Table of Contents

II

Aerial image mosaicking with online calibration - A feasibility study

Gellért Máttyus[1] and Friedrich Fraundorfer[2]

[1] German Aerospace Center, Oberpfaffenhofen, Germany
gellert.mattyus@dlr.de
[2] Graz University of Technology, Graz, Austria
fraundorfer@icg.tugraz.at

Abstract. Compiling individual aerial images to a larger mosaic image is important for many remote sensing tasks, e.g. mapping. The standard way to address this problem is to orthorectify the image and later compile it together. For the orthorectification the orientation and the location of the camera has to be measured accurately and a Digital Elevation Model (DEM) is needed or ground control points have to be set manually. In this paper we present a feasibility study on an approach which works purely on original camera images without any GPS and/or IMU. The intern and extern camera parameters and the 3D feature points are calculated based on Structure from Motion (SfM). The ground surface is assumed to be flat and a plane is fitted on the 3D feature points. A virtual camera is calculated then which is perpendicular to the plane and the images are transformed into each other by a homography. We test this approach on image sequences captured by a standard DSLR camera.

1 Introduction

With the spread of Micro Aerial Vehicles (MAVs) and light weight cameras more and more image sequences can be captured for a lower cost. This can be done by simply mounting a DSLR camera on an airplane, air glider or a MAV. This is a cost effective solution but compared to a professional aerial imaging system the external camera parameters (i.e. the camera location and orientation) are not measured, while the internal parameters (i.e. the focal length and radial distortion) are less stable and thus need more frequent calibration. To compile the images together to a mosaic multiple problems have to be solved: (i) the internal camera parameters have to be estimated, (ii) the relative external camera parameters have to be calculated and (iii) the images have to be warped to a common coordinate system. In this paper we test the feasibility of a solution which makes these 3 steps automatically online, without the need for an off-line internal camera parameter estimation. First the internal and external camera parameters are calculated jointly by Structure from Motion (SfM), then the scene is modeled by a plane and the images are warped together by homography. The figure 2 shows an example for the inputs and output of this process.

We test this approach on 4 image sequences captured by an off-the-shelf DSLR camera mounted on an airplane flying approximately 1000 m above the ground without calibration. The standard way to process these images is to use a highly accurate (and

expensive) IMU and GPS and project the images on a Digital Elevation Model (DEM). On the test image sequences we show that the alternative solution using only the camera images provides also an accurate mosaic image for areas without considerable relief.

Original frame images

Mosaic image

Fig. 1. Original aerial images at the top and the mosaic created with online camera calibration in the bottom.

2 Related work

The problem of registering aerial and MAV images was addressed in different domains.

Mattyus et al. [6] use a simple homography to register MAV images and track multiple objects images. This method doesn't address the problem of camera calibration and can not provide an orthophoto like mosaic.

The traditional photogrammetric software can be used for compiling aerial and mav images to a mosaic but they need manual interaction. In [4] a dense stereo matching is computed from MAV images over an area.

Structure from motion methods target to reconstruct 3D scenes from images without any additional information. Wu [11] gives an advanced solution for the SfM problem targeting an unordered image collection. This results in high computational time since feature matches are calculated across all the images. In case of image sequences this can be limited to consecutive images, reducing the complexity of matching from $O(n^2)$ to $O(n)$. An other disadvantage is the 1 parameter radial distortion camera model. This might lead to reconstructing plane surfaces as spheres [12].

3 Image mosaicking method

Here we describe our method, first the computation of camera parameters and 3D feature points (i.e. the Structure from Motion). Then the warping of images to a common coordinate system.

3.1 Structure from Motion

The method to recover the camera parameters and the 3D sparse structure of the scene is called Structure from Motion. A detailed description and mathematical basics can be found in the book of Hartley and Zissermann [5].

We apply the Mavmap software [10] which is available online [1]. This is a brief description of this SfM pipeline, for more detailed explanation please read [10].

Feature matching The speed of feature detection and calculation is crucial, therefore the SURF features [1] are applied which provide both fast speed and robust performance. It is assumed that the images overlap and are in a sequence. This allows to calculate the feature matches only between images with a given acquisition time difference. This assumption allows to reduce the complexity of feature matching from $O(n^2)$ to $O(n)$ where n is the number of images.

Initial reconstruction Given an image pair with an initialization for the intrinsic camera parameters the essential matrix $\mathbf{E}$ is calculated using the five-point algorithm of Nister [8].

Sequential reconstruction A new camera is registered to the previously calculated 3D model from 2D-3D correspondences. This is also known as the PnP problem [3]. For the robust estimation RANSAC [2] is applied for four correspondences between known 3D coordinates and image points. The retrieved camera parameters and 3D points are refined by Bundle Adjustment.

[1] https://github.com/mavmap/mavmap

Bundle adjustment Bundle adjustment is a process to refine the camera parameters $\mathbf{C}_j = \{\mathbf{C}_1, \ldots, \mathbf{C}_n\}$ and 3D feature points $\mathbf{P}_i = \{\mathbf{P}_1, \ldots, \mathbf{P}_m\}$ by non-linear optimization [5] based on the 2D image feature points $\mathbf{X}_k = \{\mathbf{X}_1, \ldots, \mathbf{X}_o\}$ as uncertain measurements. The projection function $\mathbf{Q}(\mathbf{P}_i, \mathbf{C}_j)$ maps the 3D feature points to 2D image points. k is the index of a 2D feature of the ith 3D feature point in the image j. A cost function S is defined as the overall reprojection error and this is minimized during the optimization.

$$S = \frac{1}{2} \sum_{k=1}^{o} \rho_k(||\mathbf{X}_k - \mathbf{Q}(\mathbf{P}_i, \mathbf{C}_j)||_2^2) \tag{1}$$

where $\rho_k()$ is the loss function.

Loop detection Only the images with a given acquisition time difference are matched for faster speed. If the camera visits the same location later again, the overlapping images are not matched. Since the camera registration has a drift this can lead to registration errors. To overcome this problem a fast loop detection is applied. This is done by fast image retrieval based on SURF features and a visual word based vocabulary tree [9]. If the loop is detected the overlapping sequences are merged together.

3.2 Plane fitting and virtual reference camera extraction

The SfM step gives a feature point cloud of the scene and the camera parameters. In a general case with 3D structure, this doesn't allow to compute a transformation from one image to the other, because a full 3D reconstruction would be needed. However, if we can make assumptions about the scene, this can be overcome. In flat geographical areas the scene can be approximated by a single plane surface. This allows to calculate a homography between the images and warp them to a common coordinate system.

Fitting a plane and choosing a reference camera A plane can be described by 4 parameters $ax + by + cz + d = 0$. The 3D points can contain outliers, e.g. trees, buildings, etc. Therefore we use Ransac [2] to fit the plane on the 3D points by considering only the points within a distance to the plane as inliers.

Aerial images are typically used in an orthorectified form because in this case the scale is uniform in the whole image. It is preferable to have the same feature for our mosaic image. This can be done by choosing a camera as reference which looks perfectly perpendicular on the plane. In this case, all the 3D surface points have a fixed z_d depth coordinate and the perspective projection becomes an orthographic projection. Such an ideal camera is not available but a virtual one can be calculated. This virtual camera is calculated as one which is at the same distance from the plane as an arbitrary chosen but looking perfectly perpendicular on it. See the Figure 2 for illustration.

Mosaicking by homography If the 3D scene is plane, the transformation between the images can be described by a $\mathbf{H}_{3\times3}$ Homography matrix.

$$\mathbf{H} = \mathbf{K}(\mathbf{R} + \frac{1}{d}\mathbf{t}\mathbf{n}')\mathbf{K}^{-\mathbf{1}} \tag{2}$$

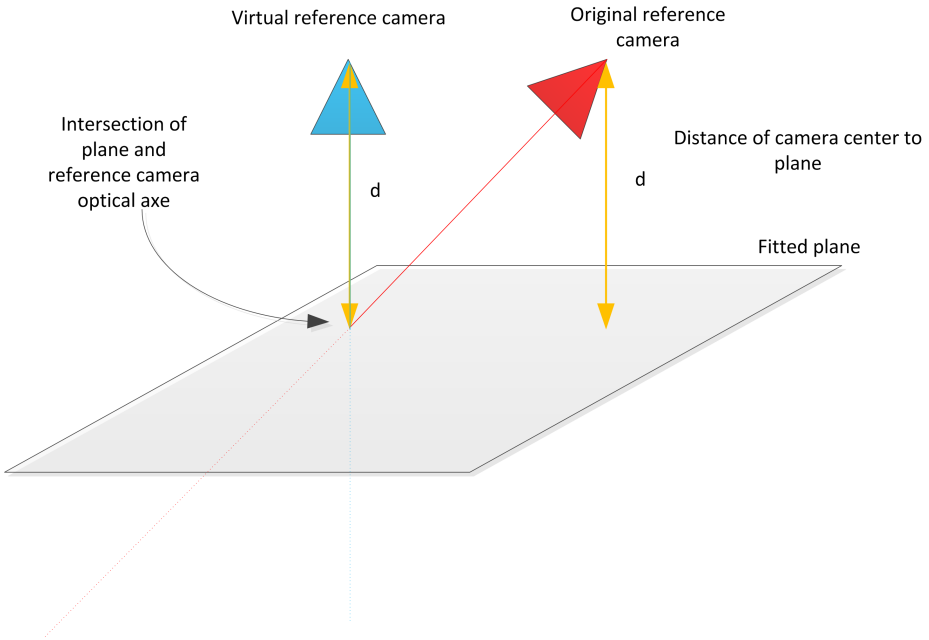


Fig. 2. The virtual reference camera looks perpendicular on the plane and is in the same distance as an arbitrary chosen camera.

where d is the distance of the virtual reference camera to the plane, $\mathbf{t}$ is the translation between the cameras, $\mathbf{n}$ is the normal vector of the plane, $\mathbf{R}$ is the rotation matrix from the reference camera to the other camera and $\mathbf{K}$ is the 3×3 camera matrix (matrix of the intrinsic parameters). We rectify the images and warp them to the virtual reference camera by the homography matrix.

4 Experimental results

Camera and images We show results on 4 image sequences captured over the German city, Munich. The images were captured by the *3K* camera system of the German Aerospace Center, similar as the system described in [7]. The Camera System consists of three commercial Canon EOS 1Ds Mark III cameras with 50mm Zeiss lenses and 5616×3744 pixel resolution. One camera is mounted in straight nadir direction, while the two others are tilted to the front and rear. Two sequences (short and long) of the nadir and front cameras are tested. The shorter consists of 25 images taken with 1 Hz frequency, while the longer has 50 images captured also with 1 Hz. The Figure 3 shows examples for the input images.

We use the OpenCV camera model with two radial κ and two tangential ρ distortion coefficients. The projected camera coordinates are transformed by the following

equations.

$$x' = x(1 + \kappa_1 r^2 + \kappa_2 r^4) + 2\rho_1 xy + \rho_2(r^2 + 2x^2) \quad (3)$$

$$y' = y(1 + \kappa_1 r^2 + \kappa_2 r^4) + \rho_1(r^2 + 2y^2) + 2\rho_2 xy \quad (4)$$

where $r^2 = x^2 + y^2$, x, y are the coordinates after projection on the camera plane, but before undistortion and x', y' after the undistortion.

Image sequence	reproj. cost [pix]	f_x	f_y	κ_1	κ_2
Nadir 25	0.12	7946.6	7942.4	-0.061	0.09
Nadir 50	0.124	8046.8	8044.2	-0.064	0.094
Front 25	0.117	8009.4	8020.9	-0.063	0.088
Front 50	0.123	8020.7	8018.6	-0.063	0.089

Table 1. The results of the SfM. The reprojection error in pixels and the calculated intern camera parameters. The focus length is in pixels.

The 50 mm nominal focal length on the full frame sensor with 36 mm width gives a focal length of 7800 pixels but this might differ from the true one. As a ground truth focal length we consider the one from a calibration of the same camera but with the use of the accurate IMU and GPS. This was only available for the nadir camera and it was 8035.2 pixels. By applying a longer image sequence our focal length calculation gets closer to this value. The initial focus length is set to 8000 pixels.

The tangential coefficient has small impact, we use it only for consistency with the OpenCV camera model. The results of the SfM are shown in the table 4. A problem might be the correlation of the focus length and the distance to the plane if the camera looks in nadir direction. To tackle this problem it is better to use longer image sequences with stronger camera movement and more 3D structure.

We show result mosaic images in Figure 4. The Figure 5 shows how the individual images fit in the mosaic. An error source are tall buildings where the parallax effect becomes visible.

Comparison to simple Homography Mattyus et al. [6] use simple Homography for the registration of the images. This is not suitable for a proper ortho-image like mosaic, since the reference coordinate system is chosen arbitrary.

5 Conclusion and outlook

We have shown the feasibility of creating aerial image mosaics fully automatically by using only single off-the-shelf camera. A future work would be to make also qualitative test for the used method. We have considered areas which can be estimated by a plane. This could be extended to a general 3D structure by applying dense stereo matching and reconstruct the 3D model of the area.

nadir front

Fig. 3. The original input frame images.

6 Acknowledgement

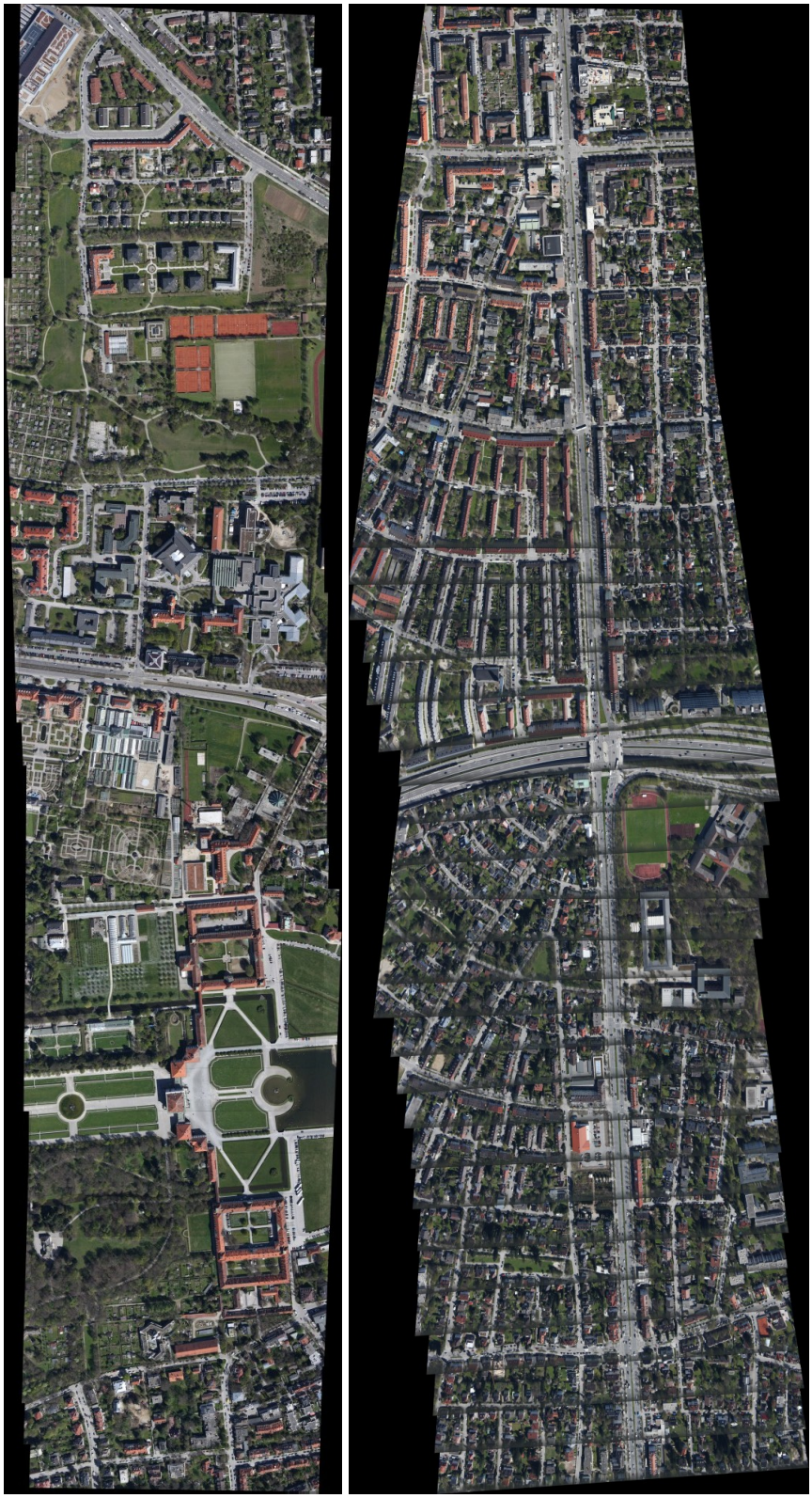
This work was financed by the DLR project Vabene++[2].

References

1. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.
2. M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.
3. J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 368–381, Berlin, Heidelberg, 2010. Springer-Verlag.
4. A. Greiwe, R. Gehrke, V. Spreckels, and A. Schlienkamp. Aspects of dem generation from uas imagery. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-1/W2:163–167, 2013.
5. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
6. G. Máttyus, C. Benedek, and T. Szirányi. Multi target tracking on aerial videos. *Proc. of ISPRS Workshop on Modeling of Optical Airborne and Space Borne Sensors*, pages 11–13, 2010.
7. G. Máttyus, F. Kurz, D. Rosenbaum, and O. Meynberg. A real-time optical airborne road traffic monitoring system. *KEPAF*, 2013.
8. D. Nister. An efficient solution to the five-point relative pose problem. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–195–202 vol.2, June 2003.
9. D. Nistr and H. Stewnius. Scalable recognition with a vocabulary tree. In *IN CVPR*, pages 2161–2168, 2006.
10. J. L. Schönberger, F. Fraundorfer, and J.-M. Frahm. Structure-from-motion for mav image sequence analysis with photogrammetric applications. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3:305–312, 2014.

[2] http://www.dlr.de/vabene/

11. C. Wu. Towards linear-time incremental structure from motion. In *3DTV-Conference, 2013 International Conference on*, pages 127–134, 2013.
12. C. Wu. Critical configurations for radial distortion self-calibration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.



nadir front

Fig. 4. The result mosaic images for the *nadir 25* and *front 25* image sequences. Note that for the front sequence the plane body obscures the image, this causes the dark edge between the individual images.

Fig. 5. Cropped part of the mosaic image showing the borders between two images. The images are stitched together without smoothing purposely to better show the accuracy. The black line shows the image border. The (a) and (b) show front camera images, while the rest are nadir images. The (f) shows the parallax on a taller building.