16^{th} *Computer Vision Winter Workshop*
Andreas Wendel, Sabine Sternig, Martin Godec (eds.)
Mitterberg, Austria, February 2-4, 2011

A Structure Based Mosaicking Approach for Aerial Images from Low Altitude of Non-Planar Scenes

Daniel Wischounig-Strucl Markus Quartisch
Bernhard Rinner
Institute of Networked and Embedded Systems
Klagenfurt University, Austria
<firstname.lastname>@uni-klu.ac.at

Abstract. *In this approach we estimate the depth structure of sceneries in aerial images captured by small-scale UAVs to improve the mosaicking of an orthographic overview image. Initial image transformations derived from inaccurate position and orientation data of UAVs are enhanced by the camera pose obtained using Structure from Motion. Corresponding points are then selected on a common ground plane to find accurate image transformations. The resulting mosaick preserves distances and minimizes distortions. A rough placement is immediately presented and optimized incrementally if more images are considered.*

1. Introduction

For many applications, such as disaster response, monitoring accident scenes and building sites, up to date and spatially accurate overview images are required. In particular, after severe disasters such as earthquakes or floodings wide area overviews are of special interest and importance to guide first-time responders.

We are investigating an approach to generate a wide area overview image from single images, preserving spatial distances as seen in orthophotos. To cover wide areas we favor aerial images from unmanned aerial vehicles (UAVs), because images from static cameras are hardly available due to the lack of infrastructure in typical scenarios.

For taking the essential aerial images small-scale UAVs, flying autonomously at low altitudes, are preferred to human operated planes or helicopters because of their advantages in availability, safety, robustness, ease of use and cost efficiency. Applying standard image registration algorithms to images from low altitudes, often lead to perspective distortions.

We achieve an overview image that can be perceived as orthophoto, if we only consider the planar ground and neglect objects on the ground. To keep the uniform scale in an orthophoto we have to optimize the image transformations accordingly. We do not aim to generate true orthophotos, which would require dense 3D models. Hence, images are taken with a nadir view, i.e., orthogonal to the earth's surface, to reduce the perspective influences of the non-planar scene and to allow a simplified orthorectification.

The ideal solution would be, of course, a full 3D reconstruction of the scene. But this is not feasible on small scale UAVs due to limitations of payload, battery capacity and computational performance. Furthermore, the resulting overview image should be presented iteratively as quick as possible. Thus, the images are processed already during flight of our networked small-scale UAVs and interim results are transmitted over the wireless channel with limited bandwidth.

In our approach, rough image transformations based on the metadata are refined by structure data from overlapping images. The Structure from Motion technique is used to compute the scene structure within overlapping regions to specifically match areas on the ground plane. For selecting corresponding points only on the ground plane it is necessary to apply a plane fitting algorithm to the structure data. With the resulting points an image transformation is computed that preserves distances while mosaicking.

Furthermore, the position and orientation data from the UAV's sensors is merged with the data extracted from images by Structure from Motion to es-

timate the real camera orientation and position. This allows a more accurate spatial referencing of points on the ground plane and refined orthorectification of single images.

The remainder of this paper is organized as follows: Section 2 gives a short overview on related work. Section 3 elaborates challenges and research questions of mosaicking aerial images incrementally and leads to Section 4, that proclaims our approach for mosaicking by means of the scene structure. Section 5 presents mosaicking results and finally Section 6 concludes the paper and gives some outlook on future work.

2. Related Work

In many cases single transformations applied to one image are sufficient to achieve appealing mosaicks. Recent works from Xing et al. [12] show satisfactory results when applying perspective transformations estimated by RANSAC (Random Sample Consensus) [1] and optimized SIFT (Scale Invariant Feature Tracker) features, taking images from airplanes.

Wang et al. combines orthorectified images with panorama images in [11] for 3D reconstruction of buildings, where the user has to select lines on the ground plane in the panorama images. The camera pose is computed from these lines on the ground plane, which represent footprints of buildings. The camera is kept at the same position and rotated to build a panorama image. After processing and manual optimization the ground images are projected on the proposed 3D model. For a larger area many panorama images are taken and processed one by one and are finally combined using bundle adjustment.

When considering hundreds of images with little overlaps, the initial image transformation is estimated by the metadata as proposed in [13]. The authors assume an exact nadir view of the camera onto a planar scene and neglect perspective distortions. Images annotated with metadata, i.e., altitude, global position and camera pose, are aligned by their global position. These transformations are refined afterwards by processing image correspondences.

In [14] the authors describe an effective method for combining data from images, taken from an airplane, with data from inertial sensors to achieve a seamless and geo-referenced mosaic. For the mosaicking the data from the inertial sensors and position sensors are combined with image features without 3D reconstruction or complex global registration. Aerial images from airplanes are made with telephoto lenses and from high distances to objects do not show perceptible perspective distortions.

Manually selected reference points on the ground are the base for a mosaicking approach presented in [9] that first extracts and matches feature points by Multi-Scale Oriented Patches (MOPs), clusters images, and finally uses RANSAC-initialized bundle adjustment to optimize all constraints over the entire image set. A simultaneous optimization balances the requirements of precise mosaicking and absolute placement accuracy on an overview image.

In our work we go one step further and introduce a basic structure and scene reconstruction with Structure from Motion to improve the metadata and image based mosaicking to deliver high resolution and frequently updated overview images.

3. Problem Definition

The goal is to mosaick a high resolution overview image from single aerial images and at the same time keep the uniform scale in the scenery. In order to generate this orthographic overview image, high resolution images are taken from multiple UAVs. Each image is annotated with metadata that contains position and orientation information, among others, from the UAV's sensors.

Creating a mosaick by simple placing images based on their metadata will lead to bad results, because this data is associated with uncertainty due to inaccuracy from the low cost and light weight design of small-scale UAVs. To cover wide areas from low altitudes, typically up to $150\,m$ above ground, with a minimum number of images it is obvious to use wide angle lenses. The tolerance of the image boundaries, projected on the ground, is in the range of $10\,\%$ of the image size, explored in detail in the work [13].

Hence, the challenge is to compute image transformations in the orthographic mosaick, while the non-planar scenery induces significant perspective distortions at individual images compared to aerial images taken from high altitudes. Moreover, a detailed 3D model of the scenery is not available.

We have to cope with several constraints, most prominent are the resource limitations. We cannot compute the whole overview image on the UAV nor transmit all high resolution images to the ground or other UAVs. For an online mosaicking a distributed processing is of interest, considering that high res-

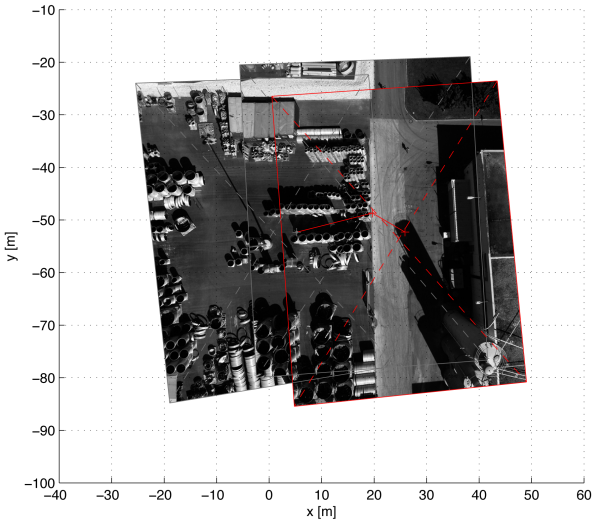


Figure 1. Initial image placement of images I_i by raw metadata where $i \in \{1, 2, 3\}$. Mosaicking errors can be explored on the ground plane. The trajectory of the UAV is shown in red.

olution images are not available immediately at the ground station.

To achieve a correct image placement that preserves distances within the overview image we need to estimate the camera position more accurately.

4. Structure Based Matching for Image Mosaicking

Our approach of mosaicking nadir aerial images annotated with metadata can be split into two main components:

I For the required online mosaicking the image transformations can be done with raw metadata without considering image contents.

II In parallel, these transformations can be refined as soon as more accurate camera extrinsics (position and orientation data), are estimated.

To improve the accuracy of the camera extrinsics from the metadata the Structure from Motion is used.

Hence, we model an optimization problem extending the two-step approach presented in [13] to find appropriate image transformations for each image in the set of aerial images. To avoid the accumulation of local perspective errors the metadata from cameras and the structure of the scene is taken into account.

4.1. Refined Estimation of Camera Extrinsics

In parallel to the rough placement, only by exploiting metadata, a refinement of the image transformation is executed as outlined in the following. Due to resource limitations the processing pipeline considers distributed execution; some processing steps can be executed directly on the UAV.

1. Determine a pair of images with sufficient overlap.
2. Match extracted feature points within the overlapping areas.
3. Use Structure from Motion to compute camera position and 3D structure for the matched feature points.
4. Merge the resulting camera extrinsics with the raw extrinsics and orthorectify both images.
5. Use plane fitting in the 3D structure to select feature points on the common ground plane and estimate the final image transformation.

Find a pair of images with sufficient overlap.

First the overlapping image areas O are determined by projecting the raw camera extrinsics from the metadata P_{IMU}, cf. Equation 17, onto the estimated ground plane. In Figure 1 the projection by the metadata and initial state for three images is presented before computing the refined transformations.

From all available pairs that overlap, a pair of images $\{I_i, I_j\}$ is selected to have the maximum overlapping area. Furthermore, for each image the features are extracted and the feature descriptor vectors δ_i and feature coordinates f_i are stored. For the following processing steps only the features, a few kilobyte in size, are necessary, instead of the whole image of up to 4 megabytes (compressed). This allows the reduction of the communication bandwidth significantly. In this approach we currently use the SIFT (Scale Invariant Feature Tracker) features [2], because it is has been proven to be very powerful [6].

$$\{\delta_i, f_i\} = \mathit{SIFT}_{\mathit{extract}}\left(I_i\right), \quad \{\delta_i, f_i\} \in \boldsymbol{F}_i \tag{1}$$

Match extracted feature points within the overlapping areas.

Only features within the overlapping area $O_{i,j} = I_i \cap I_j$ are considered for the matching. This reduced feature set $\boldsymbol{F}'_i \subseteq \boldsymbol{F}_i$ for image I_i and $\boldsymbol{F}'_j \subseteq \boldsymbol{F}_j$ for image I_j in the overlapping image area $O_{i,j}$ are matched simply by a nearest neighbor search. The

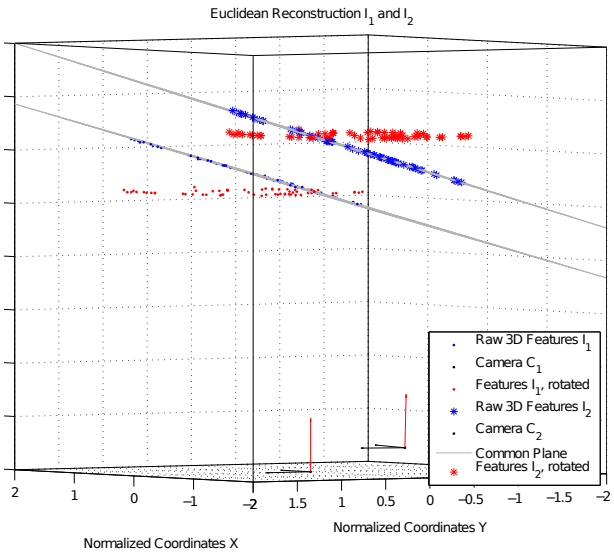


Figure 2. Matched features in the Euclidean scene reconstruction. Note, only the inliers on the same plane are plotted for a better visualization.

minimum Euclidean distance for the invariant feature descriptor vector δ'_i of feature $f'_i \in \boldsymbol{F}'_i$ is compared to a descriptor vector δ'_j of $f'_j \in \boldsymbol{F}'_j$ to find correspondences as suggested by Lowe [5].

$$\{\delta'_i, f'_i\} \in \boldsymbol{F}'_i, \quad \{\delta'_j, f'_j\} \in \boldsymbol{F}'_j \tag{2}$$

$$\boldsymbol{M} = \{\boldsymbol{F}'_i, \boldsymbol{F}'_j | f'_i, f'_j \in O_{i,j}\} \tag{3}$$

$$\{\hat{f}_i, \hat{f}_j\} = match\left(\boldsymbol{F}'_i, \boldsymbol{F}'_j\right) \tag{4}$$

Use Structure from Motion to compute camera position and 3D structure for the matched feature points.

From the matched features $\hat{f}_i, \hat{f}_j$ in the overlapping image area $O_{i,j}$ we compute the scene structure of these points by triangulation. Thus, the 3D structure, cf. Figure 2, i.e., elevation levels and the camera pose, is reconstructed by an estimation of the epipolar geometry [4]. The epipolar geometry, defined by the fundamental matrix F, essential matrix E, and the epipoles e_1 and e_2 is computed by Structure from Motion [3, 7]. Since we are using calibrated cameras, the camera calibration matrix K is known, the camera extrinsics P_{SfM_i}, cf. Equation 8, are determined by a singular value decomposition (SVD) from the essential matrix and epipoles [8].

$$E = [\hat{t}]_\times \hat{R} = U\,\Sigma\,V, \quad F = K^{-T}\,E\,K^{-1} \tag{5}$$

$$\hat{\boldsymbol{x}}_i^T E \hat{\boldsymbol{x}}_i = \boldsymbol{x}_i^T K^{-T} E K^{-1} \hat{f}_i = \boldsymbol{x}_i^T F \hat{f}_i \tag{6}$$

The essential matrix is estimated by using RANSAC within the matched features $\hat{f}_i$ and $\hat{f}_j$ to reduce outliers that do not match the approximated resulting essential matrix, cf. Equation 6.

In Figure 3 the structure inliers for each image are presented in the image plane. The point coordinates of selected feature points $\hat{f}$ in image I_i and I_j are mapped to 3D point coordinates $\boldsymbol{x} = [x, y, z]^T \in \mathbb{R}^3$. With the estimated camera extrinsics, cf. Equation 5, the Euclidian coordinates of the scene points $\boldsymbol{x}_i \in \boldsymbol{X}_i$ and $\boldsymbol{x}_j \in \boldsymbol{X}_j$ are reconstructed.

Merge the resulting camera extrinsics with the raw extriniscs.

The camera pose P_{SfM} from the image data is merged with the camera orientation and position P_{IMU} from the metadata. With the relative coordinates from Structure from Motion and the scaling from the metadata, the resulting camera extrinsics P_C are computed, cf. Equation 9. P_C describes the projective view of the camera that is used to transform images to their nadir view before the mosaicking. This process is known as orthorectification.

$$P_{\text{IMU}_i} = [R_{\text{IMU}_i}, T_{\text{GPS}_i}]_{4\times3} \tag{7}$$

$$P_{\text{SfM}_i} = [\hat{R}_i, \hat{t}_i]_{4\times3} \tag{8}$$

To project and maintain the spatial coordinates and distances on the ground plane the rotation component R_{C_i} of camera pose P_{C_i} is used. The optimized camera pose P_{C_i} replaces the first estimation from the raw metadata for image I_i.

$$P_{C_i} = [R_{C_i}, T_{C_i}]_{4\times3} \tag{9}$$

Fitting a ground plane into the 3D structure

A subset of points from the 3D points $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is adjudged as optimum for the final image transformation computation by the following constraint: All points on the same elevation level, respectively plane, preserve spatial relations with the image transformation $T_{\text{match},i}$, cf. Equation 15. Hence, it is important to find those points that avoid perspective distortions and inaccurate distances in the final mosaicking stage.

Inliers on the common plane $\boldsymbol{X}_\Pi$ are determined from the structure points in $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ by fitting a plane to all available points with RANSAC. The fitting function for RANSAC is the plane function for

plane Π in Equation 10, that is further optimized to be the most perpendicular plane to the camera's principal axis. Therefore, the angle between the plane normal vector $\vec{\boldsymbol{n}}$ and the principle axis vector $\vec{\boldsymbol{p}}$, derived from P_{C_i}, is minimized, assuming a horizontal ground plane.

$$\Pi = \vec{\boldsymbol{n}} \cdot \boldsymbol{q} \qquad \arccos(|\vec{\boldsymbol{n}}| \cdot |\vec{\boldsymbol{p}}|) \leq \varepsilon \quad (10)$$

$$\vec{\boldsymbol{n}} = (\boldsymbol{x}'_2 - \boldsymbol{x}'_1) \times (\boldsymbol{x}'_3 - \boldsymbol{x}'_1) \quad (11)$$

$$\boldsymbol{X}_\Pi = \{\boldsymbol{x}'_1, \boldsymbol{x}'_2, \boldsymbol{x}'_3\} \in \boldsymbol{X} \quad (12)$$

At least the three points defining the plane are sufficient to compute the matching transformation $T_{\text{match},i}$ in the order of a similarity transformation. For an improved matching function, e.g., by estimation and fitting again with an approximation approach, additional points $\boldsymbol{x}'_i$ can be selected by their closest distance d to the plane within a certain threshold γ.

$$d = |\vec{\boldsymbol{n}} \cdot \vec{\boldsymbol{v}}| \qquad \vec{\boldsymbol{v}} = \boldsymbol{x}' - \boldsymbol{q} \quad (13)$$

$$\boldsymbol{x}'_i \in \boldsymbol{X}_\Pi \quad | \quad d \leq \gamma \quad (14)$$

The matching transformation T_{match} applied to the whole image is computed by the normalised direct linear transformation algorithm given by Hartley and Zisserman [3].

$$\boldsymbol{x}' = T_{\text{match}}\, \boldsymbol{x} = [sR, t]_{3\times 3}\, \boldsymbol{x} \quad (15)$$

4.2. Incremental Mosaicking

After refining the image transformations and camera poses with the structure base matching the inaccurate mosaicking from raw data can be improved as expressed in the following.

Raw mosaicking with camera extrinsics

Single images I_i are merged with function $\biguplus$ to the overview image I, cf. Equation 16. Hence, the merging function $\biguplus$ is an arbitrary image fusion function. For demonstration we use a simple overlay function with alpha-blending. Initially images are placed by transformations derived from P_{IMU}, cf. Equation 17, based on their annotated GPS and IMU data. The images are orthorectified by the projective transformation $\tilde{R}_i$ and placed on the overview image by the transformation $T_{\text{pos},i}$ (cf. Figure 1).

$$I = \biguplus_{i=1}^{n(t)} T_i I_i \quad (16)$$

$$P_{\text{IMU}_i} = [R_{\text{IMU}_i}, T_{\text{GPS}_i}]_{4\times 3} \Rightarrow \{\tilde{R}_i, T_{\text{pos},i}\} \quad (17)$$

Refine the mosaicking with the output from the structure based matchting

Next, the refinement of the global mosaicking is achieved by the structure based matching, as described in Section 4.1. The optimized camera extrinsics matrix P_{C_i}, now improves the orthorectification of each image, opposed to $\tilde{R}_i$. Furthermore, the initial placement by $T_{\text{pos},i}$ is enhanced to the image alignment based on the scene structure.

Finally, the images are mosaicked with neighboring images by the transformation $T_{\text{match},i}$ that is approximated to optimize the output quality within the reduced search space in the overlapping image areas.

Hence, omitting perspective distortions that may propagate over images is one benefit of using projective transformations only for single images. When aligning individual images I_i to an overview image I by $T_{\text{match},i}$ only lower order transformations like the similarity transformation are allowed.

The resulting optimized image transformation T_i applied in the final mosaicking stage, cf. Equation 18, is composed from the raw metadata position and structure based transformation. The perspective projection R_{C_i} derived from the camera's intended pose P_{C_i} orthorectifies the image into nadir view, while the global alignment is applied with the refined global position T_{C_i}.

$$T_i = R_{C_i} \cdot T_{C_i} \cdot T_{\text{match},i} \quad (18)$$

5. Preliminary results

In the current state of evaluations the method of SIFT feature extraction is used for finding correspondences. However, the used feature extraction and matching methods are exchangeable, but SIFT shows sufficiently good results for our approach. The features are extracted from a copy of each image I_i, that is downscaled to 816×612 pixels.

In Figure 2 the result of the Structure from Motion point reconstruction in the overlapping area is presented. Note, only points on the common plane $\{\boldsymbol{x}'_i, \boldsymbol{x}'_j\} \in \boldsymbol{X}_\Pi$ and the two cameras P_{C_i}, P_{C_j} are plotted for better visualization. Figure 4 shows the finally transformed image I_i on the previous overview image. Image I_i and image I_j of the current test set $\boldsymbol{I}$ where $i = 1, j = 2$ are orthorectified by R_{C_i}, R_{C_j} derived from P_{C_i}, P_{C_j} beforehand. The selected features on the common plane are marked with red and blue crosses.

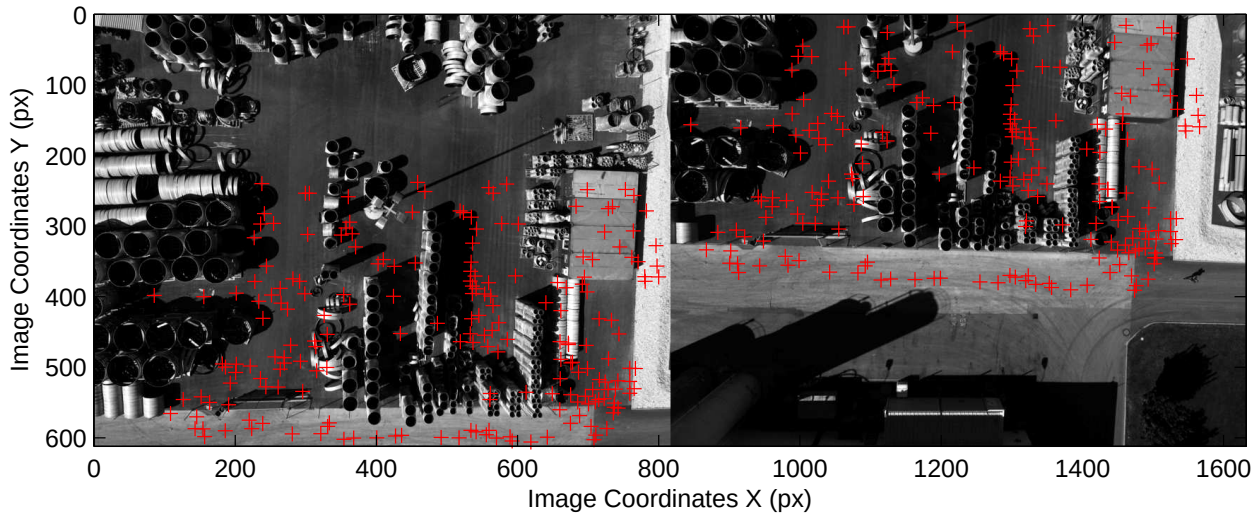


Figure 3. Image I_1 (left) and Image I_2 (right) with red markers on the remaining inliers from the Structure from Motion in the overlapping image region. These points show the input $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ for the plane fitting.

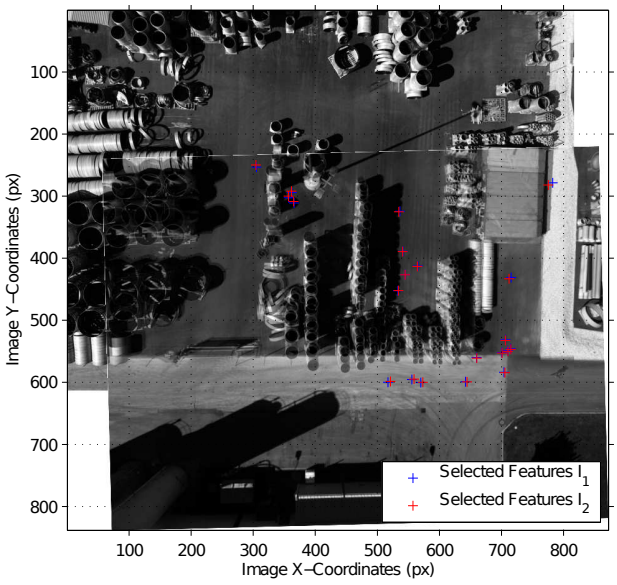


Figure 4. Matched features on the same plane in image I_1 and I_2

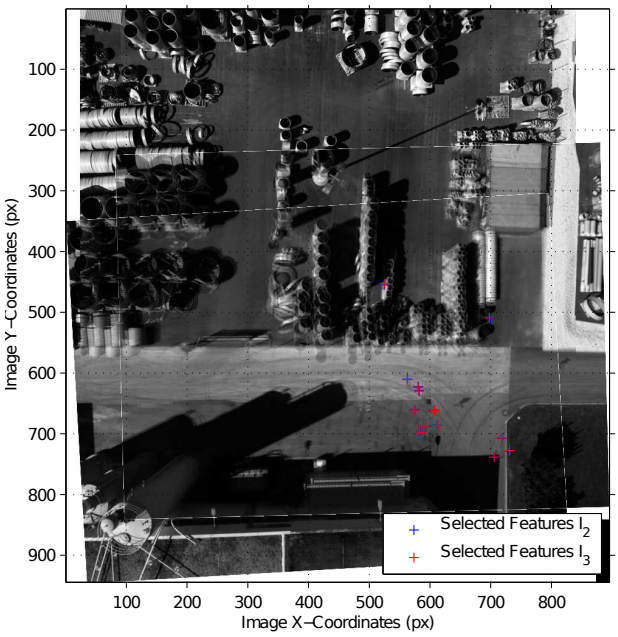


Figure 5. Matched features on the same plane in image I_2 and I_3 on top of image I_1. The correlation on the ground plane is excellent compared to the distortion effects of objects in the scene.

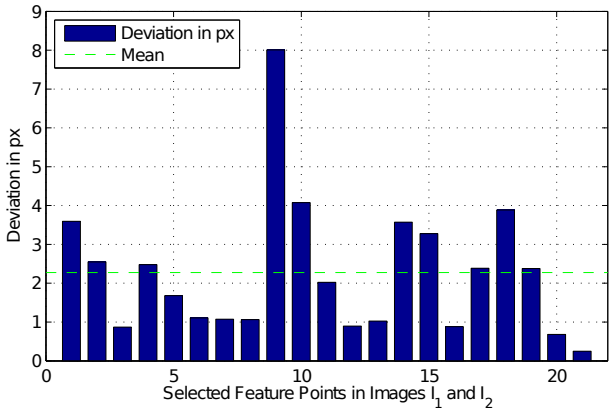
In the next iteration with the increased set of images the image I_3 has the maximum overlap with image I_2. The overview image presented in Figure 5 shows the previously mosaicked images I_1 and I_2 and the newly transformed image I_3 mosaicked on top. The red and blue markers show the common plane points from $\{\boldsymbol{x}'_2, \boldsymbol{x}'_3\} \in \boldsymbol{X}_\Pi$ again.

Moreover, in Table 1 the evolution of the features used for the final transformation optimization is presented where the significant reduction of the plane inliers to 21 in $I_1 \cap I_2$ and 13 in $I_2 \cap I_3$ can be explored.

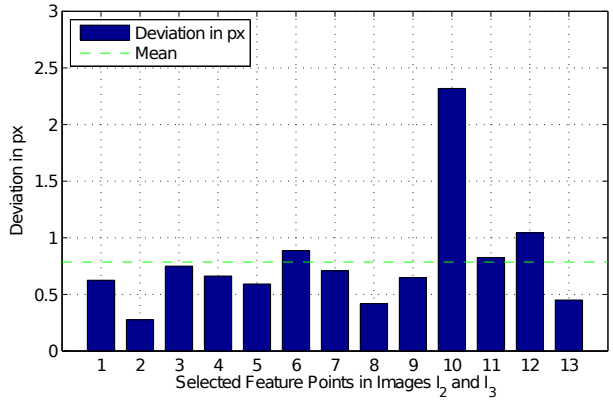
For each image and every pair of images the quality function Q is evaluated and its result is presented in Figure 6(a) for I_1, I_2 and Figure 6(b) for I_2, I_3. Figure 6 shows the pixel deviation of the inliers $\boldsymbol{x}'_i$ on the ground plane, which can be directly transformed to spatial deviations when projecting with the camera position and pose P_{C_i} and P_{C_j}. The correlation error for those ground plane inliers shows excellent results in a radius $r = 5$ pixels.

Processing Stage	I_1	I_2	I_3
Feature Extraction	1660	1492	1518
Reduced Search Range I_1, I_2	342	311	
Correlation Matching I_1, I_2	247	247	
SfM inliers I_1, I_2	201	201	
Plane Fitting I_1, I_2	21	21	
Reduced Search Range I_2, I_3		568	602
Correlation Matching I_2, I_3		548	548
SfM inliers I_2, I_3		483	483
Plane Fitting I_2, I_3		13	13

Table 1. The number of feature points can be significantly reduced from considering only overlapping regions to inliers on the same plane.



(a) Spatial distance error in pixels after transformation of image I_2 on image I_1



(b) Spatial distance error in pixels after transformation of image I_3 on image I_2

Figure 6. Distance deviations of points on the ground plane in the final mosaick.

Transformation Quality

The quality function Q weights the spatial accuracy function $G_i(I_i, I)$ and the pixel correlation function $C_i(I_i, I)$ by $\alpha, (0 \leq \alpha \leq 1)$ defined in Equation 19. The distance function of a projected feature point $\boldsymbol{x}_i$ on the ground plane of image I_i to the corresponding feature point on the overview image I is denoted by d and c measures the pixel correlation in a small neighborhood r of the feature point coordinate to the corresponding area on the overview image I.

$$Q = \sum_{i=1}^{n} (\alpha\, G_i(I_i, I) + (1 - \alpha) C_i(I_i, I)) \quad (19)$$

$$G_i(I_i, I) = \frac{1}{m} \sum_{k=1}^{m} d(\boldsymbol{x}_k \in I_i, I) \quad (20)$$

$$C_i(I_i, I) = \frac{1}{m} \sum_{k=1}^{m} c(\boldsymbol{x}_k \in I_i, I, r) \quad (21)$$
$$| r = \beta\, size(I_i)$$

6. Conclusion and Future Work

In this approach we have shown that distorted aerial images from low altitudes and taken with wide angle lenses can still be used to build an orthographic overview image that preserves a uniform scale on the ground plane. We compute the structure of the scene with Structure from Motion and optimize a rough mosaicking from annotated metadata of the images, i.e., GPS and IMU data of the UAV, to an accurate mosaick with matched correspondences on a common ground plane.

In this work, the results from Structure from Motion are only used to find a common plane and to enhance the estimation of the camera pose. This improves the spatial projection on the ground plane and delivers more accurate image transformations. We have experienced that the computational effort is significantly reduced when limiting the search range to structure inliers on the same plane and determining corresponding images from a large set by their proposed positions.

We will further analyze enhanced Structure from Motion estimation algorithms and optimization strategies for fitting common planes in adjacent images in the 3D domain. The reconstruction of the 3D structure of the scene can be further optimized by bundle adjustment [10]. We will investigate whether this method will get along with the available resources.

In future steps this additional knowledge about the scene could be used to generate a detailed depth model or mark objects in the scene.

Acknowledgment

This work was performed in the project *Collaborative Microdrones (cDrones)* of the research cluster Lakeside Labs and was partly funded by the European Regional Development Fund, the Carinthian Economic Promotion Fund (KWF), and the state of Austria under grant KWF-20214/17095/24772.

References

[1] 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*, 24:381–395, June 1981. 2

[2] P. Hansen, P. Corke, W. Boles, and K. Daniilidis. Scale invariant feature matching with wide angle images. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2007.*, pages 1689–1694, Nov. 2007. 3

[3] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 4, 5

[4] C. Loop and Z. Zhang. Computing rectifying homographies for stereo vision. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, page 131, 1999. 4

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 4

[6] O. G. Luo Juan. A Comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing IJIP*, 3(4):143–152, 2009. 3

[7] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3D vision, from images to models*. Springer Verlag, 2003. 4

[8] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, June 2004. 4

[9] P. Pesti, J. Elson, J. Howell, D. Steedly, and M. Uyttendaele. Low-cost orthographic imagery. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 24:1–24:8, New York, NY, USA, 2008. ACM. 2

[10] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372, London, UK, 2000. Springer-Verlag. 7

[11] L. Wang, S. You, and U. Neumann. Semiautomatic registration between ground-level panoramas and an orthorectified aerial image for building modeling. In *IEEE 11th International Conference on Computer Vision, ICCV 2007.*, pages 1–8, Oct. 2007. 2

[12] C. Xing and J. Huang. An improved mosaic method based on SIFT algorithm for UAV sequence images. In *Proceedings of International Conference on Computer Design and Applications (ICCDA)*, volume 1, pages V1–414–V1–417, June 2010. 2

[13] S. Yahyanejad, D. Wischounig-Strucl, M. Quaritsch, and B. Rinner. Incremental Mosaicking of Images from Autonomous, Small-Scale UAVs. *7th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Sept. 2010. 2, 3

[14] Z. Zhu, E. Riseman, A. Hanson, and H. Schultz. An efficient method for geo-referenced video mosaicing for environmental monitoring. *Machine Vision and Applications*, 16:203–216, 2005. 2