33RD CONGRESS
OF THE INTERNATIONAL COUNCIL
OF THE AERONAUTICAL SCIENCES
STOCKHOLM, SWEDEN, 4–9 SEPTEMBER, 2022

ICAS 2022 SWEDEN

VISION-BASED LOCALISATION FOR AUTONOMOUS AERIAL NAVIGATION IN GNSS-DENIED SITUATIONS

Daniel Sabel[1,2], Atsuto Maki[1], Torbjörn Westin[2] & Dag Åsvärn[2]

[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Spacemetric AB, Sollentuna, Sweden

Abstract

Flight safety is compromised when positioning with Global Navigation Satellite Systems (GNSS) is rendered unusable or unreliable. The cause of such GNSS-denied situations can be signal obstruction, multipath issues from mountains or tall buildings as well as malicious jamming or spoofing.

The Swedish company Spacemetric is working together with the Royal Institute of Technology (KTH) in Stockholm to address the challenge of accurate and reliable aerial localisation, which is a key aspect to ensure flight safety for autonomous aerial navigation. The project, called “Autonomous Navigation Support from Real-Time Visual Mapping”, focusses on vision-based methods for aerial positioning in Unmanned Aerial Vehicles navigating in natural environments.

This paper present results from the project’s ongoing research on vision-based approaches with deep learning as well as a novel approach to localisation which exploits the three-dimensional structure of vegetation and terrain.

Keywords: UAV, visual methods, GNSS-denied navigation, natural environments, deep learning

1. Introduction

The advantages of Unmanned Aerial Vehicles (UAVs) include cost effectiveness, ease of deployment and possibilities for automated operations. UAVs can vary in size from a few hundred grams to that of conventional fixed or rotary wing aircrafts. While originally developed for military purposes such as reconnaissance or aerial attacks, UAVs are today increasingly being used in the civilian and public domains. Applications include real-time monitoring of road traffic, environmental remote sensing, search and rescue operations, delivery of medical supplies in hard-to-reach areas, security and surveillance, precision agriculture, and civil infrastructure inspections [1]. For instance, fast delivery of defibrillators with UAVs is being evaluated in Sweden [2] with the first successful life-saving delivery in the beginning of 2022 [3]. In Rwanda and Ghana, autonomously navigating fixed-wing UAVs are delivering medical supplies in areas with limited infrastructure [4].

Increasing degrees of autonomy in UAVs are expected to provide immense advantages, especially in support of public safety, search and rescue operations, and disaster management [1].

The Swedish company Spacemetric is working together with the Royal Institute of Technology (KTH) in Stockholm to address the challenge of accurate and reliable aerial localisation, which is a key aspect to ensure flight safety for autonomous UAV navigation. The long-term vision is to contribute to improved flight safety for autonomous UAVs, by providing a fallback option for map-relative localisation, in the event that conventional methods of localisation become unavailable or unreliable.

The research project focusses on vision-based methods for localisation of UAVs in non-urban environments. Such methods are rooted in the domain of photogrammetry and computer vision and today have strong links with machine learning. The project builds on one hand on Spacemetric’s experience with photogrammetry, sensor modelling, 3D model generation and georeferencing of image data, stretching over more than two decades, and on the other hand on the vast experience of KTH with respect to computer vision and deep learning.

This paper presents the results of ongoing research on UAV localisation with

- deep learning approaches for UAV image matching against satellite images; and

- registration of 3D models derived from UAV images with georeferenced 3D models of terrain and vegetation.

1.1 GNSS-denied navigation

Unpiloted UAVs commonly navigate based on a combination of dead reckoning and Global Satellite Navigation Systems (GNSSs), such as GPS and GLONASS. Dead reckoning, most commonly inertial measurements, allows frequent estimation of changes in the vehicle's velocity, position and orientation. Such measurements are however affected by unbounded drift errors, as measurement errors are integrated over time. GNSS, on the other hand, provides estimates of absolute position and is used to constrain the drift errors from inertial measurements, while also providing map-relative localisation [5].

There are situations when localisation with GNSS can be rendered unusable or unreliable. The cause of this can be instrument failure, signal obstruction, multipath issues from mountains or tall buildings, and unintentional radio frequency interference as well as malicious jamming or spoofing [6], [7]. Such events can quickly render the UAV uncapable of navigating safely. The importance of being able to navigate without GNSS has been highlighted by the U.S. Army as a reaction to increased concerns of jamming and spoofing of GPS signals targeted at their unmanned aircraft systems [8].

1.2 Vision-based localisation

An alternative to GNSS for map-relative localisation is vision-based localisation, whereby data from a sensor mounted on the aerial vehicle is compared with a reference model of the environment. The sensor is often a camera, but can also be a laser scanner, a radar altimeter, or some other type of sensor. At its core, vision-based localisation is a camera pose estimation problem, where the objective is to estimate the position, i.e. the three spatial coordinates of the camera, as well as the orientation, i.e. the three angles of rotation relative to some frame of reference.

Map-relative localisation is closely related to visual odometry, in which the motion of the vehicle is estimated by matching sequences of sensor data acquired by the vehicle. The difference is that visual odometry computes position and orientation through integration of estimates of relative motion between sensor acquisitions, while map-relative visual localisation directly estimates the position and orientation relative some reference model of the world. The reference model may consist of geocoded satellite or aerial images, or 3D models, which are stored onboard the UAV. Our project is concerned with map-relative localisation.

The challenge of vision-based localisation in natural environments is the accurate registration of the UAV's observation data with the reference model, given potentially large differences in viewing perspective, scale, illumination, shadows, clouds and scene content. Furthermore, the localisation task must be carried out in real-time to be of value for the navigation system.

Most of the research on vision-based localisation has hitherto been devoted to indoor settings and outdoor urban environments. However, many applications require reliable localisation also in non-urban areas, including search and rescue missions, delivery of supplies in inaccessible locations as well as large-scale surveillance and monitoring.

Urban environments often contain many features that are distinct and stable over time and therefore are suitable image matching. The same cannot be said for forests or open terrain, as these are often affected by seasonal variations, natural growth and decay, as well as by harvesting and logging. An example of such differences in two aerial images over a rural location is shown in Figure 1. The left image was acquired on May 13, 2016, whilst the right image was acquired on April 11, 2018. Significant differences in colour and texture over the forested area can be observed, resulting from the area being imaged before and after the onset of leafing. Shadows are cast in different directions, as the images were acquired at different times of day. Such differences complicate image-based localisation. In case of low-altitude flights, the reduced imaged footprint on the ground can exacerbate the challenge, as exemplified in Figure 2.

Figure 1 - An example of significant differences in image content due to seasonal variations and illumination conditions. Left: Aerial image acquired on May 13, 2016, at 9:41AM. Right: Same area acquired on April 11, 2018, at 3:31PM. Aerial images courtesy of Lantmäteriet, Sweden.

Figure 2 – Example of a forest patch from Figure 1 with foliated (left) and bare (right) tree canopies. The extent of the images corresponds to an acquisition from a nadir-pointing camera with a 65° angular field of view 120 m above the ground. Aerial images courtesy of Lantmäteriet, Sweden.

The authors of [9] classified vision-based localisation approaches into i) template matching, ii) feature point matching and iii) deep learning. Template matching directly compares pixel intensities or metrics derived thereof, such as correlation or mutual information [10]. Feature point matching instead detects distinct image features, such as corners, and match descriptions of those features. Relatively few contributions, however, exploit deep learning. [9] attribute this to a lack of publicly available UAV data and to the current infeasibility of training Convolutional Neural Networks to learn complete reference maps accurately enough due to inherent memory limitations. The authors assess that in this field, the current state-of-art consists of approaches that use a combination of feature point matching and statistical filtering.

The use of deep learning approaches and Convolutional Neural Networks (CNNs) are of interest due to the ability of such networks to learn complex patterns and their successes for image-related tasks, such as image recognition and object detection and classification.

2. Deep Learning approaches

There is a general lack of studies on deep learning for aerial pose estimation in non-urban areas. We therefore assessed the feasibility of two different published methods that had been developed and evaluated over urban and rural areas. We evaluated the efficacy of these methods over forests and fields at several sites in Sweden.

Two different deep learning methods were evaluated, namely those of [11] and [12].

2.1 End-to-end Convolution Neural Network

[12] focussed on cross-view geolocalisation over urban areas using high resolution satellite images as reference data. It is an end-to-end deep learning approach, which means that it attempts to replace the entire conventional pose estimation pipeline of feature detection, description, matching as well as camera pose estimation by estimating the pose directly from the images.

2.1.1 Method

The method uses two separate networks, both based on the AlexNet Convolutional Neural Network architecture. An overview of the method is shown in Figure 3. The camera localisation network takes as its input a single UAV image and a patch of a satellite image, and outputs an estimated camera pose relative to the satellite image patch. The horizontal position is estimated through classification in an 8x8 cells grid with a ±200 m range along each of the two axes. The altitude and the heading and tilt angles are estimated through regression. The roll angle is assumed to be fixed at 0° and is not estimated.

A second network, here called “scene similarity network”, generates a measure of similarity between a UAV image and a satellite image patch. The similarity measure is used to combine a set of nine preliminary pose estimates from the camera localization network to obtain the final pose estimate. The nine preliminary poses are estimated based on the nine satellite image patches closest to the a priori estimate of the UAV image.

[12] used a Kalman Filter to fuse pose estimate with visual odometry to improve the overall accuracy. Our intention was to evaluate how accurate their end-to-end deep learning approach is over non-urban areas. We therefore evaluated that method in isolation, without the integration with visual odometry. For full details on the method, see [12].

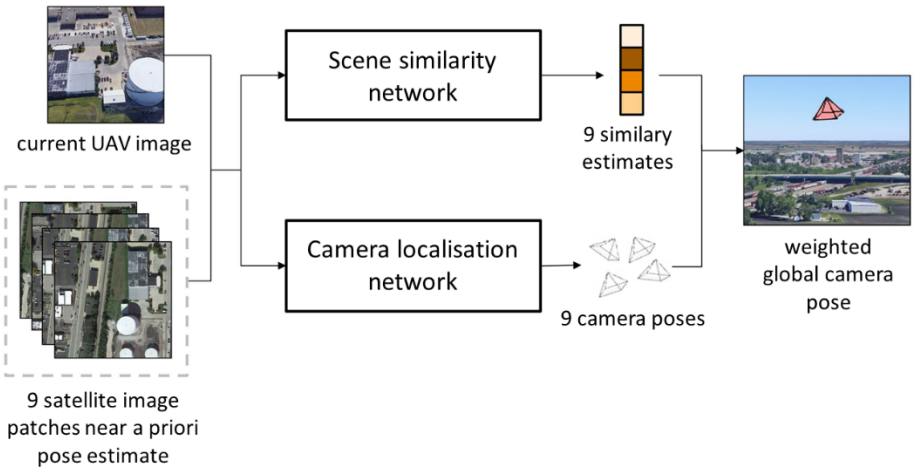

Figure 3 – Overview of the deep learning-based method proposed by [12]. Graphic adapted from [12].

2.1.2 Evaluation

This section describes the experimental setup used in our evaluation of the method proposed by [12]. We generated training data from accurately geolocated, high resolution aerial orthophotos provided by Lantmäteriet. Satellite images were simulated at 0.5 m pixel size by cropping and subsampling 480 m by 480 m patches from the orthophotos. UAV perspective images were simulated from orthophotos by varying the above ground height between 200 m and 300 m while varying the camera tilt angle between nadir (0°) and 45° off-nadir and allowing full freedom in the heading angle. Care was taken to produce a balanced dataset, such as a uniform sampling of all the classes for the horizontal classification problem.

Training was performed over two areas in Sweden, shown in Figure 4. The training areas are dominated by forest and agricultural fields and has a combined area of 100 km^2.

Figure 4 – Non-urban areas for network training. Left: “Ockelbo”, which is covered with mainly coniferous trees and agricultural fields. Right: “Tomelilla”, containing agricultural fields and a mix of deciduous and coniferous trees.

UAV and satellite image pairs were generated from orthophotos acquired several years apart, to simulate the fact that up-to-date reference data may not be available. 15000 pairs were generated and annotated with the camera pose parameters, for each of the training areas. Examples of training data are given in Figure 5 and Figure 6.

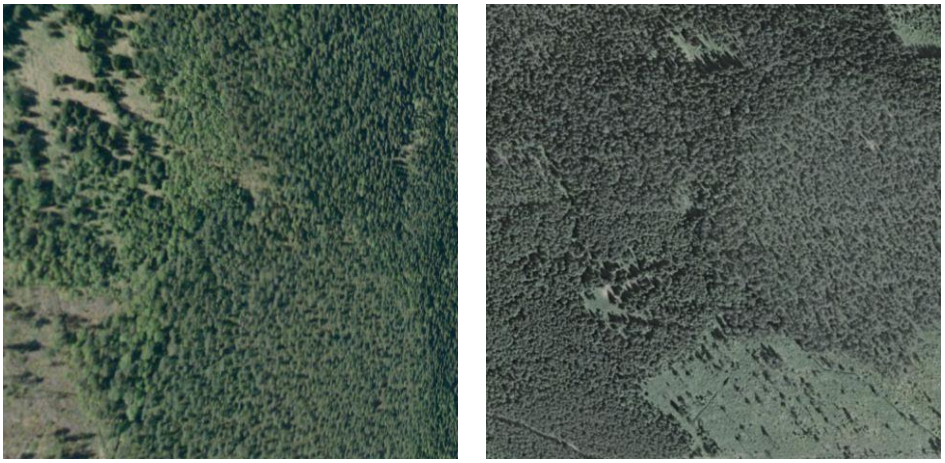

Figure 5 – Example of training data pair from the Ockelbo area. Left: simulated UAV image. Right: corresponding simulated satellite image.

Figure 6 – Example of training data pair from the Tomelilla area. Left: simulated UAV image. Right: corresponding simulated satellite image.

Figure 7 – Evaluation area “Grolanda”, containing a mix of agricultural fields and deciduous and coniferous trees.

The evaluation of the accuracy was performed with 2000 UAV images over a 42 km2 third non-urban site, shown in Figure 7.

2.1.3 Results

The mean horizontal position accuracy over the Grolanda evaluation site was 86 m. This is an unacceptable error for most applications and a 54% reduction from the error of simply uniformly guessing the horizontal position within the classification grid (which would result in a mean accuracy of 188 m). The mean altitude accuracy was 23 m.

The similarity-based weighted average of the nine initial poses was not significantly different from the corresponding unweighted average. This was caused by an inability of the scene similarity network to learn a good similarity measure for the non-urban data.

When the approach was evaluated with UAV and satellite data generated from the same aerial images, as opposed to aerial images acquired several years apart, the mean horizontal position error was 68 m. This shows that only a minor part of the position error could be contributed to differences in scene content.

2.2 Deep feature representations

The second deep learning approach evaluated in the project is the method proposed by [11]. They hypothesized that a deep CNN could learn image representations which are effective for aligning satellite imagery and UAV images containing differences caused by seasonal changes, time-of-day, perspective differences, and the addition or removal of buildings or other structures. They state that their method can generalize from urban environments to challenging low-texture rural datasets.

[11] used a deep convolutional neural network to translate the UAV image and the georeferenced satellite image into deep feature representation images before estimating the image alignment transform using a modified version of the widely used Lucas-Kanade [13] motion estimation algorithm. In contrast to [12], this method is thus not end-to-end, as the camera pose estimation itself is performed without any learning.

2.2.1 Method

The architecture consists of two fully convolutional VGG16 neural networks [14] in parallel, which share the same convolutional weights. They use pre-trained networks and fine-tune only the weights in the 3rd convolutional block, the output of which is used as feature representations. Such mid-layer features are known to have good generalisability characteristics in large image recognition neural networks. The UAV image is fed into one of the networks and the satellite image is fed into the second network. The loss functions are designed to make the networks learn to generate representations of the original images that are as photometrically similar as possible.

They used the Inverse Compositional Lucas-Kanade (ICLK) algorithm [15] for homography estimation with the use of feature representations of the UAV image and the satellite image. The homography encodes the camera pose of the UAV relative to the satellite image. The homography estimation makes a planar assumption, i.e., that the scene that is imaged is flat. This assumption will introduce errors when height variations in the scene are significant in comparison with the viewing distance.

The authors integrated visual odometry by joint optimization of the photogrammetric errors between representations of a set of overlapping UAV images and the photogrammetric errors between representations of those UAV images and a representation of a satellite image. We evaluated their deep features approach to camera pose estimation in isolation, without the integration with visual odometry, in order to compare the results with our evaluation of the method proposed by [12].

For complete details on the method, see [11].

2.2.2 Evaluation

As in our evaluation of [12], we generated training data based on high resolution aerial orthophotos. In addition to the two non-urban training sites shown in Figure 4, we included corresponding training data over Stockholm, Sweden and Woodbridge, New Jersey, USA, shown in Figure 8. The Woodbridge data was provided by the authors of [11].

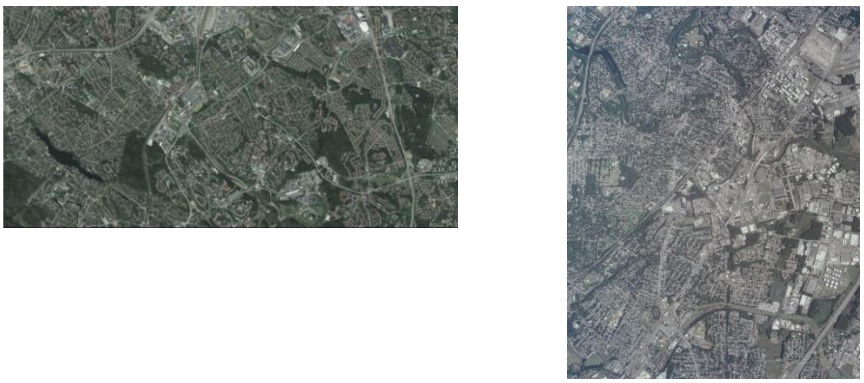

Figure 8 - Urban areas for network training. Left: “Stockholm”. Right: “Woodbridge”.

The roll and pitch angles are assumed to be fixed at 0°. The parameters estimated with this method are the horizontal offset, the scale, and the heading angle of the UAV image relative the satellite image.

To create a pair of training images, a location was randomly selected and patches around the location extracted from orthophotos acquired at two different dates. The patches were subsampled to 1 m pixel size. One of the patches represented the satellite image, while the other represented the UAV image, which was generated by warping the patch with a homography parameterised by randomized horizontal offset, heading angle and scale. The horizontal offset was varied with ±20 m along each of the two axes, the heading angle with ±20°, and the scale with ±25%. These ranges of variation correspond to the range of parameters that the network will be capable of estimating. Therefore, these ranges also correspond to the requirements on the accuracy of the a priori camera pose.

Each training pair was annotated with the parameter settings used to simulate the UAV image. In total 30000 image pairs were used during training.

Our evaluation was performed assuming a UAV at a flight height of 150 m above ground equipped with a downward facing camera with a 60° angular field of view.

2.2.3 Results

The resulting mean horizontal position accuracy for the Grolanda evaluation area was 5.5 m. Simple uniform guessing of the horizontal position would result in a mean accuracy of 18.9 m, relative to which the achieved mean accuracy is a 71% reduction.

When UAV and satellite image were generated from the same orthophoto the mean horizontal accuracy over the Grolanda area improved to 0.9 m. This indicates that a majority of the horizontal position error can be contributed to differences in image content. The mean altitude accuracy was 9.5 m.

3. Exploiting vegetation structure

Vision-based methods relying on colour or texture information in images will fail given sufficiently large differences between the image acquired by the UAV and the reference image. An example of such differences, caused by different vegetation states over a deciduous forest, was shown in Figure 2. Less severe discrepancies caused by differences in illumination, shadows and observation perspective also compromise the reliability of vision-based method. The risk of localisation failure increases with decreasing flight altitude, as the area on the ground imaged by the camera decreases and therefore contains fewer features suitable for image matching. How, then, can the reliability of vision-based methods be improved in such challenging scenarios? The project attempts to answer this question by exploiting the 3D structure of the scene, instead of its colour and texture.

3D models can be generated with stereo triangulation in overlapping images and is routinely generated from images acquired by airplanes and UAVs. Such models could theoretically be generated in real-time onboard the UAV and registered against a georeferenced 3D model stored onboard in order to determine the location and orientation of the UAV.

The project explores the idea that the 3D structure of forests and terrain may be more persistent over time than image texture and intensity, given significant differences in illumination, shadows, vegetation state or observation perspective and to some extent also snow cover. This idea is illustrated in Figure 9, in which the differences between the aerial images (top row) intuitively appear less reconcilable than the corresponding height maps (bottom row).

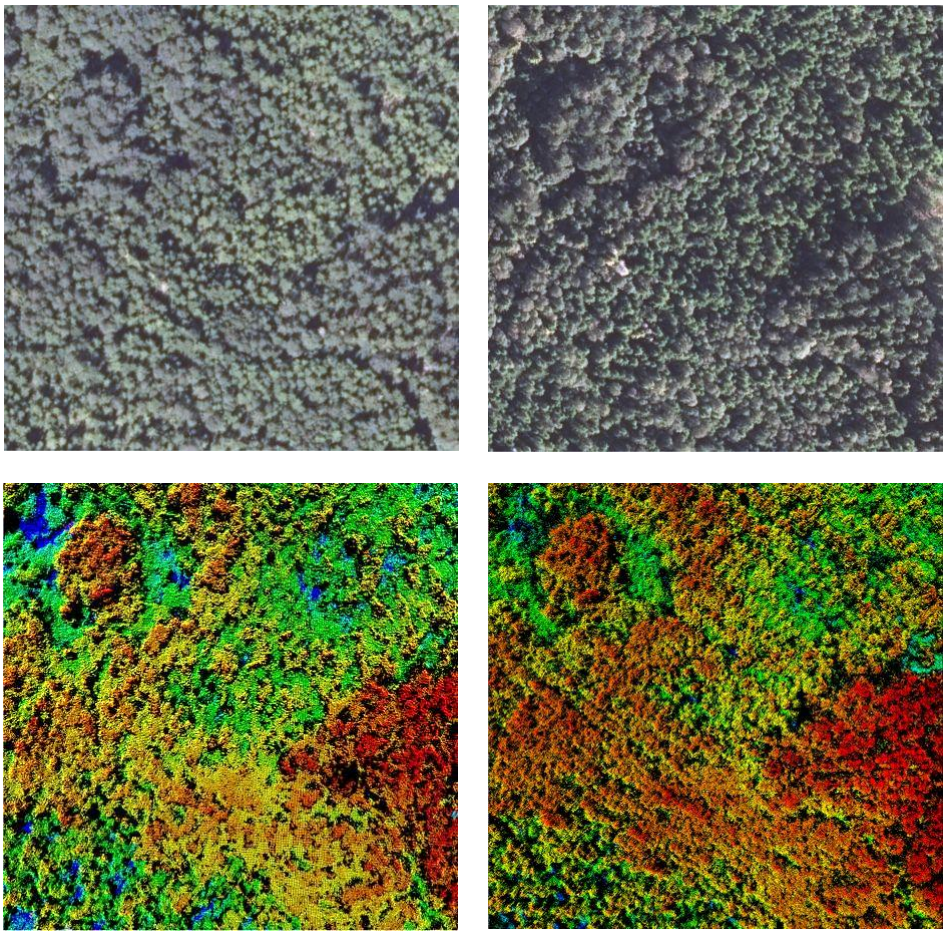

Figure 9 – Comparison between aerial images (top row) and corresponding height maps triangulated from motion stereo (bottom row) for a 135 m by 135 m forested area in the Grolanda test site. Left column: data from Aug. 14th, 2015. Right column: data from Aug. 25th, 2019.

Related to our approach is the collection of methods referred to as Terrain Referenced Navigation (TRN), variants of which has been developed since the 1970s [16][17]. However, to the best of our knowledge, our approach of explicitly relying on vegetation for UAV localisation in non-urban areas has not been studied before.

3.1 Related work

A noteworthy study on TRN is [18], who sought a method robust to changes in vegetation appearance and assumed that the use of height information was more robust to seasonal variations than an image-based approach. They matched dense local height patches derived from stereo motion against a georeferenced 3D model. They evaluated their method on aerial images acquired from a helicopter at 200 m altitude and on imagery acquired from an aeroplane at an altitude of 600 m. Over a residential area, they reported a convergence zone for the a priori pose estimate of approximately 30 m in diameter, outside of which their optimisation method would be trapped in a local minimum, resulting in a large horizontal positioning error.

Another noteworthy study is [19], who matched height patches from LiDAR measurements representing the ground topography against a Digital Elevation Model (DEM). They used the results from the height patch matching to correct drift errors of inertial measurements in an error-state Kalman Filter. For a 218 km manned helicopter flight, [19] achieved successful matches of the height patches in 74.1% of the cases and reported that the localization errors were below 20 m laterally and 5 m vertically for most of the flight.

A major difference between our approach and those of [18] and [19] is that we avoid rasterising the 3D models. Instead, both the 3D model generated on the UAV and the geocoded reference model are matched in their raw form represented as irregular point clouds, with each point holding its respective estimated easting, northing, and height coordinates. Our main reason for avoiding rasterization is that it reduces structural detail.

3.2 Method

Our 3D registration pipeline was implemented with the use of established methods for point cloud processing available in the open-source software Point Cloud Library [20]. The registration pipeline takes as input two points clouds, which we call the reference point cloud (PC_{ref}) and the UAV point cloud (PC_{UAV}). PC_{ref} contains accurate world coordinates for each point. Some a priori knowledge of the UAV's pose and its uncertainty is required so that a search space can be defined in the reference data. The a priori pose may be provided by inertial measurements, through visual odometry or from manual initiation, in case GNSS measurements are not available.

The extent of PC_{ref} is computed based on the extent of PC_{UAV} and the uncertainty in the a priori pose, with larger uncertainty requiring a larger PC_{ref} to be considered to ensure that it contains PC_{UAV}.

The point clouds are processed with general-purpose point cloud registration algorithms, resulting in a rigid body transform that aims at aligning PC_{UAV} with PC_{ref}, effectively geocoding each point in PC_{UAV}. The estimation of the camera pose can then be formulated as the well-known Perspective-n-Point problem using the correspondences between the world coordinates in PC_{UAV} and image pixel coordinates.

Note that the colour information available in the aerial images is discarded, as this information may contribute to false matches – only the 3D structure is used in the processing.

3.3 Evaluation

The point clouds used in our experiments were generated with Spacemetric's software Keystone, based on monocular motion stereo triangulation in aerial images. The aerial images were acquired as part of two different flight campaigns carried out by order of The Swedish Mapping, Cadastral and Land Registration Authority (Lantmäteriet). Specifications of the acquisitions are provided in Table 1. Two images from each date were used to generate the point clouds. The resulting point clouds, which cover a 1 km^2 area in the Grolanda test site, are shown in Figure 10. The aerial images used to generate the point clouds were acquired four years apart. This is a reasonable time span considering the challenges involved in obtaining up-to-date reference data. As can be seen in the height difference image in the rightmost graphic in Figure 10, significant changes took place within this time span, mainly in terms of vegetation growth.

Purpose	Acquisition date	Acquisition local time	Flight altitude	Spatial resolution
Generation of PC_{ref}	Aug. 14th 2015	12:28	3900 m	26 cm
Simulation of PC_{UAV}	Aug. 25th 2019	08:42	3200 m	16 cm

Table 1 – Specification of aerial images that were used for generation of our experimental data.

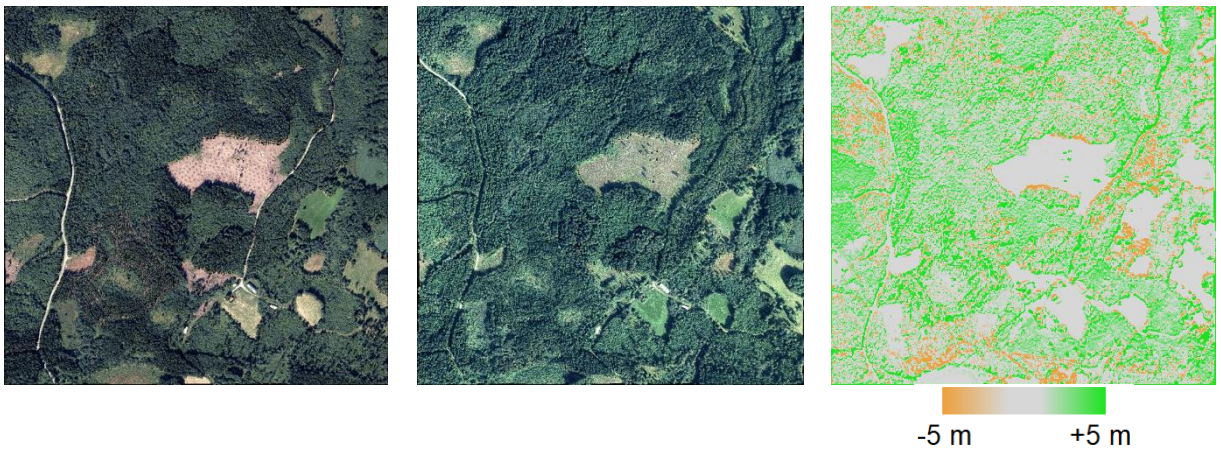

Figure 10 – Left: Point cloud PC_{2015} used for generation of PC_{ref}. Centre: Point cloud (PC_{2019}) used for simulation of PC_{UAV}. Right: Height difference map (PC_{2019} - PC_{2015}).

The assumed configuration in our experiments was a UAV at an altitude of 90 m above ground, with a downward facing camera with an angular field of view of 73.9°. An overlap between successive images of 50% along and across the flight path was assumed, resulting in a point cloud extent of 67 m x 67 m.

Here, we present initial results for two different scenarios for the a priori displacement errors of PC_{UAV}. These displacement scenarios (Table 2) correspond approximately to uncertainties in the a priori knowledge of the position and orientation of the UAV. The extent of PC_{UAV} together with the displacement scenario was used to compute the extent of PC_{ref} required to ensure full overlap with PC_{UAV}.

	Horizontal error	**Heading angle error**	**Pitch and roll angle errors**	**Scale error**	**Corresponding PC_{ref} extent**
Scenario A	60 m	5°	0.5°	2%	180 m x 180 m
Scenario B	10 m	25°	5°	2%	109 m x 109 m

Table 2 – Scenarios for a priori displacement errors in our experiments.

Registration experiments were performed at locations evenly distributed over the test area shown in Figure 10. The UAV point cloud for each particular location ($PC_{UAV,i}$) was simulated by cropping a 67 m by 67 m subset from PC_{2019} centred at the location, and subsequently translating, rotating and scaling that subset according to scenario A or scenario B from Table 2. The corresponding reference point cloud $PC_{ref,i}$ was generated by cropping a section from PC_{2015} with an extent (see Table 2) that ensured that the entire $PC_{UAV,i}$ was contained within $PC_{ref,i}$. Examples of $PC_{UAV,i}$ and $PC_{ref,i}$ for scenario A and B are shown in Figure 11 and Figure 12.

We measure the achieved registration accuracy with the mean target registration error (MTRE) [21], which is the average Euclidian distance in meters between all points in the aligned UAV point cloud and the ground truth. The ground truth in our case corresponds to the original subset in PC_{2019}, from which $PC_{UAV,i}$ was simulated. The MTRE includes the three-dimensional translation error as well as rotation errors.

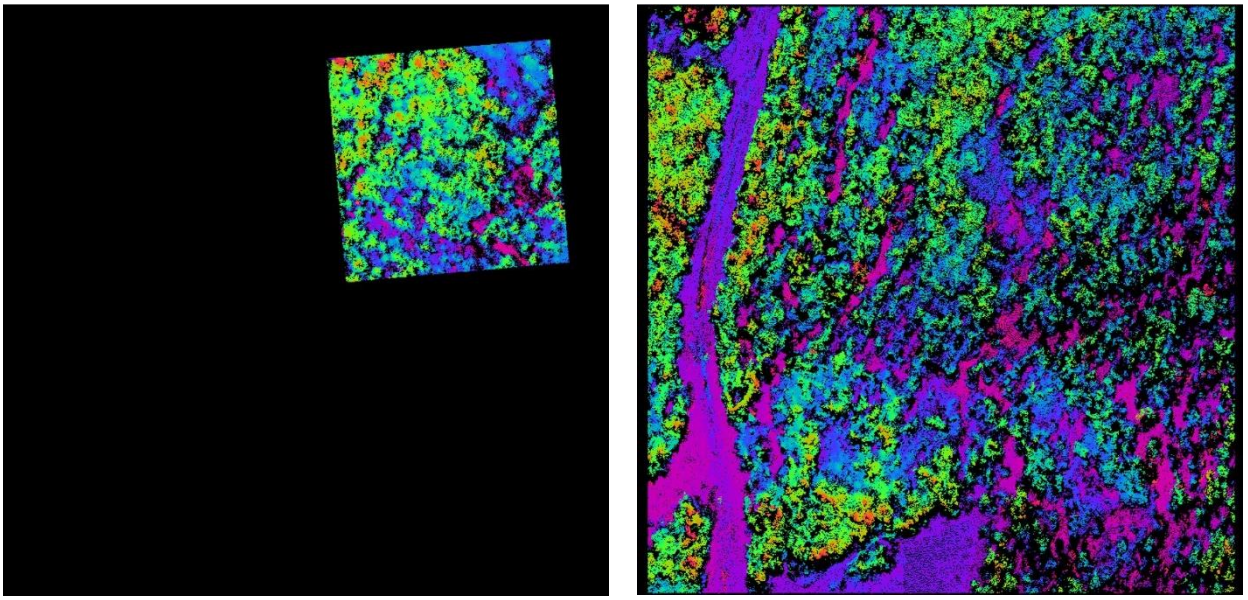

Figure 11 – Example of point clouds for scenario A. Color represents height. Nadir view of a UAV point cloud (left) and the corresponding 180 m wide reference point cloud (right). The registration accuracy (MTRE) achieved in this experiment was 3.3 m.

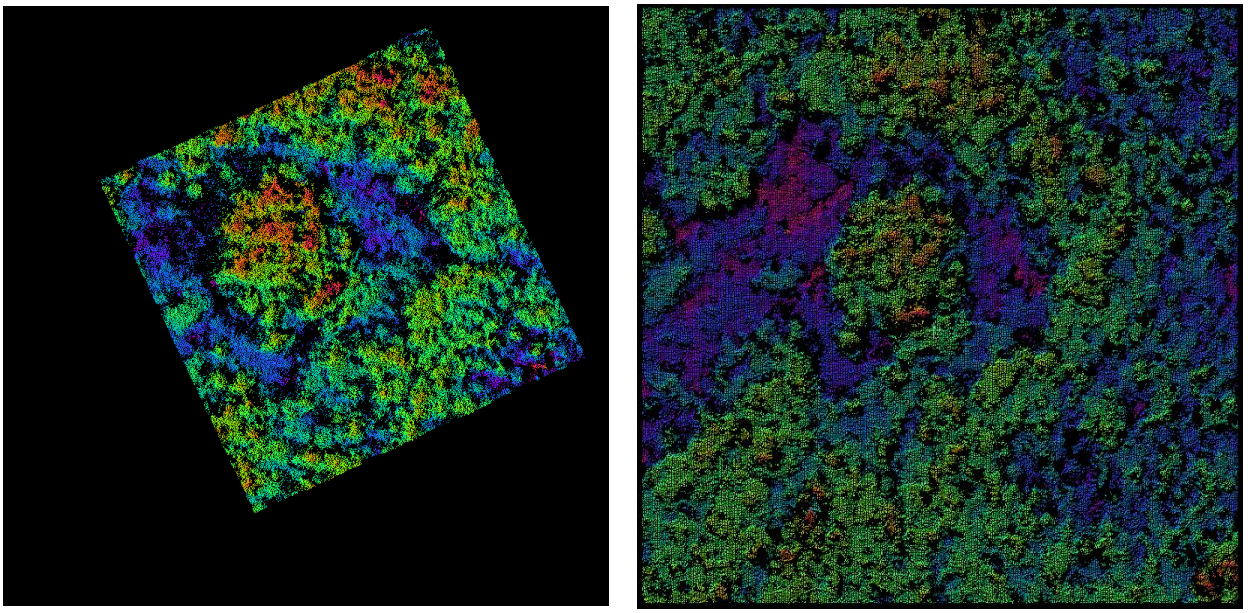

Figure 12 – Example of experiment data for scenario B. Nadir view of a UAV point cloud (left) and the corresponding 109 m wide reference point cloud (right). The registration accuracy (MTRE) achieved in this experiment was 1.8 m.

3.4 Results

We evaluated the registration accuracy rather than the camera pose accuracy in these experiments. However, the roll and pitch angle errors were consistently small, which means that the horizontal registration accuracy closely correspond to the obtainable horizontal localisation accuracy of the camera pose.

The results for scenario A are summarised in Figure 13. The average MTRE over the test site was 13.1 m. As can be seen in the histogram in Figure 13, the distribution of MTRE values is highly skewed with 73% of the experiments achieving an MTRE <5 m. The mean horizontal registration accuracy for scenario A was 8.9 m. This is approximately an 86% reduction compared with a theoretical guesswork algorithm. It can be noted that the median horizontal registration accuracy is significantly lower at only 1.9 m.

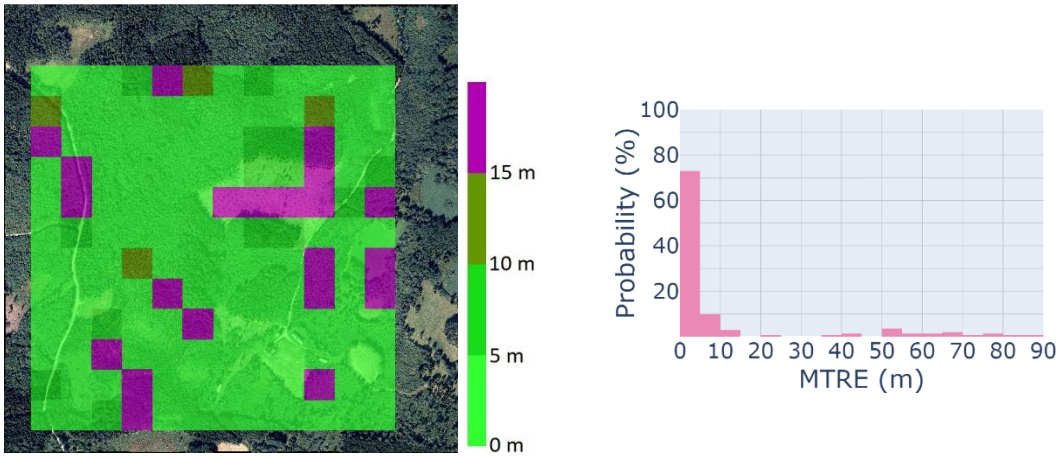

Figure 13 – Point cloud registration errors (MTRE) for scenario A. Left: spatial distribution of registration errors (MTRE) overlayed on the test area. Right: corresponding histogram.

The registration results for scenario B are shown in Figure 14. As in the case with scenario A, the distribution of MTRE values is highly skewed, with a mean MTRE of 8.2 m but with 76% of the experiments achieving an MTRE <5 m. The mean horizontal accuracy is 5.1 m, which is approximately 87% improvement compared to guesswork.

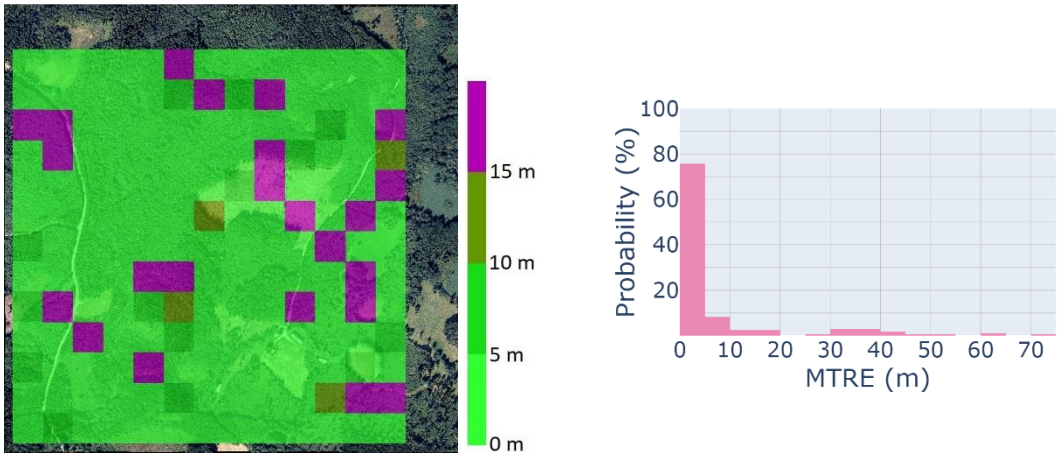

Figure 14 – Point cloud registration errors (MTRE) for scenario B. Left: spatial distribution of registration errors (MTRE) overlayed on the test area. Right: corresponding histogram.

4. Conclusions

The interpretation of the achieved pose estimation accuracies must take into account the uncertainty in the a priori knowledge of the pose. Simply put: what accuracy can be achieved with mere guesswork, given the a priori knowledge of the pose?

The mean horizontal position accuracy achieved with the end-to-end deep learning approach of [12] over the Grolanda evaluation area was 86 m. This is not sufficient accuracy for most applications and is only a 54% improvement over a theoretical guesswork-algorithm. Improvements may be achieved by increasing the number of classes used for the position classification task, or by replacing the classification with a regression approach. The approach performs better in urban areas, for which it was initially developed, than in non-urban areas. For further details on our evaluation of [12], see [22].

The mean horizontal position accuracy achieved with the deep features approach of [11] over the Grolanda evaluation area was 5.5 m. This is an order of magnitude better than our accuracy achieved with [12]. However, considering the higher accuracy of the a priori knowledge of the pose, the 5.5 m accuracy correspond to 71% improvement relative pure guesswork. Thus, while the accuracies of the two approaches differed by more than an order of magnitude in absolute terms, their improvement in relative terms is less dramatic. However, the results motivate further research with deep features representations. For full details on our evaluation of [11], see [23].

The initial evaluation of our novel approach of exploiting vegetation structure showed promising results, with successful localisation (<5 m horizontal accuracy) in more than 73% of the experiments. This success rate is at levels similar to those achieved by [19]. Furthermore, the convergence zone of our approach is at least 120 m in diameter, which can be compared with the 30 m convergence zone for the related method proposed by [18].

5. Outlook

We will further study and develop our approach for exploiting vegetation structure for UAV localisation.

We have carried out image acquisition campaigns with a small commercial UAV over several forested areas in Sweden and are currently cooperating with the Swedish company I-CONIC Vision who will provide rapid generation of highly detailed 3D models from these images.

We consider deep feature representations to be a promising approach, which we also intend to explore further. The project is expected to conclude during 2023.

There is no single vision-based method that is optimal for all possible types of discrepancies between observation data and reference data. Furthermore, there will be situations when most vision-based methods will fail, such as over large water bodies, in fog or above cloud cover. Therefore, vision-based systems should integrate several different types of vision-based methods together with inertial measurements, in order to maximise localisation accuracy and overall reliability.

6. Acknowledgements

This article presents work carried out within the project "Autonomous Navigation Support from Real-Time Visual Mapping". The project receives funding from Sweden’s National Strategic Innovation Programme for Aeronautics (Innovair) through Sweden’s Innovation Agency (Vinnova) (project nr. 2019-02746).

7. Contact Author Email Address

For enquiries regarding the contents of this publication, please contact Daniel Sabel at dosabel@kth.se.

8. Copyright Statement

References

[1] H. Shakhatreh *et al.*, “Unmanned Aerial Vehicles (UAVs): A Survey on Civil Applications and Key Research Challenges,” *IEEE Access*, vol. 7, pp. 48572–48634, 2019, doi: 10.1109/ACCESS.2019.2909530.

[2] A. Claesson *et al.*, “Time to Delivery of an Automated External Defibrillator Using a Drone for Simulated Out-of-Hospital Cardiac Arrests vs Emergency Medical Services,” *J. Ind. Eng. Manag.*, vol. 317, no. 22, pp. 2332–2334, 2017.

[3] R. Adin, “Första gången hjärtstartare från drönare används,” *SVT Nyheter*, Jan. 10, 2022.

[4] A. Roberts, “The blood is here: Zipline’s medical delivery drones are changing the game in Rwanda,” *IEEE Spectrum*, vol. 56, no. 5, pp. 24–31, May 2019.

[5] K. Nonami, “Present state and future prospect of autonomous control technology for industrial drones,” *IEEJ Trans. Electr. Electron. Eng.*, vol. 15, no. 1, pp. 6–11, 2020, doi: 10.1002/tee.23041.

[6] F. Dovis, *GNSS Interference Threats and Countermeasures*. Norwood, MA, United States: Artech House Publishers, 2015.

[7] A. Jafarnia-Jahromi, A. Broumandan, J. Nielsen, and G. Lachapelle, “GPS vulnerability to spoofing threats and a review of antispoofing techniques,” *Int. J. Navig. Obs.*, vol. 2012, 2012, doi: 10.1155/2012/127072.

[8] R. Talmadge, L. Jenkins, D. Hidalgo, J. Olivas, and R. Burk, “Alternatives for Navigating Small Unmanned Air Vehicles without GPS,” *Ind. Syst. Eng. Rev.*, vol. 4, no. 2, pp. 96–113, 2016, doi: 10.37266/iser.2016v4i2.pp96-113.

[9] A. Couturier and M. A. Akhloufi, “A review on absolute visual localization for UAV,” *Rob. Auton. Syst.*, vol. 135, 2021, doi: 10.1016/j.robot.2020.103666.

[10] P. Viola and W. M. Wells III, “Alignment by Maximization of Mutual Information,” *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[11] H. Goforth and S. Lucey, “GPS-denied UAV localization using pre-existing satellite imagery,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2974–2980, doi: 10.1109/ICRA.2019.8793558.

[12] A. Shetty and G. X. Gao, “UAV pose estimation using cross-view geolocalization with satellite imagery,” *Proc. - IEEE Int. Conf. Robot. Autom.*, vol. 2019-May, pp. 1827–1833, 2019, doi: 10.1109/ICRA.2019.8794228.

[13] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *DARPA Image Understanding Workshop, April 1981*, 1981, pp. 121–130, doi: 10.5772/5895.

[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.

[15] S. Baker and I. Matthews, “Lucas-Kanade 20 Years On: An Unifying Framework,” *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.

[16] M. Cowie, N. Wilkinson, and R. Powlesland, “Latest development of the TERPROM Digital Terrain System (DTS),” in *2008 IEEE/ION Position, Location and Navigation Symposium*, 2008, pp. 1219–1229, doi: 10.1109/plans.2008.4570042.

[17] S. Temel and N. Unaldi, “Opportunities and Challenges of Terrain Aided Navigation Systems for Aerial Surveillance by Unmanned Aerial Vehicles,” in *Wide Area Surveillance Real-time Motion Detection Systems*, no. 15, V. K. Asari, Ed. Springer, 2014, pp. 68–69.

[18] B. Grelsson, M. Felsberg, and F. Isaksson, “Efficient 7D aerial pose estimation,” *2013 IEEE Work. Robot Vision, WORV 2013*, pp. 88–95, 2013, doi: 10.1109/WORV.2013.6521919.

[19] G. Hemann, S. Singh, and M. Kaess, “Long-range GPS-denied Aerial Inertial Navigation with LIDAR Localization,” *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, no. May, pp. 1659–1666, 2016, doi: https://doi.org/10.1109/IROS.2016.7759267.

[20] D. Holz, A. E. Ichim, F. Tombari, R. B. Rusu, and S. Behnke, “Registration with the point cloud library: A modular framework for aligning in 3-D,” *IEEE Robot. Autom. Mag.*, vol. 22, no. 4, pp. 110–124, 2015, doi: 10.1109/MRA.2015.2432331.

[21] E. Saiti and T. Theoharis, “An application independent review of multimodal 3D registration

methods," *Comput. Graph.*, vol. 91, pp. 153–178, 2020, doi: 10.1016/j.cag.2020.07.012.

[22] A. Rohlén, "UAV geolocalization in Swedish fields and forests using Deep Learning," KTH Royal Institute of Technology, 2021.

[23] M. Mäkelä, "Deep Feature UAV Localization in Urban Areas and Agricultural Fields and Forests," Royal Institute of Technology, 2021.