Online Annotation of Airborne Surveillance and Reconnaissance Videos

Peter Solbrig, Dimitri Bulatov, Jochen Meidow, Peter Wernerus, Ulrich Thönnessen
FOM - Research Institute for Optronics and Pattern Recognition
FGAN - Research Establishment for Applied Science
Ettlingen, Germany
{solbrig, bulatov, meidow, wernerus, thoe}@fom.fgan.de

A useful enrichment of videos captured by unmanned aerial vehicles (UAVs) is the annotation of image data or determining coordinates for objects in the video. This requires georeferencing of the video* frames*. For surveillance or reconnaissance applications, we propose georeferencing of UAV video frames with an orthophoto. In this process, the challenge is to register temporally different images from different sensors that capture objects with quite different visual appearances. After an initial registration of the first video frame to the orthophoto, the subsequent frames are registered by image mosaicking. Whenever necessary, the mosaic is anchored by registration to the orthophoto. We explain the image-based registration approach and discuss its potential real-time processing capability. Internal performance evaluation and GPS/INS support is considered. The feasibility of the approach is demonstrated with real data sets captured by UAVs for georeferencing, video frame annotation, and motion detection.

Keywords: georeferencing, image annotation, motion analysis, motion detection, video registration

1 Introduction

1.1 Motivation

Sensors carried by unmanned aerial vehicles (UAVs) are valuable for surveillance and reconnaissance tasks. They provide up-to-date information for motion or change detection and can deliver a fast overview of the situation by image mosaicking. For time-critical applications, online annotation of the captured video frames with, e.g., street names or outline information, is beneficial for enabling quick response. This implies the need for georeferencing to transfer, for instance, scene coordinates.

Frame-to-frame registration for a monocular video stream suffers from the accumulation of transformation errors and the increasing uncertainty of the overall transformation. This inevitable drift is due mainly to observational errors and invalid model assumptions concerning the assumed image transformation. To cope with continuous video streams, one has to compensate for this effect. This can be done by registering small video subsequences to a large-scale, global reference image. If this image is an orthophoto, the desired georeference is established by these anchorages. In this process, the challenge is to register temporally different images captured by different sensors from different perspectives, which potentially leads to considerable variance in the appearance of the observed scene.

Appearance-based matching can be found in the literature ([7], [8], [9], [18]). The consideration of digital elevation maps proposed in [3] paves the way toward a 3D approach. This is balanced by the need for telemetry data for proper initialization of the expensive optimization process.

1.2 Contribution

We assume a monocular video stream from a straight-line-preserving camera. Our approach is similar to the one proposed in [9] but differs in the following several conceptual and technical aspects:

- We strive for *real-time capability* of the system. That is, we want to react to sensor data in a required time. In our context, this means that the results must be available within a reasonable time in order to fulfill the specific task. To save computational resources, we perform anchorages of current subsequences only when required.
- To decide about the need for a new anchorage during the processing, we perform an *internal performance evaluation* which provides self-diagnostics of the system. The uncertainty of the transformation parameters can be considered by checking the acceptability of the precision for the accumulated overall transformation. The validation of the mapping model, i.e., homography, can be carried out by checking alternative multi-view constraints for corresponding image structures.
- For exploitation of available GPS/INS data, we show how this *navigation information* can be incorporated seamlessly into the registration process to obtain an easy initialization and, subsequently, better results.

This paper is organized as follows. In Section 2, we introduce preliminaries and explain the approaches for frame-to-frame mapping (intrasensorial registration) and image georeferencing (intersensorial registration). We show how to perform motion detection and how GPS/INS support can advantageously be introduced. Section 3 covers topics in implementation, especially those related to the aspired real-time capability. In Section 4, we demonstrate the feasibility and performance of the system for real data sets by treating georeferencing, annotation, and motion analysis. Finally, in Section 5, we end with conclusions and an outlook.

2 Approach

Registering a video frame to an orthophoto is a challenging task, due to, e.g., illumination differences or changing vegetation. Methods capable of coping with this are computationally expensive. Fortunately, image mosaicking methods are less expensive and can be used to extrapolate an initial registration. Due to accumulating errors during mosaicking, an alternating process of image mosaicking and registration of video frames to the orthophoto is necessary and is described in this section.

In Subsection 2.1, preliminaries and the methods used to solve the registration problem are described. Subsection 2.2 contains the description of the registration process. First, the image-based registration is described. Second, a method to support this registration by GPS/INS is introduced. After that, an approach to perform self-diagnostics on the registration is proposed. The procedure used to perform object motion analysis on the video frame is described in Subsection 2.3. Finally, the annotation of image data is explained in Subsection 2.4.

2.1 Preliminaries

We assume that the depth of the scene is negligibly small compared to the sensor's altitude and that the camera used is straight-line-preserving. In this case, homography is a suitable model to describe the frame-to-frame mapping [5]. We denote the homography between the frame I_t captured at time t of the sequence and the orthophoto by H_t and the transformation between the frames I_t and I_j by $H_{j,t}$. Interest points in the frame I_t will be denoted by $\boldsymbol{x}_t$ and in the orthophoto by $\boldsymbol{X}$.

In order to estimate the homography from point correspondences contaminated with outliers, robust methods, such as *Random Sample and Consensus* (RANSAC) [4] are used. We modify the standard RANSAC-procedure with $\mathrm{T}_{1,1}$*-test* (as described in [12]) in order to speed up the processing.

Finally, two different methods are used in order to determine point correspondences. For two consecutive frames, the assumptions of rather short baseline and little change in illumination are reasonable in the majority of cases. Therefore, standard *KLT-tracking* [11] is robust enough for tracking $\boldsymbol{x}_t$ into the next frames. In order to relate interest points between orthophoto and video, scale-, rotation-, and/or illumination-invariant feature descriptors are computed and compared. Examples of this kind of detectors can be found in [1], [10]. In our experiments, we achieved best results using interest points detected by the SIFT-approach (*Scale Invariant Feature Transform*) [10]. The description of how to augment both speed and quality of these rather time-consuming steps will be given in Section 3.

2.2 Georeferencing of a Video

2.2.1 Image-Based Georeferencing

To provide a successful initial registration, we must obtain a sufficient number of correspondences. Searching for the shortest Euclidean distances of SIFT-features in the orthophoto and the video frame does not always lead to a correct homography, especially in high-textured areas. In these areas, the correct correspondences will not necessarily be given by the shortest Euclidean distance between the hypothesized correspondences, but possibly by the second- or third-shortest. Therefore, we reverse the procedure and use the initialization of the homography H_1 in order to obtain enough correspondences. The initial homography is estimated using interest points localized by the SIFT-algorithm, by using GPS/INS as described in Subsection 2.2.2, or by manually selected correspondences. The probability of obtaining the correct correspondence $\boldsymbol{x}_1 \to H_1\boldsymbol{X}$ inside a small circular area, 50 to 100 pixels in diameter, around each $\boldsymbol{x}_1$ increases considerably while the searching time decreases. This process is called guided matching [5]. After guided matching, the $\boldsymbol{x}_1$ are tracked from frame to frame by KLT-tracking and the homography between frame I_t and the orthophoto is computed incrementally:

$$H_t = H_{t,t-1}H_{t-1}. \qquad (1)$$

This is a common approach to perform image mosaicking. We call it intrasensorial registration. Unfortunately, this process usually leads to severe aberrations because of the following reasons:

- *Failing of the local homography:* Tracking may fail because of the video quality or because the points detected in the first frame are lost after several frames. In this case, we talk about failing of the local homography $H_{t,t-1}$ -estimation.
- *Failing of the global homography:* Due to accumulation of errors in the calculation of H_t by (1), the result becomes unreliable for some t. As an example, consider a mosaic from rotating cameras: In this case, the application of a

homography is justified. Still the stripe gets broader from frame to frame and, as the camera meets the initial position again, severe registration errors are observed. Here we say that the estimation of the global homography H_t failed.

- *Failing of the initialization:* The identity obtained in (1) has a disadvantage: if H_1 was not obtained correctly due to unknown reasons, then H_t will also be wrong even if all homographies $H_{t,t-1}$ were obtained correctly. Here we say that the initialization failed.

To cope with the problems mentioned above, the system must become aware of them. In the first case, either RANSAC will yield a small number of inliers or the inliers, even though sufficient in number, will not be favorably distributed within the frame. For a point, we compute the reprojection error and compare it with a fixed threshold s, which was chosen to be 1 to 2 pixels in most experiments. If the number of inliers (points with registration error below s) is low or if there is no inlier in any quarter of the frame, we reject the homography $H_{t,t-1}$. By the second condition, a good distribution of the points in the image is achieved. Then we check the global homography H_t by computing registration errors in the orthophoto and frame I_t (second case). Here we use similar criteria for rejecting H_t. In order to cope with the problem of bad initialization, we simply reject H_t in periodic lags (every 15 to 20 frames in our experiments). If either $H_{t,t-1}$ or H_t was rejected, we extract new interest points in the current frame and use the last reliable value H_{t-1} as an initialization for H_t. Then, as described above, we use descriptor-based matching, RANSAC with $T_{1,1}$-test, and guided matching. This is called intersensorial registration. The entire process is illustrated in Figure 1.

2.2.2 GPS/INS-Supported Georeferencing

Algebraic parameterization of the homography is attractive, since the corresponding mapping $\boldsymbol{x}_t = H_{t,t-1}\boldsymbol{x}_{t-1}$ (see Subsection 2.2.1) is linear in the transformation parameters, leading to direct estimations via singular value decomposition. However, for the incorporation of GPS/INS data, a 3D geometric parameterization of the mapping is required.

A starting point is the *Faugeras decomposition* [3] of the homography matrix for the image-to-image mapping

$$\boldsymbol{x}_t = K\left(R_{t,t-1} + \frac{1}{d_{t-1}}\boldsymbol{t}_t\boldsymbol{n}_{t-1}^T\right)K^{-1}\boldsymbol{x}_{t-1} \qquad (2)$$

of a point $\boldsymbol{x}$, where K is the homogeneous camera calibration matrix, $R_{t,t-1}$ the relative rotation between two consecutive frames, $\boldsymbol{t}_t$ the translation, and ($\boldsymbol{n}_{t-1}, d_{t-1}$) the world plane π, represented by its normal vector and its distance to the origin of the coordinate system.

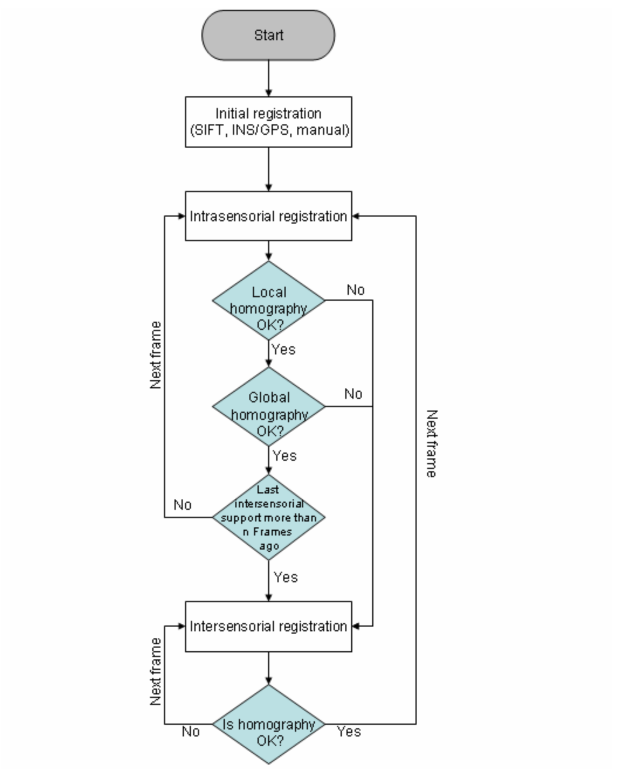
Start
Initial registration (SIFT, INS/GPS, manual)
Intrasensorial registration
Local homography OK?
No
Yes
Global homography OK?
No
Yes
Next frame
Last intersensorial support more than n Frames ago
No
Yes
Next frame
Intersensorial registration
Next frame
Is homography OK?
No
Yes

Figure 1 : Flow-chart of the mosaicking and video-to-orthophoto registration process

The GPS/INS observations for the projection centers $\boldsymbol{C}_{t-1}$ and $\boldsymbol{C}_t$ and the sensor orientations R_{t-1} and R_t are given in a world or object coordinate system. They can be introduced by the additional constraints

$$\frac{\boldsymbol{t}_t}{d_{t-1}} = R_t\left(\boldsymbol{C}_t - \boldsymbol{C}_{t-1}\right) \quad \text{and} \quad R_{t,t-1} = R_t R_{t-1}^T \qquad (3)$$

for the parameters and the observations. Note that observations of different types are involved. Therefore, one has to consider the covariances of the observations for proper relative weighting of the image and GPS/INS observations in the parameter estimation process. In situations with little image information available, the GPS/INS information weights more for the determination of the parameters sought.

The geometric 3D parameterization (2) allows the determination of the sensor trajectory and the sensor orientation on the basis of image information [15]. Furthermore, it enables the additional use of GPS/INS data in the analysis. By observing that the normal vector $\boldsymbol{n}$ is a global parameter, one can introduce further stability by using the prior information $\boldsymbol{n}_t = R_{t,t-1}\boldsymbol{n}_{t-1}$ in a filtering process.

For the georeferencing task, the image-to-world homography is immediately given by consideration of the projection matrix $P_t = K\left[R_t \middle| -\boldsymbol{C}_t\right]$, which projects an object point $\boldsymbol{X}$ into the image plane via $\boldsymbol{x}_t = P_t\boldsymbol{X}$ and includes the GPS/INS observations explicitly. If we assume a horizontal plane and set the *z*-coordinates of the camera centers to be the altitude above ground, then object points have zero height. Therefore, the third column of the projection matrix can be dropped, yielding the homography matrix.

2.2.3 Self-Diagnostics

For autonomous systems, performing self-diagnostics is essential. For a successful system, the current state of the process, the reliability and the precision of the results have to be monitored. System failures are possible due to random errors, e.g., observational errors, or gross errors (outliers) – possibly because of unfounded model assumptions. One has to recognize these failures and treat them accordingly, for instance, by establishing a new anchorage of the video subsequence to the orthophoto.

Model violations can be detected by considering alternative multi-view relations for consecutive video frames or between the current video frame and the orthophoto. A common approach is the evaluation of the *Geometric Robust Information Criterion* (GRIC) introduced in [17]. By doing so, one is able to decide whether a 2D or a 3D relation is more appropriate and to react accordingly.

While outliers can be eliminated by RANSAC or its derivatives, random errors remain and affect the precision of the estimates. The desired and required precision of the transformation parameters $\boldsymbol{p}$, where $\boldsymbol{p}$ is a vector containing the entries of *H*, can be specified by a criterion matrix *G*. We require that any result of the function $f(\boldsymbol{p})$ be more precise when determined with the empirical covariance matrix *C* than with the criterion matrix *G* [17]. This leads to the requirement $\boldsymbol{e}^T C\boldsymbol{e} \geq \boldsymbol{e}^T G\boldsymbol{e}$ with the Jacobian $\boldsymbol{e}^T = \partial f / \partial \boldsymbol{p}$. Thus, the maximal eigenvalue λ of the generalized eigenvalue problem $C\boldsymbol{e} = \lambda G\boldsymbol{e}$ needs to be less than one.

2.3 Object Motion Analysis

In surveillance and reconnaissance applications, moving objects tend to be of special interest: motion detection during disaster rescue actions can mean there are survivors and motion during military operations may signify possible threats. In both cases, motion indicates situations that are worth closer examination.

The procedures for motion detection can be classified into object-based and signal-based methods. In the first group, a procedure will recognize the shape of an object, e.g., based on its signature, follow its outline in the images and determine its track. This is the most suitable method when dealing with applications in spectral ranges in which objects can be discriminated from the background, e.g., persons in front of building walls. Video recordings with daylight cameras rarely yield such results. Signal-based methods determine differences in intensity values. Some methods search for shifted connected regions from frame to frame. Special evaluation processes of optical flow are used here; they yield estimated shift vectors as output. Another possible signal-based method is to search for noticeable regions in difference images from consecutive video frames.

In this paper, we deal solely with methods of the latter group, since object-based solutions do not promise much success in the visible spectral range. However, signal-based solutions with estimated shift vectors turn out to be much too computationally intensive for real-time applications (e.g., global estimation of the optical flow, [6] or [11]). Furthermore, only moving sensor platforms are considered here. Therefore, the motion of the sensor itself requires continuous registration of consecutive frames. This task need not be solved separately, since homographies from the registration process described in Subsection 2.2 can be used here.

Two procedures will be examined more closely, MoDe (*Motion Detection*) [2] and *Time-Recursive High-Pass Filtering* [19]. In MoDe, frames I_{t-n} and I_{t+n} are registered onto the frame I_t and difference images D_{t-n} and D_{t+n} are computed. A difference image is evaluated at every single pixel according to:

$$D_t = \begin{cases} \min\left(D_{t-n}, D_{t+n}\right), & \text{if } \operatorname{sign}\left(D_{t-n}\right) = \operatorname{sign}\left(D_{t+n}\right) \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

The choice of the parametervalue *n* determines the quality of the result. The optimal choice depends on the velocity of an object passing through the observed scene. During Time-Recursive High-Pass Filtering, *n* frames $I_{t-1}, I_{t-2}, \ldots, I_{t-n}$ are registered onto the frame I_t by means of homographies $H_{t,t-1}, \ldots, H_{t,t-n}$. The difference image is computed according to:

$$D_t = \left| I_t - \frac{1}{n}\sum_{m=1}^{n} w_m \cdot H_{t,t\text{-}m}\left(I_{t\text{-}m}\right) \right|, \quad (5)$$

where w_m regulates the weight of the frames and $H(I)$ is an image obtained after every pixel of *I* was resampled by *H*. Due to sensor motion, it makes sense in our case to set $4 \leq n \leq 6$. For small and large *m*, it turned out to be useful to choose w_m smaller than for medium *m*. The separation of the object from its background is in both cases the result of either a threshold decision or a segmentation of stable regions according to the procedure of *Maximally Stable Extremal Regions* (MSER) [13]. MoDe has the advantage of yielding acceptable contours. A disadvantage is that objects with different velocities are scanned in identical time intervals. The Time-Recursive High-Pass Filtering method is a better choice because of its comprehensive inclusion of motion history – a benefit to be paid for with more extensive computational work and half-outlines in the form of sickles. These fragmented contours need to be closed using a chain of morphological operations. To avoid high computational costs, we scaled down the difference images before segmentation.

To eliminate detected image areas that have not changed due to object motion, motion has to be verified. This can be carried out by tracking the segmented regions in intensity images or by evaluating the tracks determined using model knowledge. Verification processes will be left out of consideration here since we focus on geometric issues.

2.4 Video Frame Annotation

Annotation means enrichment of video frame content with data that is not visually perceivable and can be considered as augmented reality. Motion detected by the analysis of a video is represented on a map, and spatially relevant information from different sources is shown in the video. The procedure comprises the transformation of the image coordinates into world coordinates and vice versa. The homography H_t, computed in order to register the frame I_t onto an orthophoto, is now used as transformation matrix. For annotation of georeferenced data into the video, we use the inverse mapping with the inverse homography H_t^{-1}.

3 Concept for Real-Time Processing

Our method can be applied online using parallelization provided by modern computer architectures with multicore CPU and graphics hardware. The mosaicking process as well as motion detection are already suited for real-time applications. We discuss the most time-consuming part of the algorithm: the registration of a video frame to the orthophoto. The runtime of SIFT to localize interest points in the video frame and to compute the corresponding descriptors is at least 1 second for 720×576-resolution videos with several thousand interest points per frame. The matching of the descriptors takes additional time. On current hardware, the sequential approach does not permit online application. Fortunately, the following conceptual considerations and hardware accelerations can be used to overcome this problem:

Since there is no need to manipulate the orthophoto while the UAV is flying, the SIFT-features in the orthophoto can be precomputed for a section covering the expected operation area. As a result, the computation of SIFT-features in the orthophoto does not take any time during the registration process at all. Furthermore, orthophotos taken under different climatic (e.g., snow) or temporal (e.g., illumination) conditions can be included. Thus the orthophoto best suited to the current conditions can be chosen. Moreover, not all interest points in the orthophoto need to be considered to match them with the points in the current frame. It is sufficient to choose only points within a window around the image domain of the last frame, reprojected by the last homography into the orthophoto.

In order to filter out outliers, robust methods, such as RANSAC, must be used. During intersensorial matching, the outlier rate may be up to 90%. The use of the $T_{1,1}$-test accelerates standard RANSAC by the factor

$$\frac{p^4 + (1-p^4)\varepsilon + \tau/(pN)}{1+\tau/N}, \qquad (6)$$

where τ is the quotient of the time needed to calculate a homography and the time to evaluate the homography at a single point, N is the number of points in the samples, p is the inlier probability and ε is the probability of a point being randomly compatible with a random model. For a typical set of values, $\tau = 60$, $N = 2000$, $p = 0.1$, $\varepsilon = 0.001$, using the $T_{1,1}$-test accelerates the computations by a factor of 3.

To avoid interruptions of the mosaicking process during the calculation of the intersensorial registration, this process runs on a separate processor core. When the registration to the orthophoto is finished, the mosaicking is interrupted to correct the mosaic-to-orthophoto mapping. As shown in [16], the computation of the interest points and descriptors can be done on the GPU of a modern graphics card at a rate of 10 fps in the situation mentioned above. By running the mosaicking on one CPU-core, the SIFT-computation on the GPU, and the registration to the orthophoto on another CPU-core (Figure 2), it can be expected that the mosaic is supported every 3 to 10 frames.

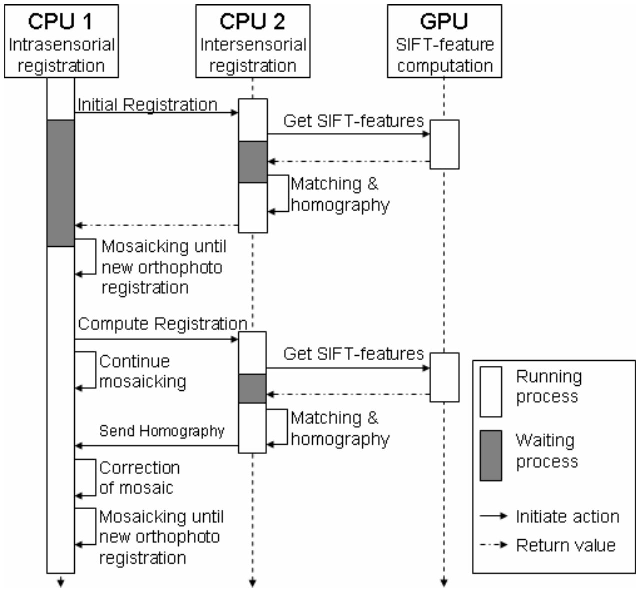

Figure 2: Once initiated, only the mosaicking process runs on the main CPU. All other computations are carried out independently on other CPUs and the GPU.

4 Experimental Tests

To provide experimental validation of the approach, we used two data sets. The first one was taken by a mini-UAV and shows the village of Bonnland. The second was taken in Karlsruhe by a camera mounted on a Cessna 172. The results of the georeferencing, motion analysis, and video frame annotation are shown using these videos.

4.1 Georeferenced Mosaics

The results of the registration of the first test sequence “Bonnland”, recorded by a small camera mounted on a mini-UAV, are shown in Figure 3. The mosaic consists of 200 frames and was successfully registered to an orthophoto and projected to a map (Figure 4). The combined approach of KLT-tracking and descriptor-based matching was also successful in the case of a frame corrupted by radio interference, as shown in Figure 5.

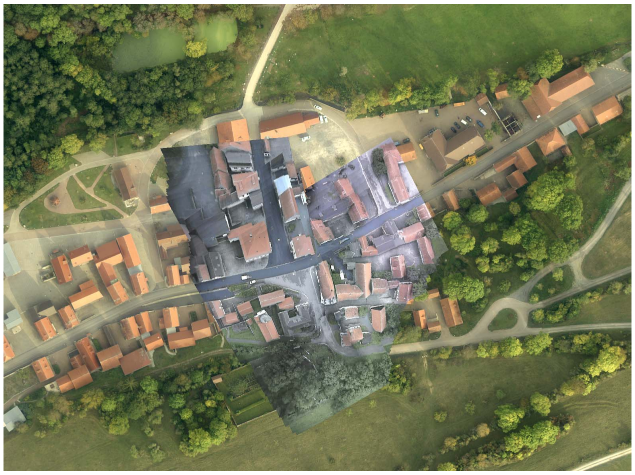

Figure 3 : Registration of a video taken during a flight over the village Bonnland.

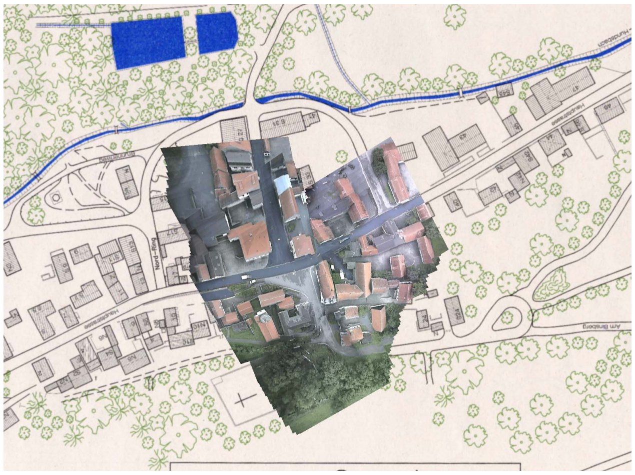

Figure 4 : Projection of the mosaic to a map of Bonnland.

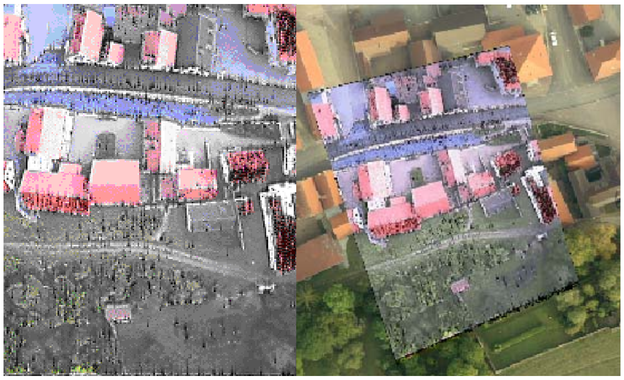

Figure 5 : Registration of a corrupted frame (left) to the orthophoto (right).

For the recording of the second sequence, “Karlsruhe, industrial harbor”, a Cessna was used, which did not offer the possibility to capture the video in nadir view. The results of registration suffer from the different view aspects (roofs of buildings are seen in the orthophoto, walls in the video) and the resulting model violation. Nevertheless, it was possible to successfully register some 100 frames taken beyond the urban area. The algorithm is robust with respect to scale changes, as depicted in Figure 6.

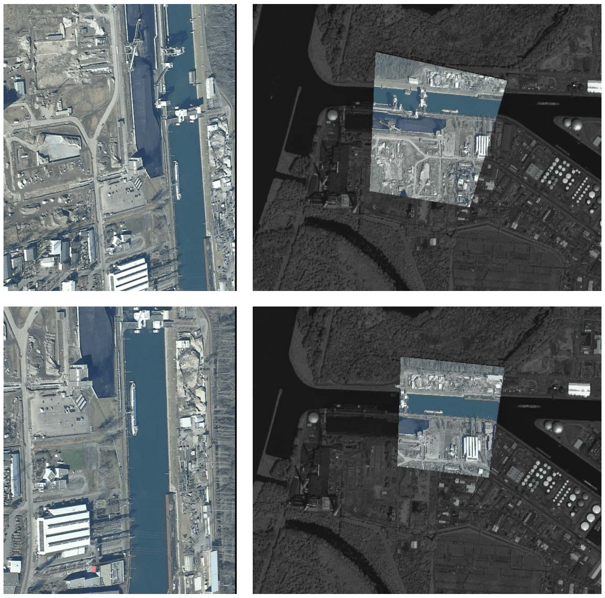

Figure 6 : Registration of two different video frames to an orthophoto of Karlsruhe. As can be seen here, registration was successful, even after zooming the camera.

4.2 Object Motion Analysis

During experiments, Time-Recursive High-Pass Filtering was applied on the test sequence "Bonnland". Vehicles passing through the scene were thus recognized as moving objects (Figure 7) after a time period that was necessary to build up a motion history. There were not more than 2 seconds available to track the objects and they sufficed to obtain an estimated absolute velocity and heading (Figure 8). Vertical building surfaces exposed due to aspect changes were wrongly detected as movements. They were successfully eliminated by means of motion analysis.

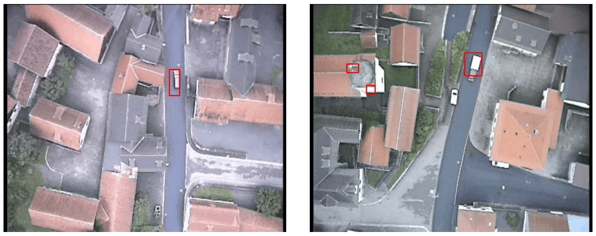

Figure 7 : Detected moving objects are marked with bounding boxes. Due to violation of the 2D-model assumption, the tower of the church was wrongly detected (right).

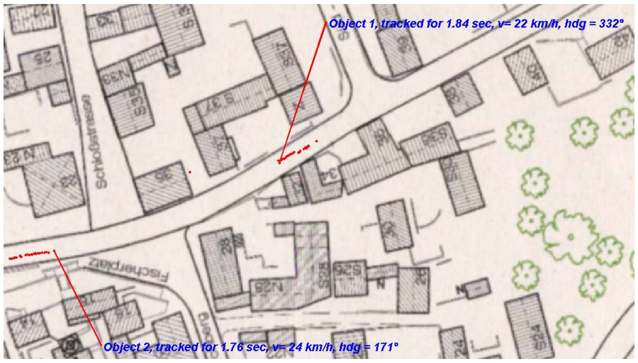
Object 1, tracked for 1.84 sec, v= 22 km/h, hdg = 332°
Schloßstrasse
Fischerplatz
Object 2, tracked for 1.76 sec, v= 24 km/h, hdg = 171°

Figure 8 : The tracks of the moving cars are marked on the map. Measured velocity and heading are displayed as well.

4.3 Video Frame Annotation

To simulate mapping of georeferenced data, we took UTM coordinates of noticeable objects (e.g., duct covers), transformed them into some video frames and marked the position. This simple mapping enables performance evaluation by estimating the geometric errors that we can expect in case of annotation. In 58% of all samples, the distance between real and annotated position was smaller than 1.5 meters; in 25%, greater than 2.0 meters (Figure 9).

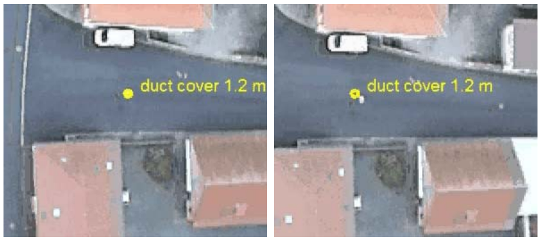

Figure 9 : Annotation of a duct cover in Bonnland (displayed detail of Figure 7). The result of a good and average registration is shown.

5 Conclusions and Outlook

We have introduced here a system for registration of UAV-videos to an orthophoto. We showed that it is possible, by use of a georeferenced orthophoto, to annotate situation information in video as well as to map selected objects in the video to maps. Thus, prompt reaction to information gained during a UAV-mission is supported. By combining a mosaicking procedure and slower registration to the orthophoto, real-time capability is acquired along with simultaneous good georeferencing quality. To enable this, observation of registration quality is included to support intrasensorial registration only when needed. In contrast to [9], no warping of the orthophoto or video frame is needed. Our experiments demonstrated that the resulting mapping was adequate for the tasks at hand. This is true even for the search of difficult frame-to-orthophoto mappings, yielding inlier rates of only 10% for the correspondences.

In the future, our method has to be transferred to the architecture depicted above. The actual run-time performance has to be evaluated on a broader set of video sequences.

To enhance the robustness of the intersensorial registration, advancements of the SIFT-algorithm by including color information will be a topic of our future work.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, *SURF: Speeded Up Robust Features*, In Proceedings of the 9th European Conference on Computer Vision, 2006

[2] Dirk Borghys et al., *Multi-sensor Image Exploitation*, NATO RTO technical Report, AC323(SET)TP/10, December 2001, NATO UNCLASSIFIED

[3] Olivier Faugeras and Francis Lustman, *Motion and Structure from Motion in a Piecewise Planar Environment*, International Journal of Pattern Recognition in Artificial Intelligence, vol. 2, pp. 485-508, 1988

[4] Martin A. Fischler and Robert C. Bolles, *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*, Communications of the ACM, vol. 24, no. 6, pp. 381-395, 1981

[5] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000

[6] Berthold K. P. Horn and Brian G. Schunck, *Determining Optical Flow.* Artificial Intelligence, 17, pp. 185-203, 1981

[7] Todd Jamison, Courtland Davis, Chong-ket Chuah, and Mark Lucas, *Automated Georeferencing of Videos Using Registration with Controlled Imagery and Video-Photogrammetry.* Proceedings of the APSRS, 2000

[8] Wolfgang Krüger, *Robust and Efficient Map-to-image Registration with Line Segments,* Machine Vision and Applications, vol. 13, no. 1, pp. 38-50, 2001

[9] Yuping Lin, Qian Yu, and Gerad Medioni, *Map-Enhanced UAV Image Sequence Registration.* IEEE Workshop on Applications of Computer Vision (WACV'07), 2007

[10] David G. Lowe, *Distinctive Image Features from Scale-Invariant Key-points,* International Journal on Computer Vision (IJCV), vol. 60, no. 2, pp. 91-110, 2004

[11] Bruce D. Lucas and Takeo Kanade, *An Iterative Image Registration Technique with an Application to Stereo Vision*, Proceedings of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679, 1981

[12] Jiří Matas and Ondřej Chum, *Randomized RANSAC with $T_{d,d}$ test***,** Proceedings of the British Machine Vision Conference (BMVA), vol. 2, pp. 448-457, 2002

[13] Jiří Matas, Ondřej Chum, Martin Urban, and Tomáš Pajdla, *Robust Wide Baseline Stereo from Maximally Stable Extremal Regions*, Proceedings of the British Machine Vision Conference (BMVA), vol. 1, pp. 384-393, 2002

[14] J. Chris McGlone, Edward M. Mikhail, and James Bethel, *Manual of Photogrammetry*, American Society of Photogrammetry and Remote Sensing, 5th edition, 2004

[15] Jochen Meidow and Michael Kirchhof, *Continuous Self-Calibration and Ego-Motion Determination of a Moving Camera Observing a Plane*, Photogrammetric Image Analysis (PIA07), International Society for Photogrammetry and Remote Sensing, vol. XXXVI, part 3/W49A of IAPRS, pp. 37-42, 2007

[16] Sudipta N. Sinha, Philippos Mordohai, and Marc Pollefeys, *GPU-Based Video Feature Tracking and Matching*, EDGE 2006, workshop on Edge Computing Using New Commodity Architectures, Chapel Hill, May 2006

[17] Philip S. Torr, Andrew W. Fitzgibbon, and Andrew Zisserman, *The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences.* International Journal of Computer Vision, vol. 32, no. 1, pp. 27-44, 1999

[18] Richard P. Wildes et al., *Video Georegistration: Algorithm and Quantitative Evaluation.* Proceedings of. 8th International Conference on Computer Vision, vol. 2, pp. 343-350, 2001

[19] Christopher R. Wren, Ali Azabayejani, Trevor Darell, and Alex Pentland, *Pfinder: Real-Time Tracking of the Human Body*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, 1997