TRANSLATING AERIAL IMAGES INTO STREET-MAP-LIKE REPRESENTATIONS FOR VISUAL SELF-LOCALIZATION OF UAVS

Michael Schleiss

Fraunhofer FKIE, Wachtberg, Germany - michael.schleiss@fkie.fraunhofer.de

KEY WORDS: Visual Positioning, Unmanned Aerial Vehicle, Image Translation, GAN, Geo-Referencing

ABSTRACT:

Unmanned aerial vehicles (UAVs) rely on global navigation satellite systems (GNSS) like the Global Positioning System (GPS) for navigation but GNSS signals can be easily jammed. Therefore, we propose a visual localization method that uses a camera and data from Open Street Maps in order to replace GNSS. First, the aerial imagery from the onboard camera is translated into a map-like representation. Then we match it with a reference map to infer the vehicle's position. An experiment over a typical sized mission area shows localization accuracy close to commercial GPS. Compared to previous methods ours is applicable to a broader range of scenarios. It can incorporate multiple types of landmarks like roads and buildings and it outputs absolute positions with higher frequency and confidence and can be used at altitudes typical for commercial UAVs. Our results show that the proposed method can serve as a backup to GNSS systems where suitable landmarks are available.

1. INTRODUCTION

Typical UAVs today are reliant on a combination of GPS and inertial navigation system (INS) for accurate position estimation. GPS and INS work complementary. INS are prone to drift but provide frequent state updates which are less noisy. GPS gives an absolute position value, which in contrast can be very noisy. Combined through an algorithm like the Kalman filter they provide a drift-free and precise position estimation.

INS alone, especially those that are part of light-weight, commercial drones, can only maintain an accurate position for a short time. If the GPS signal is lost, which can happen due to malicious attacks like jamming or unavailability of the signal (canyons, mountains, high buildings, electromagnetic interferences, weather) this can lead to a loss of the vehicle (Carroll, 2003).

Visual Odometry has been shown to alleviate the problem of short term GPS outages and can maintain a precise position estimation over a few hundred meters flight path (Sa et al., 2018) but is, similar to INS, also prone to drift over time.

In the past methods have been proposed that solely rely on a monocular video camera and are intended as backup or replacement of GPS (Conte and Doherty, 2009, Cesetti et al., 2011, Shan et al., 2015, Grönwall et al., 2017). Images taken from the onboard camera are matched with georeferenced aerial imagery providing absolute global positions.

These methods have different shortcomings. They fail on a regular basis when the reference material is not representative of the actual situation while flying or can be used only at very high altitudes. But most importantly, they have a low rate of good matches.

Their localization performance is somewhat hidden. In the experiments, they are evaluated as a component of a larger navigation system usually combined with visual odometry, which delivers high-frequency state updates. They only serve to reduce drift in case they are able to find a good match. But it is unclear how these components would perform on their own.

Figure 1. Method Overview: First, an aerial image is collected, then it is segmented into a map-like representation, finally it is matched within a reference map for localization of the vehicle.

An exception is the work by Mannberg and Savvaris. The authors investigate the standalone performance of an absolute localization component (Mannberg and Savvaris, 2014). It achieves a high matching rate but only considers buildings as landmarks to match. This performs well over cities. However, in some situations it might be useful to utilize other types of landmarks like roads, forests or rivers.

The aim of this paper is to present a method that is robust to temporal changes in the reference material, delivers a high matching rate and is able to incorporate multiple types of landmarks.

Inspired by recent improvements in satellite and aerial image segmentation (Demir et al., 2018) we extend the template matching approaches mentioned above by segmenting the onboard imagery into map-like representations and then matching them to a reference map. We use two types of landmarks, buildings and roads, to present the possibility of using multiple landmarks.

Our experiments show that the proposed method can obtain accurate absolute positions with a high matching rate. We evaluate our method on real flight data and abstain from combining it with relative localization methods like visual odometry because we want to examine its standalone absolute localization performance.

2. RELATED WORK

One of the first attempts to retrieve a global absolute position from UAV onboard imagery was proposed by Conte and Doherty. They combine their visual odometry with an algorithm which registers the onboard video to a geo-referenced satellite image to reduce drift (Conte and Doherty, 2009). The matching approach is based on normalized correlation of image intensities. They report reasonable results for their whole localization pipeline, but this is mostly due to the performance of the visual odometry. The image registration module outputs only very sparse absolute positions because most matches are discarded due to high uncertainty. In their experiment, only two successful matches could be made.

In contrast Cesetti et al. use pattern matching to georeference the onboard imagery using SIFT features (Cesetti et al., 2011). It requires the air vehicle to fly at very high altitudes in order to extract meaningful features from natural landmarks in the onboard imagery. The images used in their experiment have a ground footprint of at least one square kilometer. This restricts its use only to certain scenarios.

Grönwall et al. extend (Conte and Doherty, 2009) by adding lidar measurements for visual odometry (Grönwall et al., 2017). However, the basic problem of low matches is still present. On the other hand (Shan et al., 2015) translate the onboard images and the reference map into HOG space (Histogram of oriented Gradients). The authors report a high matching rate in a small scale scenario. But HOG features have been shown to be too lossy for many challenging object detection tasks (Vondrick et al., 2013) and can be easily outperformed by deep neural networks.

Lindsten et al. segment the imagery into different classes like streets, buildings, grassland, rivers, etc. using superpixels and then compare the histogram of classes to a reference map (Lindsten et al., 2010). However, by using a histogram spatial information is lost and the resulting position estimates can be very ambiguous in areas with similar class distributions.

Mannberg and Savvaris are using object detectors to identify the position of buildings in aerial imagery and reduce the detections to a representation where each building is represented by a point on a map (Mannberg and Savvaris, 2014). A fingerprint that considers the geometric allocation of the points is calculated and matched against a reference database. The authors claim that their framework could be used with other landmark types, not only buildings. But it is unclear how landmarks that cannot be reduced to points like rivers or roads would be incorporated.

We use the same template matching technique like Conte and Doherty. But by segmenting the onboard imagery we transform them into a more robust representation which achieves higher matching rates. The segmentation process is similar to Lindsten et al. but we fit our images into a reference map instead of using a histogram, hence maintaining their spatial arrangement. Our method works well at typical altitudes of commercial drones and we are able to incorporate multiple types of landmarks.

3. METHOD

Our localization framework consists of three steps. First, an image is collected by the onboard camera. Then the image is translated into a map like representation. Finally, it is matched within a reference map of the mission area (see Figure 1).

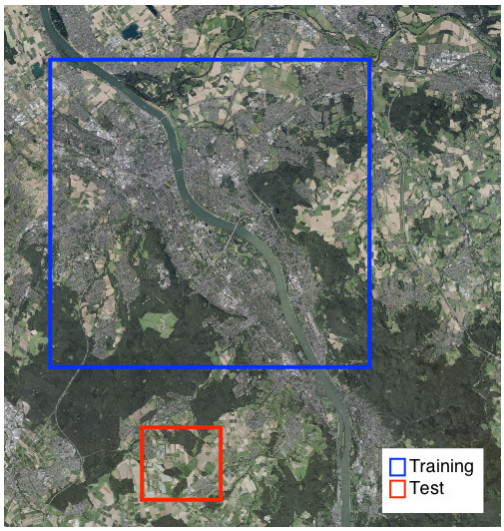

Figure 2. Aerial footage that is used for training and evaluating the image segmenter

The images are taken by a downward looking camera. This can be achieved by a gimbal that automatically levels the camera. We transform the images into a class map representation with the help of an image segmenter based on a conditional generative adversarial network (cGAN). In this paper, we consider three types of classes, buildings, roads, and background.

cGANs have been shown to perform very well for various image segmentation tasks (Luc et al., 2016). cGANs consist of a generator and a discriminator network. Conventional CNNs try to minimize a predefined loss function like the per-pixel mean squared error. But this disregards global consistency. On the other hand, the discriminator of a cGAN learns a loss that tries to distinguish if a generated image is fake or real. It considers not only pixel-wise correctness but if output is consistent as a whole (Isola et al., 2017), which is useful in our image segmentation setting.

In the next step, the segmented image is fit into a map of the mission area. It is assumend that the scale and rotation of the image are known. These parameters are provided by the compass and altimeter of the vehicle's inertial navigation system. The images are then scaled and aligned to conform to the reference map's ground resolution and orientation.

We use as suggested by Conte (Conte and Doherty, 2009) a template correlation method. We slide the segmented image over the map and at each position we calculate the sum of normalized squared difference between all pixel intensities. The position with the lowest difference is a match and the center position of the match corresponds to the estimated position of the vehicle.

There are a few assumptions made in our method. We assume a flat earth model. This is reasonable for high altitudes and most landscapes but could be problematic in specific situation (e.g. skyscrapers or mountains at low altitudes). We also assume that the intitial position of the vehicle is known so it can be equipped with a suitable reference map.

4. TRAINING IMAGE SEGMENTER

We train an image segmenter using a large collection of publicly available aerial images and Open Street Map data before being able to apply the trained network within our localization framework on our own dataset of aerial images.

The image segmenter is trained with images from the city of Bonn and its surrounding area (see Figure 2). The training and the evaluation set consist of images from 125km^2 and 9km^2 respectively. The area is split into tiles of 512x512 pixels with a ground resolution of 0.2m per pixel. We use RGB channels only. For more details on the training procedure and network architecture please refer to (Isola et al., 2017).

Labels were created with data from Open Street Maps which provides annotations of building footprints and roads amongst other landmarks. For roads only the centerline is available. The road width is estimated through context information provided by Open Street Maps like the rank of the road (highway, major road, residential street, etc.).

We train the network for 100 epochs and achieve an Intersection over Union (IoU) of 69% for building footprints and 58% for roads on the evaluation set. The Intersection over Union metric, also known as Jaccard Index, quantifies the percent overlap between the ground truth labels and the prediction output.

The relatively low accuracy can be interpreted in the following way: As can be seen from Figure 3 building footprints are sometimes very coarse. This might lead the image segmenter to pick up features from the buildings' surroundings like pavements or lawns and consequently misclassify fields of grass or streets as buildings. Same is true for roads since we just estimate the width of the road while creating the labels. This issue was partially compensated by using a very large training set to compensate for noisy labels.

Also sometimes buildings are present in the aerial images but a corresponding label from Open Street Maps is missing. Therefore the network might predict objects correctly that are not contained in the ground truth, e.g. newly constructed buildings, garages, garden sheds. This means that the accuracy number stated above might underestimate the quality of the predictions to some degree.

We show an exemplary input and prediction pair in Figure 4. We can clearly see that the image is segmented into a reasonable representation of the roads and buildings. A small part of the road is missing in the lower part of the image. However, the next section will show that these segmentation masks are good enough for good localization results as long as there are enough features to match.

5. EXPERIMENTS

The localization experiment is conducted on an entirely separate dataset from the one that was used during training. Instead of using publicly available aerial images we collected our own data with a plane that was flying over a small stretch of land south of Bonn. Besides the spatial separation, it was taken with a different camera at a different time of the year.

We can therefore show that our method is generalizable to different mission areas and across different cameras. It is also

Figure 3. Building footprints obtained through Open Street Maps are sometimes misaligned or missing. The highlighted footprint on the left for example overlaps with some pavement and lawn. The one on the right is too narrow.

Figure 4. On the left aerial image that serves as input to the image segmenter. On the right segmented image. Building footprints in brown and roads in yellow.

robust to seasonal and temporal changes. However, we do not test for extreme changes in appearance, i.e. landscapes covered in snow or day and night change. The image segmenter would need to be trained specifically for such challenges.

The dataset covers an area that is 560m long and 680m wide. This area was covered by the plane multiple times in a checkerboards manner (see Figure 5). It consists of 4121 overlapping images depicting a rural area with industrial buildings in the center, a small village at the top, a forest to the east and farmland to the south and west. Some parts of the area do not contain buildings or roads to match.

The plane was flying at an altitude of approximately 300m above ground. It was equipped with a downward looking camera taking RGB images at 10Hz. The high altitude compared to a typical civil drone flight is offset by the camera's narrow field of view of 39.1°. The images cover a footprint of approximately 216m x 144m on the ground at a ground resolution of less than 0.1m per pixel. It was also equipped with an inertial navigation system and a GPS receiver with Real Time Kinematics (RTK) capabilities providing very accurate positioning information for the ground truth to which our method is compared.

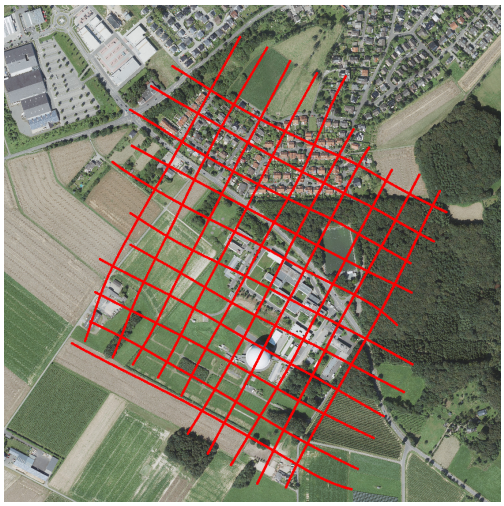

Figure 5. GPS tracks belonging to the whole dataset of 4121 images.

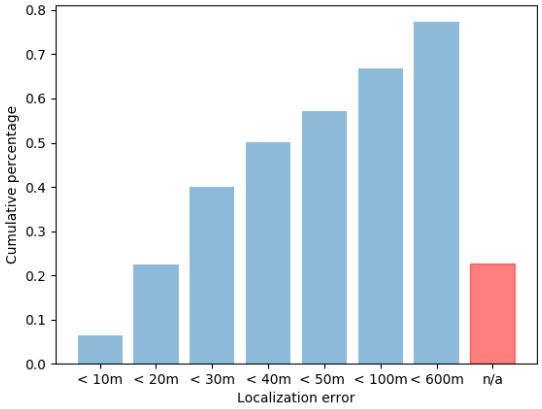

Figure 6. The cumulative error distribution for the whole dataset. Images that could not be matched due to missing landmarks are described as n/a and plotted separately in red.

We analyze the localization performance on a subset of the data first. This allows us to plot the localization results in a clear and uncluttered manner and visually assess the quality of our method. Afterward, we report the results for the whole dataset.

The data points for the subset were chosen such that they create a continuous, hypothetical flight path and each footprint would always contain features to match. Therefore areas without any buildings or roads like grass fields and farmland were omitted while creating the flight path. The flight path is 1.61km long, consists of 471 images and is plotted in Figure 7 among the predicted flight path.

It can be seen that for the most part the predicted path follows the ground truth. Only a few points to the left side of Figure 7 are clearly mislocated. We measure the localization error as the euclidian distance between the ground truth and the prediction.

The median error for this flight path is 22.7m. 19% of the predictions have an error less than 10m and only 9% have an error above 50m. In comparison, consumer grade GNSS receivers achieve an accuracy that ranges from about 5 to 10m (Tiberius, 2004) under ideal conditions. Of course, it has to be taken into account that we chose the flight path to omit areas without suitable landmarks.

The results for the whole dataset (see Figure 6) are more differentiated since it also contains images without buildings or roads. The matching rate, in this case, is lower. 20.5% of the images could not be matched. And the median error is around 40m.

Figure 8 depicts a heatmap that demonstrates where our method performs well and where it fails. Not surprisingly, areas with a lot of buildings and streets exhibit a higher potential for good localization results than the forest to the east or the farmland to the south and west.

6. CONCLUSION

We have shown that matching aerial images that were translated into a map-like representation can be used for global localization. We achieved a high matching rate with a standalone method for absolute localization. Experiments on real flight data exhibited that a localization error that is close to that of consumer-grade GNSS systems is possible.

The results could further be improved by incoporating additional knowledge. For example a motion model could rule out implausible movements and visual odometry could provide state updates in situations where suitable landmarks for absolute localization are not available.

Also, future directions of research will examine how natural landmarks could be used to increase the capabilities over unstructured areas without roads and buildings.

REFERENCES

Carroll, J.V., 2003. Vulnerability assessment of the U.S. transportation infrastructure that relies on the Global Positioning System. *Journal of Navigation*, 56, 185–193.

Cesetti, A., Frontoni, E., Mancini, A., Ascani, A., Zingaretti, P., Longhi, S., 2011. A visual global positioning system for unmanned aerial vehicles used in photogrammetric applications. *Journal of intelligent & robotic systems*, 61(1-4), 157–168.

Conte, G., Doherty, P., 2009. Vision-based unmanned aerial vehicle navigation using geo-referenced information. *EURASIP Journal on Advances in Signal Processing*, 2009, 1–18.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 172–181.

Grönwall, C., Rydell, J., Tulldahl, M., Zhang, E., Bissmarck, F., Bilock, E., 2017. Two imaging systems for positioning and navigation. *2017 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS)*, 120–125.

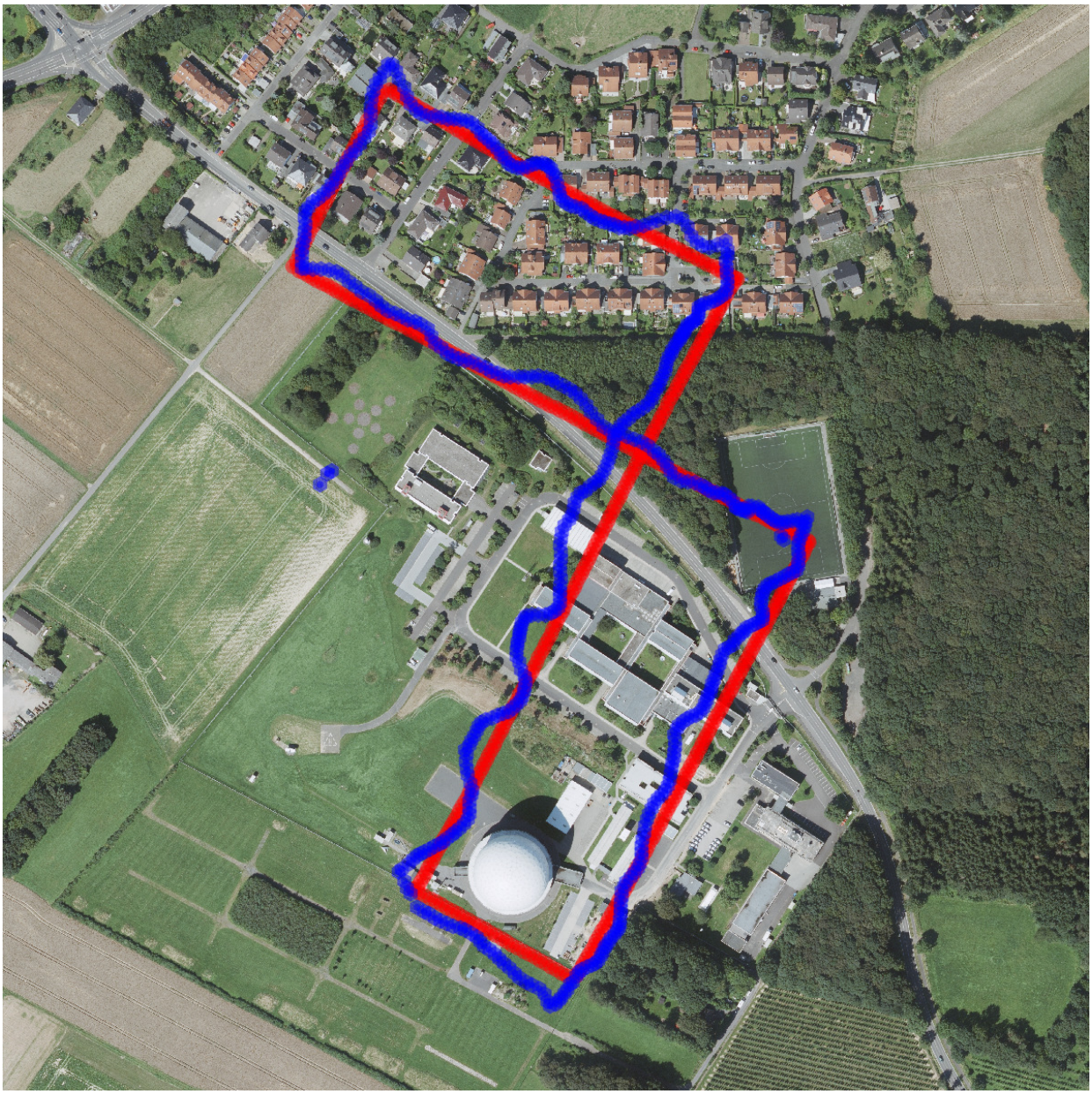

Figure 7. The GPS track of the subset is plotted in red. The corresponding predicted locations are blue. We can see a few outliers to the left of the image.

Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976.

Lindsten, F., Callmer, J., Ohlsson, H., Törnqvist, D., Schön, T.B., Gustafsson, F., 2010. Geo-referencing for UAV navigation using environmental classification. *2010 IEEE International Conference on Robotics and Automation*, 1420–1425.

Luc, P., Couprie, C., Chintala, S., Verbeek, J., 2016. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408.*

Mannberg, M., Savvaris, A., 2014. Landmark fingerprinting and matching for aerial positioning systems. *Journal of Aerospace Information Systems*, 11(3), 131–139.

Sa, I., Kamel, M., Burri, M., Bloesch, M., Khanna, R., Popovic, M., Nieto, J., Siegwart, R., 2018. Build your own visual-inertial drone: A cost-effective and open-source autonomous drone. *IEEE Robotics & Automation Magazine*, 25(1), 89–103.

Shan, M., Wang, F., Lin, F., Gao, Z., Tang, Y.Z., Chen, B.M., 2015. Google map aided visual navigation for UAVs in GPS-denied environment. *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 114–119.

Tiberius, C., Verbree E., 2004. GNSS positioning accuracy and availability within location based services: The advantages of combined GPS-Galileo positioning. *2nd ESA/Estec workshop on Satellite Navigation User Equipment Technologies*, 1–12.

Vondrick, C., Khosla, A., Malisiewicz, T., Torralba, A., 2013. Hoggles: Visualizing object detection features. *Proceedings of the IEEE International Conference on Computer Vision*, 1–8.

Figure 8. This heatmap visualizes the localization quality over the whole dataset. Green areas correspond to low localization errors, red areas correspond to high localization errors.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W13-575-2019