MaCal – Macro Lens Calibration and the Focus Stack Camera Model

Xiangyu Weng[*], Mengkun She[*], David Nakath, Kevin Köser
GEOMAR Helmholtz Centre for Ocean Research Kiel
Wischhofstr. 1-3, 24148 Kiel, Germany
xweng@geomar.de

Abstract

Macro photography is characterized by a very shallow depth of field, which challenges classical structure from motion and even camera calibration techniques, since images suffer from large defocussed areas. Computational photography methods such as focus stacking combine the sharp areas of many photos into one, which can produce spectacular images of insects or small structures. In this contribution we analyse the camera model to describe such focus stacked images in photogrammetry and computer vision and derive a camera calibration pipeline for macro photography to enable photogrammetry and 3D reconstruction of tiny objects. We demonstrate the effectiveness of the approach on ray-traced images with ground truth and real images.

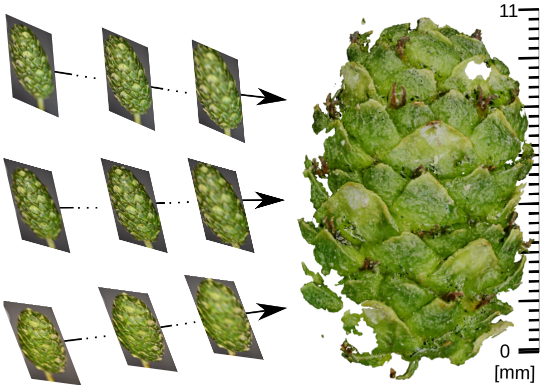

Figure 1. Calibrated focus stacked 3D macro-reconstruction of a tiny bud from series of shallow depth of field macro photos.

1. Introduction

Macro photography is defined as extreme close-up photography [7], in which the object's image on the sensor is larger or equal than the object's size in 3D [15]. In a broader sense, macro photography is concerned with photographing tiny objects in the micrometer to millimeter range and has a multitude of applications in life sciences and technology, but is also popular in art and among photographers, since it typically produces images from unconventional perspectives (e.g. close-up of insects). Due to the relative proximity of the object to the aperture, the special setting suffers from optical effects such as diffraction and defocus that influence lateral resolution and depth resolution and make 3D computer vision and photogrammetry in the micro world very challenging. When using large apertures (small F-numbers), lenses collect a lot of light, but only a very limited depth range is ”in focus”, which lets large parts of the image appear unsharp (”out of focus”). This limited *depth of field* (DOF) makes it difficult to find correspondences between different object poses (see Fig. 1 for a challenging reconstruction setting). Reducing the aperture size limits the amount of light collected by the sensor, but increases the depth of field. However, due to the effect of diffraction, using a small aperture leads to reduced image resolution. The *diffraction-barrier* is governed by fundamental laws of physics, and thus can not easily be overcome by changing the lens or aperture design. It can already be observed in *normal* scale images, however, its impact is significantly higher in the macro-scale, where a careful trade-off has to be made between image resolution and DOF (cf. Fig. 2).

A common way to deal with this trade-off is to combine the pixels of multiple shallow DOF images into one by focus stacking [2]. However, when combining pixels from different images into one virtual image, it remains unclear what the virtual camera model for such an image is (cf. to [8]) and how camera intrinsics such as lens distortion can be properly considered in photogrammetric and 3D vision applications. In this paper, we propose a way to calibrate macro lenses and analyse the common focus stacking camera geometry as originally introduced in [2]. Specifically, we contribute the following: We (i) show that a focus-stacked image can be described by an affine camera model, and how the pinhole camera parameters of the original camera are related to the affine camera model for a focus stacked image. We (ii) propose a novel cascaded chessboard

[*] These authors contributed equally to this work.

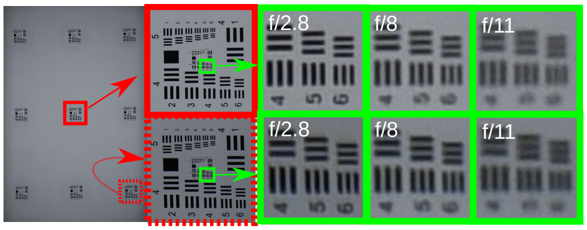

Figure 2. Impact of the diffraction effect shown on the AF-1951 chart: As the f-number increases, the aperture decreases. The resolution reaches 114 lp/mm at aperture f/8. With even smaller apertures, the resolution becomes insufficient. **Left two columns:** show the evaluation setup; the remaining **right three columns** show the green patch magnified in the different conditions.

corner localization approach on image stacks. It utilizes a focus stacked image (with extended DOF) for detecting the chessboard and propagates the corners into the source images, where we apply sub-pixel refinement on the sharp regions of the particular shallow DOF images. We further use finding (i), to (iii) extract parameters for the macro camera model from the affine camera parameters of a single stacked image in closed form. We show that those can be used as initial values for maximum-likelihood estimation (MLE) in a multi-view perspective camera calibration approach, operating on the corner observations from (ii) in the largely defocused images. To facilitate further research, (iv) calibration code and datasets will be made available online. [1]

2. Previous Work

There is a huge body of related work with respect to macro photography and we refer the reader to [20, 7, 15] for an overview and insights into the main ideas. In general, cameras need to have a large enough aperture to collect enough light, but the larger the aperture, the smaller the depth of field will become. In this context Hasinoff et al. [14] provide a detailed analysis of the exposure time tradeoffs. While in principle it is possible to photograph with a very small aperture to maximize the depth of field [20], diffraction effects [3] limit the resolution (see Fig. 2), so to maximize spatial resolution a limited depth of field is typically accepted. Typically, macro photography is used in certain 2D applications that do not require a large DOF or 3D applications that require less precision [18, 4].

When using techniques such as depth-from-focus [12] or depth-from-defocus [23] the limited DOF can be even turned into an advantage. In depth-from-focus approaches multiple images are taken with different parts focused in order to infer their distance (see e.g. [9] for an early comparison of different methods). Such a stack of images with different parts in focus can also be used to create one merged image that contains the sharp parts of all the ”sub-images” by using a sharpness measure (e.g. the modified Laplacian [16]) on each input image [2]. Several improvements exist to avoid artefacts when compositing multiple images into one focus-stacked image (see e.g. [8, 24]). To bring different object parts into focus, the two main principles are to move (or modify) the lens with respect to the chip, or to move the entire camera relatively to the object. In this paper we consider only the second case, which keeps the intrinsics of the camera, which we want to calibrate, constant. Due to the extremely small DOF in the close-up focusing, manual or automatic micrometer platforms are generally used for *extended depth of field* (EDOF) techniques [5].

In order to use cameras for 3D vision and photogrammetry applications, the relation between pixels in the image and rays in the camera coordinate system must be known, which can be solved by calibration [6] and the classical calibration approach involves presenting a known target to a camera multiple times, and then solving for the fixed intrinsics and varying extrinsics of all poses [25]. Calibration parameters can also be obtained by self-calibration from an unknown scene (see e.g. [10, 19] for early works). When using reconstructions of many images with nicely distributed features, state-of-the-art structure-from-motion systems such as [21] can also read approximate calibration information from image meta data and optimize calibration parameters in bundle adjustment [1]. Therefore, it is possible in principle to pass macro photography images directly to 3D reconstruction systems [22, 11] using self-calibration. However, there are degenerate cases for self calibration e.g., when scene structure, camera motion or distribution of observations is not general enough, which leaves some ambiguities between the parameters and can lead to skewed or biased reconstructions [13].

For shallow DOF images that sometimes show only a narrow corridor, or ring structure, of features which are in focus, we argue that one should avoid these ambiguities and pre-calibrate the camera beforehand to avoid potentially biased or skewed 3D models. To the best of our knowledge, calibrating a shallow DOF camera that cannot take an all-in-focus image of the calibration target has not been done before. Additionally, the authors are unaware of work that analyzes the geometric camera model for focus-stacked images, which is what we will derive in the next section.

3. Focus Stacking Revisited

Notation We will write elements of projective space upright with serifs (e.g. $\mathrm{x} \in \mathbb{P}^2$, or $\mathrm{X} \in \mathbb{P}^3$), the corresponding Euclidean vectors (e.g. $\boldsymbol{x} \in \mathbb{R}^2$, or $\boldsymbol{X} \in \mathbb{R}^3$) are displayed in *italics*.

Let us assume we have a chessboard, and we attach a coordinate system such that the x- and y-axis coincide with

[1] https://www.geomar.de/en/omv-research/macro-and-micro-photogrammetry

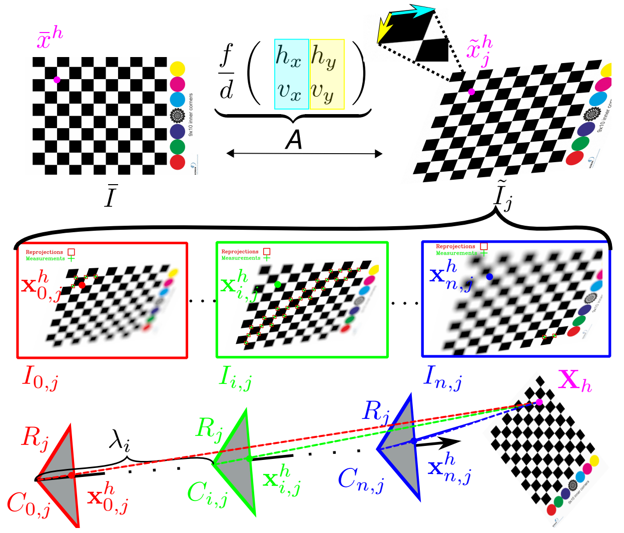

Figure 3. Geometry, notation and transformations of the jth stack. We observe a 3D point $\mathbf{X}_h$ on a chessboard with the jth stack, comprised of sub-images $\mathcal{I}_{0,j}, \cdots, \mathcal{I}_{i,j}, \cdots, \mathcal{I}_{n,j}$. They can be focus-stacked into a stack specific image $\tilde{\mathcal{I}}_j$. It serves as the basic interface of our approach by allowing for an affine transformation A between the stacked and a perpendicularly taken image $\bar{\mathcal{I}}$.

the row and column directions on the board, and all points on the board share the property $z = 0$, i.e. the $z-$axis is perpendicular to the chessboard. We position a pinhole camera in front of the board, viewing from some oblique, non-aligned angle. The pinhole camera's projection matrix $\mathbf{P}_0$ can be described by $\mathbf{P}_0 = \mathbf{K}\left(\mathsf{R}^\mathsf{T} \mid -\mathsf{R}^\mathsf{T}\boldsymbol{C}_0\right)$. The camera's orientation in the above chessboard coordinate system is characterized by the rotation matrix $\mathsf{R} = (\boldsymbol{h}\ \boldsymbol{v}\ \boldsymbol{a})$, with $\boldsymbol{h}, \boldsymbol{v}, \boldsymbol{a} \in \mathbb{R}^3$ being the direction of the horizontal image axis, the vertical image axis and the optical axis respectively. In case we have a shallow DOF, we now move the camera stepwise forward (typically on a motorized or manual linear stage) to take a series of images with different distances to the object, while keeping orientation and intrinsics constant:

$$\mathbf{P}_i = \mathbf{K}\left(\mathsf{R}^\mathsf{T} \mid -\mathsf{R}^\mathsf{T}\boldsymbol{C}_i\right), \tag{1}$$

where $\boldsymbol{C}_i = \boldsymbol{C}_0 + \lambda_i \boldsymbol{a}$ and λ_i encodes the magnitudes of the forward step from the start position $\boldsymbol{C}_0$ [2]. We call each of these images a *sub-image* of the stack of images for a particular object pose. Since intrinsics and focus of the lens are kept constant, in all these images, only those points appear sharp which are inside[3] the focus plane at distance d with respect to the camera. We can imagine that the focus plane sweeps over the object when we move the camera forward. A point $\mathbf{X}$ in space is projected into the ith image as $\mathbf{x}_i$:

$$\mathbf{x}_i = \mathbf{P}_i\,\mathbf{X} = \mathbf{K}\left(\mathsf{R}^\mathsf{T} \mid -\mathsf{R}^\mathsf{T}\boldsymbol{C}_i\right)\,\mathbf{X} \tag{2}$$

Substituting R and $\boldsymbol{C}_i$, and since $\boldsymbol{h}, \boldsymbol{v}, \boldsymbol{a}$ are orthogonal:

$$\mathbf{x}_i = \mathbf{K}\begin{pmatrix} \boldsymbol{h}^\mathsf{T} & -\boldsymbol{h}^\mathsf{T}\boldsymbol{C}_0 \\ \boldsymbol{v}^\mathsf{T} & -\boldsymbol{v}^\mathsf{T}\boldsymbol{C}_0 \\ \boldsymbol{a}^\mathsf{T} & -\boldsymbol{a}^\mathsf{T}\boldsymbol{C}_0 - \lambda_i \end{pmatrix}\mathbf{X} \tag{3}$$

We now synthesize a new focus-stacked image $\tilde{\mathcal{I}}$ of the same dimension as the original image, where we pick from each of the input images $\mathcal{I}_i$ only the sharp pixels, i.e. those that were in the focus plane at distance d relative to the camera[4]:

$$\tilde{\mathcal{I}}(\mathbf{x}) := \mathcal{I}_i(\mathbf{x}) \quad \text{with} \quad (0\ 0\ 1)\,\mathbf{P}_i\,\mathbf{X} = d \quad \text{and} \quad \mathbf{x} = \mathbf{P}_i\,\mathbf{X} \tag{4}$$

for some 3D scene point $\mathbf{X}$ (with Euclidean representation $\boldsymbol{X}$). The corresponding image pixel is therefore picked from the ith image, *iff*

$$\boldsymbol{a}^\mathsf{T}\boldsymbol{X} - \boldsymbol{a}^\mathsf{T}\boldsymbol{C}_0 - \lambda_i = d \tag{5}$$

Plugging this into equation 3 provides its position in the focus-stacked image:

$$\tilde{\mathbf{x}} = \mathbf{K}\begin{pmatrix} \boldsymbol{h}^\mathsf{T}\boldsymbol{X} - \boldsymbol{h}^\mathsf{T}\boldsymbol{C}_0 \\ \boldsymbol{v}^\mathsf{T}\boldsymbol{X} - \boldsymbol{v}^\mathsf{T}\boldsymbol{C}_0 \\ d \end{pmatrix} \tag{6}$$

Consequently, the focus stacked image is represented by the affine camera

$$\tilde{\mathbf{P}} = \mathbf{K}\left(\begin{array}{c|c} \boldsymbol{h}^\mathsf{T} & -\boldsymbol{h}^\mathsf{T}\boldsymbol{C}_0 \\ \boldsymbol{v}^\mathsf{T} & -\boldsymbol{v}^\mathsf{T}\boldsymbol{C}_0 \\ (0\ 0\ 0) & d \end{array}\right) \tag{7}$$

The focus stacking therefore removes the perspective effects and creates a parallel projection instead.

Transformation Between Chessboard and Image Plane

We now observe a chessboard (in the $z = 0$ plane, as defined earlier) and again capture a series of sub-images, given a particular camera pose. We later refer to the ith sub-image of the jth stack (chessboard orientation) as image $\mathcal{I}_{i,j}$. The chessboard contains a number of 3D points $\mathbf{X}_h$ at the corners, whose projection to the ith sub-image of the jth stack

[2] Throughout the paper, we assume that the forward stacking motion axis is aligned with the optical axis, which is reasonable in a carefully crafted macro camera setup. However, it can also be included easily in the optimization. Please see Sect.1 of the supplementary material for an in-detail discussion of the misalignment issue

[3] In practice, not only exactly one plane is in focus, but the DOF covers a range with decreasing image sharpness when moving away from the focus plane. As long as the unsharpness effects are not noticable, e.g. far below one pixel, also points at those distances appear sharp. However, for the sake of deriving the model we concentrate on the sharpest points.

[4] Note that focus stacking algorithms for creating beautiful images can do extra blending, de-ghosting, warping or other non-linear operations, but here we consider only basic focus stacking for our camera calibration.

is $\mathbf{x}_{i,j}^h$ (cf. Fig. 3). For the sake of readability we omit stack indices j, as well as point indices h, wherever we discuss generally valid relations that do not refer to a particular point or stack. For instance, coordinates of the chessboard are mapped to image coordinates as follows

$$\tilde{\mathbf{x}} = \mathbf{K} \left(\begin{array}{c|c} \boldsymbol{h}^\mathsf{T} & -\boldsymbol{h}^\mathsf{T}\boldsymbol{C}_0 \\ \boldsymbol{v}^\mathsf{T} & -\boldsymbol{v}^\mathsf{T}\boldsymbol{C}_0 \\ (0\ 0\ 0) & d \end{array} \right) (k\ l\ 0\ 1)^\mathsf{T} \tag{8}$$

Omitting the $z = 0$ column, we obtain a 3×3 homography matrix, which is actually an affine mapping:

$$\tilde{\mathbf{x}} = \mathbf{K} \begin{pmatrix} h_x & h_y & -\boldsymbol{h}^\mathsf{T}\boldsymbol{C}_0 \\ v_x & v_y & -\boldsymbol{v}^\mathsf{T}\boldsymbol{C}_0 \\ 0 & 0 & d \end{pmatrix} (k\ l\ 1)^\mathsf{T} \tag{9}$$

Substituting $\mathbf{K}$ with

$$\mathbf{K} = \begin{pmatrix} f & & c_x \\ & f & c_y \\ & & 1 \end{pmatrix}, \tag{10}$$

we obtain

$$\tilde{\boldsymbol{x}} = \underbrace{\frac{f}{d}\begin{pmatrix} h_x & h_y \\ v_x & v_y \end{pmatrix}}_{A} (k\ l)^\mathsf{T} + \underbrace{\frac{f}{d}\begin{pmatrix} -\boldsymbol{h}^\mathsf{T}\boldsymbol{C}_0 \\ -\boldsymbol{v}^\mathsf{T}\boldsymbol{C}_0 \end{pmatrix} + \begin{pmatrix} c_x \\ c_y \end{pmatrix}}_{\boldsymbol{o}} \tag{11}$$

This equation describes how a point (k, l) on the chessboard is mapped into the focus stacked image $\mathcal{I}$ by an affine transformation. The first part A is responsible for the shape, orientation and scale of the pattern, the second part $\boldsymbol{o}$ affects only the position in the image (offset).

Affine cameras are independent of translations if the relative coordinates, with respect to a given reference point, are used both in space and in image [19] and we can simply subtract a reference point $\tilde{\boldsymbol{x}}_{\boldsymbol{r}}$ (e.g. top left corner) to obtain an equation without $\boldsymbol{o}$. Therefore, points on the chessboard relative to the reference point in the focus stacked image are

$$\hat{\boldsymbol{x}} = \underbrace{\frac{f}{d}\begin{pmatrix} h_x & h_y \\ v_x & v_y \end{pmatrix}}_{A} \left(\hat{k}\ \hat{l}\right)^\mathsf{T}, \tag{12}$$

with $\hat{\boldsymbol{x}} = \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_{\boldsymbol{r}}$ and $(\hat{k}, \hat{l}) = (k - k_r, l - l_r)$. The matrix A can be readily measured from the image: Its columns are simply the image space vectors for a horizontal chessboard step and a vertical step, respectively (see also Fig. 3).

4. Macro Lens Calibration

After having analyzed the macro setting, we now turn to estimating the camera's parameters. The overall approach lends from classical camera calibration [25] and takes several images of a chessboard for which the internal camera parameters are optimized, jointly with the individual camera poses. However, since in each image we can observe only a fraction of the chessboard due to limited DOF, we capture an entire series of *sub-images* for the pose that differ only in a camera forward movement by a few micrometers each (as can be seen in Fig. 3). We stack those images into an all-sharp focus stacked image, in which we can detect the board, but which suffers from parallel projection as outlined in the previous section, nevertheless this is useful for predicting corner positions in the sub-images. The focus-stacked image also provides partial calibration information such as the magnification f/d and information about the chessboard orientation $(\boldsymbol{h}, \boldsymbol{v})$, which we extract and use as initialization for optimizing all parameters.

4.1. Focus Stacking of Chessboard Sub-images

For each individual camera pose, in order to obtain the focus stacked image containing all the sharp corners positions, we proceed like this: Since our objects are black and white chessboards, we do not need to reason about depth discontinuities and can stick to a simple approach: We compute the modified Laplacian [16] at each pixel position in each sub-image. Afterwards, at each image position we select the pixel with the highest Laplacian response at this position among all sub-images and copy its intensity value into the focus-stacked image. We also store from which image we took the respective pixel.

4.2. Detection of Chessboard Corners

Each chessboard sub-image contains a lot of blur in the unsharp areas which hinders direct detection of the chessboard. As detection of partial chessboards from the sub-images seems complicated we utilize the focus-stacked image as a prediction: First we run chessboard detection on the focus stacked image. For difficult cases such as strong artifacts due to our simple focus stacking, this step can be supported by clicking the four outermost corners and predicting, followed by a local search in the focus-stacked image. Once all corners are found in the focus-stacked image we leverage the fact that each pixel originated from one of the sub-images, and we predict the corner positions in the resp. sub-images. In order not to change the position of the corner points, no post-processing is performed on the sub images. Rather we do a local corner search around the predicted position. We also predict the corner in the neighboring sub-images and search for the precise position. This way we obtain a set $\mathcal{S}_{ij}$ of detected corners for each sub-image in each chessboard pose.

4.3. Initial Camera Calibration from a Single Stack

Focal Length and Rotation Matrix from Affine Camera From equation 11 it can be seen that the apparent deformation and scale of the chessboard is caused by the matrix

$$\mathsf{A} = \frac{f}{d}\begin{pmatrix} h_x & h_y \\ v_x & v_y \end{pmatrix} \tag{13}$$

The image offset vector obtained when moving a horizontal step $(1\ 0)$ on our chessboard, is therefore $\frac{f}{d}(h_x\ v_x)$ (and for a vertical step: $\frac{f}{d}(h_y\ v_y)$). Let $\mathsf{V} = \mathsf{A} \cdot \mathsf{A}^{\mathsf{T}}$:

$$\mathsf{V} = \left(\frac{f}{d}\right)^2 \begin{pmatrix} h_x^2 + h_y^2 & h_x v_x + h_y v_y \\ h_x v_x + h_y v_y & v_x^2 + v_y^2 \end{pmatrix} \tag{14}$$

Since $\boldsymbol{h}, \boldsymbol{v}$ are orthonormal we obtain

$$\mathsf{V} = \mu \begin{pmatrix} 1 - h_z^2 & -h_z v_z \\ -h_z v_z & 1 - v_z^2 \end{pmatrix}, \tag{15}$$

where $\mu = (f/d)^2$. When solving equation 15 for μ, results in a 4th-degree polynomial, where 4 solutions are obtained, but 2 of them are duplicated. Once f/d is known, h_x, h_y, v_x, v_y can be readily computed from the affine transformation A as shown in equation 13. Next we have

$$h_z = \pm\sqrt{1 - h_x^2 - h_y^2} \quad \text{and} \quad v_z = -\frac{V_{10}}{h_z \mu}, \tag{16}$$

where V_{10} is the lower-left element of matrix V. Therefore, the third rotation vector can be computed as $\boldsymbol{a} = \boldsymbol{h} \times \boldsymbol{v}$. The so-computed matrix $\mathsf{R} = (\boldsymbol{h}, \boldsymbol{v}, \boldsymbol{a})$ may be noisy and hence not be a proper rotation matrix, but orthonormality can be enforced e.g. through singular value decomposition. The rotation obtained only serves as a start value for optimization in the next section. We assume the focus distance d is roughly known (e.g. a few centimeters) as it is practically needed to position the object at a reasonable distance. Then the focal length can be computed as $f = d\sqrt{\mu}$. This also only serves as an initial guess for the subsequent optimization.

So far, we have solved the focal length of the camera and the rotation matrix from one focus stacked image. However, as can be seen, we obtained 2 solutions when solving μ and each solution leads to another 2 rotation matrix solutions when solving h_z from equation 16. We hence prune the solutions which are geometrically infeasible, which leaves us with a two-fold pose ambiguity that can be resolved using the coarse relative depth information of the corners (depth-from-focus). In the next subsection, we will use all solutions to solve for the principal point of the camera and the translation vector.

Solving Principal Point and Translation from Homography Since the affine camera removes the perspective effects, the camera translation is not fully constrained by equation 11. Therefore, we turn to use one of the sub-images to recover the camera translation $\boldsymbol{t}$ as well as the principal point (c_x, c_y). Since the ith sub-image is taken by the perspective camera at a forward step of distance λ_i, it satisfies equation 3. And also the 3D point which the camera is observing has $z = 0$, implying that we can remove the $z = 0$ column of the projection matrix. Now, as the focal length f and the rotation matrix R have been computed, we can expand equation 3:

$$\mathbf{H}_i = \begin{pmatrix} \frac{fR_{00}+c_xR_{20}}{t_z-\lambda_i} & \frac{fR_{01}+c_xR_{21}}{t_z-\lambda_i} & \frac{ft_x}{t_z-\lambda_i} + c_x \\ \frac{fR_{10}+c_yR_{20}}{t_z-\lambda_i} & \frac{fR_{11}+c_yR_{21}}{t_z-\lambda_i} & \frac{ft_y}{t_z-\lambda_i} + c_y \\ \frac{R_{20}}{t_z-\lambda_i} & \frac{R_{21}}{t_z-\lambda_i} & 1.0 \end{pmatrix} \tag{17}$$

where $\boldsymbol{t} = (t_x, t_y, t_z)^{\mathsf{T}} = -\mathsf{R}^{\mathsf{T}}\boldsymbol{C}_0$. For initialization purposes, we assume that the step size of the camera forward stacking movement is known (as it is usually provided by the motorized or manual linear stage) and therefore λ_i is given. Now, the camera translation and the principal point can be solved sequentially.

Note that, when solving v_z from equation 16, h_z should not be equal to 0 (optical axis of camera perpendicular or parallel to board), but these special cases can be easily identified and addressed.

4.4. Maximum-Likelihood Estimation of All Calibration Parameters

Since we assume that the coarse focus distance d and the forward stacking step size is known when solving for camera calibration parameters, they can only be considered as starting values. In this section, we describe a corresponding optimization procedure to optimize all calibration parameters using all sub-images at the same time.

Different from the classical camera calibration approach [25], this paper not only estimates the intrinsics of the camera but also the forward stacking movement of each sub-image. Assume the camera observes the chessboard at m different poses, then the camera takes n sub-images in a forward stacking motion at each pose.

Suppose there are l chessboard corners identified in the sharpest region of the image, then we seek to minimize the following energy function:

$$E(\Theta) = \sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{h\in\mathcal{S}_{ij}} \|\boldsymbol{x}_{i,j}^h - \boldsymbol{\pi}(\boldsymbol{X}_h, \mathcal{K}, \mathsf{R}_j, \boldsymbol{C}_{i,j})\|^2 \tag{18}$$

where $\boldsymbol{x}_{i,j}^h$ the observation in ith sub-image of jth stack of the hth corner $\boldsymbol{X}_h$ of the 3D calibration board. As we do not observe all corners in all sub-images, $\mathcal{S}_{ij}$ denotes the set of sharp observations in sub-image ij. Next, $\boldsymbol{\pi}(\boldsymbol{X}_h, \mathcal{K}, \mathsf{R}_j, \boldsymbol{C}_{i,j})$ is the perspective projection equation and $\boldsymbol{C}_{i,j} = \boldsymbol{C}_{0,j} + \lambda_i \boldsymbol{a}_j$, where $\boldsymbol{a}_j$ is the 3rd column of rotation matrix R_j. Note that instead of the simple calibration matrix $\mathbf{K}$, obtained from the closed form solution, we

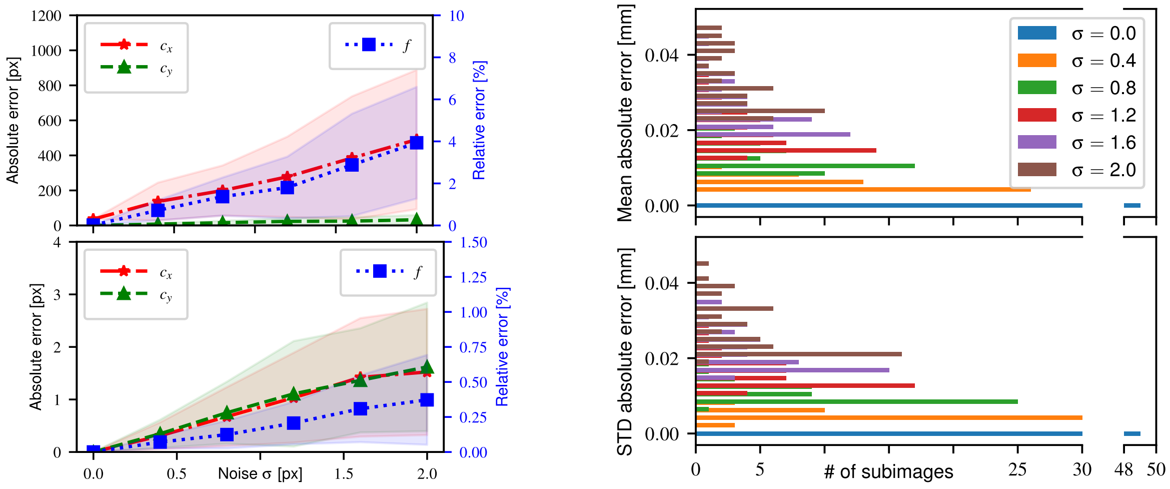

Figure 4. Camera model parameter estimation on synthetic data avg'd over 100 runs and subject to increasing noise. **Left:** AVG./STD. error of f relative to GT, and AVG./STD. absolute error of c_x and c_y. **Left, top:** Close-form initialization result. **Left, bottom:** Optimized result. The colorfill indicates one STD. Please note the *different scales on the y-axis*. **Right, top:** mean estimation error of λ_i binned into 0.005 mm error ranges. **Right, bottom:** STD. of the λ_i estimates with same binning.

here optimize the parameters of a more generic function $\mathcal{K}$ with the same meaning as $\mathbf{K}$ (mapping from rays to image positions), but utilizing more generic parameters such as distortion (with distortion parameters initialized as 0).

The MLE is obtained using the Levenberg-Marquardt algorithm (Ceres-Solver [1]). [5]

5. Evaluation

The evaluation is conducted on synthetic data and real-world images acquired by a macro lens camera system. The synthetic evaluations are twofold: First, numerically simulated projections of 3D points by a macro lens camera with perfectly known ground truth and noise models are employed. This helps us to validate the proposed macro lens camera calibration approach and to evaluate the accuracy and robustness against noise. The second type of synthetic data is synthesized by ray-tracing, simulating a perspective camera with a thin lens. Thus, we can investigate the parallel projection effect in the focus stacked images and evaluate camera calibration in the presence of uncertainty in corner detection and a simulated DOF.

Evaluation on Numerical Simulations First, we simulate a macro-scale chessboard with square size of $1mm \times 1mm$ observed by a perspective camera with a macro lens (focal length $f = 6450.0$ pixels, principal point $(c_x, c_y) = (1032.0, 688.0)$ pixels, focus distance $d = 45mm$) under forward focus stacking movement. The step size of each forward movement is $0.02mm$ and only the 3D calibration targets that are inside the DOF of the camera will be projected to the corresponding sub-image. Later on, zero-mean Gaussian noise with $\sigma \in \{0.0, 0.2, \ldots 2.0\}$ pixels is added to each simulated projection.

We perform 100 calibration trials on each noise level using the closed-form solution as described in Sect. 4.3. Afterwards, we perform optimization on the parameters and obtain the optimal values. For each trial, the step size of the forward stacking movement is initialized as 20 % off from the ground truth, and we add Gaussian noise to the focus distance d with noise level of $\sigma = 20$ % of ground truth d. Estimation errors for each calibration parameter are recorded and results are displayed in Fig. 4. As can be seen we obtained a good estimate of f and c_y from the closed-form solution, however, c_x deviates from the GT solution with increasing noise. Nevertheless, the optimization brings all the parameters back to the optimal state. It is evident that the calibration approach produces high accuracy camera intrinsic calibration as well as focus stacking movements despite a high noise level.

Evaluation on Ray-Tracing Data Then we basically keep all the simulation setups the same but render photo-realistic images with de-focus effect using the ray-tracing software Mitsuba2 [17]. Here, the thin-lens perspective camera with aperture of $1.425mm$ is employed to create a de-focus effect. Sample sub-images are shown in Fig. 3 (middle row). The estimated focal length is $f_x = 6448.79$ pixels, the principal point is $(c_x, c_y) = (1031.92, 687.395)$ pixels, which is nearly the ground truth, and the reported reprojection error is 0.051 pixels. Two sub-images with

[5] Note that in the perspective projection equation, we also include the lens distortion parameters, for more details about distortion parameters, we kindly refer interested readers to [13].

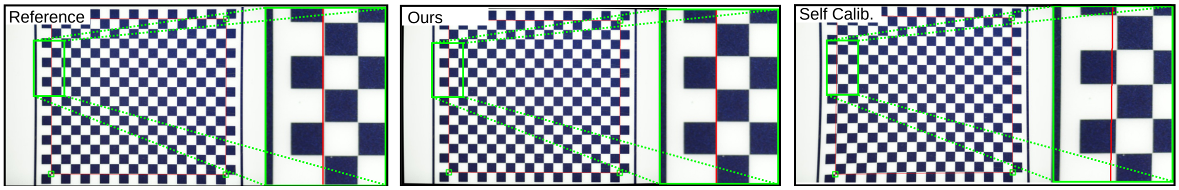

Figure 5. To verify the calibrated intrinsics, we draw 4 straight lines connecting the outer 4 corners to evaluate the straightness of lines after undistortion. **Left:** the original reference image; **Middle:** the undistorted image using the calibrated intrinsics; **Right:** undistorted image using self-calibration during bundle-adjustment. Each **right** column shows the zoomed details of the respective condition.

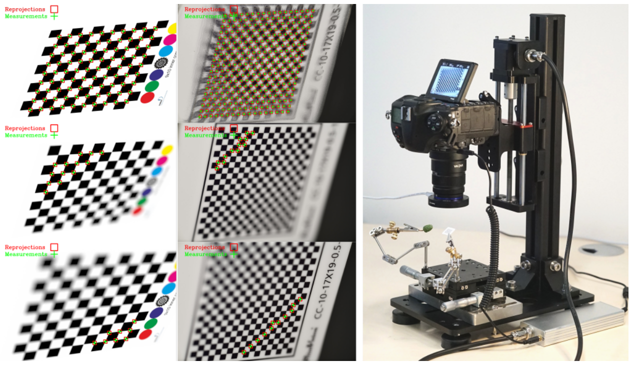

Figure 6. **First two columns:** Reprojection results are shown in red squares and measured corners are shown in green crosses. **Left:** Ray-tracing synthetic data. **Right:** Real-world data . **Top row:** Reprojection from the chessboard to the focus stacked image using reconstructed affine transformation. **Bottom two:** Reprojection from the chessboard to two of the sample sub-images. **Rightmost column:** Real-world evaluation experimental platform. The camera is mounted on a vertical focus stacking rail. A ring LED is fixed on the front section of the lens. XY-micrometer and mechanical arms constitute the loading platform. A drive box controls the movement of the rail and triggers the camera.

drawn reprojection results and measured chessboard corners are shown in Fig. 6 (left, bottom). Next, to evaluate the estimated camera intrinsic parameters and also to verify the affine transformation model for the focus stacked image, we reconstruct the affine transformations using the estimated parameters and project 3D chessboard corners onto all focus stacked images (one of them is shown in Fig. 6 left, top). The resulting mean reprojection error is 1.179 pixels, which is in agreement with the focus stacked images.

Real-World Evaluation To verify the effectiveness of the method, a real-world evaluation including resolution tests under different apertures, a real macro lens camera calibration and an SfM evaluation was carried out.

The image acquisition platform is shown in Fig. 6. This platform (WeMacro) can acquire vertical-up-to-down focus stack image sequences and triggers the shutter automatically. The rail is driven by a stepping motor, and the minimum step size of the movement is $0.001mm$, with no accumulative error, which can ensure that the relative position of each stack movement can be repeated after a calibration of the backlash error of the rail. A Nikon D850 camera with a Laowa 25mm f/2.8 2.5-5X Ultra Macro lens is used as the macro lens camera system. The object platform combines an XY-micrometer with mechanical arms that ensure that the specimen is located in the field of view of the camera. The employed macro-scale chessboard has a square size of $0.5mm \times 0.5mm$ and is printed using photolithography technology with a printing accuracy of $0.001mm$.

Method	f_x	f_y	c_x	c_y
Ours	8087.03	8083.17	1162.1	764.99
Self Calib.	10142.81	10111.64	1032.00	688.00
Method	k_1	k_2	p_1	p_2
Ours	0.935	-0.576	0.013	0.021
Self Calib.	-1.915	63.405	-0.068	0.114

Table 1. Calibration results used in the evaluation in Fig. 5.

Resolution Evaluation First, an AF-1951 chart is used to evaluate the center and boarder resolution of the image at different apertures (cf. Fig. 2). Reproduction ratio on the lens is set to be 2.5. As a compromise between number of sub-images needed and spatial resolution, we chose aperture f/8 with a reported DOF of $0.257mm$.

Calibration Evaluation Next, we evaluate the proposed macro lens camera calibration approach using the above described system. We first set the step size of the motorized linear stage to $0.15mm$, and then place the macro-scale chessboard approximately at $54mm$ in front of the camera, which is also the focus distance of the camera system. Afterwards, we take 6 stacks of images where each stack contains 47 sub-images. For each stack, we change the chessboard pose and we try our best to maximally rotate the chessboard while keeping it inside both the field of view and the extended DOF of the camera. Two sample sub-images from one of the stacks can be found in Fig. 6 (right, bottom), and the focus stacked image from this stack is shown in the top row. Finally, the proposed calibration approach is applied to the captured dataset, which yields a

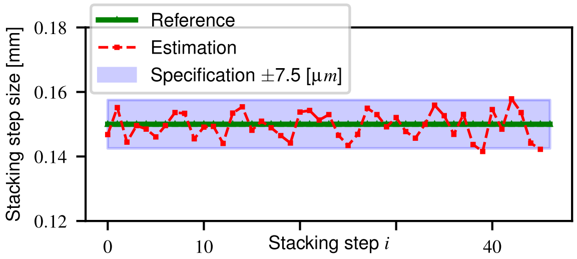

Figure 7. Calibration of the forward stacking step sizes for the real macro lens camera. The reference stacking step size is $0.15mm$, and the motorized stepping accuracy is 5% of the step size ($7.5\mu m$ in this case). The latter is indicated by the blue error fill.

mean reprojection error of 0.20 pixels. The resulting intrinsics are shown in Tab. 1. In addition to that, the calibrated forward stacking step sizes are shown in Fig. 7, the estimated step sizes are almost perfectly in agreement with the manufacturer's specification.

To verify the calibrated camera intrinsic parameters, we place the chessboard frontal parallel to the camera lens, and take an independent photo once the chessboard is entirely brought in focus. Then we undistort the photo with estimated intrinsic parameters. As shown in Fig. 5, the original image is shown with distortion and the chessboard corners in between each of the two corners do not align along the line. After undistortion using our parameters it is evident that straight lines are depicted straight. Next, we use one of the focus stacked images to re-estimate the affine transformation between the 3D chessboard plane and the corners on the focus stacked image, and this time we re-calculate the focus distance d out of the affine transformation. The re-calculated focus distance is $54.51mm$ which is in agreement with the hand-measurement. Afterwards, we use the re-calculated focus distance, the estimated camera intrinsics together with the estimated stack poses to reconstruct the affine transformations for the other stacks (the stack used for re-calculating focus distance is excluded). With the reconstructed affine transformations, we project 3D chessboard corners onto each of the focus stacked image and compute the mean reprojection error over all stacks. The reprojection error is measured as 3.15 pixels and one sample focus stacked image with reprojection results is shown in Fig. 6 (top, middle). This is a reasonable result given that lens distortion and the forward stacking motion offset are not considered in the affine transformation.

For comparison, we also run self-calibration using a state-of-the-art SfM [21] software on 22 poses each containing 46 sub-images (1012 images in total) of the bud of Fig. 1. Due to the shallow DOF only 15% of the images were registered when producing a 3D model of the bud (not shown). We obtain the intrinsic parameters shown in Tab. 1, which we also use for undistorting the chessboard photo. The undistorted image is displayed in Fig. 5 (right), which shows that the estimated radial distortion is quite off, as compared to our calibration-target-based method, although registration succeeded for many images. It also indicates that self-calibration-SfM based reconstruction might be skewed, since it is not using correct intrinsics.

Application: Structure from Motion Finally, we rerun the reconstruction, but this time enforcing to use our calibration, and we obtain the 3D model as presented in Fig. 1. In this paper, we focus on the calibration leaving a detailed evaluation and comparison of 3D macro SfM to future work. However, this result already shows that our calibration is useful for 3D reconstruction of tiny objects.

6. Conclusion

When maximizing spatial image resolution as often desired in macro photography, the shallow DOF hinders classical camera calibration. We have shown that when combining multiple images into one by focus stacking, this means that this focus stacked image actually obeys an affine camera model. We have derived a closed-form solution to extract focal length, principal point and board orientation from an image of a chessboard and proved on synthetic data that the algorithms are robust to noise and that the results are suitable to initialize a maximum-likelihood camera parameter estimation using partial chessboard observations in multiple views. For this we have devised a chessboard corner detection strategy that also employs the focus-stacked image as a prediction. We have then calibrated a real macro lens and show consistent and plausible results comparing to the manufacturer's data sheet, and image undistortion with the obtained distortion parameters produces straight lines, whereas parameters obtained through self-calibration using generic SfM software did not. Future work will investigate applicability of the system to shallow DOF microscopy, for which the corner detection strategy seems already readily usable. We think that using calibration information from known targets can strongly improve 3D reconstruction of micro objects in cases (close to) degenerate for self-calibration and where currently no other calibration approach exists.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101000858 (TechOceanS) and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) Projektnummer 396311425, through the Emmy Noether Programme. We are also grateful for support from the Chinese Scholarship Council (CSC) for M. She.

References

[1] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. http://ceres-solver.org. 2, 6

[2] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. Interactive digital photomontage. *ACM Trans. Graph.*, 23(3):294–302, Aug. 2004. 1, 2

[3] George Biddell Airy. On the diffraction of an object-glass with circular aperture. *Transactions of the Cambridge Philosophical Society*, 5:283, 1835. 2

[4] LM Angheluță and R Rădvan. Macro photogrammetry for the damage assessment of artwork painted surfaces. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019. 2

[5] Jonathan Brecko, Aurore Mathys, Wouter Dekoninck, Maurice Leponce, Didier VandenSpiegel, and Patrick Semal. Focus stacking: Comparing commercial top-end set-ups with a semi-automatic low budget approach. a possible solution for mass digitization of type specimens. *ZooKeys*, (464):1, 2014. 2

[6] D. C. Brown. Close-range camera calibration. champ, 1971. 2

[7] Thomas Clark. *Digital Macro and Close-Up Photography For Dummies*. John Wiley & Sons, 2011. 1, 2

[8] Marc Levoy David E. Jacobs, Jongmin Baek. Focal stack compositing for depth of field control. Technical Report 2012, Stanford Computer Graphics Laboratory, 2012. 1, 2

[9] J. Ens and P. Lawrence. An investigation of methods for determining depth from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(2):97–108, 1993. 2

[10] O. D. Faugeras and S. J. Luong, Q. T.and Maybank. Camera self-calibration: Theory and experiments. In G. Sandini, editor, *Computer Vision — ECCV'92*, pages 321–334, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg. 2

[11] Alessandro Gallo, Maurizio Muzzupappa, and Fabio Bruno. 3d reconstruction of small sized objects from a sequence of multi-focused images. *Journal of Cultural Heritage*, 15(2):173–182, 2014. 2

[12] Paul Grossmann. Depth from focus. *Pattern recognition letters*, 5(1):63–69, 1987. 2

[13] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision (Second Edition)*. Cambridge University Press, second edition, 2004. 2, 6

[14] Samuel W. Hasinoff and Kiriakos N. Kutulakos. Light-efficient photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2203–2214, 2011. 2

[15] Erez Marom. Macro photography: Understanding magnification. *Digital Photography Review*, 28, 2011. 1, 2

[16] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(8):824–831, 1994. 2, 4

[17] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, 38(6):1–17, 2019. 6

[18] Hugues Plisson and Lydia V Zotkina. From 2d to 3d at macro-and microscopic scale in rock art studies. *Digital Applications in Archaeology and Cultural Heritage*, 2(2-3):102–119, 2015. 2

[19] Long Quan. Self-calibration of an affine camera from multiple views. *International Journal of Computer Vision*, 19(1):93–105, 1996. 2, 4

[20] Graham Saxby. *The science of imaging*. CRC Press, 2016. 2

[21] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 8

[22] Bernhard Ströbel, Sebastian Schmelzle, Nico Blüthgen, and Michael Heethoff. 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, 2018. 2

[23] Y. Xiong and S.A. Shafer. Depth from focusing and defocusing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 68–73, 1993. 2

[24] Chao Zhang, John Bastian, Chunhua Shen, Anton Van Den Hengel, and Tingzhi Shen. Extended depth-of-field via focus stacking and graph cuts. In *2013 IEEE International Conference on Image Processing*, pages 1272–1276. IEEE, 2013. 2

[25] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 2, 4, 5