Robust Long-Term Aerial Video Mosaicking by Weighted Feature-Based Global Motion Estimation

Holger Meuel(✉), Stephan Ferenz, Florian Kluger, and Jörn Ostermann

Institut für Informationsverarbeitung,
Leibniz Universität Hannover, Hannover, Germany
{meuel,ferenz,kluger,office}@tnt.uni-hannover.de

Abstract. Aerial video images can be stitched together into a common panoramic image. For that, the global motion between images can be estimated by detecting Harris corner features which are linked to correspondences by a feature tracker. Assuming a planar ground, a homography can be estimated after an appropriate outlier removal. Since Harris features tend to occur clustered at highly structured 3D objects, these features are located in various different planes leading to an inaccurate global motion estimation (GME). Moreover, if only a small number of features is detected or features are located at moving objects, the accuracy of the GME is also negatively affected, leading to severe stitching errors in the panorama.

To overcome these issues, we propose: Firstly, the feature correspondences are weighted to approximate a uniform distribution over the image. Secondly, we enforce a fixed number of correspondences of highest possible quality. Thirdly, we propose a temporally variable tracking distance approach to remove outliers located at slowly moving objects.

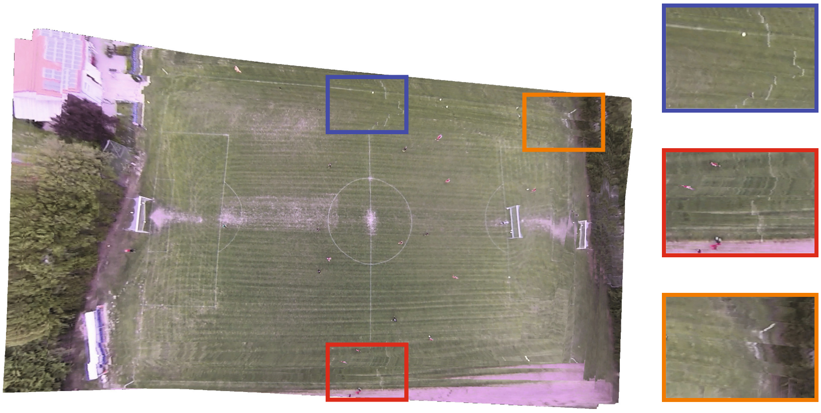
As a result we improve the GME accuracy by 10% for synthetic data and highly reduce the structural dissimilarity (DSSIM) caused by stitching errors from 0.12 to 0.035.

1 Introduction

For the visualization of aerial videos, *e.g.* captured from Unmanned Aerial Vehicles (UAVs) in a nadir view (orthorectified video), one common approach is to stitch the video images together to a panoramic image by mosaicking. For the generation of this panorama, each video image is registered into a common coordinate system. Since GPS/IMS systems can not provide a satisfactory accuracy, the global motion has to be estimated from the video images. One common approach is the detection of features, *e.g.* Harris Corner features [4] in one video image and its correspondence in the preceding image (feature correspondence) by a KLT feature tracker [16]. Assuming a planar ground and thus a uniform motion of detected feature points, RANSAC [2] can be used to remove feature correspondences not matching the global motion (outliers). From the remaining feature correspondences (inliers), a homography can be estimated. However, for a

M. Felsberg et al. (Eds.): CAIP 2017, Part I, LNCS 10424, pp. 135–147, 2017.
DOI: 10.1007/978-3-319-64689-3_11

small number of detected features – *e. g.* due to unstructured, blurry or low quality content – and small local displacements of moving objects between images (*e. g.* for pedestrians), RANSAC is not able to remove wrong correspondences anymore. Thus, a reliable estimation of a projective transform representing the global motion of the surface of the earth in the video is not possible. Moreover, features are often detected on non-planar structures, *e. g.* houses or trees whose motion does not match the motion of the ground plane of the scene. Furthermore, those features tend to be spatially clustered, which is known to negatively influence the quality of the global motion estimation [3]. Figure 1 shows an example of a wrong stitching based on the global motion estimation (GME) from [8] and using a standard mosaicking approach like [7,10].



(a) Entire panorama

(b)

Fig. 1. Panoramic image from 3000 images of the self-recorded *Soccer* sequence and magnifications in (b).

In this paper we propose different methods to increase the quality of the global motion estimation, which are mainly based on the usage of weighted features. To prevent an over-proportional weighting of feature clusters at highly structured areas in the image (like 3D objects), we propose to approximate a uniform distribution of the features in the entire image, considering the detected feature positions (Subsect. 3.1). In order to provide enough features for a reliable motion estimation, we propose to use a high, fixed number of features of highest possible quality (Subsect. 3.2). To further improve the quality of the resulting estimation, we rely on tracking over long temporal distances in order to remove features positioned at (slowly) moving objects which are not detected as outliers by a common RANSAC in case of small motion (Subsect. 3.3).

The remaining paper is organized as follows: Sect. 2 gives a short overview of global motion estimation for aerial videos. In Sect. 3 we describe our proposed robust long-term mosaicking approach. Our weighting algorithm for RANSAC which approximates a uniform distribution of the features in the image is introduced in Subsect. 3.1. Furthermore, we introduce a straight forward approach for detecting sufficient high quality features in the image in Subsect. 3.2. The tracking over long temporal distances is explained in Subsect. 3.3. In Sect. 4 we present experimental results for synthetic as well as real-world data, using the structural dissimilarity DSSIM [12] as quality metric. Finally, Sect. 5 concludes the paper.

2 Related Work: Global Motion Estimation for Aerial Videos

A lot of research has been done for the reliable estimation of the global motion in video sequences. Typical approaches are based on defining discriminative features like SIFT/SURF [1], Harris corners [4], MSER [6] etc. in one video image [9,15,20,22], the generation of trajectories for these features (*e. g.* by feature relocation [16], dense [14] or sparse optical flow [11]), and finally the estimation of the global motion according to an assumed scene model, *e. g.* using RANSAC [2].

In this work we extend the global motion estimation framework from [9] which is designed for the usage onboard of UAVs with limited energy and processing power. We also rely on KLT tracking of Harris corners, which are highly efficient to be computed compared to other features like SIFT or SURF. Whereas the common approach consisting of feature detection, RANSAC and least-square-minimization works well for a lot of applications, it fails for certain conditions as outlined above based on the example from Fig. 1. Thus, we aim at the improvement of the global motion estimation using RANSAC for videos captured from UAVs with low translational movement and slowly moving objects in the scene, *e. g.* in an aerial police surveillance scenario for soccer games.

3 Robust Long-Term Global Motion Estimation for Aerial Videos

Assuming the surface of the earth to be planar – which is valid for flight altitudes of several hundred meters – we can project one camera image I_k into the previous image I_{k-1} using a homography $\mathbf{H}_k^{k-1}$ which is described by a projective transform with 8 parameters $\vec{a}k = (a_{1,k}, a_{2,k}, \ldots, a_{8,k})^\top$:

$$\mathbf{H}_k^{k-1} = \begin{pmatrix} a_{1,k} & a_{2,k} & a_{3,k} \\ a_{4,k} & a_{5,k} & a_{6,k} \\ a_{7,k} & a_{8,k} & 1 \end{pmatrix}. \tag{1}$$

We can calculate the transformed pixel coordinates (x_{k-1}, y_{k-1}) in image $k-1$ from the image coordinates (x_k, y_k) in image k:

$$x_{k-1} = \frac{a_{1,k}x_k + a_{2,k}y_k + a_{3,k}}{a_{7,k}x_k + a_{8,k}y_k + 1}, \quad y_{k-1} = \frac{a_{4,k}x_k + a_{5,k}y_k + a_{6,k}}{a_{7,k}x_k + a_{8,k}y_k + 1}. \tag{2}$$

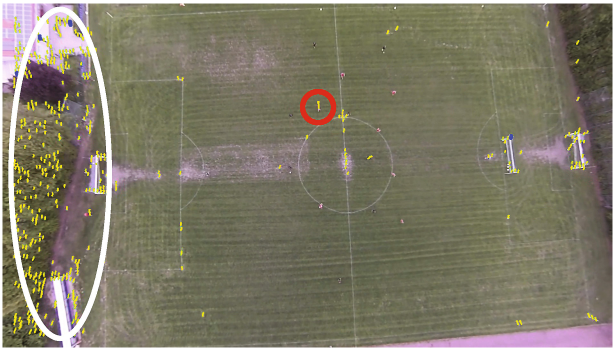


Fig. 2. Video image from the *Soccer* sequence with inliers and their trajectories (yellow lines) after KLT & RANSAC. The inliers are highly clustered at 3D structures (trees/houses) on the left (white ellipse). Moreover, a correspondence located at a player was errouneously considered as inlier (red circle). (Color figure online)

However, for a reliable homography estimation, the detected feature correspondences have to be located in one plane which becomes even more important for the projection of several video images into one common panoramic image. This plane optimally should be the ground plane, i. e. the feature correspondences have to be located on the surface of the earth. Whereas RANSAC is often capable of removing correspondences not matching the global motion, it may fail in removing correspondences not matching the global motion of the ground plane, if from the set of all correspondences C the amount of correspondences located on the ground $J \in C$ (inliers) is small compared to the amount of correspondences located on various different planes $O \in C$ (outliers). As a consequence, the estimated plane does not reflect the real ground plane which leads to an estimated global motion not reflecting the true motion of the surface of the earth. If $O \gg J$ (Fig. 2, white ellipse), the ground plane estimation becomes instable, resulting in stitching errors (Fig. 1).

3.1 Weighted Feature-Based Global Motion Estimation

Since only a few high quality features are typically located in unstructured areas (*e. g.* on the lawn in our example) compared to the number of features located at 3D structures (*e. g.* trees or houses), the former features have to be considered stronger within the least-square optimization in order to retain a homography representing the real global motion. Based on this idea, we formulate the least-squared minimization problem for the set of inliers J as:

$$\min \sum_{j \in J} \Big((\tilde{x}_{j,k-1} - x_{j,k-1})^2 + (\tilde{y}_{j,k-1} - y_{j,k-1})^2 \Big) \cdot (W_{j,k})^2, \qquad (3)$$

where $(\tilde{x}_{j,k-1}, \tilde{y}_{j,k-1})$ are the estimated coordinates and $W_{j,k}$ is a weighting function in dependence of $x_{j,k}$ and $y_{j,k}$. Based on Eqs. (3) and (2) we build a linear equation system which can be solved with a least-squares approach.

The weighting function $W_{j,k}$ is modeled with an instance reweighting approach, such that a uniform distribution $p_e(x,y)$ of the feature correspondences is approximated over the entire image.

The real feature distribution $p_{\text{feat}}(x,y)$ in the image for the (discrete) feature positions with the kernel function K is given as:

$$p_{\text{feat},k}(x,y) = \frac{1}{J}\sum_{i=1}^{J} K(x - x_{i,k}, y - y_{i,k}). \tag{4}$$

We approximate K by a Gaussian probability density function (pdf) $p_g(x,y)$ to model the neighborhood of each feature [18]:

$$p_g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right] \tag{5}$$

As suggested in [13], we define the variances σ_x and σ_y being the mean value of the pairwise distances of all feature correspondences and κ being a scaling factor:

$$\sigma_x = \sigma_y = \kappa \cdot \frac{2}{J^2}\sum_{j=1}^{J}\sum_{i=1}^{j-1}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{6}$$

The weighting function $W_{j,k}$ finally is calculated by dividing p_t by p_{feat} [17,19], i.e. the weighting for each feature is the reciprocal of the real feature distribution:

$$W_{j,k} = \frac{p_e}{p_{\text{feat},k}(x_{j,k}, y_{j,k})} = J \cdot \frac{2\pi\sigma_x\sigma_y}{\sum_{i=1}^{J}\exp\left[-\frac{1}{2}\left(\frac{(x_{j,k}-x_{i,k})^2}{\sigma_x^2} + \frac{(y_{j,k}-y_{i,k})^2}{\sigma_y^2}\right)\right]} \tag{7}$$

3.2 Increase of the Number of Features with Highest Possible Quality ("More Features")

The approximation of a uniform distribution of the feature correspondences over the entire image as described in the last subsection leads to highly improved global motion results. However, if only a small number of features can be detected *e. g.* due to bad input image quality or unstructured areas, an accurate solution for the global motion can not be determined.

Therefore, we propose to include a predefined minimum number of Harris features in the global motion estimation, always using the best available detected features. First, we calculate the Jacobian matrix and its lowest eigenvalue for each image pixel and sort them in a list. As a second step, we select the n-best features from the sorted list, with n being a predefined number of features. These n features are fed into subsequent motion estimation steps (RANSAC and homography estimation).

3.3 Variable Tracking Distance

Whereas we focused on the improvement of feature correspondences based on their spatial position in the image in Subsect. 3.1 and on the number of detected features in Subsect. 3.2, feature correspondences located at slow moving objects may not be recognized as wrong correspondences and thus not be removed as outliers by RANSAC (Fig. 2, red circle). As a consequence, these correspondences negatively influence the accuracy of the homography estimation. To overcome this issue, we propose to increase the temporal distance d between the images used for the homography estimation. Thereby, local motion tends to be larger and RANSAC is more likely able to remove features located on moving objects as outliers. Furthermore, to reduce drift as it may occur in image-to-image-based approaches, we aim at tracking against one specific image (reference image) as long as possible. Whereas in general it is beneficial to have a larger temporal tracking distance d, it may be disadvantageous, if the temporal distance between the images becomes too large. In such a case, KLT may not be able to reliably find correspondences due to shape changes or rotations which impairs the feature correspondence accuracy. Thus, we propose to use a constraint variable tracking distance d between the images. Summarizing, we aim at using one specific reference image for the estimation of homographies of several consecutive video images, whereas we limit the temporal distance to a predefined maximum value $d_{\max}$ and try to prefer large tracking distances. For each image k, we first calculate the distance d:

$$d = (k \mod \frac{d_{\text{ref}}^{\text{curr}}}{2}) + 1 + \frac{d_{\text{ref}}^{\text{curr}}}{2}, \tag{8}$$

with $d_{\text{ref}}^{\text{curr}}$ being an intermediate tracking distance (initialized to $d_{\max}$ for each image). The first term of Eq. (8) selects the same reference image as long as possible, whereas the last term enforces high tracking distances. Assuming a linear global motion, we approximate an estimated homography $\tilde{\mathbf{H}}_k^{k-d} = \mathbf{H}_{k-1}^{k-d} \cdot \mathbf{H}_{k-1}^{k-2}$ from already known homographies and transform all features from the current image using this $\tilde{\mathbf{H}}_k^{k-d}$. Then we check, if the following conditions are fulfilled:

1. Are enough transformed features located within the area of image I_{k-d}?
2. Is the intersection area of images I_k and I_{k-d} large enough?

If at least one of these conditions is violated, we halve $d_{\text{ref}}^{\text{curr}}$ and restart again with the computation of d. If all conditions are fulfilled, we use a guided tracking for the generation of accurate feature correspondences. For that, we apply the extrapolated homography $\tilde{\mathbf{H}}_k^{k-d}$ to all features in image I_k and use the result as seed position for the KLT search, resulting in accurate correspondences. The latter are used for the subsequent outlier removal and for the estimation of the improved, final homography $\mathbf{H}_k^{k-d}$.

4 Experiments

We present results for synthetic data in the Subsect. 4.1 before we evaluate our approach in detail for camera captured (real world) data in Subsect. 4.2.

4.1 Synthetic Data

In order to show that our method reliably improves the homography estimation, we generated a synthetic scene. We defined an array containing 30 × 17 blocks, each of size 64 × 64 pixels, which is approximately the size of one HDTV resolution image. For each block we randomly defined if it is supposed to be a block containing 3D structure ("house block") or not, and limited the amount of house blocks to 25%. In order to simulate a unequal feature distribution, we randomly draw a predefined mean number of feature positions $n_h = [0 \ldots 50]$ for the house blocks (green) and for the non-house blocks (blue) $n_n = 4$ (Fig. 3).

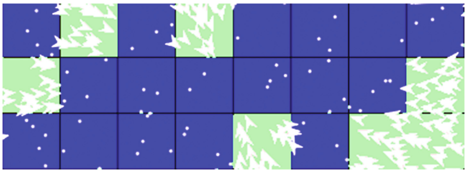


Fig. 3. Visualization of a synthetic image with "house blocks" (green), non-house blocks (blue) and randomly drawn features (white dots) and their simulated movement (white arrows). (Color figure online)

Furthermore, we manually generated homography parameters $\boldsymbol{a}_{\mathrm{syn}_k}$ similar to those which we observed in real multicopter videos (Table 1).

Table 1. Example synthetic homography parameters $\boldsymbol{a}_{\mathrm{syn}_k}$.

k	$a_{k,1}$	$a_{k,2}$	$a_{k,3}$	$a_{k,4}$	$a_{k,5}$	$a_{k,6}$	$a_{k,7}$	$a_{k,8}$	$a_{k,9}$
1	1	0	0.6	0.0001	1	−0.5	0	0	1
2	1	0	−0.8	0	1	−2.8	0	0	1
3	1	0	−0.1	0	0.9999	0.1	0	0	1
4	1	0	0.6	0	1	0.7	0	0.0001	1
⋮					⋮				
30	1	0	−0.6	0	1	0	0	0	1

The feature points from the current image I_k were transformed according to the synthetic homographies. We simulated motion parallax effects by moving all features on house blocks after the global motion compensation in the direction of the image center by m pixels. Since m should correspond to the motion parallax which can be observed in real scenes, we linearly increase m dependent on its distance to the image center up to a maximum of $m = 50$ pixels (which is a realistic motion parallax to be observed for high 3D structures and relatively low flight altitudes). Afterwards we applied zero-mean Gaussian noise with a variance of $\sigma^2 = 2\,\text{pel}$ to all feature positions.

Finally we used the synthetic scene as input for the motion estimation system, one time without and one time with our proposals, and compared the accuracy of the estimated homographies. For the improvement measure we applied each estimated homography to the corner pixels of the image and calculate the errors compared to the projected point position using the real homography parameters $\boldsymbol{a}_{\text{syn}}$. We varied the mean number of features n_h located in each house block between $10 \ldots 50$. The average error at the corner points was decreased from 10.1 to 9.0 pel which corresponds to 10.6% for $n_h = 10$ and from 18.1 to 16.4 pel for $n_h = 50$ (9.4%).

4.2 Camera Captured Videos

In this subsection we present results for real world data. Since the amount of test sequences providing a nadir view of the camera and containing 3D structured areas as well as plain areas is limited (although it may be the predominant view for aerial surveillance missions from UAVs), we recorded a test sequence of a soccer game (*Soccer* sequence) and present detailed results for this sequence. To underline the versatility of our proposals, we also provide results for the *1500* m sequence from the *TNT Aerial Video Testset* (TAVT) [5,9]. We will show that we can improve the homography estimation leading to subjectively highly improved results in panoramic images, especially in terms of line consistency.

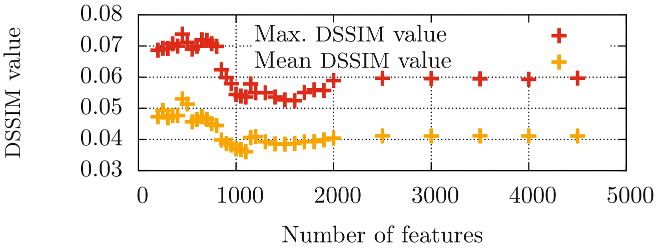


Fig. 4. Structural dissimilarity (DSSIM) [12] values (smaller is better) of reconstructed video images from panoramic image for different numbers of features for the *Soccer* sequence.

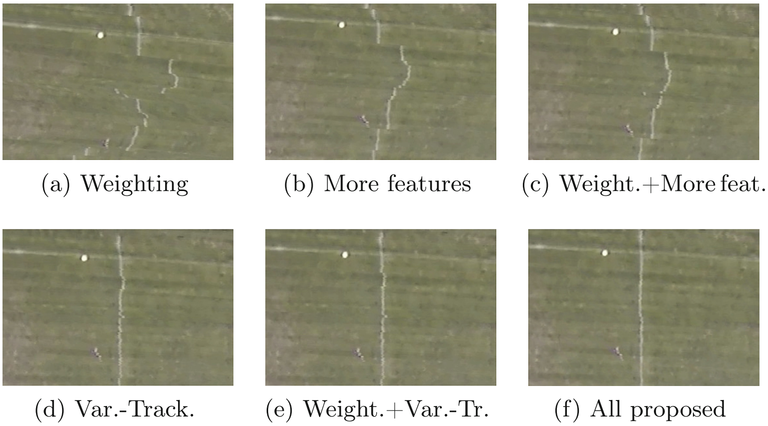
We generate a mosaic from the videos based on the estimated homographies. From this, we reconstruct video images again as described in [7,10]. For the quality measure we reconstruct video images from the mosaic and compare them image-wise with the input sequence. Due to the image reconstruction from the mosaic, no motion parallax is contained in the reconstructed video images. Thus, we cannot rely on a PSNR-based quality evaluation but use the structural dissimilarity (DSSIM) [12] instead. The structural dissimilarity is based on the well-known structural similarity (SSIM) [21] and lies between 0 (identical images) and ∞ (no similarity). It reflects the subjective impression in terms of cross-correlation between both images (structure), luminance similarity as well as contrast similarity.

Quality measures for the self-recorded *Soccer* sequence and the *1500 m* sequence from the data set TAVT [5,9] are presented in Table 2 and in Fig. 5 for each proposed method alone and all combinations.

Table 2. Results of different methods for the *Soccer* sequence, 3000 images ((∗): manual reference only for 100 images) and the *1500* m sequence from TAVT [5,9].

Sequence	*Soccer* seq. DSSIM		*1500* m seq. DSSIM	
Method	Mean	Max	Mean	Max
Manual reference	$0.036^{(*)}$	$0.060^{(*)}$	—	—
Baseline (w/o proposed methods)	**0.120**	**0.146**	**0.067**	**0.156**
Weighting of correspondences	0.123	0.151	0.066	0.155
More features	0.094	0.129	0.065	0.133
Weighting & more features	0.094	0.128	0.064	0.133
Variable tracking	0.054	0.079	0.062	0.094
Weighting & variable tracking	0.045	0.071	0.063	0.099
Weighting & more feat. & var. track.	**0.035**	**0.051**	**0.061**	**0.088**

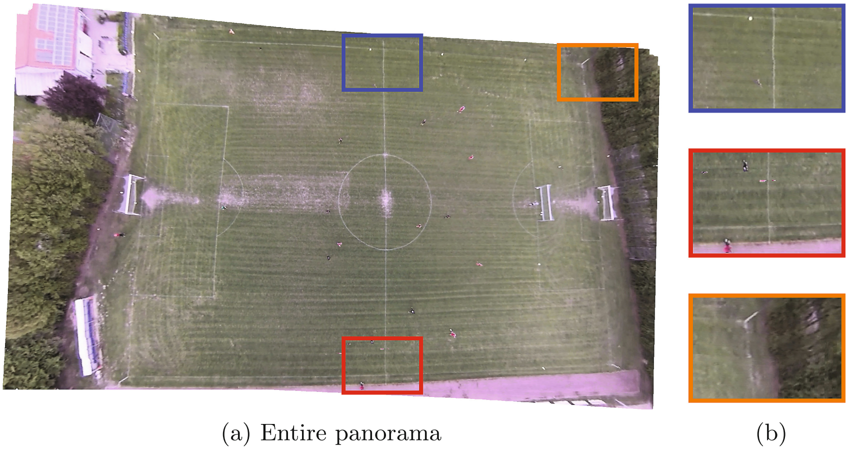
From the detailed results it is obvious, that our proposed weighting algorithm can improve the quality of the global motion estimation, if *enough* features are in the image (*Weighting & more features* in the tables: 0.120 to 0.094 for the *Soccer* sequence, Fig. 5c, 0.067 to 0.064 for the *1500* m sequence). Simulations for the HDTV resolution *Soccer* sequence lead to an optimal value of about $n = 1050$ features (Fig. 4), which is in the range of $n = [900 \ldots 1200]$ we found as optimal number of features also for other sequences we tested. If the number of features is too small, we only can observe small average gains (0.067 to 0.066 for the *1500* m sequence) or even small (average) losses (0.120 to 0.123 for the *Soccer* sequence, Fig. 5a) if – like in the latter case – not enough features of high quality are contained due to a low image quality. Thus, the combination of weighting and more features is always beneficial for low as well as for high quality videos. The usage of a variable tracking distance is recommendable in any case, since it improves the line accuracy by enforcing tracking against one reference image for several video images. Thus, drift is highly reduced and the objective and subjective results are improved on average (0.120 to 0.054 for the *Soccer* seq., 0.067 to 0.062 for the *1500* m sequence) as well as for the maximum DSSIM



(a) Weighting (b) More features (c) Weight.+More feat.

(d) Var.-Track. (e) Weight.+Var.-Tr. (f) All proposed

Fig. 5. Subjective comparison of different proposed methods and combinations for the self-recorded *Soccer* sequence.

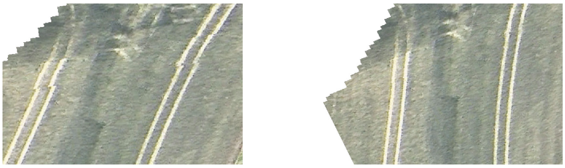


(a) Entire panorama (b)

Fig. 6. Final panorama using all proposed improvements for global motion estimation with uniform distribution and weight of $\kappa = 0.575$. (b) magnifications.

values (0.146 to 0.079 for the *Soccer* sequence, Fig. 5d, 0.156 to 0.094 for the *1500* m sequence). This holds also true for the combined approaches with the variable tracking (Figs. 5e and f).

Combining our approaches, we observe that we highly improve the DSSIM from 0.12 to 0.035 for the *Soccer* sequence. Our combined methods even slightly outperform a manually generated reference, which matches the subjective impression. For the *1500* m sequence we achieve an improvement from 0.067 to 0.061 in terms of mean DSSIM. Although the average gain for the latter sequence is smaller than for the *Soccer* sequence, the maximal structural dissimilarity was drastically reduced (*Soccer* seq.: 0.146 to 0.051; *1500* m seq.: 0.156 to



(a) Baseline (b) Combined proposals

Fig. 7. Subjective results for the *1500* m sequence from the TAVT data set [5,9].

0.088) which results in smaller maximal distortions leading to subjectively much more pleasing results, especially in terms of line accuracy (Figs. 5f and 7b). In Fig. 6 we present the final long-term panoramic image after the fully automatic processing of 3000 images. A subjective impression for the *1500* m sequence is shown in the magnifications from the panoramic image in Fig. 7.

5 Conclusions

In this paper, we aim at a robust global motion estimation for UAV captured ortho-videos which contain distinct 3D structures (*e. g.* houses, trees) as well as real ground.

We propose to tackle the problem of a unequal feature correspondence distribution over the image by introducing a weighting function which approximates a uniform distribution over the image. In order to provide enough features also in scenarios with only a small number of high-quality features, we additionally propose to use a high but fixed number of features based on the feature quality. Finally, our third contribution is to track over long temporal distances with a variable tracking distance. The benefits of this approach are twofold: firstly, we use the same reference image for several images which reduces drift. Secondly, the motion of small and slow moving objects can more likely be removed by an outlier removal (RANSAC).

We show, using synthetic data, that our feature correspondence weighting proposal improve the estimation accuracy by up to 10% for realistic assumptions. For camera captured data, the resulting panoramic images which were generated based on the estimated global motions were improved and provide much better and virtually drift free reconstruction of linear structures (*e. g.* lines at a Soccer play ground). The structural dissimilarity (DSSIM) for reconstructed images from the panoramic image was highly reduced, *e. g.* from 0.120 to 0.035 on average for the self-recorded *Soccer* sequence.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comp. Vis. Image Underst. **110**(3), 346–359 (2008). http://dx.doi.org/10.1016/j.cviu.2007.09.014

2. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
3. Han, Y., Choi, J., Byun, Y., Kim, Y.: Parameter optimization for the extraction of matching points between high-resolution multisensor images in urban areas. IEEE Trans. Geosci. Remote Sens. **52**(9), 5612–5621 (2014)
4. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceeding of the Fourth Alvey Vision Conference, pp. 147–151 (1988)
5. Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover: TNT Aerial Video Testset (TAVT) (2010–2014). https://www.tnt.uni-hannover.de/project/TNT_Aerial_Video_Testset/
6. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of the British Machine Vision Conference, pp. 36.1–36.10. BMVA Press (2002)
7. Meuel, H., Kluger, F., Ostermann, J.: Illumination change robust, codec independent low bit rate coding of stereo from singleview aerial video. In: 10th IEEE International 3DTV Conference, pp. 1–4, July 2014. http://ieeexplore.ieee.org/document/7548961/
8. Meuel, H., Munderloh, M., Ostermann, J.: Low bit rate ROI based video coding for HDTV aerial surveillance video sequences. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW), pp. 13–20, June 2011
9. Meuel, H., Munderloh, M., Reso, M., Ostermann, J.: Mesh-based piecewise planar motion compensation and optical flow clustering for ROI coding. In: APSIPA Transactions on Signal and Information Processing, vol. 4 (2015). http://journals.cambridge.org/article_S2048770315000128
10. Meuel, H., Schmidt, J., Munderloh, M., Ostermann, J.: Region of interest coding for aerial video sequences using landscape models. In: Advanced Video Coding for Next-Generation Multimedia Services. Intech, January 2013. http://tinyurl.com/ntx7u29
11. Munderloh, M., Meuel, H., Ostermann, J.: Mesh-based global motion compensation for robust mosaicking and detection of moving objects in aerial surveillance. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), pp. 1–6 (2011)
12. Pornel: RGBA Structural Similarity (2016). https://kornel.ski/dssim/
13. Reddi, S.J., Ramdas, A., Póczos, B., Singh, A., Wasserman, L.: On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In: Proceeding of the AAAI Conference on Artificial Intelligence, pp. 3571–3577 (2015)
14. Reso, M., Jachalsky, J., Rosenhahn, B., Ostermann, J.: Temporally consistent superpixels. In: Proceeding of the IEEE International Conference on Computer Vision (ICCV), pp. 385–392, December 2013
15. Shi, G., Xu, X., Dai, Y.: SIFT feature point matching based on improved RANSAC algorithm. In: International Conference on Intelligent Human-Machine Systems and Cybernetics, vol. 1, pp. 474–477, August 2013
16. Shi, J., Tomasi, C.: Good features to track. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, June 1994
17. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Planning Infer. **90**(2), 227–244 (2000). http://www.sciencedirect.com/science/article/pii/S0378375800001154

18. Silverman, B.W.: Density Estimation for Statistics and Data Analysis, vol. 26. CRC Press (1986)
19. Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., Kawanabe, M.: Direct importance estimation with model selection and its application to covariate shift adaptation. In: Proceeding of the Conference on Neural Information Processing Systems (NIPS), pp. 1433–1440 (2007). http://tinyurl.com/jt5rdz8
20. Wang, Y., Fevig, R., Schultz, R.R.: Super-resolution mosaicking of UAV surveillance video. In: IEEE International Conference on Image Processing, pp. 345–348, October 2008
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
22. Xu, Y., Li, X., Tian, Y.: Automatic panorama mosaicing with high distorted fisheye images. In: International Conference on National Computation, vol. 6, pp. 3286–3290 (2010)