CorNet: Unsupervised Deep Homography Estimation for Agricultural Aerial Imagery

Dewi Endah Kharismawati[1,2,3,4], Hadi Ali Akbarpour[1,2], Rumana Aktar[1,2], Filiz Bunyak[1,2], Kannappan Palaniappan[1,2], and Toni Kazic[1,2,3,4(✉)]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, USA
{aliakbarpourh,bunyak,pal,kazict}@missouri.edu
[2] MU Plant Science Foundry, University of Missouri, Columbia, USA
[3] Missouri Maize Center, University of Missouri, Columbia, USA
[4] Interdisciplinary Plant Group, University of Missouri, Columbia, USA
{dek8v5,rayy7}@mail.missouri.edu

Abstract. Efficient and accurate estimation of homographies among images is the first step in mosaicking crop fields for phenotyping. The current strategy uses sophisticated vehicles that have excellent telemetry to hover over a grid of waypoints, imaging each one. This approach simplifies homography estimation, but precludes more flexible, adaptive protocols that can collect richer information. It also makes aerial phenotyping impractical for many researchers and farmers. We are developing an alternative strategy that uses consumer-grade vehicles, freely flown over a variety of trajectories, to collect video. We have developed an unsupervised deep learning network that estimates the sequence of planar homography matrices of our corn fields from imagery, without using any metadata to correct estimation errors. The vehicle was freely flown using a variety of trajectories and camera views. Our system, *CorNet*, performed faster than and with comparable accuracy to the gold standard ASIFT algorithm in many challenging cases.

Keywords: Unsupervised convolutional neural network · Homography estimation · UAV · Freely flown · Maize · Mosaicking

1 Introduction

The small size, high image resolution, and low cost of unmanned aerial vehicles (UAV) offer researchers and farmers a powerful tool for capturing the raw data of plant phenotypes. Drones are emerging as phenotyping and agricultural tools because of their low cost, convenient use, and high image resolution [19,22,32,36]. So far, the prevailing strategy for transforming the UAV sensor data into quantitative biological information has relied on canopy reflectances at different frequencies that are collected from nadir views at precomputed, fixed

A. Bartoli and A. Fusiello (Eds.): ECCV 2020 Workshops, LNCS 12540, pp. 400–417, 2020.
https://doi.org/10.1007/978-3-030-65414-6_28

waypoints. Though information collected this way has been successfully used to phenotype crop plants [14,16,20,24,43,46], this strategy has three significant limitations. The first is that canopy observations require extensive, often line-specific, validation before they can be used as surrogates for morphological and developmental phenotypes [10]. The second is the heavy reliance on nadir views at fixed waypoints. These simplify image registration by constraining the possible camera poses and trajectories and ensuring high overlap between images, and allow one to pre-compute the matrices that map the spatial transformations between frames. However, nadir views inherently lose significant morphological information that is visible in oblique views and fixing many waypoints requires significant effort and longer missions. Finally, grid-based trajectories lose information for other reconstructions, including three-dimensional ones [6,42]. In principle, researchers and farmers should not need to "engineer up" to exploit UAV technologies.

An alternative strategy is to use cheap equipment to freely fly any trajectory with any camera pose, then compute the orthomosaics and three-dimensional reconstructions directly from video imagery. This strategy has several advantages, especially in vehicle cost and complexity, ease of deployment, speed of data acquisition, the inherent sequencing of video, and information richness. It discards vehicle telemetry and georegistration, which are usually only marginally accurate for consumer-grade vehicles, and avoids the labor-intensive placement and maintenance of ground control points. Perhaps most importantly, this strategy is tailor-made for developing adaptive protocols that would dynamically survey large areas at low resolution, identify phenotypic outliers, and reimage those for more and different details. However, this alternative entails significant computational costs, beginning with registering successive pairs of images that vary much more in their spatial relationships than those from fixed, constant altitude trajectories. Without prior knowledge of these relationships or metadata from which to derive them, the matrices that transform the local coordinate system of one image into that of its successor, and these into the global coordinate system of the mosaic, must be computed *de novo*.

Image registration has benefitted from many years of research in computer vision [3,4,9,11,13,29,30,39,45,51]. Its essential problem is to compute the homography matrix that optimally maps the local coordinate system of one image and its population of recognizable features onto another's, accounting for the relative translations, rotations, and scalings between the two images' features. Not all members of the first image's feature population will exhibit identical motions relative to the second image's. This is obvious for rotations: the greater a feature's displacement from the axis of rotation, the greater its Cartesian motions. But it can also occur in translations and scalings, notably when a vehicle experiences minor air currents that push it horizontally or vertically during flight. Camera lenses invariably distort the image around the lenses' peripheries, and minor errors in gimbal pointing foreshorten one side of the image relative to another.

Algorithmic approaches to registration rely on identifying *key points*—highly localized, statistically distinct features in one image—and finding their correspondents in the next [41]. Sparse features such as lines, corners, and edges are used to increase the spatial separation among them and to exploit standard feature recognition algorithms and feature descriptor functions [8,33,35,40]. The features are then matched between images under the assumption that the field of interest is sufficiently planar and the matrix estimated using, for example, direct linear transform, singular value decomposition, or RANSAC [17,21,28,34]. Mosaics built this way significantly benefit from metadata to correct the drift that accumulates as the homography matrices are multiplied together during mosaicking [31,34,50]. This two-dimensional approach can perform well on built environments, but it is problematic for imagery with fuzzy features that repeat frequently, such as plants in crop fields. Significant improvements in algorithmically mosaicking maize fields imaged by free flight from imagery alone depend on the statistical feature descriptor, on selecting the strongest features, and on minimizing the chain of matrix multiplications that build the mosaic by first mosaicking subsequences of frames [5]. Nonetheless, Aktar *et al.*'s VMZ pipeline can still geometrically distort the mosaic and mismatch repetitive features. Structure from Motion (SfM) finds the third dimension for unconstrained trajectories without metadata at higher accuracy, so it would seem perfect for our alternative strategy [2,18]. However, its computational cost grows exponentially with the number of images unless the imagery can be pre-processed to partition it or to select key frames. This makes SfM prohibitive for routine use on the large data sets needed for agricultural landscapes [2,49]. More modern incremental methods, such as iSAM2, may produce better geometric models more efficiently for agricultural imagery without so much intervention [25].

In contrast, convolutional neural networks (CNN) estimate homography by learning the regression of the displacement of patches between two frames. DeTone *et al.*'s supervised HomographyNet estimated the homography of 128×128 pixel patches bounded by four points using a VGG8-like architecture [44]. They augmented the original images by synthesizing new ones using random projective transformations to the images. They minimized L_2 loss between the four-point patch locations in the synthesized training data and those predicted by the network [15,44]. The average corner error was 9.2 pixels compared to the 11.2 pixels average corner error computed by ORB and RANSAC [17,40]. Kang *et al.* proposed a hybrid framework that incorporates deep learning and energy minimization of photometric refinement for an accurate and efficient homography estimation [26]. It simplifies HomographyNet to reduce the computational time significantly and performs well on images of room ceilings interspersed with large fluorescent light fixtures using 32×32 pixel patches. The light fixtures have very sharp edges, and it is unclear how the patches are chosen. Without available code to test, it seems unlikely this approach would be suitable for agricultural imagery. Nguyen *et al.* extended HomographyNet to an unsupervised CNN they called Unsupervised Deep Homography (UDH). These investigators used the estimated homography matrix to warp one image and compared the warped image to the successor image, minimizing the L_1 photometric loss function [37]. They trained two models using UDH: one with synthetic data and the other with aerial imagery. Throughout this paper, we will refer to the aerial model as UDH. While UDH did as well or better on built landscapes than HomographyNet, our experiments showed it does not generalize well to crop field imagery, especially fields with a high vegetation index (see Sect. 3.1).

Rapid, accurate image registration for freely flown trajectories is the first step in making it possible for researchers and farmers to routinely phenotype their fields with minimal effort and cost. Ideally, a flexible, efficient framework that could accommodate any trajectory and camera pose could serve as the back-end for a public computational resource. Here we present our experiments extending the unsupervised CNN of Nguyen *et al.* to video of maize nurseries imaged with a freely flown consumer-grade vehicle. Our trained network, *CorNet*, ran faster and produced results comparable to ASIFT, the gold standard feature descriptor for algorithmic homography estimation. No vehicle or camera telemetry or ground control points are used in matrix estimation, but accurate mosaicking requires training data that cover a range of vehicle movements and camera poses. Once trained, *CorNet* performs well on a wide variety of landscapes.

2 Materials and Methods

2.1 Maize Nurseries

Imagery was collected from four different fields during the summer, 2019 field season in Columbia, MO. Three were planted with 36 in. row spacing. Two genetic nurseries, fields 33 and 22, were planted in 20 ft rows with 4 ft alleys; the central file of rows was left unplanted to help orient the pilots. Field 33 was hand-planted with inbred lines and disease lesion mimic mutants using an Almaco jab planter at 12 in. spacing and machine-planted with an elite line using a Jang TD-1 push planter at 8 in. spacing [48]. Field 22 was planted late in the season with inbred lines (hand-planted) and the elite line (Jang-planted). Field 30 was machine-planted with sweet corn at $\approx$ 6 in. spacing without alleys. A field of maize varieties for drought trialing planted at 30 in. row spacing with $\approx$ 6 in. spacing between plants was imaged after fertilization and before sensecence.

2.2 *CorNet*'s Architecture

DeTone *et al.* used a supervised regression to predict the four point homography and used the L_2 loss as a cost function relative to the ground truth. Nguyen *et al.* began with a set of unsupervised convolution layers, passed the result to DeTone *et al.*'s regression layer, and converted the four point homography to a 3×3 matrix. This was used to warp the frames and produce a prediction. The L_1 photometric loss, used as the cost function, was back-propagated to the regression model. *CorNet* uses UDH's basic architecture, but adds four layers to the input, as shown in Fig. 1. These layers allow us to use a larger patch size

(512 × 512 × 2) compared to Nguyen *et al.* (128 × 128 × 2). The larger patch size prevents many estimation errors arising from the repetitive nature of the imagery and information loss. Using half the number of filters used by UDH prevented loss of the fuzzier plant features (data not shown). Figure 2 shows *CorNet*'s workflow, which ranged from field operations to mosaicking.

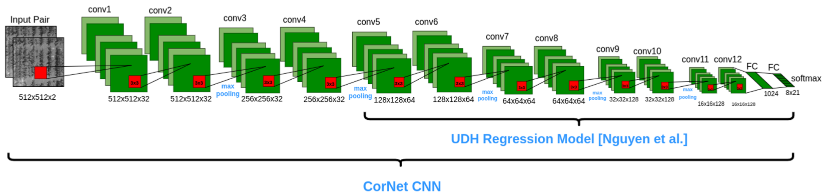


Fig. 1. Architecture of *CorNet*'s CNN.

2.3 Data Collection

We used a DJI Phantom 4 manufactured by Da-Jiang Innovations, Shenzhen, China, which carries an RGB camera to image our fields, forests, and buildings (https://www.dji.com/phantom-4). No structural alterations to the vehicle or its camera were made; by 2019, the Phantom 4 had accumulated errors in its gimbal position. All flights were flown manually using the Autopilot mobile app (https://autoflight.hangar.com/autopilot/flightschool). Flight trajectories included all four translations (forward, backward, right slides, and left slides); scaling (changing altitude); rotations (clockwise and counter-clockwise); and combinations of two or more movements. In forward and backward movements, the camera is pointed parallel to the direction of vehicle motion; in "slides", the camera is pointed perpendicular to the vehicle's motion. In rotations, the position of the camera is fixed and the vehicle rotates, sometimes with other translations and scalings. Six different types of trajectories were used: straight to target; serpentine, parallel or perpendicular to the rows, with slides or rotations on each pass; and orbits around the field. Camera poses included both nadir

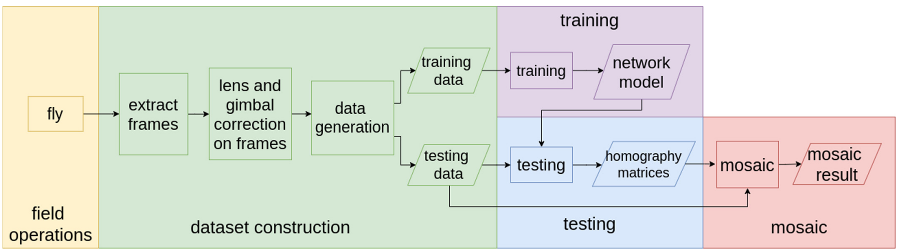


Fig. 2. *CorNet* workflow.

and a variety of oblique orientations. All flights were at relatively low altitudes, *ca.* 25–100 ft above ground level (AGL). The vehicle was flown in relatively light winds, but occasionally experienced horizontal and vertical displacements due to air currents.

Video was collected at 24 frames/sec in 24-bit color depth on Secure Digital High Capacity (SDHC) cards inserted in the vehicle and manually copied to the computer. The Phantom 4 can record a maximum of about 19 min 42 s of video, providing about 30,000 frames of 4069×2160 each for a total of about 850 GB. All code was run on a Lambda Labs Intel Core i9-9920X, 2 NVIDIA RTX 2080Ti, and 128 GB memory.

2.4 Training and Testing

We used Nguyen *et al.*'s published code and model to retune their UDH aerial model. We used a similar approach to training *CorNet*, with the necessary modifications for network architecture and emphasizing a large data set of diverse motions, rather than cross-validating a smaller sample. All new code was written in the TensorFlow framework using stochastic gradient descent, a batch size of 128 for rUDH and 32 for *CorNet* [1]. The Adam Optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and a learning rate of 0.0001 [27]. The mean intensity of all pixels was used to normalize images during training. We fine-tuned rUDH for 250,000 iterations in approximately 36 h. *CorNet* was trained from scratch using the lower half of Table 1 for 150,000 iterations in 47 h. Training *CorNet* was essentially complete by $\approx 30,000$ iterations. For testing, the pairwise sucessive frames are fed to the network, which returns the four points' homographic prediction and the L_1 photometric loss. We then compute the 3×3 homography.

2.5 Mosaicking

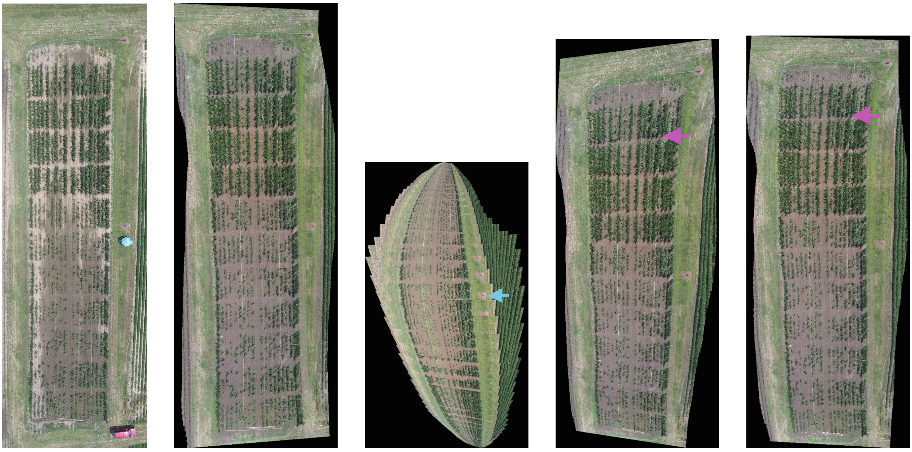
The output of homography estimation is a 3×3 matrix that projects the current frame to the next frame ($H_{n \rightarrow n+1}$). We used the first frame as a base frame, projecting the inverse of $H_{n \rightarrow n+1}$ for each successive frame back to the first, $H_{1 \leftarrow n}$:

$$H_{1 \leftarrow n} = H_{1 \leftarrow 1} \times H_{1 \leftarrow 2} \times H_{2 \leftarrow 3} \times \ldots \times H_{n-1 \leftarrow n} \tag{1}$$

We estimated the global canvas by transforming the four corners of each frame to the accumulated $H_{1 \leftarrow n}$, shifting any projections with negative values to $(0, 0)$ with a translation offset matrix, M_o. Thus, for every n^{th} frame, the total projection is

$$I_n^w = (M_o \times H_{1 \leftarrow n}) \times I_n, \tag{2}$$

where the original image I_n is projected and translated to produce the warped image I_n^w. We used a simple pixel replacement blending to produce the final mosaic [23].



(a) field (b) ASIFT (c) UDH (d) rUDH (e) *CorNet*

Fig. 3. Homography accuracy by the four methods, visualized as mosaics, compared to an image of the whole field. Panel (a), the whole field. Panels (b)—(e), mosaics by the four methods. The blue arrow in panel (c) marks the single irrigation stand pipe that is repeated five times. The pink arrows in panels (d) and (e) mark the same field location in the two mosaics. (Color figure online)

2.6 Model and Mosaic Evaluation

To evaluate the quality of the estimated homography matrices, we picked the four corners of a random rectangular patch in the first frame and constructed the corners of a ground truth patch in the second frame with ASIFT [35], then compared these constructed corners to the ones predicted by CorNet. The root mean squared error (RMSE) between the n corners of the estimated, *Pred*, and ASIFT ground truth Gt, patches was computed using

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Pred_i - Gt_i)^2}{n}}. \tag{3}$$

We estimated geometric distortion of the mosaics relative to a whole field image as $Area(Field)/Area(OrientedBoundingBox)$ [47]. We also visually evaluated the homography by comparing the outlines of the *CorNet* and ASIFT patches in the second frame. Finally, we evaluated the quality of the mosaics by visual inspection, since the different errors are quite characteristic. Run time performance was also benchmarked.

3 Results

3.1 Motivation

The acid test of homography estimation is how well it maps between images when mosaicked. Errors that seem insignificant when comparing between two frames accumulate during mosaicking, since the matrices are multiplied together (Eq. 1). Figure 3 shows how errors in homography estimation can accumulate. For these frames, the trajectory was backwards over the center of a nearly rectangular field, with the camera pointing downwards in nadir view. Lens and gimbal errors were corrected.

ASIFT (panel (b)) produces the best results, with only minor geometric distortion of the field, seen as bending of the two irrigation pipes on either side. UDH's difficulties in computing the homographies are apparent from the extreme geometric distortion; the area around the single irrigation stand pipe marked by the blue arrow is repeated five times (panel (b)). While the field is not perfectly rectangular (0.9612), the geometric distortion of each mosaic relative to an image of the field, measured as deviation from a perfect rectangle, are 0.9613, 0.7034, 0.8428, and 0.9169, for ASIFT, UDH, rUDH, and *CorNet*, respectively. *CorNet*'s performance is superior to those of UDH and rUDH, though not quite as good as ASIFT.

If the problem of UDH was simply a breakdown in transfer learning, then retraining it with our own imagery should improve matters (panel (d)). It does, but geometric distortion is visible moving from the top to the bottom of the mosaic. The left edges of the mosaics are shorter than the right edges and the overall length is too small. Registration errors between frames are also apparent when looking along the pipes and at the maize itself (pink arrow). These results led to the development of *CorNet* (Sect. 2.2). Panel (e) shows *CorNet* reduced the false joins and produces a better overall field geometry than rUDH, though errors are still visible at the extreme ends of the field. The pink arrow in panel (e) marks the same position as in panel (d) so that the relative magnitude of the errors between the two models can be compared. Compared to ASIFT, *CorNet* compresses the field slightly from top to bottom; for example, the left border rows wiggle slightly and the overall row lengths are slightly shorter in panel (e).

3.2 Training and Testing Data

Table 1 summarizes the data sets used to retrain UDH, which we denote rUDH, and *CorNet*. Frames grouped under translation include five different movements (forward, backward, left slide, right slide, and both types of serpentines) and both orientations relative to the rows. We retrained UDH for maize with 13,614 pairs of frames covering both seedling and adult maize. No lens or gimbal corrections were applied (Table 1, upper portion). The training data for *CorNet* only slightly overlapped those for UDH (Table 1, lower portion). Seedling imagery was excluded, more movements were added, and lens and gimbal errors were corrected using the procedure of [5]. Both data sets were augmented for movements by frame rotation and swapping, doubling the total frames; and each frame had some random illumination changes [37].

To evaluate the accuracy of homography estimation and mosaicking, run time, and generalizability, we used a wide variety of imagery that was corrected for lens and gimbal errors (Table 2). Maize testing data were drawn from multiple videos shot over the growing season on seedlings, before flowering, and mature maize, again with multiple trajectories and camera poses. The same vehicle imaged farm buildings, patches of forest in south-central Missouri in summer and fall, and the surface of a pond. We used the telemetrized versions of the soil, urban, and countryside imagery of UMDC, but without using the telemetry [7].

3.3 Homography Estimation for Complex Movements

Table 3 compares the run time and errors of ASIFT, UDH, our rUDH, and *CorNet* for eight different movements. UDH consistently has the best run time performance, with the exception of an orbital trajectory. *CorNet*'s run time lies between that of ASIFT and either version of UDH, running between ≈ 0.1 to ≈ 0.5 of ASIFT. *CorNet* is the most accurate for all movements except the right slide, consistently outperforming UDH and rUDH. Examination of the predicted homographies supports these quantitative comparisons. Figure 4 compares homographies for a pair of frames for UDH, rUDH, and *CorNet*, using ASIFT as the ground truth. UDH and rUDH both produce more errors in homography estimation compared to *CorNet*: the RMSEs for UDH, rUDH, and *CorNet* are 7.656, 3.190, and 2.096 for these pairs, respectively.

Table 1. Training data for retraining UDH (upper half) and *CorNet* (lower half). The translation data include forward, backward, left slide, right slide, and serpentine movements.

Movement	Camera pose	Growth stage, sampling	Raw frame pairs
Translation	Nadir	Seedling, 3 fps	2817
		Adult, 1/4 – 3 fps	5333
	Oblique	Seedling, 3 fps	2038
		Adult, 1/2 fps	185
Rotation	Nadir	Adult, 1/4– 3 fps	653
	Oblique	Seedling, 3 fps	931
Scale	Nadir	Adult, 1/4 – 3 fps	561
Orbit	Oblique	Adult, 1–2 fps	1096
Total rUDH			13,614
Translation	Nadir	Adult, 3 fps	2130
	Oblique	Adult, 1/3 – 3 fps	2132
Rotation	Nadir	Adult, 1/3 – 5 fps	2152
Scale	Nadir	Adult, 1/5 – 5 fps	1367
Orbit	Oblique	Adult, 1/3 – 3 fps	1430
Total *CorNet*			9211

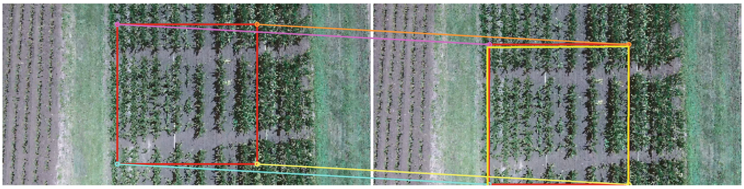
Table 2. Test data. Target is maize unless otherwise indicated.

Video no., data set	Sampling rate (fps)	No. pairs
174, forward	4	164
167, backward	3	83
201, right slide	3	101
201, left slide	3	140
362, scaling	3	180
413, dense maize	3	179
466, forest, summer	2	120
517, forest, fall	2	180
202, rotation	4	120
360, oblique parallel to row	3	96
274, orbit	3	180
179, maize seedlings	3	180
urban, UMDC	3	141
soil, UMDC	2	150
countryside, UMDC	3	119
368, farm buildings	1	166
520, water	3	90

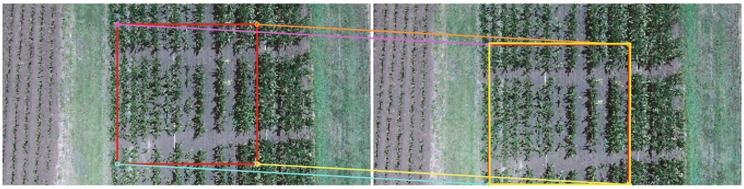
Table 3. Comparison of the methods' performance over different aerial movements. Run time is in seconds, RMSE is in pixels; the best values for each are shown in **bold font**.

	ASIFT	UDH		retrained UDH		*CorNet*	
Movement	Run time	Run time	RMSE	Run time	RMSE	Run time	RMSE
Forward	2101.5102	**73.7880**	38.3918	80.0166	3.2817	181.9988	**1.7370**
Backward	750.5475	**67.8707**	132.6135	74.9763	2.7282	130.0924	**2.3456**
Right slide	941.6449	**69.7518**	120.3709	76.8579	**4.5235**	140.5475	8.1995
Left slide	1609.9070	**74.6321**	108.6612	78.8836	4.0709	166.6295	**2.3374**
Scale	347.4869	**89.2801**	91.8952	91.8022	2.3649	161.6219	**0.6875**
Rotation	1954.2377	**90.6798**	114.5680	93.4896	3.4074	164.3279	**2.0111**
Oblique	653.4463	**85.7609**	47.6286	89.0532	11.4312	143.9081	**9.5610**
Orbit	2106.3459	103.8676	110.0210	**98.7048**	4.3955	220.8464	**2.9083**

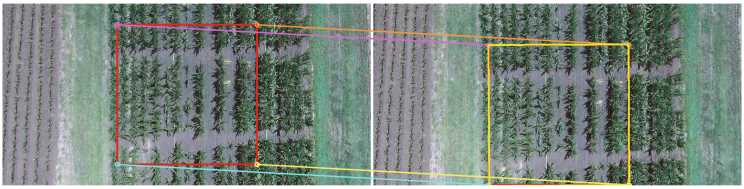
Combinations of simple movements can produce even greater mosaicking errors. Figure 5 shows a very stringent test: the UAV was flown forward and backward without rotating the camera. Minor errors in homography estimation accumulate over twice as many frames, and are readily apparent as mismatches between the beginning and end of the sequence. In this type of trajectory, the perspective changes as the UAV moves. ASIFT (panel (a)) is the least sensitive



(a) UDH



(b) rUDH



(c) *CorNet*

Fig. 4. Sample four point homography estimation by ASIFT (ground truth), UDH, rUDH, and *CorNet* for a frame and the adjacent sampled frame. The red squares in the left frames are the randomly chosen patches; in the right frames, these are the matching patches predicted by ASIFT. The yellow frames are predicted by each of the CNNs. (Color figure online)

to this change: the tilt of the top edge and the imperfect registration at the bottom of the image illustrate this error accumulation. None of the CNNs perform as well as ASIFT, but *CorNet* (panel (d)) registers each pass fairly well individually. The worse case is visible in the figure; the forward pass lying underneath is better, similar to Fig. 3, panel (d). Misalignment of the two passes increases moving down the field than in ASIFT (panel (a)). Consistent with this, starting the mosaic in the center of the sequence produces a better overlay of the two passes because there is less error to accumulate (data not shown). The perspective change combines with these errors to compress the mosaic. UDH and rUDH illustrate more extreme versions of these errors (panels (b) and (c)). UDH cannot register the backward pass at all, and rUDH exaggerates the backward pass errors so much they form a small tail at the top of the mosaic.

(a) ASIFT (b) UDH (c) rUDH (d) *CorNet*

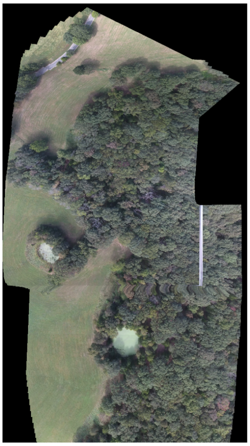
Fig. 5. Mosaics by the four methods for forward and backward motion, with a slightly oblique camera pose. (Color figure online)

Table 4. Run time and RMSE for the four methods on challenging data.

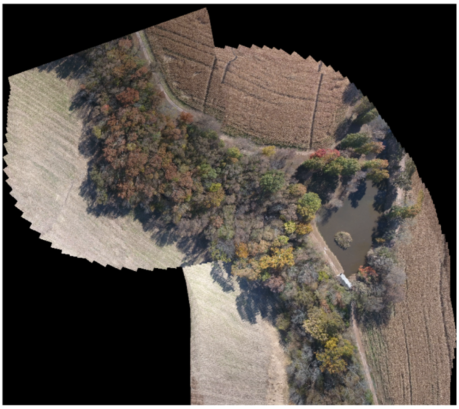
Movement	ASIFT	UDH		Retrained UDH		*CorNet*	
	Run time	Run time	RMSE	Run time	RMSE	Run time	RMSE
Seedling maize	1899.9222	**77.1553**	41.8223	80.9572	**4.0044**	192.3239	16.0899
Dense maize	2393.3591	**77.0296**	92.7713	81.3696	4.7668	194.0659	**1.7360**
Farm buidings	595.6757	**78.0988**	100.4067	80.5361	3.2728	183.6212	**3.2293**
Summer forest	1114.0191	**74.3189**	54.9672	77.2994	3.2922	151.8185	**1.6663**
Fall forest	1457.5384	**79.2406**	66.6672	81.7336	5.6429	195.0878	**3.3883**
UMDC, urban	632.2847	**73.8767**	58.7524	77.9679	4.9085	166.0596	**1.6666**
UMDC, dirt	555.7996	**78.6077**	39.9764	83.8195	5.2164	166.3824	**3.2201**
UMDC, countryside	590.4839	**75.0135**	69.2046	77.0026	6.2166	150.5176	**1.5583**
Irrigation pond	200.3115	**83.6731**	555.6703	86.0839	535.4424	136.5369	**529.0653**

3.4 *CorNet*'s Generalizability to Different Landscapes

CorNet performs well on a variety of landscapes. Table 4 summarizes the performance of the four methods. Apart from seedling maize, *CorNet* has the best RMSE; as before, UDH has the best run time. Figure 6 gives some examples of *CorNet*'s mosaics. The major difficulty it experiences is when the vehicle hovers for some time. This is visible in panels (a) and (d). In hovering, the vehicle's motion becomes more comparable to the error in homography estimation. This can produce very large errors, even in built landscapes (data not shown).

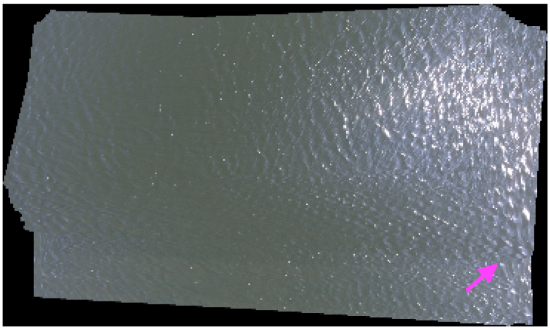


(a) Summer forest.



(b) Fall forest.

(c) UMDC dirt.



(d) Irrigation pond in the late morning.

Fig. 6. *CorNet* generalizes to other landscapes. Panel (a): the white object is the landing gear when the UAV was tipped slightly by an updraft. Just below that is an area of poor mosaicking due to the vehicle hovering. Panel (b): a different forest. Panel (c): the blue arrow marks a misalignment in the road's edge. Panel (d): minor misalignments of frames are visible in the highlighted ripples (pink arrow). (Color figure online)

4 Discussion

CorNet is successful for several reasons. First, to a homography estimator, all corn looks alike. A simple thought experiment illustrates why: in a relatively uniform stand of maize, a patch the size of a corn plant would lack the adjacent features that help to distinguish one plant from another. Increasing the patch size allows more of these features in the patch and the combinations of those features helps distinguish patches. Second, *CorNet* was trained with data from many different movements. Our initial model used only forward movement in training and failed miserably on backwards movements. As we added more movements, more mosaics of different trajectories were computed successfully. Since the training data used for rUDH and *CorNet* overlapped only slightly, it is possible the former was trained with more right slides than the latter. Diversifying the training imagery also helps: the only time another model was more accurate than *CorNet* was rUDH on seedlings, imagery of which was included in that training set. Third, the trajectories used to train rUDH and *CorNet* varied considerably. The majority of UDH's predictions lie on the same location regardless of camera movements. Though no details of camera movements are given in [37], we surmise their data were collected from a vehicle flown automatically, resulting in approximately uniform overlap. This partially explains why their model did not transfer to maize imagery.

Several areas need improvement. Four classes of movement need additional training data: forwards and backwards over long serpentine trajectories; turns, especially by rotation; scaling due to altitude changes; and hovers. We also need to find substitutes for checkerboard-based lens correction and empirical gimbal correction, since few researchers and farmers will have calibration checkerboards. We are examining ways to exploit imagery of farm buildings and other structures with perpendicular corners to automatically determine gimbal error.

Several algorithmic changes to homography estimation and mosaicking could also improve our results. Adaptive frame sampling to assure sufficient overlap between frames without bloating the data would resolve some of *CorNet*'s issues with turns and scale changes. Comparison of *CorNet*'s mosaics to those produced by AutoStitch and LF-Net show the strengths and weaknesses of each approach [12,38]. Though outside the scope of the present paper, some preliminary conclusions suggest several improvements. All the methods perform best with nadir, as opposed to oblique, camera poses, and all have difficulty mosaicking serpentine trajectories. AutoStitch and LF-Net's use of more features over the entire frame, as opposed to the random image patches of *CorNet* and UDH, sometimes produce superior mosaics. However, *CorNet* often generalizes better, especially when the features are more uniform, as in the pond imagery. Finally, our mosaics would benefit from a more sophisticated mosaicking algorithm, such as the mini-mosaicking algorithm in VMZ [5]. This is particularly true for complex motions, such as those in Sect. 3.3.

5 Conclusions

In this paper we have presented a new unsupervised homography estimation network, sets of videos captured by a consumer-grade UAV, and a set of training strategies for generating mosaics of crop fields that suffer from repetitive structure and fuzzy features. *CorNet* homography estimation outperforms both

ASIFT and other CNNs over a variety of UAV movements and landscapes. While UDH and our retrained version of it are much faster, their results are far less accurate than *CorNet* in nearly all cases. Finally, *CorNet* generalizes well to non-maize landscapes.

Acknowledgments. We are grateful for the support of the NSF Midwest Big Data Hub Digital Agriculture Spoke, the Missouri Maize Center, the National Corn Growers Association, the U.S. Army Research Laboratory (cooperative agreement W911NF1820285), the Army Research Office (DURIP W911NF1910181), an Executive Women's Forum doctoral fellowship through the University of Missouri College of Engineering (to R.A.), and an anonymous gift. We thank Shizeng Zhou, Behirah Hartranft, Steven Suddarth, and Vinny for helpful discussions. We are especially grateful to the referees for their very helpful comments. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors' and do not necessarily reflect the views of the U. S. Government or any agency thereof. The authors declare no conflict of interest.

References

1. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous distributed systems (2016). https://arxiv.org/abs/1805.09662
2. Agarwal, S., et al.: Building Rome in a day. Commun. ACM **54**(10), 105–112 (2011)
3. Aktar, R., Aliakbarpour, H., Bunyak, F., Seetharaman, G., Palaniappan, K.: Performance evaluation of feature descriptors for aerial imagery mosaicking. In: Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–7. IEEE (2018)
4. Aktar, R., Prasath, V.S., Aliakbarpour, H., Sampathkumar, U., Seetharaman, G., Palaniappan, K.: Video haze removal and Poisson blending based mini-mosaics for wide area motion imagery. In: Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–7. IEEE (2016)
5. Aktar, R., et al.: Robust mosaicking of maize fields from aerial imagery. Appl. Plant Sci. **8**, e11387 (2020). https://doi.org/10.1002/aps3.11387, https://bsapubs.onlinelibrary.wiley.com/doi/full/10.1002/aps3.11387
6. Aliakbarpour, H., Palaniappan, K., Seetharaman, G.: Stabilization of airborne video using sensor exterior orientation with analytical homography modeling. In: Sergiyenko, O., Flores-Fuentes, W., Mercorelli, P. (eds.) Machine Vision and Navigation, pp. 579–595. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-22587-2_17
7. Avola, D., Cinque, L., Foresti, G.L., Martinel, N., Pannone, D., Piciarelli, C.: A UAV video dataset for mosaicking and change detection from low-altitude flights. IEEE Trans. Syst. Man Cybern. Syst. **PP**, 1–11 (2018)
8. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_32
9. Bentoutou, Y., Taleb, N., Kpalma, K., Ronsin, J.: An automatic image registration for applications in remote sensing. IEEE Trans. Geosci. Remote Sens. **43**(9), 2127–2137 (2005)
10. Blancon, J., et al.: A high-throughput model-assisted method for phenotyping maize green leaf area index dynamics using unmanned aerial vehicle imagery. Fron. Pl. Sci. **10**, 685 (2019). https://doi.org/10.3389/fpls.2019.00685

11. Brown, L.G.: A survey of image registration techniques. ACM Comput. Surv. (CSUR) **24**(4), 325–376 (1992)
12. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. Int. J. Comput. Vis. **74**(1), 59–73 (2006). https://doi.org/10.1007/s11263-006-0002-3
13. Chen, H.M., Arora, M.K., Varshney, P.K.: Mutual information-based image registration for remote sensing data. Int. J. Remote Sens. **24**(18), 3701–3706 (2003)
14. Condorelli, G.E., et al.: Comparative aerial and ground based high throughput phenotyping for the genetic dissection of NDVI as a proxy for drought adaptive traits in durum wheat. Frontiers Plant Sci. **9** (2018). https://doi.org/10.3389/fpls.2018.00893
15. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. arXiv preprint arXiv:1606.03798 (2016)
16. Enciso, J., et al.: Validation of agronomic UAV and field measurements for tomato varieties. Comput. Electron. Agric. **158**, 278–283 (2019). https://doi.org/10.1016/j.compag.2019.02.011
17. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In: Readings in Computer Vision, pp. 726–740. Elsevier (1987)
18. Gao, K., AliAkbarpour, H., Palaniappan, K., Seetharaman, G.: Evaluation of feature matching in aerial imagery for structure-from motion and bundle adjustment. In: Geospatial Informatics, Motion Imagery, and Network Analytics VIII, vol. 10645, p. 106450J. International Society for Optics and Photonics (2018)
19. Gnädinger, F., Schmidhalter, U.: Digital counts of maize plants by unmanned aerial vehicles (UAVs). Remote Sens. **9**(6), 544 (2017)
20. Gracia-Romero, A., Kefauver, S.C., Fernandez-Gallego, J.A., Vergara-Díaz, O., Teresa Nieto-Taladriz, M., Araus, J.L.: UAV and ground image-based phenotyping: a proof of concept with durum wheat. Remote Sens. **11**(10), 1244 (2019). https://doi.org/10.3390/rs11101244
21. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2004)
22. Huang, Y., Thomson, S.J., Hoffmann, W.C., Lan, Y., Fritz, B.K.: Development and prospect of unmanned aerial vehicle technologies for agricultural production management. Int. J. Agric. Biol. Eng. **6**(3), 1–10 (2013)
23. Jain, P.M., Shandliya, V.: A review paper on various approaches for image mosaicing. Int. J. Comput. Eng. Res. **3**(4), 106–109 (2013)
24. Johansen, K., Morton, M.J.L., Malbeteau, Y.M., Aragon, B., Al-Mashharawi, S.K., Ziliani, M.G., Angel, Y., Fiene, G.M., Negrão, S.S.C., Mousa, M.A.A., Tester, M.A., McCabe, M.F.: Unmanned Aerial Vehicle-based phenotyping using morphometric and spectral analysis can quantify responses of wild tomato plants to salinity stress. Frontiers Plant Sci. **10**, 370 (2019). https://doi.org/10.3389/fpls.2019.00370
25. Kaess, M., Johannsson, H., Roberts, R., Ila, V., Leonard, J.J., Dellaert, F.: iSAM2: incremental smoothing and mapping using the Bayes tree (2011). https://doi.org/10.1177/0278364911430419, https://arxiv.org/abs/1908.02002v1
26. Kang, L., Wei, Y., Xie, Y., Jiang, J., Guo, Y.: Combining convolutional neural network and photometric refinement for accurate homography estimation. IEEE Access, pp. 109460–109473 (2019)
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Kriegman, D.: Homography estimation. Lecture Computer Vision I, CSE a 252 (2007)

29. Le Moigne, J., Netanyahu, N.S., Eastman, R.D.: Image Registration for Remote Sensing. Cambridge University Press, Cambridge (2011)
30. Li, Q., Wang, G., Liu, J., Chen, S.: Robust scale-invariant feature matching for remote sensing image registration. IEEE Geosci. Remote Sens. Lett. **6**(2), 287–291 (2009)
31. Lin, Y., Medioni, G.: Map-enhanced UAV image sequence registration and synchronization of multiple image sequences. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007, pp. 1–7. IEEE (2007)
32. López-Granados, F., Torres-Sánchez, J., De Castro, A.I., Serrano-Pérez, A., Mesas-Carrascosa, F.J., Peña, J.M.: Object-based early monitoring of a grass weed in a grass crop using high resolution UAV imagery. Agron. Sustain. Dev. **36**(4), 67 (2016)
33. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
34. Molina, E., Zhu, Z.: Persistent aerial video registration and fast multi-view mosaicing. IEEE Trans. Image Process. **23**(5), 2184–2192 (2014)
35. Morel, J.M., Yu, G.: ASIFT: a new framework for fully affine invariant image comparison. SIAM J. Imaging Sci. **2**(2), 438–469 (2009)
36. Nasir, A.K., Tharani, M.: Use of Greendrone UAS system for maize crop monitoring. In: ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 263–268 (2017)
37. Nguyen, T., Chen, S.W., Shivakumar, S.S., Taylor, C.J., Kumar, V.: Unsupervised deep homography: a fast and robust homography estimation model. IEEE Rob. Autom. Lett. **3**(3), 2346–2353 (2018)
38. Ono, Y., Trulls, E., Fua, P., Yi, K.M.: LF-Net: learning local features from images. arXiv preprint arXiv:1805.09662v2 (2018)
39. Pohl, C., Van Genderen, J.L.: Review article multisensor image fusion in remote sensing: concepts, methods and applications. Int. J. Remote Sens. **19**(5), 823–854 (1998)
40. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to sift or surf. In: 2011 International Conference on Computer Vision, pp. 2564–2571. IEEE (2011). https://doi.org/10.1109/ICCV.2011.6126544
41. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)
42. Seetharaman, G., Palaniappan, K., Akbarpour, H.A.: Method for fast camera pose refinement for wide area motion imagery (2019). U.S. Patent 9,959,625
43. Shi, Y., et al.: Unmanned Aerial Vehicles for high-throughput phenotyping and sgronomic research. PLoS ONE **11**(7), e0159781 (2016). https://doi.org/10.1371/journal.pone.0159781
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
45. Teters, E., AliAkbarpour, H., Palaniappan, K., Seetharaman, G.: Real-time geoprojection and stabilization on airborne GPU-enabled embedded systems. In: Geospatial Informatics, Motion Imagery, and Network Analytics VIII, vol. 10645, p. 106450H. International Society for Optics and Photonics (2018)
46. Wang, X., et al.: Dynamic plant height QTL revealed in maize through remote sensing phenotyping using a high-throughput unmanned aerial vehicle (UAV). Sci. Rep. **9**(1) (2019). https://doi.org/10.1038/s41598-019-39448-z
47. Wirth, M.A.: Shape Analysis & Measurement. University of Guelph (2004). http://www.cyto.purdue.edu/cdroms/micro2/content/education/wirth10.pdf

48. Woodward Crossings: Jang TD-1 Push Planter. Woodward Crossings (2019)
49. Wu, C.: Towards linear-time incremental structure from motion. In: 2013 International Conference on 3D Vision, pp. 127–134 (2013)
50. Zhu, Z., Riseman, E.M., Hanson, A.R., Schultz, H.: An efficient method for georeferenced video mosaicing for environmental monitoring. Mach. Vis. Appl. **16**(4), 203–216 (2005)
51. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)