Neural Computing and Applications (2021) 33:7421–7440
https://doi.org/10.1007/s00521-021-05926-7

S.I. : DICTA 2019

Fixed-Lens camera setup and calibrated image registration for multifocus multiview 3D reconstruction

Shah Ariful Hoque Chowdhury[1] · Chuong Nguyen[2] · Hengjia Li[3] · Richard Hartley[1]

Received: 4 May 2020 / Accepted: 11 March 2021 / Published online: 6 April 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract
Image-based 3D reconstruction or 3D photogrammetry of small-scale objects including insects and biological specimens is challenging due to the use of a high magnification lens with inherently limited depth of field, and the object's fine structures. Therefore, the traditional 3D reconstruction techniques cannot be applied without additional image preprocessing. One such preprocessing technique is multifocus stacking/fusion that combines a set of partially focused images captured at different distances from the same viewing angle to create a single in-focus image. We found that the image formation is not properly considered by the traditional multifocus image capture and stacking techniques. The resulting in-focus images contain artifacts that violate the perspective projection. A 3D reconstruction using such images often fails to produce accurate 3D models of the captured objects. This paper shows how this problem can be solved effectively by a new multifocus multiview 3D reconstruction procedure which includes a new Fixed-Lens multifocus image capture and a calibrated image registration technique using analytic homography transformation. The experimental results using the real and synthetic images demonstrate the effectiveness of the proposed solutions by showing that both the fixed-lens image capture and multifocus stacking with calibrated image alignment significantly reduce the errors in the camera poses and produce more complete 3D reconstructed models as compared with those by the conventional moving lens image capture and multifocus stacking.

Keywords Fixed-lens multifocus capture · Macroimaging · Multifocus stacking · Multifocus image fusion · Image registration · Perspective image formation · Multiview stereo · 3D reconstruction · Multifocus multiview reconstruction · Image-based 3D reconstruction · Insects · Small objects · Small specimens

1 Introduction

3D reconstruction of small objects including insects and biological specimens is challenging due to the use of a high magnification lens with limited depth of field, fine features, and complex surface properties. Recent advancements [15, 39, 47, 49] show that the photogrammetry or image-based multiview 3D reconstruction can be applied with some success to create the true-colour 3D models of small specimens.

The solutions enabling the images of small specimens (a few centimeters or smaller) to be reconstructed in 3D generally include a two-axis turntable combined with macrorail and macrophotography to capture multifocus multiview images [39], calibration target [39, 49], multi-focus image fusion/stacking [57], scale-shift calibration and automatic background masking [49], and a multiview 3D reconstruction [1–3, 45, 46].

Despite multiple techniques to tackle different issues of image-based 3D reconstruction of small specimens, obtaining an accurate 3D model is still difficult. Our preliminary study [26] showed that there is a neglected source of the error caused by the multifocus image stacking that

✉ Shah Ariful Hoque Chowdhury
shah.chowdhury@anu.edu.au

Chuong Nguyen
chuong.nguyen@data61.csiro.au

Hengjia Li
hengjia.li.19@ucl.ac.uk

Richard Hartley
richard.hartley@anu.edu.au

[1] Australian National University, Canberra, Australia

[2] CSIRO Data61, Canberra, Australia

[3] University College London, London, England

needed to be accounted for. To tackle this problem, [26] proposed a Fixed-Lens multifocus image capture technique with fitted homography-based image alignment/registration.

This paper is an extension of our preliminary study [26]. We describe the fixed-lens multifocus capture setup. We further improve the image registration with a new analytic formulation of homography transformation for the multifocus images. Finally, we perform a more detailed comparison with state-of-the-art techniques.

The remainder of this paper is organized as follows. Section 2 contains the review of the relevant literature. An overview of the concept of the pinhole camera model, the problem of perspective inconsistency, depth of focus and multifocus stacking is introduced in Sect. 3. Section 4 describes our proposed method fixed-lens image acquisition setup. Formulations of the required shift and scale to construct homography transformation is presented in Sect. 5. In Sect. 6, a brief overview of the Laplacian pyramid-based image fusion algorithm to fuse the aligned images of a single stack into an in-focus image and background masking to remove the background of the fused image is discussed. Section 7 contains the 3D reconstruction results showing the effectiveness of the proposed method versus the conventional method. Finally, Sect. 8 concludes the paper with our findings. The codes and data will be released on GitHub[1].

2 Related work

Image-based multiview 3D reconstruction procedure [18, 46, 52] utilizes the images of a scene/object captured from multiple viewpoints to reconstruct the scene/object in 3D. However, 3D reconstruction of small-scale objects is different from the conventional 3D reconstruction mainly in terms of image acquisition and preparation. Instead of conventional image acquisition, multifocus image acquisition is required for small-scale objects because of the shallow depth of field of high magnification lens (macro-lens) used to capture the images. Consequently, an additional image preprocessing step, namely multifocus stacking, is also required to combine the multifocus images before feeding them to a conventional 3D reconstruction algorithm.

Multifocus images can be synthesized from light field images for dense light field reconstruction [43, 44]. Multifocus images are obtained by weighted summation of shifted multiview images captured by a light field/plenoptic camera. These synthesized multifocus images can also be used to estimate 3D scene flow [14]. For dense light field synthesis, Kodama and Kubota [23] propose reconstructing an in-focus image from the multifocus images obtained by shifting the pinhole. Light field images can be used directly for 3D reconstruction as well [36, 48, 60, 62].

However, light field reconstruction is not the same as multifocus multiview reconstruction. Light field images are captured from slightly different (shifted) viewpoint but from the same direction, whereas in multifocus multiview reconstruction, multiview images are captured from different directions and different distances between the camera and the object.

Several investigations have been performed to reconstruct 3D models of small specimens from multifocus multiview images. For detailed 3D surface/image reconstruction, structured light scanning is used by Geng [16], whereas Ritz et al. [41] used lens-shifted structured light. Real-time 3D image reconstruction from multifocus images has been proposed by Kodama et al. [24] using efficient linear filtering with multidimensional symmetry. These methods did not consider 3D reconstruction from multiple views.

Gallo et al. [15] and Silvester and Hillson [47] used structure-from-motion (SfM)-based multifocus multiview 3D reconstruction for small biological specimens. Nguyen et al. [39] and Ströbel et al. [49] investigated 3D reconstruction of small insects from multifocus multiview images applying SfM with shape-from-silhouette and multiview-stereo, respectively. However, all these works utilize conventional multifocus image capture and stacking technique. Hence, these studies do not properly address the perspective distortion caused by the conventional multifocus image capture and fusion technique.

Multifocus image fusion is very crucial for multifocus multiview 3D reconstruction. A large number of multifocus fusion algorithms have been proposed so far, and several review articles can be found in the literature [29, 33, 35, 57]. The fusion methods can be classified into two groups: (a) the spatial domain methods [6, 7, 11, 27, 28, 31, 32, 61] and (b) the transform domain methods [9, 21, 23, 25, 38, 53, 56, 59].

Most of the spatial domain image fusion methods are fast and easy to implement compared with transform-based methods. But spatial domain methods depend heavily on the accuracy of the optimal weight/decision map estimation. In general, transform domain methods achieve superior performance compared to spatial domain methods at the cost of computational complexity.

To overcome the complexity of manually designing the feature extraction and fusion rules, recently deep learning-based image fusion approaches have been introduced [4, 17, 20, 37, 40, 55, 58]. All the deep learning-based fusion techniques take a pair of partially focused images as

[1] https://github.com/chuong/multifocus_multiview_stereo_reconstruction.

input to produce an in-focus image as the output and not useful when the number of images in the stack is higher than two.

As the input images could contain some misalignments, image registration before image fusion is necessary. A common practice is to compute the homography transformations between the images with a reference image by direct registration or feature-based registration [5, 51, 63]. Recent state-of-the-art multifocus fusion techniques with misalignment consideration include the methods proposed by Liang et al. [30] with feature-based registration and Ji et al. [22] with optimization robust to misalignment. However, these techniques do not account for the image scaling effect, therefore limiting their use for the more general imaging conditions.

Most multifocus fusion algorithms aim to produce good looking images without considering further applications such as 3D reconstruction. These good looking stacked in-focus images often contain artifacts due to the parallax error between the images in a stack. This leads to unreliable camera pose estimation for 3D reconstruction and consequently, an inaccurate 3D model.

To deal with unreliable estimation of the camera poses, Ströbel et al. [49] precomputed the camera poses using a textured sphere for 3D reconstruction. However, this workaround solution restricts the predetermined camera positions. To tackle the parallax error, Ströbel et al. [49] proposed to use a calibration target to estimate the scale and shift components for image registration. Although the scale-shift image registration partially corrects the parallax error, the perspective projection in a stacked in-focused image is not guaranteed. This violates the pinhole projection model used by the most multiview 3D reconstruction algorithms, and it causes poor 3D reconstruction.

The Fixed-Lens multifocus image capture technique with fitted homography-based image alignment proposed by Li and Nguyen [26] aims to preserve the perspective consistency. The fixed-lens image capture provides the images with the same perspective projection. The fixed-lens multifocus images can be registered perfectly using a homography transformation as there is no parallax error. A calibration target is used to precompute the homography transformations for the image registration.

In this paper, as an extension of the method proposed by Li and Nguyen [26], we combine the fixed-lens image capture with the proposed image registration using the analytic homography transformation formulated in Sect. 5 to further improve the accuracy of the estimated camera poses and reconstructed 3D models.

3 Perspective image formation and depth of focus effect

3.1 Pinhole camera and perspective image formation

The first camera invented was the pinhole camera as shown in the top of Fig. 1 where the pinhole C is the center of image formation allowing the light rays to pass straight through. Modern cameras use lenses instead of the pinhole to improve the image quality, but it adds the depth of focus where a film or image sensor is located to capture a clear in-focus image. Due to the image capture process of a camera, a transformation between the 3D world to the 2D image happens where the depth is missing in the captured image. Furthermore, different parts in the scene appear at different scales or magnifications depending on their distances to the camera lens. This is demonstrated by the bottom of Fig. 1 where the images of the same object are captured at two different distances. As the camera lens moves, the center C of image formation moves. This leads to the changes in the relative distances (and magnifications) between parts of the scene to the camera center. As a result, the change of the relative scales of different parts of the scene causes different perspective distortions.

3.2 Depth of focus, depth of field and multifocus stacking

Due to the use of lenses in modern cameras, image quality is significantly better, but the depth of focus is introduced in the image. As shown in the top of Fig. 2, an object at distance d_0 forms an image at distance $d'_0 = \frac{d_0 f_L}{d_0 - f_L}$, where f_L is the lens focal length. If an image sensor is placed at this distance (i.e. at the back focal plane), the image will be clear and in focus.

In reality, an image is considered in focus if the image of a point remains smaller than a circular dot called the circle of confusion (COC) with an empirical diameter ϕ_{coc}. From [19], ϕ_{coc} is chosen approximately 0.1% of the mean of the width and height of the image sensor. For example, for a 35mm sensor format, ϕ_{coc} is chosen to be 0.025mm. Strictly speaking, ϕ_{coc} should be selected as the larger value of the size of an image pixel and the optical resolution of the lens.

The distance range where the image remains in focus is called the depth of focus DoFocus. Using the proportional relationship, one can obtain the depth of focus as:

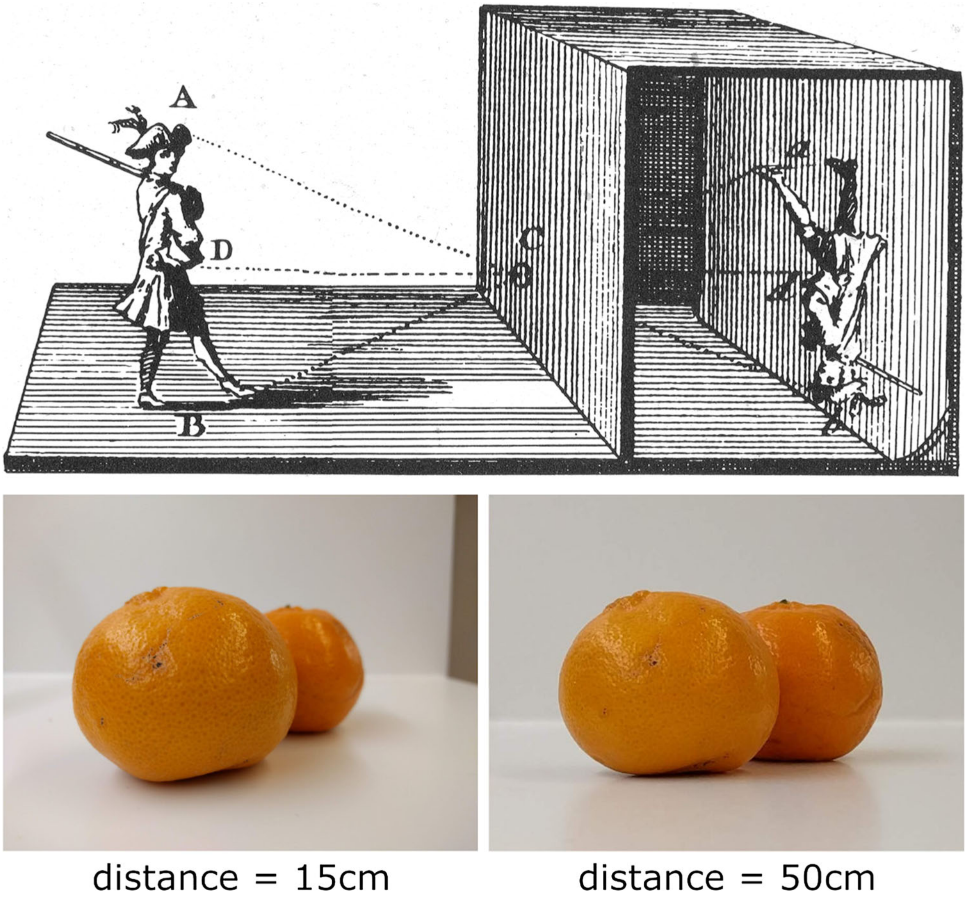


Fig. 1 Top: pinhole camera principle and perspective projection [50] where all the straight light rays go through this hole (image in the public domain). The pinhole represents the camera center for image formation. Bottom: the pair of oranges captured by the same camera of 5mm focal length at different distances leading to different perspective image formations (or distortions). (Best viewed in colour on the screen)

$$\frac{\text{DoFocus}}{2d'_0} = \frac{\phi_{\text{coc}}}{\phi_a}$$

$$\Rightarrow \text{DoFocus} = 2\frac{\phi_{\text{coc}}}{\phi_a}d'_0 \tag{1}$$

$$\text{or DoFocus} = 2\frac{f_{\text{number}}\phi_{\text{coc}}}{f_L}d'_0$$

where ϕ_a is the diameter of the lens aperture, and $f_{\text{number}} = f_L/\phi_a$.

This depth of focus translates to the depth of field DoField where the objects stay within so that their images are in-focus. Assuming the circle of confusion ϕ_{coc} is backprojected into the scene to the size ϕ_{coc}/M where $M = d'_0/d_0$ is the lens magnification, the proportional relationship gives:

$$\frac{\text{DoField}}{2d_0} = \frac{\phi_{\text{coc}}}{M\phi_a}$$

$$\Rightarrow \text{DoField} = 2\frac{\phi_{\text{coc}}}{M\phi_a}d_0 \tag{2}$$

$$\text{or DoField} = 2\frac{d_0^2\phi_{\text{coc}}}{d'_0\phi_a} = 2\frac{d_0(d_0 - f_L)\phi_{\text{coc}}}{f_L\phi_a}$$

To capture the scenes at different distances, the position of the lens and/or the image sensor needs to be adjusted to put the image into focus.

For an object with the parts at different distances, only some parts are in focus and the other parts are out of focus. This is especially prevalent in macroimaging where the size of the scene is tens of centimeter or smaller and a high magnification lens is used. Figure 3 shows the images captured as a camera (and lens) moves toward the specimen at constant incremental depth at the same viewing angle. Multifocus image stacking produced an in-focus image representing that single viewing angle.

The conventional multifocus capture is to move the camera and lens together. The step size of the macrorail movement is equivalent to the depth of field. However, the recommended macrorail step size is 50% of the depth of field in this case to allow 50% overlapping between the in-focus regions of the successive images to help guide the image registrations for the optimization-based multifocus stacking algorithms.

By using such multifocus stacked images captured at different angles, the 3D reconstruction of small objects becomes possible as reported by [15, 39, 47, 49]. However, such in-focus stacked images do not correctly represent the perspective image formation, leading to reconstruction errors or failures. This becomes more severe when the size of the objects becomes smaller, i.e. a few centimeters long or smaller.

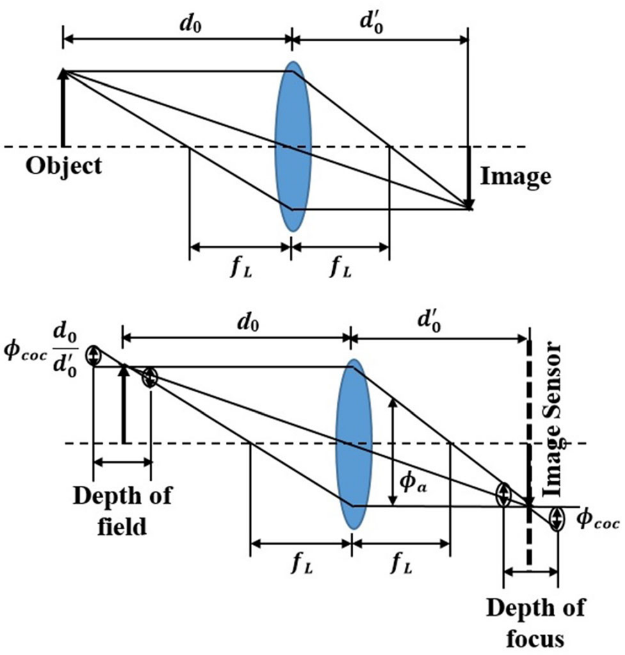


Fig. 2 Top: the image formation through a lens of focal length f_L. After passing the lens, the rays parallel to the optical axis go through the focal point and the rays going through the focal point become parallel to the optical axis. Bottom: the image sensor is used to capture an image of an object at a position within the depth of focus where a point source grows into a circle of confusion [19] of diameter ϕ_{coc}. A captured image is considered as out of focus if the image sensor is placed outside the depth of focus or the object is located outside of the depth of field

4 Fixed-Lens Multifocus Capture

The conventional camera setup captures the multifocus images by moving the camera and lens together as shown in the top of Fig. 4. As the whole camera moves, the camera center moves, and the perspective view moves with it. To avoid moving the perspective view of the camera, we propose that the lens is fixed and only the image sensor moves during scanning as shown in the bottom of Fig. 4. In this case, the center of perspective image formation does not change. Although the image size of the whole scene becomes larger when the sensor is moving away from the lens, the relative scales or magnifications of different parts of the scene stay the same.

It is worth mentioning that, for the fixed-lens multifocus capture, the step size of the image sensor movement (macrorail step size) is equivalent to the depth of focus, and similar to the conventional multifocus capture, the recommended macrorail step size is 50% of the depth of focus to allow 50% overlapping between the in-focus regions of the successive images.

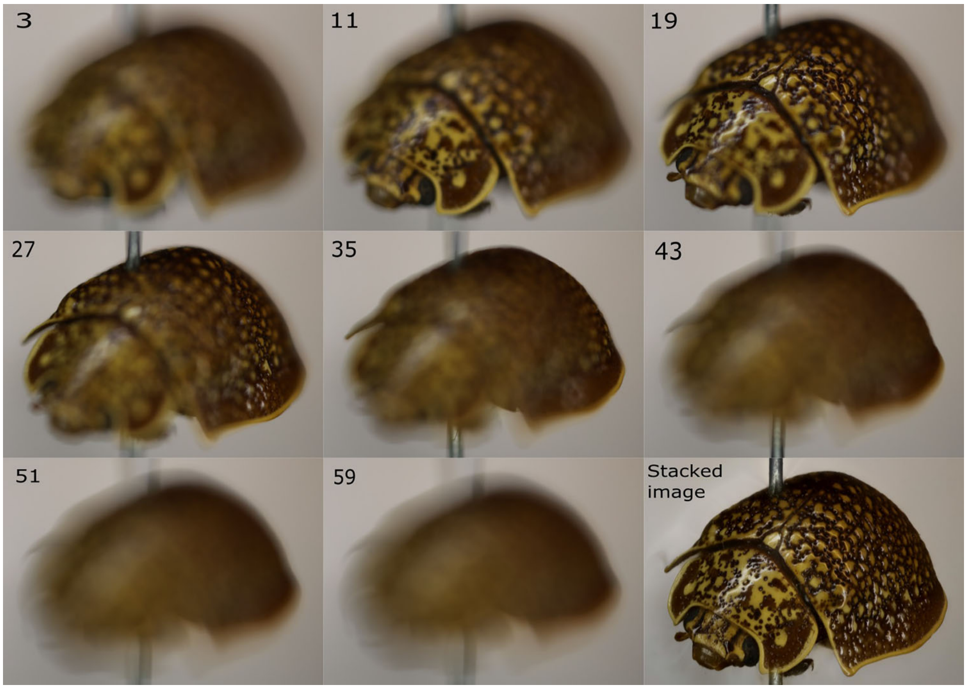


Fig. 3 A set of 61 partially out-of-focus images (only eight of them shown here) captured at different camera positions are stacked to produce an in-focus image. The stacked image, however, does not represent a single perspective image formation due to the moving lens. (Best viewed in colour on the screen)

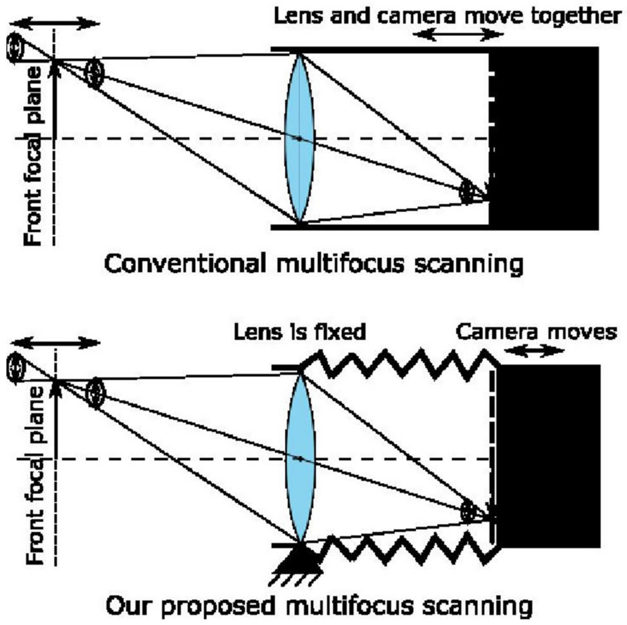

Lens and camera move together
Front focal plane
Conventional multifocus scanning
Lens is fixed
Camera moves
Front focal plane
Our proposed multifocus scanning

Fig. 4 The conventional multifocus capture (top) and the proposed Fixed-Lens Multifocus Capture (bottom) for multifocus image capture. For the conventional scanning, the movement of the camera and lens is the same as that of the front focal plane. For the proposed scanning, the movement of the camera is smaller than that of the front focal plane if the lens magnification is smaller than 1, and vice versa

An example of the camera lens setup for the proposed fixed-lens multifocus capture is shown in the right of Fig. 5 as compared to the conventional Moving-Lens Capture (left). For the Fixed-Lens type, the lens is fixed to the upper frame while the camera body is moved by a macrorail to capture multiple partially focused images. A rubber duct connects the lens and the camera body to prevent ambient light and dust from entering the camera sensor chamber. A small flower is mounted on a two-axis turntable to capture the images at different pan-tilt angles.

5 Image alignment for multifocus stacking

Images captured at different distances have different magnifications or scales. Furthermore, the direction of the camera's movement is not generally aligned with the lens optical axis, causing a relative shift in the image. To account for such changes, camera calibration is performed to measure the relative scale change and shift from one image to another. This can be carried out by capturing multifocus images of a known calibration target placed perpendicular to the optical axis of the lens. The relative scale and shift can be measured from the relative positions of the control points (circles) between the images. Once the relative scales and shifts of the scene in the partially focused images are determined, the image mapping is applied to remove the scale change and shift between the images captured from the same viewing angle.

Unlike the methods proposed by [26, 49] where the calibration images are used to extract the scale shift or homography transformation directly from the multifocus images of a calibration target, we propose in the following sections that analytic homography transformations can be obtained with additional constraints from the macrorail displacement. Our method, therefore, generates more consistent transformations as it is less affected by the out-of-focus effect of the multifocus images of a calibration target.

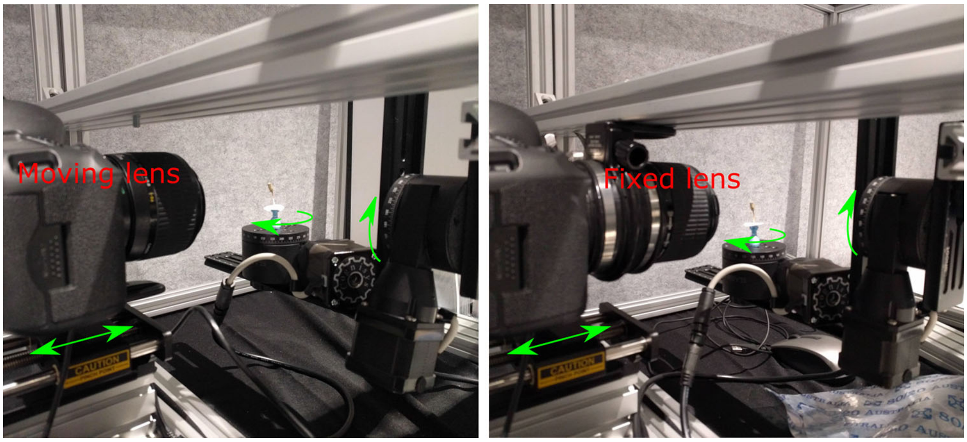


Fig. 5 The lens and pan-tilt setups capture the multifocus multiview images of small objects. The camera is mounted on a macrorail fixed to the lower frame for the multifocus capture. A macrolens is attached to a camera for the conventional moving lens setup (left) or to a stationary frame for our proposed fixed-lens setup (right). For the fixed-lens setup, an expandable rubber duct connects the macrolens and the camera body. The object/specimen is pinned to a fridge magnet on a pan-tilt turntable for the multiview capture. (Best viewed in colour on the screen)

5.1 Analytic formulation of homography transformation for the moving lens

We used a laser printed dot pattern as the calibration target. For the scale-shift calibration, multiple images of the same target are captured at different depths, d_j where $j = 0$ to $j_{\max}$ as shown in Fig. 6. For simplicity, it is assumed that all distances (including depth) are measured in the pixel coordinates throughout the manuscript unless mentioned otherwise. The linear step size of the camera movement along the macrorail is denoted as Δd. An image, I_j of the calibration target is captured from a camera position S_j located at depth d_j from the target.

The aim of the calibration is to determine the camera intrinsics (camera matrix and distortion coefficients) as well as the shifts and scales between the images captured at different depths. Also, note that zero distortion is assumed in the image alignment stage to keep the scale-shift formulation simpler. However, the 3D reconstruction stage takes the distortion into account.

5.1.1 Estimation of the magnification and camera matrix

As a laser printed dot pattern is used, the physical distances between the dots of the calibration target can be measured in millimeters and these distances can then be converted into pixels using the pixel size of the camera.

Let ΔX and ΔY be the horizontal and vertical distances between the centers of the dots in the world coordinates. We assume that ΔX remains the same between any two neighboring centers of the dots of the same row which also holds for all rows as well and ΔY remains the same between any two neighboring centers of the dots of the same column which also holds for all columns. In other words, the object (calibration target in this case) plane is assumed to be perpendicular to the optical axis and parallel to the image plane.

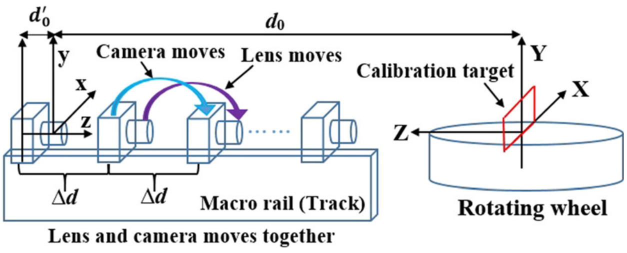


Fig. 6 Image acquisition for calibration. Images of a dot pattern target are captured at different depths using the moving lens camera setup

Let, $I_{j=0}$ be the in-focus image of the calibration target captured with the above-mentioned assumption where the horizontal and vertical distances between the centers of the dots in the image coordinates become Δu and Δv, respectively. Now assume that the calibration target is located at depth d_0 and imaged at depth d'_0. The relation between d_0 and d'_0 can be expressed using the thin-lens equation as follows

$$\frac{1}{d_0} + \frac{1}{d'_0} = \frac{1}{f_L} \tag{3}$$

where f_L is the focal length of the lens which is known from the camera specification.

Now, the magnification M can be defined by

$$M = \frac{d'_0}{d_0} = \frac{\Delta u}{\Delta X} = \frac{\Delta v}{\Delta Y} \tag{4}$$

Putting the expression $d'_0 = Md_0$ into (3), we get

$$\frac{1}{d_0} + \frac{1}{Md_0} = \frac{1}{f_L}$$

Hence, the depth d_0 becomes

$$d_0 = \frac{f_L(M+1)}{M} \tag{5}$$

So, d'_0 can be written as

$$d'_0 = f_L(M+1) \tag{6}$$

Here, d'_0 represents the focal length of the camera, f_C. For the conventional image captures, the object is relatively far from the camera, i.e. $d_0 \longrightarrow \infty$, making $d'_0 \longrightarrow f_L$. However, for the macrophotography, d_0 is relatively small and f_L cannot be used as an approximation for d'_0.

So, the camera matrix, K can be expressed assuming the camera center is at the image center as

$$K = \begin{bmatrix} f_C & 0 & C_x \\ 0 & f_C & C_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} d'_0 & 0 & w/2 \\ 0 & d'_0 & h/2 \\ 0 & 0 & 1 \end{bmatrix}$$

where $f_C = d'_0$ is the focal length of the camera, $(C_x = w/2, C_y = h/2)$ is the principal point of the camera assuming that the optical center coincides with the image center, w and h are the width and height of the image.

Thus, an initial estimate of the camera matrix is obtained. Later, the estimated camera matrix can be refined by minimizing the reprojection error through an optimization procedure that seeks to optimize the magnification, M, the principal point (C_x, C_y) and other calibration parameters.

5.1.2 Shift and scale formulation for analytic homography

Since d_0 represents the distance between the target and the camera located at S_0, we can calculate the distances d_j between the target and the other camera positions S_j as

$$d_j = d_0 - j\Delta d$$

where $j = 0, 1, \ldots, j_{\max}$ and Δd is the linear displacement per step of the camera toward the target. The negative sign indicates the decreasing distances between the camera and the target as the camera is moving toward the target.

Now, the coordinates of a pixel p_0 (of the image I_0 taken from the camera located at d_0 distance from the target) with respect to the target coordinate system can be expressed as

$$\begin{aligned} T_p &= [X, Y, Z, 1]^T \\ &= \left[\frac{d_0(u_0 - C_x)}{f_C}, \frac{d_0(v_0 - C_y)}{f_C}, d_0, 1\right]^T \end{aligned}$$

where (u_0, v_0) is the coordinate of the pixel, p_0 in the image coordinate system.

The pixel p_0 of the image I_0 is mapped to pixel p_j of the image I_j taken from the camera located at d_j distance from the target. Let, the coordinate of the pixel p_j in the image coordinate system is (u_j, v_j). The pixel position p_j can be determined as follows

$$[u_j, v_j, 1]^T = K[R \mid t]T_p \tag{7}$$

where R is a 3×3 rotation matrix, t a 3×1 translation vector. This rotation translation matrix $[R \mid t]$ represents the extrinsic parameters of the camera relative to the target. As this calibration involves only the linear motion of the camera, the relative rotation matrix R between different camera positions can be considered as an identity matrix.

However, the translation vector t with the displacement direction vector $[\sigma, \gamma, 1]$ becomes

$$t = [-\sigma j \Delta d, -\gamma j \Delta d, -j \Delta d]^T$$

The displacement direction vector represents misalignment between the macrorail axis along which the camera moves and the camera optical axis which causes an additional lateral shift of the pixel p_j. Misalignment occurs when the macrorail axis is not parallel to the camera optical axis. Now, (7) becomes

$$\begin{aligned} &[u_j, v_j, 1]^T = K[I \mid t]T_p \\ \Rightarrow &\begin{bmatrix} u_j \\ v_j \\ 1 \end{bmatrix} = \begin{bmatrix} f_C & 0 & C_x & -(\sigma j \Delta d f_C + C_x d_j) \\ 0 & f_C & C_y & -(\gamma j \Delta d f_C + C_y d_j) \\ 0 & 0 & 1 & -j\Delta d \end{bmatrix} T_p \end{aligned}$$

The expressions of u_j and v_j can be written as

$$u_j = \frac{d_0(u_0 - C_x) - \sigma j \Delta d f_C}{d_j} + C_x \tag{8}$$

and

$$v_j = \frac{d_0(v_0 - C_y) - \gamma j \Delta d f_C}{d_j} + C_y, \tag{9}$$

respectively.

After rearranging, (8) and (9) become

$$\begin{aligned} u_j &= \alpha(j)u_0 + \beta_u(j, \sigma) \\ &= \frac{d_0}{d_j}u_0 + C_x\left(1 - \frac{d_0}{d_j}\right) - \frac{\sigma j \Delta d f_C}{d_j} \end{aligned}$$

and

$$\begin{aligned} v_j &= \alpha(j)v_0 + \beta_v(j, \gamma) \\ &= \frac{d_0}{d_j}v_0 + C_y\left(1 - \frac{d_0}{d_j}\right) - \frac{\gamma j \Delta d f_C}{d_j}, \end{aligned}$$

respectively, where $\alpha(j)$, $\beta_u(j, \sigma)$ and $\beta_v(j, \gamma)$ represent the required scaling factor, horizontal shift and vertical shift, respectively, to map the pixel p_0 of the image I_0 onto the image I_j and can be expressed as

$$\begin{aligned} \alpha(j) &= \frac{d_0}{d_j} = \frac{d_0}{d_0 - j\Delta d} \\ \beta_u(j, \sigma) &= C_x(1 - \alpha(j)) - \frac{\sigma j \Delta d f_C}{d_j} \end{aligned}$$

and

$$\beta_v(j, \gamma) = C_y(1 - \alpha(j)) - \frac{\gamma j \Delta d f_C}{d_j},$$

respectively.

For the perfect alignment case where the camera displacement axis is parallel to its optical axis, $\sigma = \gamma = 0$. In such case, the horizontal and vertical shifts become $\beta_u(j, \sigma = 0) = C_x(1 - \alpha(j))$ and $\beta_v(j, \gamma = 0) = C_y(1 - \alpha(j))$, respectively. Now, the relationship between the corresponding shifts for the misaligned and perfectly aligned cases can be written as follows

$$\begin{aligned} \beta_u(j, \sigma) &= \beta_u(j, 0) - \frac{\sigma j \Delta d f_C}{d_j} \\ \beta_v(j, \gamma) &= \beta_v(j, 0) - \frac{\gamma j \Delta d f_C}{d_j} \end{aligned}$$

Once the shifts and scale are known, mapping of the image I_0 onto I_j can be done using the following transformation

$$\begin{aligned} &I_j = HI_0 \\ \Rightarrow &\begin{bmatrix} u_j \\ v_j \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha(j) & 0 & \beta_u(j, \sigma) \\ 0 & \alpha(j) & \beta_v(j, \gamma) \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} \end{aligned} \tag{10}$$

where H is the homography transformation function between the front-parallel planes at distances d_0 and d_j from the lens of the first camera position S_0. Thus, a reference image, I_0 can be mapped onto all other images, I_j for $j = 0$ to $j_{\max}$.

However, in multifocus stacking, we are interested in the inverse homography transformation function so that every image can be mapped onto a reference image, say I_0. This inverse transformation is expressed as follows

$$I_0 = H^{-1} I_j \tag{11}$$

The scaling factor $\alpha(j)$ can be computed easily from the known focal length f_L, magnification M, and step size Δd. To estimate σ and γ, the mapping error between the calibration images is minimized.

5.2 Analytic formulation of homography transformation for the fixed lens

For the fixed-lens setup, the lens is fixed at a position while the camera moves. The camera refers to the sensor/focal plane. In this case, the distance between the lens and the calibration target is fixed, while the distance between the lens and the sensor plane varies. In the moving lens case, the lens and the sensor plane (called camera altogether) move together and the same amount. This means the distance between the lens and the sensor plane (which is called the focal length of the camera) is fixed and the distance between the lens and the target varies for the moving lens setup.

Similar to the moving lens case, multiple images of the same target are captured to determine the shift and scale by moving the sensor plane $\Delta d'$ amount per step toward the lens while keeping the lens fixed as shown in Fig. 7. An image I_j is captured with d'_j distance between the sensor plane and the lens where $j = 0, 1, \ldots j_{\max}$.

5.2.1 Estimation of the magnification and camera matrix

The magnification M can be estimated in a similar way as discussed in Sect. 5.1.1 for the moving lens case using (4). Let I_0 be the in-focus image of the target captured with d'_0 distance between the sensor plane and the lens, where the calibration target is located at depth d_0 from the lens. Here, d'_0 represents the focal length of the camera which is the distance between the sensor plane and the lens. The distances, d_0 and d'_0 can be calculated using (5) and (6), respectively, once the magnification is known.

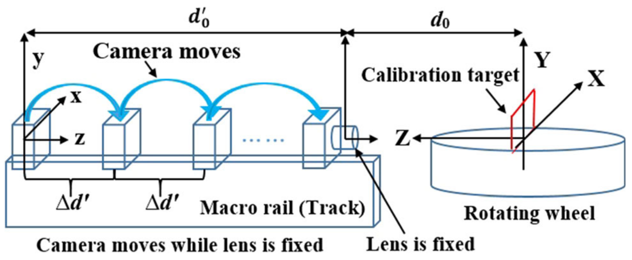


Fig. 7 Image acquisition for calibration. Images of a dot pattern target are captured at different depths using the fixed-lens camera setup

Since d'_0 represents the distance between the sensor plane and the lens, it is possible to compute the distances, d'_j between the lens and the other sensor plane positions as

$$d'_j = d'_0 - j\Delta d'$$

where $j = 0, 1, \ldots, j_{\max}$ and $\Delta d'$ is the linear displacement per step of the sensor plane toward the lens. The negative sign indicates the decreasing distances between the sensor plane and the lens as the sensor plane is moving toward the lens.

As the sensor plane moves $\Delta d'$ amount per step with the displacement direction vector $[\mu, \lambda, 1]$, it causes small lateral movement of the image center with respect to the camera center. This happens due to the misalignment between the sensor plane movement along the macrorail axis and the optical axis of the lens. This misalignment is represented by the displacement direction vector.

Thus, the camera matrix changes for each position of the sensor plane. Let, K_j be the camera matrix corresponds to image I_j captured with d'_j distance between the sensor plane and the lens. It is also assumed that the optical center of the lens is at the image center for the first sensor plane position when $j = 0$. So, K_j becomes

$$K_j = \begin{bmatrix} f_C & 0 & C_x + j\Delta C_x \\ 0 & f_C & C_y + j\Delta C_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} d'_j & 0 & (w/2) + j\mu\Delta d' \\ 0 & d'_j & (h/2) + j\lambda\Delta d' \\ 0 & 0 & 1 \end{bmatrix}$$

where $f_C = d'_j$ is the focal length of the camera, $\Delta C_x = \mu\Delta d'$ and $\Delta C_y = \lambda\Delta d'$ are the horizontal and the vertical shifts of the image center due to the misalignment.

Now, the camera matrices can be estimated by minimizing the reprojection error through an optimization procedure. The calibration parameters required to be estimated are the magnification M, the principal point (C_x, C_y), the horizontal and vertical components of the displacement direction vector μ and λ, respectively.

5.2.2 Shift and scale formulation for analytic homography

For the fixed lens, the relative rotation matrix R between different sensor plane positions can be considered as an identity matrix, while the translation vector t becomes

$$t = [0, 0, 0]^T$$

Now, (7) can be rewritten for the fixed-lens setup as

$$[u_j, v_j, 1]^T = K_j[I \mid t]T_p = \begin{bmatrix} d_j' & 0 & C_x + j\Delta C_x & 0 \\ 0 & d_j' & C_y + j\Delta C_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} T_p$$

Using the expressions of u_j and v_j, we can obtain the shifts and scale for the fixed-lens setup similarly as shown in Sect. 5.1.2. So, for the fixed-lens setup, the scaling factor, $\alpha'(j)$, the horizontal shift, $\beta_u'(j, \mu)$ and the vertical shift, $\beta_v'(j, \lambda)$ required to map the image I_0 onto the image I_j can be expressed as

$$\alpha'(j) = \frac{d_j'}{d_0'} = \frac{d_0' - j\Delta d'}{d_0'}$$
$$\beta_u'(j, \mu) = C_x(1 - \alpha'(j)) + j\mu\Delta d'$$

and

$$\beta_v'(j, \lambda) = C_y(1 - \alpha'(j)) + j\lambda\Delta d' ,$$

respectively.

Now, for the fixed-lens case, the homography transformation between the back-parallel planes at distances d_0 and d_j from the lens can be constructed as shown in (10) to map the image I_0 onto I_j. Inverse transformation can be used as shown in (11) for multifocus stacking. In our experimental setup, we carefully align the lens as well as the image sensor with the macrorail so that the camera displacement axis perfectly aligns with the camera optical axis for both the moving and the fixed-lens setups. Hence, the shifts and scale have been computed without considering the additional lateral shift caused by the misalignment.

For the perfect alignment case, shifts and scale can be computed easily without performing any optimization once the magnification, M is known and with the assumption that the optical center of the lens lies at the image center. Images of a stack captured from the same camera view are aligned using the computed shifts and scale to a reference image position in that stack. Please note that shifts and scales are the same for all camera views as long as the reference image position remains the same in the image stacks. The aligned images of the same stack are then fused together to obtain an in-focus image for that viewpoint.

5.3 Step size selection for the camera movement

It is already known that the recommended linear step size for the camera displacement for the moving lens setup is half of the depth field. However, in our experiment, the chosen step size is almost equal to the depth of field to reduce the computational burden and memory usage. The depth of field can be computed from four parameters using (2). These parameters are the magnification M, the COC ϕ_{coc}, the distance between the lens and the object d_0 and the aperture size of the lens ϕ_a. All parameters can be obtained except the COC.

For the lens (Canon MP-E 65mm f/2.8 1-5x macrophotolens) we used, the manufacturer provided a table describing the depth of field for different magnifications M and f-number. A continuous relation between magnification and depth of field can be obtained using the provided data by fitting an empirical function. COC can also be computed using that same data from the table. In this case, the computed value of the COC is 0.035 mm.

For the fixed-lens setup, the chosen linear step size for the camera displacement is almost equal to the depth of focus to have a reasonable number of captured images, although it is recommended to use half of the depth of focus as the step size. The depth of focus can be obtained using (1) from given COC ϕ_{coc} and ϕ_a. We can compute d_0' for a given magnification M using (6).

6 Image fusion and background masking

In this work, the Laplacian pyramid fusion approach [54] is combined with our proposed calibrated analytic homography for multifocus image stacking. The Laplacian pyramid fusion is based on a multiresolution signal decomposition scheme which produces a high-quality fused image by exploiting the global and local information as well as the spatial and gray information. In this approach, each source image of the stack is decomposed into multilevel images.

For the comparison, stacking with feature-based alignment using the scale-invariant feature transform (SIFT) [34] is chosen as the baseline approach which does not consider the constrained movement of the macrorail. In addition, the stacking results of the calibrated image alignment techniques proposed by Ströbel et al. [49] and Li and Nguyen [26] are also compared with the proposed stacking method. The Laplacian pyramid fusion method is used for blending in all the cases. Our proposed multifocus stacking consists of image alignment using the analytic homography transformation formulated in Sect. 5 and blending those aligned images using the Laplacian pyramid fusion.

The multiview 3D reconstruction optionally accepts the masks or object silhouettes to ignore the image region belong to the background to speed up the process and to improve the reconstruction accuracy. The automatic background segmentation can lead to significant errors and often requires manual input to correct them. To improve the automatic background segmentation, Ströbel et al. [49] proposed to use two images, one in normal front lighting and one with strong back light. The back light is a uniform light source such as a light box to produce a clear contrast between the background area and object area. As a result, the automatic segmentation of the image with the strong back light is very accurate and efficient without the need for manual correction. The drawbacks of capturing the separate images with the back light include doubling of the data storage and the time to capture and preprocess the images, and obtaining the exact same camera positions between the front light and back light. Figures 8 and 9 show the examples of the in-focus foreground images (left), the in-focus back light images (middle) and the final images with background masks (right), for different combinations of the lens setups and multifocus stacking methods.

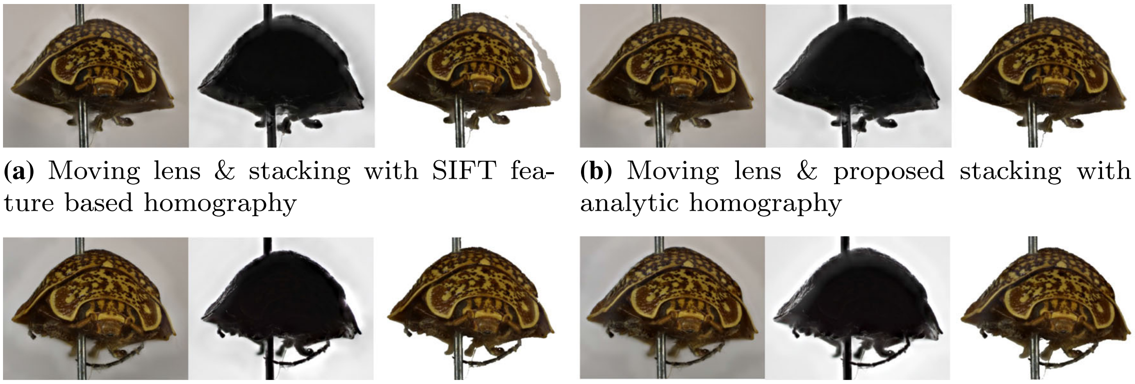
The results of Fig. 8 are shown for the moving lens with focus stacking using SIFT feature-based homography (8(a)), the moving lens with our proposed multifocus stacking approach using analytic homography (8(b)), the fixed lens with stacking using SIFT feature-based homography (8(c)) and our proposed stacking method using analytic homography (8(d)). Similar comparisons using the synthetic images of a different scene are provided in Fig. 9. More blending results with different camera views are provided in the supplementary material.

It is noted that the stacked images look very much similar (except the changes in the brightness due to slightly different lighting conditions for different runs) between the moving lens and fixed lens for different multifocus stacking methods. This shows that the artifacts and distortions are difficult to recognize and they only make a difference when used for the 3D reconstruction.

7 Multiview 3D reconstruction

An overview of the multifocus multiview 3D reconstruction experiment can be summarized as follows:

- Capture a set of multifocus single-view images of a calibration dot target facing perpendicular to the optical axis of the lens.
- Estimate the magnification M from the known size vs the imaged size of the target. For the calibrated image registration, precompute a homography matrix for each image of the calibration target in a stack relative to the reference image in the middle of the stack.
- Capture the multifocus multiview images of an object of interest using either the Moving-Lens setup or the Fixed-Lens setup. For each setup, two images were capture with the front light and again with the back light at each camera position.
- Apply the multifocus stacking methods separately for the images with the front light and the back light.
- Threshold the stacked in-focus image with the back light to create a background mask and add this as a transparent channel to the corresponding stacked image with the front light.
- Obtain the 3D model by feeding the stacked images with the background masks into the 3D reconstruction software. This paper includes the results using an open source software Meshroom of AliceVision [3]. Please refer to the supplementary material for the results using a commercial software Photoscan of Agisoft [2].



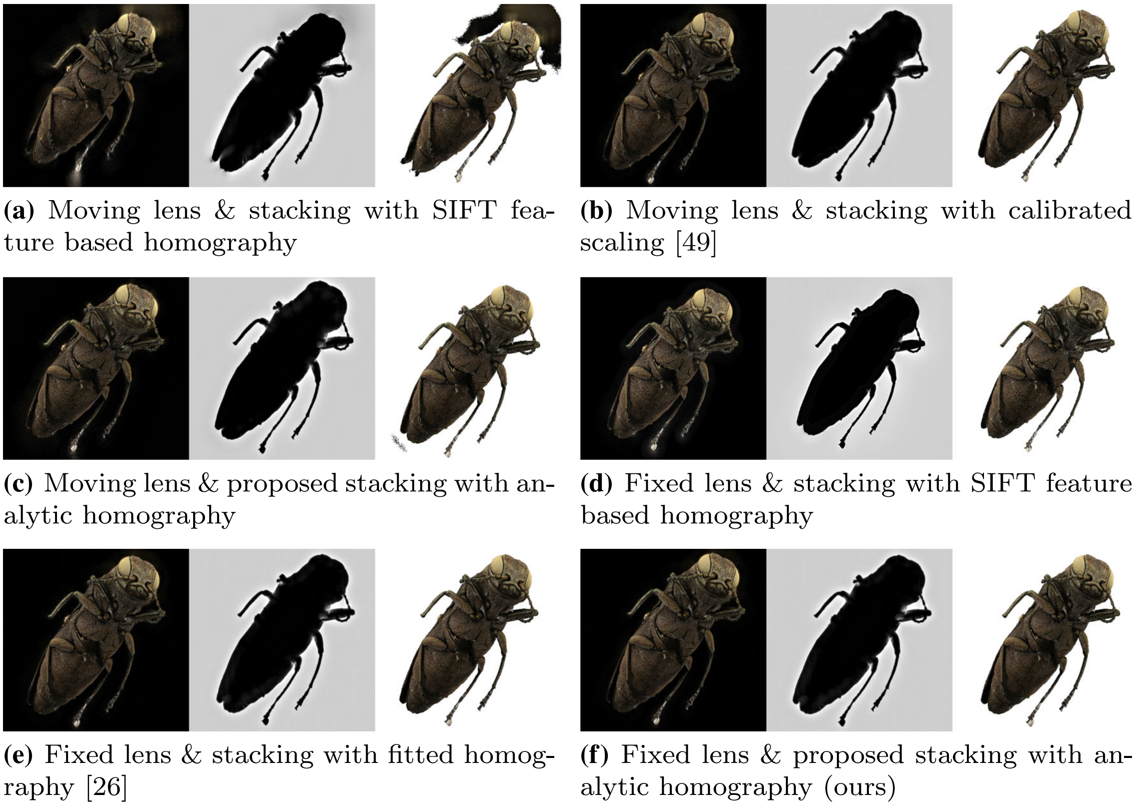
(a) Moving lens & stacking with SIFT feature based homography

(b) Moving lens & proposed stacking with analytic homography

(c) Fixed lens & stacking with SIFT feature based homography

(d) Fixed lens & proposed stacking with analytic homography (ours)

Fig. 8 Comparisons of the in-focus images for different cases. **(a)** moving lens and stacking with SIFT feature-based homography, **(b)** moving lens and proposed stacking with analytic homography, **(c)** fixed lens and stacking with SIFT feature-based homography and **(d)** fixed lens and proposed stacking with analytic homography (ours). For each case, **left:** the blended in-focus image, **middle:** the in-focus mask and **right:** the in-focus image with background mask as the transparent channel. The in-focus image by the fixed lens is sharper than that by the moving lens. All images are cropped to show mostly the specimen. (Best viewed in colour on the screen) (colour figure online)



(a) Moving lens & stacking with SIFT feature based homography

(b) Moving lens & stacking with calibrated scaling [49]

(c) Moving lens & proposed stacking with analytic homography

(d) Fixed lens & stacking with SIFT feature based homography

(e) Fixed lens & stacking with fitted homography [26]

(f) Fixed lens & proposed stacking with analytic homography (ours)

Fig. 9 Comparisons of the in-focus images of *agrilus-anxius* obtained from partially in-focus synthetic images for different cases. For each case, **left:** the blended in-focus image, **middle:** the in-focus mask and **right:** the in-focus image with background mask as the transparent channel. (Best viewed in colour on the screen) (colour figure online)

7.1 Image acquisition setup

Figure 5 shows setups for the multifocus multiview imaging experiment. The camera attached to the macrorail and the object of interest is mounted on a pan-tilt rig (Cognisys StackShot 3X). The camera Canon 5DS is coupled with a macrolens Canon MP-E65mm f/2.8 1-5X. For the moving-lens setup (left), the lens is attached to the camera body and moves together with the camera. For the Fixed-Lens setup (right), the lens is fixed to an upper frame; therefore, the lens remains stationary during capturing multifocus images while the camera body is moving. An expandable rubber duct connects the lens and the camera body to protect the camera sensors from ambient illumination and dust.

In the experiment, an insect was chosen as a scanning target. The two experimental setups were performed on the specimen to capture the multifocus multiview images and to reconstruct the 3D models. Different viewpoints have been created by the pan-tilt rotation of the insect after placing the insect on a stage in front of the camera. The chosen step size for the pan and tilt rotations was 15°. To cover the 360° pan angle view, 24 images have been captured with different pan angles. For each pan angle, five images with different tilt angles have been captured with 15° interval. So, the total number of viewpoints used in this experiment is 120.

The pan-tilt rotation was similar for both the moving and fixed-lens setups. The images were also captured in similar conditions for the two image acquisitions with different lens setups. We used the same exposure time and ISO for both setups. The chosen exposure time and ISO was 1/40s and 200, respectively. However, the magnification was different for different lens setups due to some setup constraints. The magnification for the fixed and the moving lens setup was 1.7 and 1.3, respectively. Image resolution was set to 4320×2880 pixels.

The number of camera positions or the required number of images to be captured for each view depends on the insect size and the magnification of the camera. For the fixed-lens setup, the depth of focus is estimated as 0.53mm where COC, $\phi_{\text{coc}} = 0.035$mm, magnification, $M = 1.7$, focal length of the lens, $f_L = 65$mm and aperture size, $\phi_a = f_L/2.8$. So, the recommended linear step size of the camera movement along the macrorail is 0.265 mm (half of the depth of focus). The maximum dimension (length in this case) of the insect in this experiment is approximately 13mm. Considering the movement of the insect due to the pan-tilt rotation, the total linear distance that the camera needs to move to get the insect in and out of focus for all views is 30mm. Consequently, the required number of camera positions to cover that distance is approximately 113 with the recommended step size, 0.265mm. But, to speed up the process, we chose the step size as 0.5mm which requires 61 images to be captured for each view.

Now, for the moving lens setup, the depth of field is approximately 0.267 mm where the parameters are the same as the fixed-lens setup except the magnification $M(=1.3)$. We captured 61 images with 0.25mm step size of the camera movement for each view instead of 114 images with the recommended step size of 0.133mm to reduce the required number of images to be captured. So, for both fixed and moving lens setups, the total number of images captured is 7320.

To collect more data and evaluate different configurations, the synthetic datasets are generated using Blender [8] with the known 3D models of two specimens *Agrilus Anxius* from Digital Archive of Natural History [12] and *beechnut* from Digital Archive of Natural History [13]. The Blender add-ons were used to render the images for the moving and fixed-lens configurations with the magnifications $M=0.76$ and $M=0.5$, respectively. The number of the tilt positions is 13 for complete spherical camera positions with 15^o tilt step size. The number of the pan positions varies with the tilt to maintain an equal pan angle step of 15^o along each tilt angle. Thus, a uniform distribution of the camera positions can be achieved where the angle between the two nearest camera positions is approximately 15^o, leading to 184 camera views to cover the complete spherical surface around the virtual specimen. The macrorail step size is 0.625 mm for both moving and fixed-lens setups. Similar image size and lens to those of the real camera setup are chosen.

7.2 Results and discussion

For the multifocus fusion comparison, our baseline is SIFT feature-based image registration with homography transformation computed between the multifocus images of the object. Multifocus fusion by Ströbel et al. [49] is called "stacking with calibrated scaling," where the scale and shift for image registration are precomputed using the calibration target images and included for the moving lens camera setup. Implementation by Li and Nguyen [26] is called "stacking with fitted homography," where the homography transformation is precomputed using the calibration target images, and included for the Fixed-Lens camera setup. Finally, our proposed registration technique is called "stacking with analytic homography" and included for both the Moving-Lens and the Fixed-Lens camera setups. All these techniques use the Laplacian pyramid algorithm [54] for image fusion.

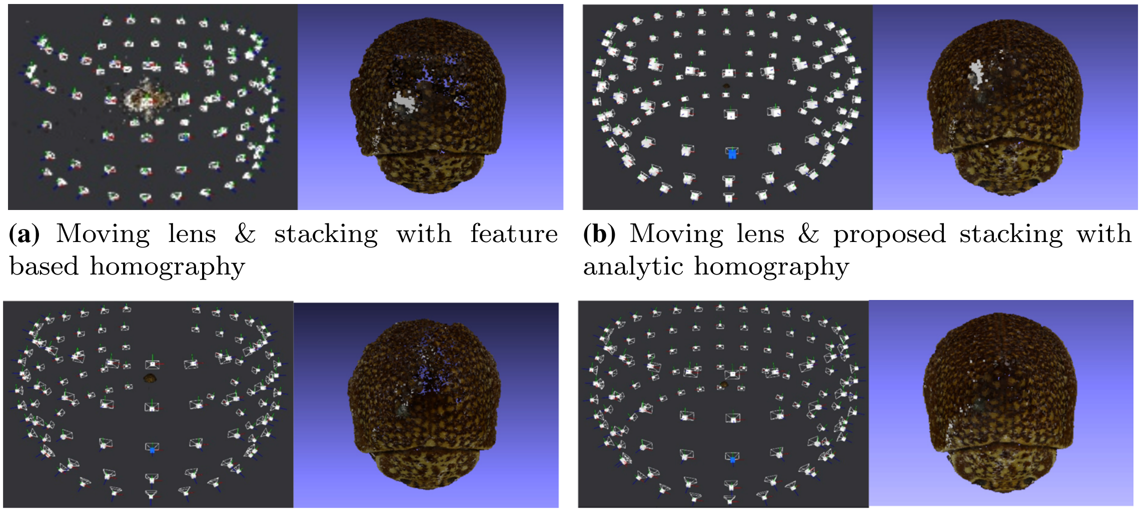
We also try two state-of-the-art algorithms proposed by Ji et al. [22] and Liang et al. [30]. However, the results are not suitable for the 3D reconstruction as shown in Sect. 1 of the supplementary material. This is because the image scale varies significantly between the multifocus images and these algorithms do not account for such variation.

For the multiview stereo reconstruction, an open-source software Meshroom [3] is used for all comparisons in this paper. Meshroom is executed with default pipeline and "Describer Preset = high" on FeatureExtraction node and "Geometric Estimator = loransac" on FeatureMatching node. The supplementary material includes the results by commercial software Photoscan Pro [2] which has "visibility-consistent" feature for mesh generation to obtain a better 3D reconstruction of fine structures which often get missing in the reconstructions by Meshroom.

Figure 10 provides a qualitative comparison of the dense clouds and camera poses obtained from the images captured by different camera setups and fused with different registration techniques. The visual accuracy and completeness of the camera poses using the fixed lens are significantly better than that of the Moving-Lens capture. From the dense clouds, the model generated from the Moving-Lens capture is sparser, and this implies that the conventional moving lens camera setup recovers fewer details of the surface geometry than that of our Fixed-Lens capture. From the estimated camera poses, we can see that the poses reconstructed by the moving lens setup are not well aligned. On the contrary, the estimated camera poses generated by the fixed lens can reproduce a completed camera trajectory, and the recovered camera positions are also well aligned as expected on a spherical surface. The combination of the Fixed-Lens capture and our proposed multifocus fusion with calibrated homography transformation provides the best results.

Figure 11 shows the snapshots of the reconstructed mesh models from the point clouds shown in Fig. 10. This figure confirms the effectiveness of the Fixed-Lens capture versus the Moving-Lens capture, and our proposed multifocus stacking scheme using analytic homography versus the stacking with SIFT feature-based homography. The reconstructed mesh model by the Fixed-Lens capture and our proposed stacking method is the most accurate and shows the best resolution.

For a quantitative comparison, we use the estimated camera positions to compute the errors in the radial positions and the pan-tilt rotation angles as compared to the expected values. Due to the pan-tilt scanning motion, the camera positions are expected to move around the rotation axis and lie on a spherical surface. The rotation axis and the center of rotation were estimated by the plane and circle



(a) Moving lens & stacking with feature based homography

(b) Moving lens & proposed stacking with analytic homography

(c) Fixed lens & stacking with feature based homography

(d) Fixed lens & proposed stacking with analytic homography (ours)

Fig. 10 A comparison between the structure-from-motion reconstructions obtained using Meshroom from the images captured using the proposed Fixed-Lens Multifocus Capture and the Moving-Lens Multifocus Capture by stacking with SIFT feature-based homography and analytic homography. The comparison cases are: **(a)** moving lens and stacking with SIFT feature-based homography, **(b)** moving lens and proposed stacking with analytic homography, **(c)** fixed lens and stacking with SIFT feature-based homography and **(d)** fixed lens and proposed stacking with analytic homography (ours). For each case, **left:** the camera poses from the reconstruction and **right:** the dense point cloud from the reconstruction. More uniform camera poses and a denser point cloud from the fixed lens indicate better image quality than that of the moving lens. (Best viewed in colour on the screen) (colour figure online)

fittings. From the fitted center and rotation axis, the radial distance and pan-tilt rotation steps can be computed from the camera positions. The mean and standard deviations of these values indicate the accuracy of the 3D reconstructions.

Tables 1 and 2 show the mean and standard deviation of the radial distances and angles of the poses for both the moving lens and the fixed lens with our proposed (using analytic homography) and standard (using SIFT feature-based homography) stacking methods using Meshroom. Table 1 shows the comparison between different image stacking methods, whereas Table 2 shows the comparison between different lens setups.

Tables 1 and 2 confirm that both the Fixed-Lens capture and our proposed multifocus stacking with analytic homography reduce the noise (standard deviation) and the difference with the expected rotation steps. The combination of the two approaches resulted in the lowest errors shown in the bottom rows of Tables 1 and 2. Note that the mean radial distances were computed based on the lens magnification as multiview reconstruction does not provide the true scale.

Table 3 shows the comparison of the expected camera focal lengths and estimated focal lengths by bundle adjustment for different settings. Again, it confirms that the fixed-lens setup and our proposed image stacking with analytic homography produces the closest focal length to the expected value. The overestimated focal length for the Moving Lens setup suggests that its stacked images contain significant perspective projection distortion, while those from the Fixed-Lens setup still maintain original perspective projection.

Note that while the estimated camera focal length of 189.84 mm by the moving-lens setup and focus stacking with SIFT feature-based homography using Meshroom is reasonably close to the expected value of 149.5mm, the reconstruction quality is very poor as shown in Figs. 10 and 11. Furthermore, there were several stacked images obtained by focus stacking with SIFT feature-based homography where bundle adjustment for 3D reconstructions failed to estimate their camera poses, leading them to be excluded. These indicate that general focus stacking without constrained image alignment is not suitable for stacking the images for 3D reconstruction of small objects.

Figure 12 shows the qualitative comparisons between the snapshots of the reconstructed 3D models of *Agrilus Anxius* beetle obtained using Meshroom with the camera positions on a complete spherical surface. Again, the fixed-lens setup and our proposed image stacking with analytic homography lead to a more complete and cleaner 3D model. Figure 12 also shows the comparison of the reconstruction error with respect to the ground truth model in terms of mean Hausdorff distance (MHD) computed using MeshLab [10].

The Hausdorff distance [42] represents how far the reconstructed 3D mesh is from the original 3D model.

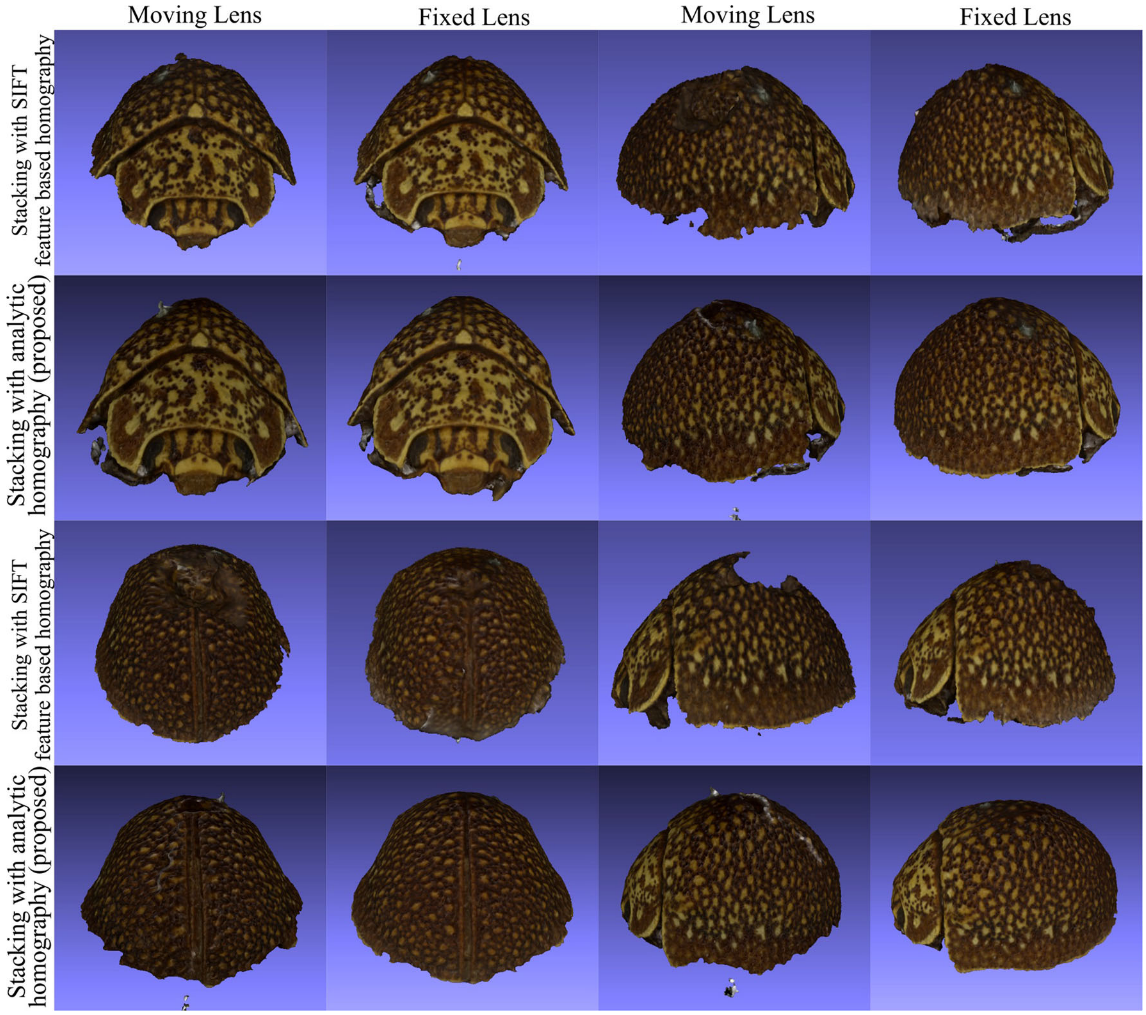

Moving Lens
Fixed Lens
Moving Lens
Fixed Lens
Stacking with SIFT feature based homography
Stacking with analytic homography (proposed)
Stacking with SIFT feature based homography
Stacking with analytic homography (proposed)

Fig. 11 A qualitative comparison between the reconstructed 3D models obtained using Meshroom. (Best viewed in colour on the screen) (colour figure online)

Table 1 A comparison between the expected and estimated camera centers obtained using Meshroom

Stacking with feature-based homography			
	Mean ± STD of radial distance (mm)	Mean ± STD of pan step (degree)	Mean ± STD of tilt step (degree)
Moving lens	115.00 ± 15.74	16.63 ± 6.89	16.31 ± 4.41
Fixed lens (ours)	103.23 ± 3.67	15.12 ± 2.07	15.11 ± 0.89
Stacking with analytic homography (proposed)			
Moving lens	115.00 ± 1.87	15.02 ± 1.14	15.21 ± 1.03
Fixed lens (ours)	103.23 ± **0.98**	**14.94** ± **0.68**	**15.09** ± **0.28**

The true pan step = 15°, and the true tilt step = 15°

Again, the fixed-lens setup and image stacking with analytic homography produce the most accurate 3D reconstruction and the smallest MHD, about half of that of the Moving Lens with image alignment using calibrated scaling proposed by Ströbel et al. [49]. Furthermore, the relative improvement of our 3D reconstruction method using analytic homography in terms of MHD is 13.9% compared with the second-best method proposed in our preliminary study [26].

The corresponding quantitative comparisons in terms of the mean and standard deviation of the radial distances and angles of the poses between different image stacking

Table 2 A comparison between the expected and estimated camera centers obtained using Meshroom

Moving lens			
	Mean ± STD of radial distance (mm)	Mean ± STD of pan step (degree)	Mean ± STD of tilt step (degree)
Stacking with feature based homography	115.00 ± 15.74	16.63 ± 6.89	16.31 ± 4.41
Stacking with analytic homography (proposed)	115.00 ± 1.87	15.02 ± 1.14	15.21 ± 1.03
Fixed lens (ours)			
Stacking with feature based homography	103.23 ± 3.67	15.12 ± 2.07	15.11 ± 0.89
Stacking with analytic homography (proposed)	$103.23 \pm \mathbf{0.98}$	$\mathbf{14.94} \pm \mathbf{0.68}$	$\mathbf{15.09} \pm \mathbf{0.28}$

The true pan step $= 15^{\circ}$, and the true tilt step $= 15^{\circ}$

Table 3 A comparison between the expected and estimated camera focal lengths obtained using Meshroom

	Estimated Focal Length (mm)		Expected Focal Length (mm)
	Stacking with SIFT feature based homography	Stacking with analytic homography (proposed)	
Moving lens	189.84	276.44	149.5
Fixed lens (ours)	155.94	**193.44**	175.5

methods and between different lens setups are given in Tables 4, 5, respectively. Additional quantitative comparison in terms of the focal length is shown in Table 6. 3D reconstructions are also performed with incomplete spherical camera positions and the results are shown in the supplementary material. In addition, the experiments were performed with another specimen *beechnut-fagus-sylvatica* and with another 3D reconstruction software Agisoft Photoscan. These results are included in the supplementary material. The results obtained from the synthetic dataset show similar trends as those of the real dataset.

8 Conclusion

In this paper, we proposed the new Fixed-Lens Multifocus Capture and the calibrated analytic homography-based image alignment process for accurate 3D reconstruction of small-scale objects. Currently, the image-based 3D reconstruction devices using the lens moving with the camera to capture the multifocus images suffer from perspective distortion that reduces the accuracy of 3D reconstruction. With the proposed Fixed-Lens Multifocus Capture setup, the lens remains stationary while the camera and the image sensor move during multifocus image capturing. The calibrated image alignment using analytic homography is performed to account for the change of relative scale and in-plane shift between the images captured at different depths with the constraint of macrorail motion. The registered images are fused to create an in-focus image that is consistent with the perspective image formation for each viewing angle.

The experiments using the real and synthetic images of different objects demonstrated the effectiveness of our proposed fixed-lens setup with analytic homography-based image alignment stacking as compared with the conventional moving lens setup and various stacking methods. The reconstruction results showed that both of our proposed Fixed-Lens capture setup and analytic homography-based image stacking improve the accuracy of 3D reconstruction significantly as compared with the conventional approaches. Particularly, the combination of our proposed approaches produced the most accurate 3D model and camera poses. Furthermore, the estimated focal lengths by bundle adjustment suggested that the perspective projection distortion was introduced in the stacked images from the Moving-Lens capture setup. Our study also shows that the multifocus stacking without calibrated scale-shift homography-based alignment likely leads to the image rejections and 3D reconstruction errors as the bundle adjustment could not estimate their camera poses.

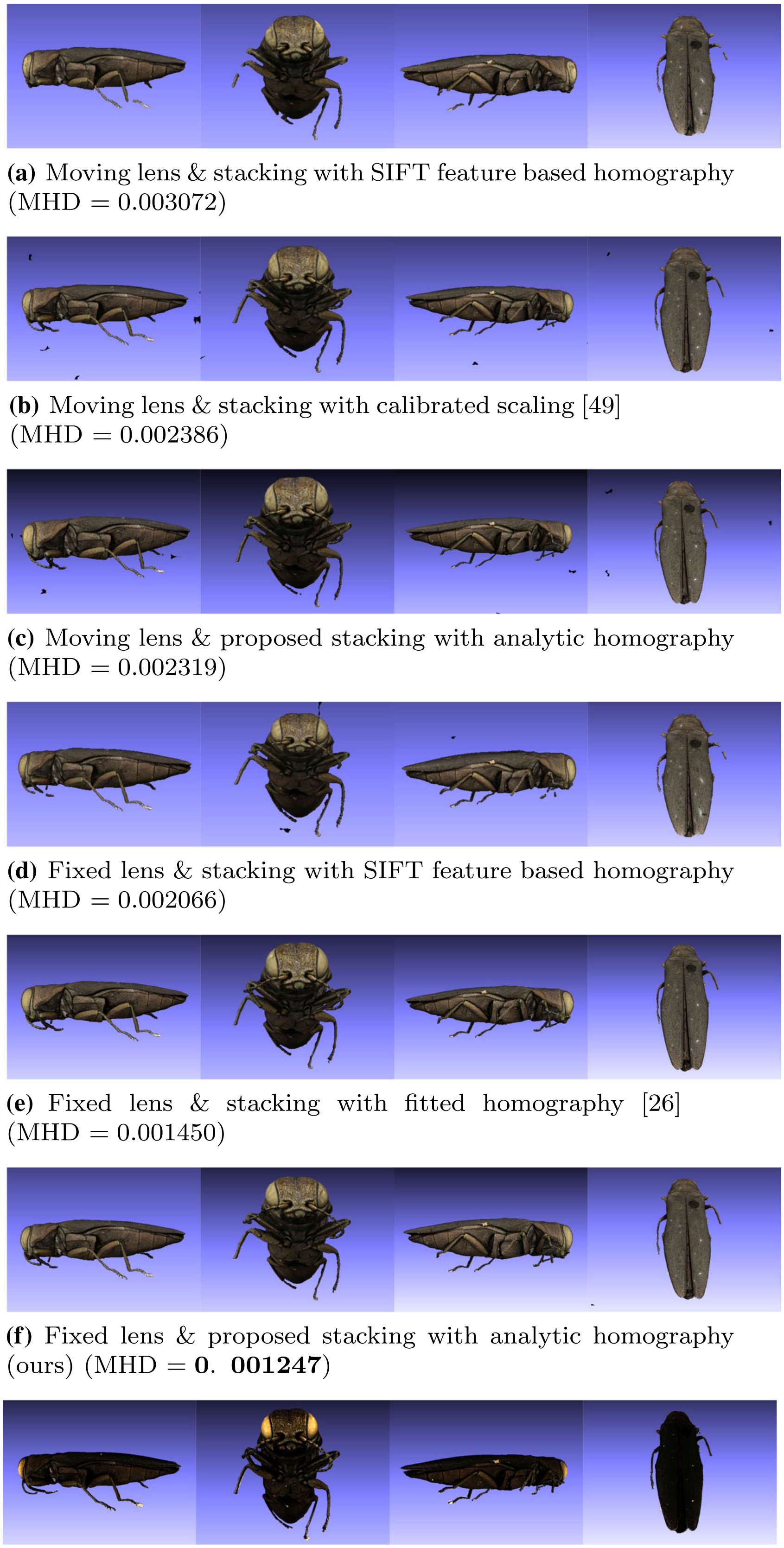


Fig. 12 A comparison between the snapshots of the reconstructed 3D models of *Agrilus Anxius* with mean Hausdorff distance (MHD) as the reconstruction error. The relative improvement of our proposed method in terms of MHD is **13.9%** compared with the second-best method [26]. (Best viewed in colour on the screen) (colour figure online)

(a) Moving lens & stacking with SIFT feature based homography (MHD = 0.003072)

(b) Moving lens & stacking with calibrated scaling [49] (MHD = 0.002386)

(c) Moving lens & proposed stacking with analytic homography (MHD = 0.002319)

(d) Fixed lens & stacking with SIFT feature based homography (MHD = 0.002066)

(e) Fixed lens & stacking with fitted homography [26] (MHD = 0.001450)

(f) Fixed lens & proposed stacking with analytic homography (ours) (MHD = **0. 001247**)

(g) Ground truth model

Table 4 A comparison between the expected and estimated camera centers obtained using Meshroom. The true pan step = 15°, and the true tilt step = 15°

Agrilus Anxius (with complete spherical camera positions)			
Stacking with feature-based homography			
	Mean ± STD of radial distance (mm)	Mean ± STD of pan step (degree)	Mean ± STD of tilt step (degree)
Moving lens	150.00 ± 1.53	14.88 ± 1.18	14.89 ± 0.32
Fixed lens	193.17 ± 0.81	14.92 ± 0.63	14.94 ± 0.17
Stacking with calibrated scaling			
Moving lens [49]	150.00 ± 0.49	14.93 ± 0.29	14.95 ± 0.32
Stacking with fitted homography			
Fixed lens [26]	193.17 ± 0.05	14.95 ± 0.12	14.99 ± 0.01
Stacking with analytic homography (proposed)			
Moving lens	150.00 ± 0.47	14.94 ± 0.33	14.95 ± 0.12
Fixed lens (ours)	193.17 ± **0.04**	**14.96** ± **0.11**	**14.99** ± **0.01**

Table 5 A comparison between the expected and estimated camera centers obtained using Meshroom. The true pan step = 15°, and the true tilt step = 15°

Agrilus Anxius (with complete spherical camera positions)			
Moving lens			
	Mean ± STD of radial distance (mm)	Mean ± STD of pan step (degree)	Mean ± STD of tilt step (degree)
Stacking with feature-based homography	150.00 ± 1.53	14.88 ± 1.18	14.89 ± 0.32
Stacking with calibrated scaling [49]	150.00 ± 0.49	14.93 ± 0.29	14.95 ± 0.32
Stacking with analytic homography (proposed)	150.00 ± 0.18	14.94 ± 0.33	14.95 ± 0.12
Fixed lens (ours)			
Stacking with feature-based homography	193.17 ± 0.81	14.92 ± 0.63	14.94 ± 0.17
Stacking with fitted homography [26]	193.17 ± 0.05	14.95 ± 0.12	14.99 ± 0.01
Stacking with analytic homography (proposed)	193.17 ± **0.04**	**14.96** ± **0.11**	**14.99** ± **0.01**

Table 6 A comparison between the expected and estimated camera focal lengths obtained using Meshroom

Agrilus Anxius (with complete spherical camera positions)					
	Estimated Focal Length (mm)				
	Stacking with SIFT feature based homography	Stacking with calibrated scaling [49]	Stacking with fitted homography [26]	Stacking with analytic homography (proposed)	Expected focal length (mm)
Moving Lens	80.84	85.58	NA	91.66	114.70
Fixed lens (ours)	89.70	NA	97.02	**97.39**	97.96

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-021-05926-7.

Acknowledgements The authors would like to thank Dr. Matt Adcock and Mr. Stuart Anderson from CSIRO Data61 CPS Quantitative Imaging, Mr. Nunzio Knerr from CSIRO National Research Collections Australia for their help with building the image capture system, Mr. Fabien Castan from Mikros Image for his help with using Meshroom for this project and Mr. Julien Haudegond and Mr. Enguerrand De Smet from Mikros Image for their Blender add-on to create synthetic images. This research is partially funded by CSIRO Julius Career Award.

Declarations

Conflict of interest The authors report no conflict of interest.

References

1. 3DSOM (2019) 3dsom - 3d models from photos. http://www.3dsom.com/, [Online; accessed 14-July-2019]
2. Agisoft (2019) Agisoft - metashape (photoscan). https://www.agisoft.com/, [Online; accessed 14-July-2019]
3. AliceVision (2019) Meshroom: A 3D reconstruction software. https://github.com/alicevision/meshroom
4. Amin-Naji M, Aghagolzadeh A, Ezoji M (2018) Fully Convolutional Networks for Multi-Focus Image Fusion. In: 2018 9th International Symposium on Telecommunications (IST), pp 553–558
5. Ascencio C (2020) Estimation of the Homography Matrix to Image Stitching bookTitle = Applications of Hybrid Metaheuristic Algorithms for Image Processing, Springer International Publishing, Cham, pp 205–230. https://doi.org/10.1007/978-3-030-40977-7_10
6. Aslantas V, Kurban R (2010) Fusion of multi-focus images using differential evolution algorithm. Exp Syst Appl 37(12):8861–8870. https://doi.org/10.1016/j.eswa.2010.06.011
7. Bai X, Zhang Y, Zhou F, Xue B (2015) Quadtree-based multifocus image fusion using a weighted focus-measure. Inf Fusion 22:105–118. https://doi.org/10.1016/j.inffus.2014.05.003
8. Blender-Foundation (2020) Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam, http://www.blender.org
9. Burt P, Adelson E (1983) The Laplacian Pyramid as a Compact Image Code. IEEE Trans Commun 31(4):532–540
10. Cignoni P, Callieri M, Corsini M, Dellepiane M, Ganovelli F, Ranzuglia G (2008) MeshLab: an Open-Source Mesh Processing Tool. In: Eurographics Italian Chapter Conference
11. De I, Chanda B (2013) Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure. Inf Fusion 14(2):136–146. https://doi.org/10.1016/j.inffus.2012.01.007
12. Digital Archive of Natural History (2018a) Agrilus anxius. https://sketchfab.com/3d-models/agrilus-anxius-2fd91429bd54423cafea7d7ba22273cc, [Online; accessed 31-October-2020]
13. Digital Archive of Natural History (2018b) Beechnut (Fagus sylvatica). https://sketchfab.com/3d-models/beechnut-fagus-sylvatica-509980aaf55746ddbe18b7f03e15c9c6, [Online; accessed 31-October-2020]
14. Fujii H, Kodama K, Hamamoto T (2016) Scene flow estimation through 3D analysis of multi-focus images. In: 2016 Visual Communications and Image Processing (VCIP), IEEE, pp 1–4
15. Gallo A, Muzzupappa M, Bruno F (2014) 3D reconstruction of small sized objects from a sequence of multi-focused images. J Cult Herit 15(2):173–182. https://doi.org/10.1016/j.culher.2013.04.009
16. Geng J (2011) Structured-light 3D surface imaging: a tutorial. Adv Opt Photon 3(2):128–160. https://doi.org/10.1364/AOP.3.000128http://aop.osa.org/abstract.cfm?URI=aop-3-2-128
17. Guo X, Nie R, Cao J, Zhou D, Qian W (2018) Fully Convolutional Network-Based Multifocus Image Fusion. Neural Comput 30(7):1775–1800
18. Hartley R, Zisserman A (2003) Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, USA
19. Hecht E (2002) Optics. Addison Wesley Longman, Reading, Massachusetts
20. Huang J, Le Z, Ma Y, Mei X, Fan F (2020) A generative adversarial network with adaptive constraints for multi-focus image fusion. Neural Comput and Applic. https://doi.org/10.1007/s00521-020-04863-1
21. Hui Li, Manjunath BS, Mitra SK (1994) Multi-sensor image fusion using the wavelet transform. In: Proceedings of 1st International Conference on Image Processing, vol 1, pp 51–55 vol.1
22. Ji Z, Kang X, Zhang K, Duan P, Hao Q (2019) A Two-Stage Multi-Focus Image Fusion Framework Robust to Image Mis-Registration. IEEE Access 7:123231–123243
23. Kodama K, Kubota A (2013) Efficient Reconstruction of All-in-Focus Images Through Shifted Pinholes From Multi-Focus Images for Dense Light Field Synthesis and Rendering. IEEE Trans Image Proc 22(11):4407–4421. https://doi.org/10.1109/TIP.2013.2273668
24. Kodama K, Wang Z, Sato M, Murakami T (2017) Real-time 3-D image reconstruction from multi-focus images by efficient linear filtering with multi-dimensional symmetry. In: 2017 IEEE International Conference on Image Processing (ICIP), pp 3575–3579, https://doi.org/10.1109/ICIP.2017.8296948
25. Lewis JJ, O'Callaghan RJ, Nikolov SG, Bull DR, Canagarajah N (2007) Pixel- and region-based image fusion with complex wavelets. Inf Fusion 8(2):119–130. https://doi.org/10.1016/j.inffus.2005.09.006 (**special Issue on Image Fusion: Advances in the State of the Art**)
26. Li H, Nguyen C (2019) Perspective-consistent multifocus multiview 3D reconstruction of small objects. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), IEEE, pp 1–8
27. Li S, Kwok JT, Wang Y (2001) Combination of images with diverse focuses using the spatial frequency. Inf Fusion 2(3):169–176. https://doi.org/10.1016/S1566-2535(01)00038-0
28. Li S, Kang X, Hu J (2013) Image Fusion With Guided Filtering. IEEE Trans Image Proc 22(7):2864–2875
29. Li S, Kang X, Fang L, Hu J, Yin H (2017) Pixel-level image fusion: A survey of the state of the art. Inf Fusion 33:100–112. https://doi.org/10.1016/j.inffus.2016.05.004
30. Liang Y, Mao Y, Tang Z, Yan M, Zhao Y, Liu J (2019) Efficient misalignment-robust multi-focus microscopical images fusion. Sign Proc 161:111–123
31. Lie WN, Ho CC (2019) Multi-Focus Image Fusion and Depth Map Estimation Based on Iterative Region Splitting Techniques. J Imag 5(9):73. https://doi.org/10.3390/jimaging5090073
32. Liu Y, Liu S, Wang Z (2015) Multi-focus image fusion with dense SIFT. Inf Fusion 23:139–155. https://doi.org/10.1016/j.inffus.2014.05.004
33. Liu Y, Wang L, Cheng J, Li C, Chen X (2020) Multi-focus image fusion: A Survey of the state of the art. Inf Fusion 64:71–91. https://doi.org/10.1016/j.inffus.2020.06.013

34. Lowe DG (2004) Distinctive Image Features from Scale-Invariant Keypoints. Int J Comput Vision 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
35. Mishra D, Palkar B (2015) Image Fusion Techniques: A Review. Int J Comput Appl 130:7–13
36. Murgia F, Giusto D, Perra C (2015) 3D reconstruction from plenoptic image. In: 2015 23rd Telecommunications Forum Telfor (TELFOR), pp 448–451, https://doi.org/10.1109/TELFOR.2015.7377504
37. Mustafa HT, Yang J, Zareapoor M (2019) Multi-scale convolutional neural network for multi-focus image fusion. Imag Vision Comput 85:26–35. https://doi.org/10.1016/j.imavis.2019.03.001
38. Nencini F, Garzelli A, Baronti S, Alparone L (2007) Remote sensing image fusion using the curvelet transform. Inf Fusion 8(2):143–156. https://doi.org/10.1016/j.inffus.2006.02.001 (**special Issue on Image Fusion: Advances in the State of the Art**)
39. Nguyen CV, Lovell DR, Adcock M, La Salle J (2014) Capturing natural-colour 3D models of insects for species discovery and diagnostics. PloS one 9(4):e94346
40. Pan T, Jiang J, Yao J, Wang B, Tan B (2020) A Novel Multi-Focus Image Fusion Network with U-Shape Structure. Sens (Basel, Switzerland) 20(14):3901. https://doi.org/10.3390/s20143901
41. Ritz M, Langguth F, Scholz M, Goesele M, Stork A (2012) High resolution acquisition of detailed surfaces with lens-shifted structured light. Computers & Graphics 36(1):16–27. https://doi.org/10.1016/j.cag.2011.10.004http://www.sciencedirect.com/science/article/pii/S0097849311001506, cultural Heritage
42. Rucklidge W (ed) (1996) The Hausdorff distance, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 27–42. https://doi.org/10.1007/BFb0015093,
43. Sakamoto T, Kodama K, Hamamoto T (2012a) A novel scheme for 4-D light-field compression based on 3-D representation by multi-focus images. In: 2012 19th IEEE international conference on image processing, IEEE, pp 2901–2904
44. Sakamoto T, Kodama K, Hamamoto T (2012b) A study on efficient compression of multi-focus images for dense light-field reconstruction. In: 2012 Visual Communications and Image Processing, IEEE, pp 1–6
45. Schönberger JL, contributors (2020) COLMAP: a general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline. https://github.com/colmap, [Online; accessed 15-November-2020]
46. Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE, vol 1, pp 519–528
47. Silvester CM, Hillson S (2020) A critical assessment of the potential for Structure-from-Motion photogrammetry to produce high fidelity 3D dental models. Am J Phys Anthropol 173(2):381–392. https://doi.org/10.1002/ajpa.24109
48. Skinner KA, Johnson-Roberson M (2016) Towards real-time underwater 3D reconstruction with plenoptic cameras. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 2014–2021, https://doi.org/10.1109/IROS.2016.7759317
49. Ströbel B, Schmelzle S, Blüthgen N, Heethoff M (2018) An automated device for the digitization and 3D modelling of insects, combining extended-depth-of-field and all-side multi-view imaging. ZooKeys 759:1
50. Sturm P (2014) Pinhole Camera Model, Springer US, Boston, MA, pp 610–613. https://doi.org/10.1007/978-0-387-31439-6_472,
51. Szeliski R (2006) Image Alignment and Stitching: A Tutorial. Found Trends Comput Graph Vis 2(1):1–104. https://doi.org/10.1561/0600000009
52. Szeliski R (2010) Comput Vision: Algorithms Appl, 1st edn. Springer-Verlag, Berlin, Heidelberg
53. Toet A (1989) Image fusion by a ratio of low-pass pyramid. Pattern Recogn Lett 9(4):245–253. https://doi.org/10.1016/0167-8655(89)90003-2
54. Wang W, Chang F (2011) A Multi-focus Image Fusion Method Based on Laplacian Pyramid. J Comput 6(12):2559–2566. https://doi.org/10.4304/jcp.6.12.2559-2566
55. Xu H, Fan F, Zhang H, Le Z, Huang J (2020) A Deep Model for Multi-Focus Image Fusion Based on Gradients and Connected Regions. IEEE Access 8:26316–26327
56. Yang B, Li S (2010) Multifocus Image Fusion and Restoration With Sparse Representation. IEEE Trans Instrum Measure 59(4):884–892
57. Yang B, Zl Jing, Ht Zhao (2010) Review of pixel-level image fusion. J Shanghai Jiaotong Univ (Sci) 15:6–12. https://doi.org/10.1007/s12204-010-7186-y
58. Zhang H, Le Z, Shao Z, Xu H, Ma J (2021) MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Inf Fusion 66:40–53. https://doi.org/10.1016/j.inffus.2020.08.022
59. Zhang Q, long Guo B (2009) Multifocus image fusion using the nonsubsampled contourlet transform. Sign Proc 89(7):1334–1346. https://doi.org/10.1016/j.sigpro.2009.01.012
60. Zhou Y, Guo H, Fu R, Liang G, Wang C, Wu X (2015) 3D reconstruction based on light field information. In: 2015 IEEE International Conference on Information and Automation, pp 976–981, https://doi.org/10.1109/ICInfA.2015.7279428
61. Zhou Z, Li S, Wang B (2014) Multi-scale weighted gradient-based fusion for multi-focus images. Inf Fusion 20:60–72. https://doi.org/10.1016/j.inffus.2013.11.005
62. Zhu D, Wu C, Liu Y, Fu D (2018) 3D reconstruction based on light field images. In: Yu H, Dong J (eds) Ninth International Conference on Graphic and Image Processing (ICGIP 2017), International Society for Optics and Photonics, SPIE, vol 10615, pp 951 – 959, https://doi.org/10.1117/12.2304504,
63. Zitova B, Flusser J (2003) Image registration methods: a survey. Image Vision Comput 21(11):977–1000

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.