High-resolution Ecosystem Mapping in Repetitive Environments Using Dual Camera SLAM

Brian M. Hopkinson
Department of Marine Sciences, University of Georgia
Athens, Georgia 30602, USA
bmhopkin@uga.edu

Suchendra M. Bhandarkar
Department of Computer Science, University of Georgia
Athens, Georgia 30602, USA
suchi@uga.edu

Abstract—**Structure from Motion (SfM) techniques are increasingly being used to create 3D maps from images in many domains including environmental monitoring. However, SfM techniques are often confounded in visually repetitive environments as they rely primarily on globally distinct image features. Simultaneous Localization and Mapping (SLAM) techniques offer a potential solution in visually repetitive environments since they use local feature matching; however, SLAM approaches work best with wide-angle cameras that are often unsuitable for documenting the environmental system of interest. We resolve this issue by proposing a dual-camera SLAM approach that uses a forward facing wide-angle camera for localization and a downward facing narrower-angle, high-resolution camera for documentation. Video frames acquired by the forward facing camera are processed using a standard SLAM approach providing a trajectory of the imaging system through the environment which is then used to guide registration of the documentation camera images. Fragmentary maps, initially produced from the documentation camera images via monocular SLAM, are subsequently scaled and aligned with the localization camera trajectory and finally processed using a global optimization procedure to produce a unified, refined map. An experimental comparison with several state-of-the-art SfM approaches shows the dual-camera SLAM approach to perform better in repetitive environmental systems based on select samples of ground control point markers.**

I. Introduction

The spatial arrangement of organisms within an ecosystem reflects fundamental underlying ecological processes such as competition, resource availability, trophic relationships, and symbioses [1], [2]. Consequently, the ability to map the abundance and distribution of organisms within an ecosystem is critical for advancing ecology. Traditionally, mapping organismal distributions entailed time-consuming, manual field work, thereby limiting the scale and frequency at which maps could be generated. In recent decades, satellite remote sensing has offered unprecedented insight into the spatial arrangement and coverage of various ecosystems, but because of the coarse resolution of satellite imaging ($\approx$1–30 m pixel size) it is typically only useful at the ecosystem level (e.g. distribution of forest vs. grassland) and cannot assess the distribution of individual species [3]. Recent advances in computer vision algorithms, most notably in Structure-from-Motion (SfM) [4], [5], [6], have begun to see their application in ecology for construction of much higher resolution (mm to cm) 3D optical maps of ecosystems. When combined with machine learning tools for automated classification, these maps are capable of delineating the distribution of individual species across landscape scales [7], [8].

SfM has been the tool of choice for map generation from images for ecologists and geographers due to its ability to use structured or unstructured image collections and the availability of high-quality commercial and open-source implementations [9], [10]. SfM relies on *globally* distinct visual features in images to register overlapping images for map generation. This requirement is typically met when images are acquired at high altitude via unmanned aerial vehicles (UAVs or drones). However, when ecosystem images are acquired closer to the scene, for example to resolve small plants or animals, they often become repetitive causing conventional SfM approaches to fail. In contrast, Simultaneous Localization and Mapping (SLAM) approaches developed in the robotics community process images sequentially in the order they are acquired using *locally* distinct visual features to map the environment and determine the camera pose [11], [12]. SLAM approaches, though promising for mapping repetitive scenes [13], work best with high-frame rate or high-speed (and consequently lower resolution), wide-angle cameras whose images are of limited use for identification of organisms [14].

In this paper, we describe a dual-camera SLAM approach to map visually repetitive environments such as grasslands, shrublands, or agricultural fields (Fig. 1). A high-speed, wide-angle camera is used for conventional visual SLAM-based localization whereas the other high-resolution, medium- to narrow-angle (video or still image) camera is used to acquire high-quality ecosystem images suitable for "documentation", i.e., identification and localization of organisms to the species or genus level. The documentation camera does not need to be tightly integrated at the hardware level with the SLAM camera, allowing the use of extremely high-quality and low-cost commercial off-the-shelf (COTS) digital cameras. The trajectory of the localization camera is used to guide detailed map generation from the documentation camera images using the proposed dual-camera SLAM approach.

The primary contributions of this work are: (a) a novel approach to ecosystem map generation that allows flexible use of high-resolution, medium- to narrow-angle COTS digital cameras to resolve smaller organisms by decoupling localization and documentation; (b) development of a multistage alignment process for the documentation camera images that

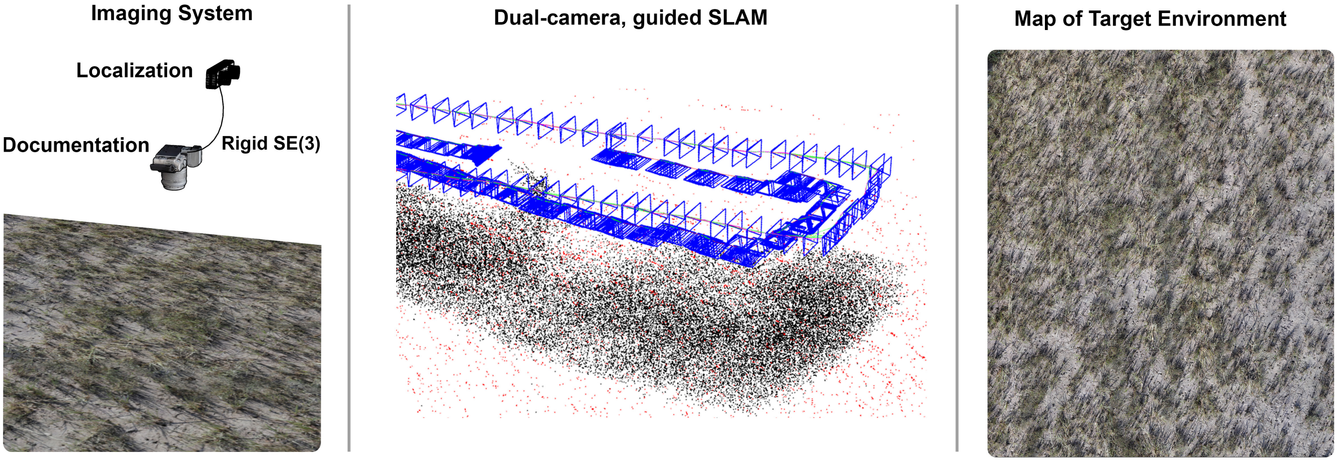

Fig. 1. **System Overview:** The dual-camera SLAM system consists of a forward facing stereo-camera for localization and a downward facing, high-resolution camera for documentation. Trajectories (magenta line), landmark maps (black and red dots), and image poses (blue frames) are generated for each camera. The image poses and landmark map from the documentation camera are used to generate a map of the targeted, visually repetitive salt marsh environment.

uses the localization camera to guide image pose determination and map point positioning; and (c) experimental demonstration of the proposed system's ability to map visually repetitive environments.

II. Related Work

A. Ecosystem Mapping

Although mapping the distribution of ecosystems (e.g. forest, coral, and grassland extent) from satellite imagery has a long history [3], [15], we confine our overview to methods that exploit higher-resolution imagery (mm-cm pixel size) enabling more fine-grained taxonomic resolution (species, genera) since they are more closely related to our work. The most common ecosystem mapping workflow is comprised of image acquisition from a structured aerial survey using a UAV at ≈10–250 m altitude followed by image positioning and map generation via SfM. This approach provides ground resolution of ≈1 cm, suitable for classifying and delineating moderate- to large-sized organisms.

Hayes et al. [16] use SfM to construct orthomosaics of seabird colonies from UAV images acquired at 60–90 m altitude followed by CNN-based object detectors to count individual birds for population tracking. Baena et al. [17] map the distribution of the keystone Algarrobo tree in Pacific Equatorial dry forests using SfM procedures on large-scale 260 m altitude UAV imagery. However, in low-visibility underwater environments, images are typically acquired closer to the scene (several meters) either manually or using underwater vehicles and then processed via SfM [18] or customized approaches [19].

B. Structure from Motion (SfM)

SfM approaches to scene reconstruction (i.e., mapping) and image pose determination are currently the primary tool for image-based ecosystem mapping. SfM attempts to map a scene from unordered images from uncalibrated and possibly multiple cameras, imposing minimal constraints on image acquisition [6], [9]. The relative pose between images is computed by extracting features and associated descriptors (e.g. SIFT) from the unordered image collection and matching these features between images, using geometric verification to remove outliers [20]. Matched feature points are triangulated to generate a sparse representation of the scene. The image poses, scene points, and camera calibration parameters are optimized via a bundle adjustment procedure [21]. Further, this sparse scene representation can be made dense using multi-view stereo methods [22], and/or converted to a triangular mesh representation.

SfM is extremely flexible and relatively easy for non-specialists to use since it imposes minimal constraints on the image acquisition process, but the allowance for unordered image sets makes SfM computationally expensive with a super-linear time complexity [6], [23]. Furthermore, because images are unordered they must be visually similar only to their true neighbors; otherwise matches between spatially disparate locations may be incorrectly accepted. Consequently, SfM is often confounded in scenes with repetitive features when this requirement is violated resulting in *visual aliasing* [24].

C. Simultaneous Localization and Mapping (SLAM)

SLAM techniques typically provide a similar final mapping solution to SfM, i.e., a sparse or dense representation of the scene and image poses [12]. However, SLAM assumes that the images are acquired sequentially, reflecting its origins in robotics. For ecosystem mapping, the sequential image acquisition constraint is generally not onerous as images are typically acquired from a single camera as it is moved over the underlying ecosystem. Since images are processed sequentially, the features need be only locally distinct, thus making SLAM better suited to handle repetitive environments. However, SLAM works best with a wide-angle, high frame rate camera whose limited image resolution is typically impractical for organismal identification [14]. This motivates our incorporation of an additional, higher-resolution camera for ecosystem documentation. Although several multi-camera SLAM systems have been developed for robotics, they assume that precise synchronization information, in the form of exact timing and precise orientation, is available for each camera [25], [26], [27], thereby preventing use of most inexpensive, high quality COTS cameras, which do not expose synchronization signals. By relaxing the precise synchronization requirement, we allow use of COTS cameras, but do not incorporate documentation camera images into the estimation of the system pose.

III. Proposed System

The proposed dual-camera SLAM-based ecosystem mapping approach uses SLAM to determine the trajectory of the localization camera, which is then used to guide map generation from the documentation camera images (Fig. 2). The system assumes the relative orientation of the two cameras is constant and approximately known and that the image streams are roughly ($<$ 0.5 s) synchronized. As SLAM is applied to the localization camera images, the documentation camera images are processed concurrently via monocular SLAM using the localization trajectory to guide generation of an initial ecosystem map and approximate image poses. However, the initial ecosystem map is generally fragmentary as tracking is frequently lost due to rapid scene movement resulting from the narrow field of view (FOV) and limited number of trackable features in the documentation camera images. After completion of SLAM processing of the localization camera image sequence, the fragmentary maps from the documentation camera are scaled and transformed to approximately align with the localization camera trajectory based on the acquisition time of documentation camera images and the pose of the documentation camera relative to the localization camera. Finally, the documentation camera poses and associated map are optimized based on constraints derived from the localization camera trajectory and landmark-to-camera correspondences in the fragmentary maps using a factor graph framework. Our strategy of composing a global map from fragments is similar to many previous approaches [28], [29], [30], but the method described here makes use of a secondary camera trajectory and allows for loose temporal coupling between the two camera systems.

A. SLAM System Core

The proposed system employs a modified version of ORB-SLAM2 [31], a keyframe-based SLAM approach, as its base since it is a comprehensive SLAM approach capable of using monocular, stereo, and RGB-D cameras and produces accurate maps through multiple rounds of optimization (i.e., bundle adjustment). ORB-SLAM2 performs the following main operations: *tracking*, which localizes each frame relative to the existing map and determines when new keyframes should be added, *mapping*, which maintains the current map and updates it via insertion of new keyframes, triangulation of new map points, and optimization of the map via bundle adjustment, and *loop closing*, which identifies revisited locations (i.e., trajectory loops) and revises the map accordingly.

We made several notable modifications to the ORB-SLAM2 system. First, the system was extended to handle multiple cameras simultaneously, with each camera maintaining a separate master map. The map data structure was converted into a recursive tree structure to handle sub-maps that are generated when tracking is lost. Sub-maps can be kept private or optionally registered with their parent to make keyframes and map points accessible to the parent. The *relocalization* state, which ORB-SLAM2 enters immediately upon loss of tracking, was eliminated; instead a new sub-map is spawned and the SLAM procedure reinitialized upon loss of tracking. For the localization camera, the new sub-map is registered with the parent assuming the camera maintained a constant velocity between the time when tracking was lost and when a new map was successfully initialized. Finally, per-frame camera trajectories are explicitly recorded as $SE(3)$ transformations relative to reference keyframes, whose positions are continuously updated via optimization. The availability of these trajectories is critical for positioning of the documentation camera images based on the localization camera poses. Our SLAM code is available online at https://github.com/bmhopkinson/hyslam.

B. Generation of the Fragmentary Maps

The first step in generating the ecosystem map and associated documentation camera/image poses is the application of the modified monocular ORB-SLAM2 procedure to the sequentially acquired documentation images. A new map is initialized by tracking ORB feature points [32] through multiple frames until there is sufficient parallax to accurately triangulate map points corresponding to the tracked features. In our system, the resulting map of arbitrary scale is brought into an approximately consistent scale with the localization camera map by estimating the absolute motion of the documentation camera over the initialization period using the motion of the localization camera. The pose of the documentation camera can be estimated at any time t as:

$$\mathbf{X}_d(t) = \mathbf{X}_l(t)\mathbf{T}_{dl} \tag{1}$$

where $\mathbf{X}_d(t)$ is the pose of the documentation camera at time t, $\mathbf{X}_l(t)$ is the pose of the localization camera at time t, and $\mathbf{T}_{dl}$ is the rigid-body $SE(3)$ transformation between the localization and documentation cameras. All poses are expressed in the camera-to-world transformation convention. The motion of the documentation camera over the initialization period can then be determined as:

$$\mathbf{V}_{d0}(t) = \mathbf{X}_{d0}^{-1}\mathbf{X}_{d1} \tag{2}$$

The new map is scaled using the estimated motion ($\mathbf{V}_{d0}$), which works well in most cases but can occasionally be inaccurate as a result of small absolute distances traveled during monocular SLAM initialization. The fragmentary maps are later rescaled using the full distance travelled during their generation providing a consistent, absolute scale among the fragmentary maps.

Once the map is initialized, the documentation images are processed using standard monocular SLAM [31], [33]

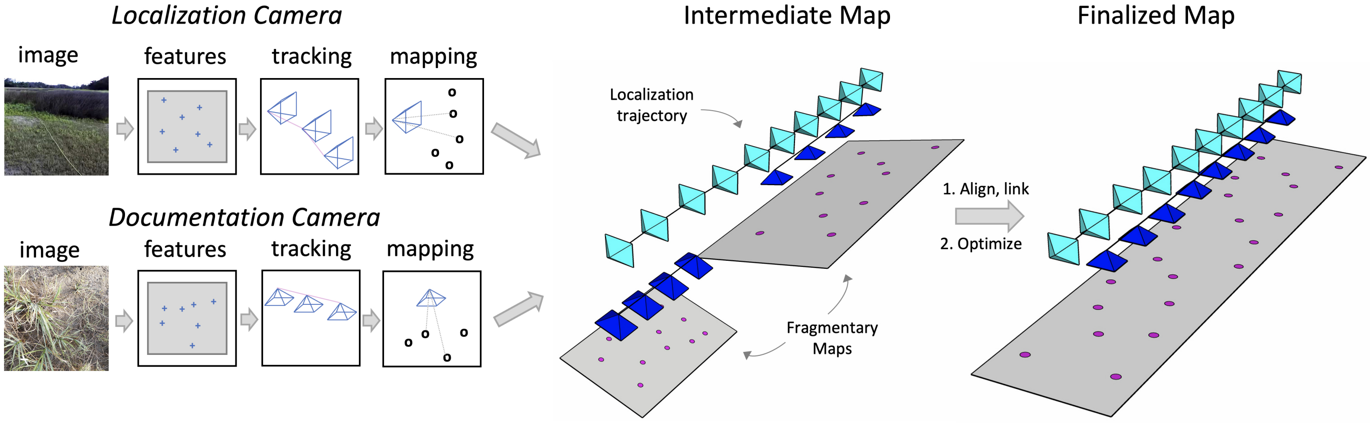

Fig. 2. **Dual Camera SLAM Procedure:** Images from the localization and documentation cameras are processed in parallel starting with feature extraction, followed by initial pose estimation through landmark tracking, and finally optimization of the pose and generation of new landmarks in the mapping phase. Processing of the entire video streams results in an intermediate map consisting of a unified map for the localization camera and fragmentary maps from the documentation camera. The localization camera's trajectory is used to align the fragmentary maps, which are then linked through commonly viewed landmarks. Finally, all poses and landmarks are globally optimized to produce a finalized map from the documentation camera.

which provides a convenient and robust way to determine the approximate relative documentation camera poses between the documentation images and identify features that are consistently matched and geometrically verified. Tracking loss, which occurs when the number of tracked map points drops below a predefined threshold, occurs regularly due to the narrow FOV and rapid relative motion of the downward facing documentation camera at which point a new fragmentary map is initialized.

C. Alignment of Fragmentary Maps

The relevant outputs of the initial processing steps consist of the full localization camera trajectory and multiple fragmentary maps from the documentation camera as depicted in the *Intermediate Map* in Fig. 2. The fragmentary maps are already approximately scaled to the localization trajectory but their orientations and positions may diverge substantially from their true values. Although a full non-linear optimization procedure is ultimately employed to provide the best estimate for the documentation camera map, proper initialization is necessary for convergence. Since the documentation camera fragmentary maps often deviate substantially from their correct configurations, a two-step procedure is used to approximately align the fragmentary maps with the localization camera trajectory. First, the camera centers for the documentation images in each fragmentary map are brought into alignment with their expected positions based on the localization camera trajectory using a $Sim(3)$ transformation estimated using Horn's method [34]. When the path traveled within any fragmentary map is approximately linear, which is often the case, there is inherent rotational ambiguity in the documentation camera pose when the alignment is performed based solely on the documentation camera center positions. To resolve this ambiguity, an optimal $SO(3)$ transform is determined, again using Horn's method [34], from the documentation camera poses in their current and expected positions augmented with points representing unit positions along the documentation camera pose axes. The $SO(3)$ transform is applied to the documentation camera poses, resulting in fragmentary maps in a coherent world coordinate system defined by the localization camera. The two-step procedure ensures that the arbitrary-length vectors taken to represent positions along the documentation camera frame axes, that are used to resolve the rotational ambiguity, do not influence the scale parameter estimated in the first step.

D. Linking Fragmentary Maps

Each fragmentary map has its own private set of landmarks, some of which represent real-world landmarks that are repeatedly viewed across fragmentary maps, thus providing critical inter-map constraints needed to weave the fragmentary maps into a unified ecosystem map. We identify these landmarks by collecting all landmarks from other fragmentary maps potentially visible in a given fragment. Correspondences between these landmarks and keypoints in the keyframes of the target fragmentary map are determined and validated using geometric and feature-based criteria. When sufficient correspondences are established to keypoints associated with a landmark in the target fragmentary map, the duplicate landmarks are merged, providing a new constraint between the fragmentary maps.

E. Global Optimization of the Ecosystem Map

The previously described procedures yield a set of fragmentary ecosystem maps from the documentation camera images approximately aligned based on the localization camera trajectory and linked via mutually visible landmarks. This ecosystem map, comprising of keyframes and landmarks, is refined using global, non-linear optimization on the $SE(3)$ manifold resulting in a consistent, unified ecosystem map, depicted as the *Finalized Map* in Fig. 2. The objective (error) function incorporates costs assigned to all landmark-to-feature point associations and constraints on the keyframe poses based on the localization trajectory (Fig. 3). The constraints and estimated variables (i.e., keyframe poses and landmark positions) are represented as a locally-connected factor graph to facilitate global optimization [35]. The optimization procedure largely follows previous work [31], [36], the novelty being the incorporation of constraints imposed by the localization trajectory.

As depicted in Fig. 3, the constraints on the documentation camera keyframe poses derived from the localization trajectory involve two variables: the documentation image acquisition time t_i and the transformation $\mathbf{T}_{dl}$ between the localization and documentation cameras, both of which are, in turn, constrained by their prior estimates. The documentation and imaging cameras are not required to be precisely synchronized in time, i.e., the acquisition time for each documentation camera image, in terms of the localization trajectory, is only approximately known. Specifically, the constraint on the documentation camera keyframe pose is expressed as a ternary edge in the factor graph wherein t_i implies a specific localization camera pose that can be obtained from the trajectory. The localization camera pose can then translated into an implied documentation camera pose via $\mathbf{T}_{dl}$ and equation (1). Since the imaging system upon which the cameras are mounted is assumed to be rigid and stable throughout the data collection process, a single value of $\mathbf{T}_{dl}$ is estimated for the entire data set. The non-linear optimization attempts to minimize the following error function:

$$\operatorname*{arg\,min}_{\mathbf{X}_d,\mathbf{L},\mathbf{T}_{dl},t} \left[\sum_{i,j} \rho_h(r_{i,j}(\mathbf{X}_{d_i}, L_j)) + \sum_{i} \rho(p_i(\mathbf{X}_{d_i}, \mathbf{X}_{l_i}, \mathbf{T}_{dl})) \right.$$
$$\left. + \sum_{i} \rho(t_i - t_{i_{prior}}) + \rho(\mathbf{T}_{dl}^{-1}\mathbf{T}_{dl_{prior}}) \right] \qquad (3)$$

where t_i denotes the documentation image acquisition time and $t_{i_{prior}}$ its prior estimate, $\mathbf{X}_{d_i}$ the documentation camera pose at time t_i, L_j the landmark position, $\mathbf{T}_{dl}$ the transformation between the localization and documentation cameras and $\mathbf{T}_{dl_{prior}}$ its prior estimate, ρ the squared-error function, ρ_h the robust Huber error function, $r_{i,j}$ the reprojection error for landmark j observed in keyframe i and p_i the pose error between the estimated pose $\mathbf{X}_{d_i}$ of documentation camera i and its pose implied by the localization camera pose $\mathbf{X}_{l_i}$ at the time the documentation image was taken via transformation $\mathbf{T}_{dl}$. Specifically:

$$p_i = \mathbf{X}_{d_i}^{-1}\mathbf{X}_{l_i}\mathbf{T}_{dl} \qquad (4)$$

IV. Experimental Evaluation

A. Data Collection

A dual-camera rig was constructed consisting of a Stereolabs ZED-mini interfaced to a Jetson Xavier computer as the localization camera and a Panasonic GH5s configured with a 14 mm prime lens as the documentation camera. The cameras were secured to a rigid frame so that their relative orientation was constant. The Panasonic GH5s was oriented downward to provide ecosystem images of the highest quality. The ZED-mini was either facing directly forward or angled slightly downward ($\approx 25°$, measured for each deployment). Aligning the localization camera with the direction of travel allows for persistence of features in the FOV and observation of features at a wide range of distances, thereby improving tracking. However, it was found to be advantageous to angle the localization camera slightly downward to observe more proximal features thereby improving motion estimation and avoiding tracking loss. The ZED-mini stereo video was recorded at 60 fps with 1280×720 resolution and the Panasonic GH5s video at 60 fps with 4096×2160 resolution. The image data was collected in two visually repetitive environments: a lawn on the University of Georgia (UGA) campus and a salt marsh on Sapelo Island, GA, USA. Patches of roughly 10 m $\times$ 10 m were imaged by traversing the area in a lawn-mower (boustrophedon) pattern. At four sites on Sapelo Island, nine AprilTag markers [37] were placed in the imaged patch to serve as ground control points for accuracy assessment. The ground-truth positions of the AprilTags were determined to $<$ 2 cm accuracy using an RTK-GPS system comprising of a Trimble R12 GNSS receiver and Trimble TSC7 controller.

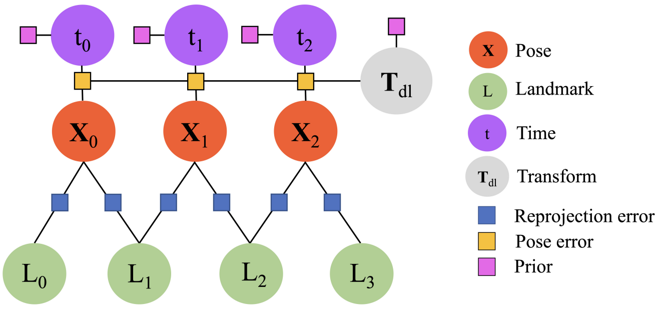

Fig. 3. **Documentation Camera Factor Graph**: Circles represent model parameters estimated via global optimization and squares represent error terms constraining the parameters.

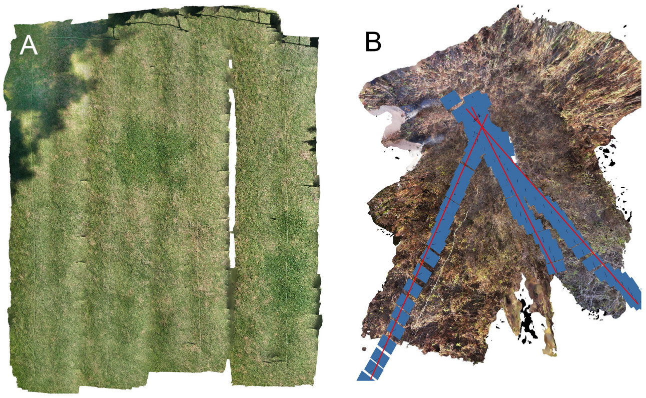

Fig. 4. **Sample Reconstructions**: A: Accurately reconstructed campus lawn using dual camera SLAM. B: SfM failure due to visual aliasing (blue squares represent aligned images). The three lines of images (highlighted in red) should be parallel but instead converge on a single point in the reconstruction.

B. Comparison with SfM

The dual-camera SLAM approach was compared to two state-of-the-art SfM programs, i.e., COLMAP [9] and Agisoft Metashape [38], since these SfM programs are commonly used in current environmental applications. We did not compare our system to pure SLAM approaches because our initial SLAM-based approaches to mapping repetitive ecosystems revealed complications, such as tracking loss and low image quality, that motivated development of the dual-camera approach. Six datasets (four from salt marshes on Sapelo Island, two from the UGA campus lawn) were processed to generate maps using our dual-camera SLAM approach and the two SfM programs. Video frames were extracted at 4 fps from the documentation camera videos (resulting in 80%-90% inter-frame overlap) and processed using the default SfM program settings that were slightly modified based on preliminary trials to improve reconstruction quality. First, the reconstructions were visually assessed to determine if the reconstruction was roughly consistent with the planar geometry of the patches and whether the inferred image locations relative to the reconstruction approximately matched the camera trajectory. Second, the completeness of the reconstructions was assessed using reconstruction metrics based on the fraction of aligned images for the SfM programs and, in the case of the dual-camera SLAM system, the fraction of visible mesh elements out of those determined to be potentially visible. For the SfM approaches all sub-maps were considered, offering a charitable representation of their performance. On these six datasets, the dual-camera SLAM approach was able to successfully reconstruct repetitive environments in cases when traditional SfM systems either failed entirely or were unable to fully reconstruct the imaged location (Table I). COLMAP was better able to generate reconstructions than Metashape, but the reconstructions were typically broken into multiple (up to 22) fragmentary maps. In contrast, our dual-camera SLAM approach was able to produce a single, unified map. As examples, we show texture mapped reconstructions of a salt marsh grassland (Fig. 2) and a UGA campus lawn (Fig. 4A). The SfM approaches often incorrectly merged subsections due to visual aliasing (Fig. 4B).

TABLE I
COMPARISON WITH SFM PROGRAMS

Sample	Metashape	COLMAP	DC-SLAM
P1	6% / Y	39% / Y	99% / Y
P2	16% / N	92% / N	100% / Y
P3	15% / N	65% / Y	99% / Y
P4	34% / Y	79% / Y	99% / Y
P5	83% / Y	99% / Y	99% / Y
P6	40% / Y	100% / Y	100% / Y

For each method the completeness metric is reported first followed by whether the reconstruction geometry was correct (Y) or not (N).

C. Accuracy of Dual Camera SLAM

At all the test sites, the dual-camera SLAM system produced reconstructions that appeared reasonable and covered the entire imaged patch (Table I). For quantitative assessment of the reconstruction quality, distances between the ground control points (AprilTags) in the reconstructions were compared with the true distances determined using RTK-GPS positions. At four locations on Sapelo Island, nine AprilTags were placed spanning the imaged patch: four at the corners forming a square defining the edges of the patch, four forming a nested square, and one at the center of the patch. Determination of AprilTag locations in the reconstructions was done as a post-processing step. After running the dual-camera SLAM program, a triangular surface mesh was fit to the landmark point cloud. AprilTags were detected in the documentation camera images and their 3D locations determined via back-projection onto the mesh using the camera poses and inverse camera model. Since most AprilTags were viewed in multiple images, the backprojected 3D positions of all views were averaged to produce a single location for each tag. Euclidean distances between all reconstructed tag pairs and their RTK-GPS positions were computed and compared (Fig. 5). The estimated inter-tag distances were generally in good agreement with the true distances (Fig. 5), though slightly overestimated ($\approx$ 3% on average). Nonetheless, the reconstructions were deemed sufficiently accurate for most ecological applications.

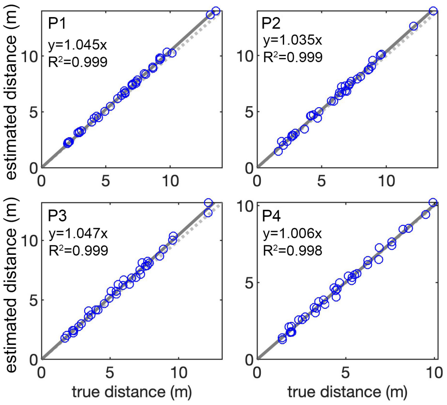

Fig. 5. Accuracy assessment via comparison of inter-tag distances measured using RTK-GPS (true distances) and from the dual-camera SLAM reconstructions (estimated distances) for four imaged patches (P1–P4). The solid line is a linear regression fit to the data forced through the origin (equation and R^2 listed on the plot) and the dotted line is the 1:1 line.

V. CONCLUSIONS AND FUTURE WORK

We proposed a dual-camera SLAM approach to map repetitive environments for use in ecological monitoring applications. While the proposed approach does entail a more complicated image acquisition setup, it offers much more reliable mapping of repetitive environments, typical of many ecosystems and agricultural fields. Furthermore, decoupling the localization and documentation cameras allows use of cameras ideally suited to each task and flexible swapping of either camera as required for the task at hand. Future improvements include incorporation of additional constraints such as IMU or GPS data [39] into the factor graph for more accurate mapping and development of fully coupled optimization strategies for the reconstructions generated by the localization and documentation cameras [26].

Acknowledgements: This work was supported by the US NSF Award DBI 2016741. We thank the University of Georgia Marine Institute on Sapelo Island, GA, USA for facilitating the field work.

References

[1] S. A. Levin, "The Problem of Pattern and Scale in Ecology," *Ecology*, vol. 73, no. 6, pp. 1943–1967, DEC 1992.

[2] C. E. Tarnita, J. A. Bonachela, E. Sheffer, J. A. Guyton, T. C. Coverdale, R. A. Long, and R. M. Pringle, "A theoretical foundation for multi-scale regular vegetation patterns," *Nature*, vol. 541, no. 7637, pp. 398+, JAN 19 2017.

[3] D. Schimel, R. Pavlick, J. B. Fisher, G. P. Asner, S. Saatchi, P. Townsend, C. Miller, C. Frankenberg, K. Hibbard, and P. Cox, "Observing terrestrial ecosystems and the carbon cycle from space," *Global Change Biology*, vol. 21, no. 5, pp. 1762–1776, MAY 2015.

[4] A. W. Fitzgibbon and A. Zisserman, "Automatic camera recovery for closed or open image sequences," in *Computer Vision — ECCV'98*, H. Burkhardt and B. Neumann, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 311–326.

[5] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM TRANSACTIONS ON GRAPHICS*, vol. 25, no. 3, pp. 835–846, JUL 2006.

[6] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a Day," in *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, 2009, pp. 72–79.

[7] P. G. Brodrick, A. B. Davies, and G. P. Asner, "Uncovering Ecological Patterns with Convolutional Neural Networks," *Trends in Ecology & Evolution*, vol. 34, no. 8, pp. 734–745, AUG 2019.

[8] B. M. Hopkinson, A. C. King, D. P. Owen, M. Johnson-Roberson, M. H. Long, and S. M. Bhandarkar, "Automated classification of three-dimensional reconstructions of coral reefs using convolutional neural networks," *PLOS ONE*, vol. 15, no. 3, MAR 24 2020.

[9] J. L. Schonberger and J. M. Frahm, "Structure-from-Motion Revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[10] T. D. Jackson, G. J. Williams, G. Walker-Springett, and A. J. Davies, "Three-dimensional digital mapping of ecosystems: a new era in spatial ecology," *Proceedings of the Royal Society B-Biological Sciences*, vol. 287, no. 1920, FEB 12 2020.

[11] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99–108, JUN 2006.

[12] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, DEC 2016.

[13] F. Shu, P. Lesur, Y. Xie, A. Pagani, and D. S. Tricker, "SLAM in the Field: An Evaluation of Monocular Mapping and Localization on Challenging Dynamic Agricultural Environment," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV 2021)*, 2021, pp. 1760–1770.

[14] A. J. Davison, Y. G. Cid, and N. Kita, "Real-time 3d slam with wide-angle vision," *5th IFAC/EURON Symposium on Intelligent Autonomous Vehicles*, pp. 868–873, 2004.

[15] C. B. Anderson, "Biodiversity monitoring, earth observations and the ecology of scale," *Ecology Letters*, vol. 21, no. 10, pp. 1572–1585, OCT 2018.

[16] M. C. Hayes, P. C. Gray, G. Harris, W. C. Sedgwick, V. C. Crawford, N. Chazal, S. Crofts, and D. W. Johnston, "Drones and deep learning produce accurate and efficient monitoring of large-scale seabird colonies," *Ornithological Applications*, vol. 123, pp. 1–16, 2021.

[17] S. Baena, J. Moat, O. Whaley, and D. S. Boyd, "Identifying species from the air: Uavs and the very high resolution challenge for plant conservation," *Plos One*, vol. 12, no. 11, 2017.

[18] C. B. Edwards, Y. Eynaud, G. J. Williams, N. E. Pedersen, B. J. Zgliczynski, A. C. R. Gleason, J. E. Smith, and S. A. Sandin, "Large-area imaging reveals biologically driven non-random spatial patterns of corals at a remote reef," *Coral Reefs*, vol. 36, no. 4, pp. 1291–1305, 2017.

[19] A. Bodenmann, B. Thornton, and T. Ura, "Generation of high-resolution three-dimensional reconstructions of the seafloor in color using a single camera and structured light," *Journal of Field Robotics*, vol. 34, no. 5, pp. 833–851, 2017.

[20] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[21] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment — a modern synthesis," in *Vision Algorithms: Theory and Practice*, B. Triggs, A. Zisserman, and R. Szeliski, Eds. Springer Berlin Heidelberg, 2000, pp. 298–372.

[22] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[23] C. Wu, "Towards Linear-time Incremental Structure from Motion," in *2013 International Conference on 3D Vision (3DV 2013)*, 2013, pp. 127–134.

[24] P.-Y. Lajoie, S. Hu, G. Beltrame, and L. Carlone, "Modeling Perceptual Aliasing in SLAM via Discrete-Continuous Graphical Models," *IEEE Robotics & Automation Letters*, vol. 4, no. 2, pp. 1232–1239, 2019.

[25] S. Urban, S. Wursthorn, J. Leitloff, and S. Hinz, "Multicol bundle adjustment: A generic method for pose estimation, simultaneous self-calibration and reconstruction for arbitrary multi-camera systems," *International Journal of Computer Vision*, vol. 121, no. 2, pp. 234–252, 2017.

[26] J. Kuo, M. Muglikar, Z. Zhang, and D. Scaramuzza, "Redesigning slam for arbitrary multi-camera systems," in *IEEE International Conference on Robotics and Automation (ICRA)*, Conference Proceedings.

[27] M. J. Tribou, A. Harmat, D. W. L. Wang, I. Sharf, and S. L. Waslander, "Multi-camera parallel tracking and mapping with non-overlapping fields of view," *International Journal of Robotics Research*, vol. 34, no. 12, pp. 1480–1500, 2015.

[28] Q.-Y. Zhou, S. Miller, and V. Koltun, "Elastic fragments for dense scene reconstruction," in *IEEE International Conference on Computer Vision*, 2013, pp. 473–480.

[29] D. Thomas and A. Sugimoto, "Modeling large-scale indoor scenes with rigid fragments using rgb-d cameras," *Computer Vision and Image Understanding*, vol. 157, pp. 103–116, 2017.

[30] C. Houscago, M. Bloesch, and S. Leutenegger, "Ko-fusion: Dense visual slam with tightly-coupled kinematic and odometric tracking," in *International Conference on Robotics and Automation*, 2019, pp. 4054–4060.

[31] R. Mur-Artal and J. D. Tardos, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.

[33] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[34] B. K. P. Horn, "Closed-form solution of absolute orientations using unit quaternions," *Journal of the Optical Society of America a-Optics Image Science and Vision*, vol. 4, no. 4, pp. 629–642, 1987.

[35] F. Dellaert, *Factor Graphs: Exploiting Structure in Robotics*, ser. Annual Review of Control Robotics and Autonomous Systems, 2021, vol. 4, pp. 141–166.

[36] R. Kuemmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: a general framework for graph optimization," in *IEEE International Conference on Robotics and Automation (ICRA)*, Conference Proceedings, pp. 3607–3613.

[37] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2011, pp. 3400–3407.

[38] "Agisoft metashape," 2021. [Online]. Available: https://www.agisoft.com/

[39] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-Manifold Preintegration for Real-Time Visual-Inertial Odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, FEB 2017.