1}{N} \sum_{i=1}^N \|T^i(\mathbf{P}) - \mathbf{p}_i\|^2}$, where T^i is the transformation from 3D space to the original image i of the feature $\mathbf{P}$.

For our method without view selection, after obtaining $\mathbf{P}'$, we project the point $\mathbf{P}'$ from $\mathcal{P}$ back to all the images that include $\mathbf{P}'$ in the sequential warping sequence using the camera poses from SfM. Finally, we compute RMSRE for each $\mathbf{P}$ using the original position of each feature in each image. For our method with view selection, we project every feature $\mathbf{P}$ back to only the image chosen after best view selection.

We use a total of 258182 inlier SIFT features to compute reprojection error in orchard dataset and totally 27277 inlier SIFT features in corn field dataset. Figure 6 and Figure 7 show reprojection error histograms for these methods. The results indicate that (1) our method improves over the baseline method significantly and (2) view selection reduces reprojection error.

B. Qualitative Evaluation

In this section, we compare our algorithm with view selection against Auto-Stitch [27], Autopano [28], Microsoft Image Composite Editor (ICE) [29], the baseline method and our method before view selection.

Figure 8 shows the comparison of the mosaic result on the orchard data for these six methods. Out of the 3500 images in the sequence, 350 images (corresponding to an interval of 50 frames for the original video) were input to the Autopano, Auto-Stitch, Microsoft ICE software. For our algorithm, we choose the same 350 images as input for the SfM, the results without view selection (i.e. sequential warping of the 350 images) and the baseline method. The final result of our algorithm uses additional 3150 images in the final view selection step. In Figures 8a, 8b and 8c, we can see that the mosaic has serious misalignment; figure 8d and figure 8e generated by Microsoft ICE and sequential warping from SfM also induce visible misalignment and skewness. Figure 8f generated by our method with best view selection removes misalignment and skewness.

Figure 9 shows the results for the corn field dataset using the same 6 methods. Out of the 540 images, 90 images (corresponding to 30-frame intervals for the original video) are input to Autopano, Auto-Stitch and Microsoft ICE and baseline methods. Our method uses the same set of 90 images for SfM and to generate the mosaic result without view selection. We generate the result with view selection using

additional 450 images. From figure 9a, 9b and 9c, it can be seen that the corn rows in mosaic suffer from serious perspective distortion. Figure 9d generated by Microsoft ICE has visible misalignment due to scale change of the video. Microsoft ICE automatically discards several images because it fails to align them with the final mosaic. In addition, there is motion blur in the middle of the mosaic for the first 5 mosaics. Figure 9f generated by our method with the best view selection eliminates motion blur and perspective distortion, creating the final mosaic with the highest resolution.

Fig. 6: Histogram of reprojection error (RE) on orchard dataset. Top: RE for baseline method. Middle: RE for our method without view selection. Bottom: RE for our method with view selection

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel approach which can generate high-quality mosaics even when the images are obtained by a UAV flying at low altitudes and performing complex motions. The two key features of our approach are the explicit construction of the ground geometry using a small fraction of input images, and a novel view selection routine to obtain the highest resolution views of scene patches. Experiment results clearly indicate that the approach can handle a large number of images and generate mosaics with reduced misalignment and improved resolution.

We are releasing our data-sets and code including SfM results and camera calibration matrices. These are posted at http://rsn.cs.umn.edu/index.php/Image_Mosaic_Generation.

Our method can be extended to generate mosaics even when the ground is composed of multiple planes. We plan to further explore this feature in our immediate future work. Another venue for future research is to extend the view

Fig. 7: Histogram of reprojection error (RE) on corn field dataset. Top: RE for baseline method. Middle: RE for our method without view selection. Bottom: RE for our method with view selection

selection algorithm to include additional criteria such as image sharpness.

REFERENCES

- [1] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.
- [2] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.
- [3] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2006.
- [4] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.
- [5] M. Brown and D. Lowe, "Recognising panoramas," in *Proceedings of the 9th International Conference on Computer Vision*, vol. 2, Nice, October 2003, pp. 1218–1225.
- [6] N. Gracías and J. Santos-Victor, "Underwater video mosaics as visual navigation maps," *Computer Vision and Image Understanding*, vol. 79, no. 1, pp. 66–91, 2000.
- [7] —, "Underwater mosaicing and trajectory reconstruction using global alignment," in *OCEANS, 2001. MTS/IEEE Conference and Exhibition*, vol. 4. IEEE, 2001, pp. 2557–2563.
- [8] N. Gracías, J. P. Costeira, and J. Victor, "Linear global mosaics for underwater surveying," in *5th IFAC Symposium on Intelligent Autonomous Vehicles*, vol. 1, 2004.
- [9] A. Elibol, R. García, O. Delaunoy, and N. Gracías, "A new global alignment method for feature based image mosaicing," in *Advances in Visual Computing*. Springer, 2008, pp. 257–266.
- [10] A. Elibol, R. García, and N. Gracías, "A new global alignment approach for underwater optical mapping," *Ocean Engineering*, vol. 38, no. 10, pp. 1207–1219, 2011.
- [11] N. Gracías and S. Negahdaripour, "Underwater mosaic creation using video sequences from different altitudes," in *OCEANS, 2005. Proceedings of MTS/IEEE*. IEEE, 2005, pp. 1295–1300.
- [12] J. Ferrer, A. Elibol, O. Delaunoy, N. Gracías, and R. García, "Large-area photo-mosaics using global alignment and navigation data," in *MTS/IEEE OCEANS Conference, Vancouver, Canada, 2007*, pp. 1–9.
- [13] E.-Y. Kang, I. Cohen, and G. Medioni, "A graph-based global registration for 2d mosaics," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 1. IEEE, 2000, pp. 257–260.
- [14] A. Fusiello, M. Aprile, R. Marzotto, and V. Murino, "Mosaic of a video shot with multiple moving objects," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 2. IEEE, 2003, pp. II–307.
- [15] M. Pfingsthorn, A. Birk, S. Schwertfeger, H. Bülow, and K. Pathak, "Maximum likelihood mapping with spectral image registration," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4282–4287.
- [16] B. Steder, G. Grisetti, C. Stachniss, and W. Burgard, "Visual slam for flying vehicles," *Robotics, IEEE Transactions on*, vol. 24, no. 5, pp. 1088–1093, 2008.
- [17] T. Botterill, S. Mills, and R. Green, "Real-time aerial image mosaicing," in *Image and Vision Computing New Zealand (IVCNZ), 2010 25th International Conference of*. IEEE, 2010, pp. 1–8.
- [18] Y. Lin and G. Medioni, "Map-enhanced uav image sequence registration and synchronization of multiple image sequences," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–7.
- [19] A. Marburg and M. P. Hayes, "Smartpig: Simultaneous mosaicking and resectioning through planar image graphs," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5767–5774.
- [20] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment modern synthesis," in *Vision algorithms: theory and practice*. Springer, 2000, pp. 298–372.
- [21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [22] P. Moulon, P. Monasse, and R. Marlet, "Global fusion of relative motions for robust, accurate and scalable structure from motion," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3248–3255.
- [23] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3057–3064.
- [24] A. Kim and R. Eustice, "Pose-graph visual slam with geometric model selection for autonomous underwater ship hull inspection," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 1559–1565.
- [25] D. Steedly, C. Pal, and R. Szeliski, "Efficiently registering video into panoramic mosaics," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1300–1307.
- [26] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An invitation to 3-d vision: from images to geometric models*. Springer Science & Business Media, 2012, vol. 26.
- [27] "Autostitch," <http://matthewalunbrown.com/autostitch/autostitch.html>.
- [28] "Kolor autopano," <http://www.kolor.com/autopano/>.
- [29] "Microsoft research image composite editor," <http://research.microsoft.com/en-us/um/redmond/projects/ice/>.

(a) Mosaic using baseline method

(b) Mosaic using Auto-Stitch software

(c) Mosaic using Autopano software

(d) Mosaic using Microsoft Image Composite Editor (ICE)

(e) Mosaic using SfM without view selection algorithm

(f) Mosaic using SfM with view selection algorithm

Fig. 8: Mosaic results (orchard dataset)

(a) Mosaic using baseline method

(b) Mosaic using Auto-Stitch software

(c) Mosaic using Autopano software

(d) Mosaic using Microsoft Image Composite Editor (ICE)

(e) Mosaic using SfM without view selection algorithm

(f) Mosaic using SfM with view selection algorithm

Fig. 9: Mosaic results (corn field dataset)