2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

On training datasets for machine learning-based visual relative localization of micro-scale UAVs

Viktor Walter*, Matouš Vrba* and Martin Saska*

***Abstract*— By leveraging our relative Micro-scale Unmanned Aerial Vehicle localization sensor *UVDAR*, we generated an automatically annotated dataset MIDGARD, which the community is invited to use for training and testing their machine learning systems for the detection and localization of Micro-scale Unmanned Aerial Vehicles (MAVs) by other MAVs. Furthermore, we provide our system as a mechanism for rapidly generating custom annotated datasets specifically tailored for the needs of a given application. The recent literature is rich in applications of machine learning methods in automation and robotics. One particular subset of these methods is visual object detection and localization, using means such as Convolutional Neural Networks, which nowadays enable objects to be detected and classified with previously inconceivable precision and reliability. Most of these applications, however, rely on a carefully crafted training dataset of annotated camera footage. These must contain the objects of interest in environments similar to those where the detector is expected to operate. Notably, the positions of the objects must be provided in annotations. For non-laboratory settings, the construction of such datasets requires many man-hours of manual annotation, which is especially the case for use onboard Micro-scale Unmanned Aerial Vehicles. In this paper, we are providing for the community a practical alternative to that kind of approach.**

I. Introduction

The growing amount of processing resources sufficiently portable for deployment onboard lightweight MAVs has made it possible to run machine learning-based image processing on these devices in real time. This development is a crucial step towards visual relative localization of unmarked MAVs by other MAVs. This kind of localization is primarily useful for two applications: first, for detecting of cooperating MAVs in a swarm, a formation or otherwise operating friendly units, without the need to equip them with explicit markers or transmitters.

A second, more significant application is for detecting non-cooperating units, where marking them is not possible. These applications include, but are not limited to, reporting or eliminating MAVs intruding into a protected area, avoiding collisions in areas with unrelated active MAVs, or establishing cooperation with foreign units that the observer MAV encounters, if these are open to such a rapport.

For these reasons, markerless detection and relative localization of nearby flying aerial vehicles are topics that have been recently attracting interest of the robotic community

Fig. 1: Examples of annotated footage from the proposed dataset from a disparity of environments. The bounding boxes were generated automatically, using our relative localization system *UVDAR*.

[1]–[8]. Deep-learning based detectors have surpassed traditional detection methods in detection precision and robustness in general detection problems [9]–[11]. However, not many researchers are working on the use of deep-learning methods for the detection and relative localization of MAV. MAVs can be used in large quantities due to their low cost and their greater safeness than large UAVs. A large, varied and labeled dataset is a prerequisite for using any deep-learning based methods to train the classifier or regressor. These datasets are usually meticulously labeled manually, which is an arduous task. In this paper, we will address what we see as a significant reason why deep-learning is not used more often for MAV detection and relative localization, *i.e.*, the lack of suitable datasets and the lack of a simple and automatic way to generate such datasets.

In addition to machine learning (ML)-based vision, relative localization can be retrieved from absolute positions, obtained by a Real-time kinematic (RTK) global navigation satellite system (GNSS) [12] or by Motion capture (mo-cap) systems [13]. We deem this approach entirely unsuitable for field deployment, since both systems require lengthy setup and direct access of the operators to the operational space, which limits the size of this space. Other methods include measuring the relative strength of a radio signal, such as in [14], which however requires multiple observers or specific motion [15] to retrieve the full relative position. In addition, these systems are susceptible to interference.

Another relevant technology is LIDAR, which captures the surrounding surfaces as angularly distributed sample points, represented as a point cloud. These points are obtained by rapidly reorienting laser range sensor or sensors in a rotary manner and combining their measurements with the known current orientation. Using such system for relative localization of MAVs is problematic due to the small size and thin structures of these targets compared to the typical angular density of LIDAR rays, as well as due to the need to distinguish the small clusters representing MAVs

*Faculty of Electrical Engineering, CTU in Prague, Technická 2, Prague 6, {viktor.walter|matous.vrba}@fel.cvut.cz martin.saska@fel.cvut.cz
This research was supported by the by the Grant Agency of the Czech Republic under grant no. 20-10280S

from noise and background. Approaches that are better suited for full onboard operation are based on vision. If a stereoscopic system with a sufficiently wide baseline [1] is available, flying MAVs may be retrieved on the basis of their distance from the surroundings. These systems are unfortunately expensive, large and require a great amount of processing power. A simpler approach is to mark MAVs with easy-to-detect passive visual markers, as in [16], [17]. These are inexpensive and are easily manufactured, but for the typical mutual distances of flying MAVs they need to be impractically large and are susceptible to adverse lighting conditions, particularly bright outdoor sunlight and shadows.

Active markers can also be used, as in [18], or in our own system UltraViolet Direction And Ranging (*UVDAR*) specialized for use in large compact swarms of cooperating MAVs. In this project, *UVDAR* is used for bootstrapping ML systems (see below). The obvious disadvantage of using artificial markers is that these methods are unsuitable for the detection of unmarked units, in addition to the fact that they are additional equipment that needs to be applied. These drawbacks do not apply at all to ML.

We hold that it is imperative to mitigate the main drawback of ML vision - the need for complicated dataset building.

While multiple benchmarking datasets for ML vision are available [19]–[21], these are primarily aimed at evaluating various qualities of a given system, and to the best of our knowledge no dataset has been released with the explicit goal of real-world deployment of MAV that can localize other MAV units. In particular, no system for automatic generation of such a dataset is currently available. An automated approach for this task can significantly expand the usability of ML-based vision, where some source of the ground-truth relative positions of targets has to be used for pointing out their image positions.

With MAVs, it is very difficult to retrieve reliable and precise orientation measurements, that are needed for correct projection of their relative positions into the image space of an on-board camera. The main challenges in this are the combination of the hysteretic properties of magnetometric sensors and their susceptibility to metal in environments, and also the insufficient scale of MAVs for orientation measurements from two absolutely localized body points using GNSS. Since our task primarily requires good relative bearing estimates, other vision-based systems appear to be the best candidate for the source of ground-truth.

Large passive visual markers, however, alter the appearance of MAVs too much to be suitable for training ML systems for detecting unmarked units, as the system would tend to specialize in detection of these markers, which are not expected to appear in final deployment. To address this challenge, we have developed an innovative system that allows for automated generation of annotated datasets for training ML systems that can be deployed for detecting MAVs from other MAVs. The system consists of a color camera attached to an observer MAV, as the source of the images for ML training, and also our specialized sensor for mutual relative localization *UVDAR* (see section II),

Fig. 2: Experimental platform views, with a passive visual marker (left) [22] and active ultraviolet markers, on and off, (center – left). Without the passive marker, the resemblance to arbitrary MAV is higher, making it more suitable for use in training ML algorithms.

an onboard computer and blinking ultraviolet LED markers attached to one or more target MAVs. Thanks to their small size and the fact that they radiate predominantly outside the visible wavelengths, these markers have minimal effect the visual appearance of the target (see Fig. 2, 3). This makes it possible to apply the trained ML system subsequently for detecting similar MAVs that do not carry these markers. These elements allow for easy and fast detection, localization and subsequent annotation of the areas in the camera images where the target MAVs can be found.

A unique feature of this system is that the above is done while circumventing requirements such as communication - liable for interference, blocking, congestion and other issues - or any source of absolute localization, such as RTK GNSS or a mo-cap setup. Our system addresses these limitations by applying relative measurements from one camera-based sensor into another, thus exploiting their known mutual orientation and distance, and also the difference in the wavelength ranges. Additionally, the presented system does not rely on any pre-existing infrastructure (such as a base-station and satellite visibility in the case of RTK GNSS or expensive, pre-calibrated camera setup in the case of mo-cap) in the deployment area, enabling fast and easy creation of labeled datasets in new environments. Additionally, the targets only require small LED markers as opposed to large antennas of RTK systems, making this approach more suitable even for very small MAVs. Here, we provide a large open-ended dataset MAV Identification Dataset Generated Automatically in Real-world Deployment (MIDGARD) for use by the community, to enable various ML systems to be trained and tested. The datasets comprise sets of images from continuous color camera footage of MAVs in a wide range of environments and backgrounds, together with annotations in the form of bounding boxes containing the MAVs in question, as well as their approximate distances.

II. UVDAR

For relative localization of surrounding MAVs used as a ground truth for labeling pictures in ML datasets, we propose to apply our system called *UVDAR*, described in detail in [23]–[25]. The system is based on computer vision in the ultraviolet (UV) range of radiation. This exploits the observation that sunlight is significantly weaker in UV than in the visible spectrum, allowing for easy detection of active UV markers in an effectively arbitrary indoor or outdoor environment by removing most other data from the image with the use of simple optical filtering. *UVDAR* sees active UV LED markers attached to cooperating MAVs as

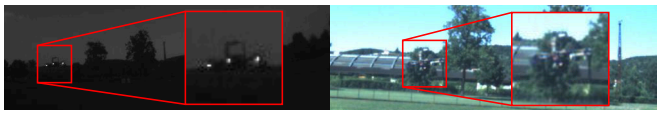


Fig. 3: Example of the view from the UV camera used in our *UVDAR*, system compared with a simultaneous view from the color camera. Note the apparent relative brightness of the markers in UV, while they are invisible in the color image. ML trained systems on such a dataset will be able to detect MAVs without markers.

small bright points. If multiple such points, belonging to the same MAV, are seen, geometrical considerations are used to estimate the distance of the vehicle from the sensor.

Since the markers would otherwise appear identical to each other, we enriched their information content by setting them to blink at a defined frequency, and we used a specialized algorithm to retrieve these frequencies as identifiers, in addition to retrieving the image positions of the markers even in the off phase of the blinking, when they would otherwise be invisible. With such information, we can either distinguish between individual MAVs, as is done in this paper, or we can retrieve their relative orientations by distinguishing different sides of the MAV with different frequencies [25].

The *UVDAR* system provides accurate bearing information on the neighbors, as well as estimates of their distances from the sensor. The sensor has a $180°$ field of view, with bearing errors of *approx.* $0.3°$, and a detection range of *approx.* $15\,\mathrm{m}$, with typical error of 10-20 % of the target distance. Detailed experimental and analytical evaluation of the precision of relative position estimation by *UVDAR* is provided in [23]. Since the specific goal here is to annotate the image of another camera affixed to the body of the observer MAV carrying *UVDAR*, the precision of the bearing information is more significant. This is because a camera is essentially a device that converts the bearings of the points in its surroundings into pixel positions in its output image.

The targets are equipped with the UV LEDs, mounted on their exreme points, *i.e.*, typically the ends of their arms. These markers should have their output power set to account for the maximum expected distance from the observer. With our configuration, we use LEDs with a peak at $395\,\mathrm{nm}$ and a Lambertian radiation pattern. We drive these at $170\,\mathrm{mA}$ to produce $230\,\mathrm{mW}$ of radiated power. Since these markers radiate predominantly in the near-UV wavelength outside the visible spectrum, they have limited influence on the image of a color camera, which typically separates its color channels with miniature color band-pass filters applied to its imaging elements. As stated above, these active markers have to blink with specified signals. There are three reasons for this: 1) The signals serve for identifying a specific marker, which enables multiple targets to be distinguished. 2) It makes the system more robust to specular reflections of the sun *e.g.*, from metallic corners, based on the observation that these do not blink as expected. 3) Specifically for this project, the blinking reduces the influence of the markers on the appearance of the targets. Depending on the exposure rate of the color camera, this is either due to apparently dimming them, or by producing frames where these markers are invisible because they were in the off-phase of the blinking.

III. Dataset generator

The aerial platforms that were used for generating the MIDGARD dataset also serve as an example of the equipment needed for other users to deploy the proposed system. Two types of MAV are involved - targets and observers. Targets are units serving as templates that the ML algorithms will train to detect. Observers are MAVs equipped with our special vision-based suite that generates image streams and annotates the positions of the targets within these images.

The observer units are equipped with two cameras attached to the same holder, the first being the *UVDAR* sensor (see section II) for relative position measurements, while the second is the camera producing the images for ML training. Any properly calibrated camera can be used, ideally of the same type as will be used for ML-based MAV detection.

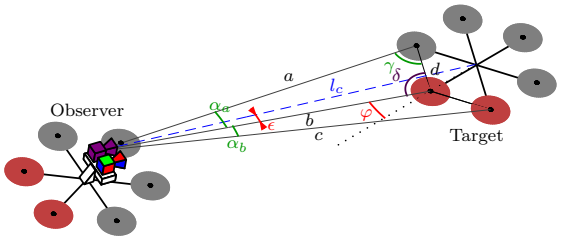
In our datasets, we used *mvBlueFOX MLC200wC* camera sensors with a global shutter, with different lenses for different parts of the dataset. To calibrate them, we used the *OCamCalib* [26] omnidirectional calibration suite, since this toolbox accounts very well for lens distortions near the edges of the images, allowing for correct annotation across the whole image plane of the color camera. The two sensors are attached $8\,\mathrm{cm}$ apart along the horizontal plane, perpendicular to both of their approximately parallel optical axes. Due to such compact installation, their mutual relative poses will only affect the projection of the measured positions into the image by their orientation component.

The processing of the *UVDAR* sensor data and also storage of the color camera stream and the performance of flight-essentials computations, are done with *Intel NUC* onboard computer. A short video demonstration of the system can be found at `mrs.felk.cvut.cz/midgard`.

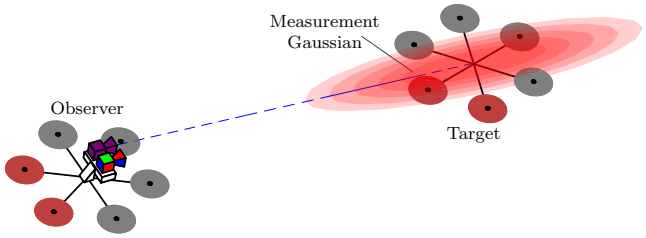
A. Data acquisition

The UV camera is set to capture views at least at $70\,\mathrm{FPS}$, to allow for the retrieval of a blinking signal of up to $30\,\mathrm{Hz}$, below the Nyquist frequency. The raw input images do not need to be stored at this full frame-rate, as performing such rapid storage access operations tends to reduce the actual frame-rate of the detection system significantly. Instead, only detected active maker positions are stored in each frame.

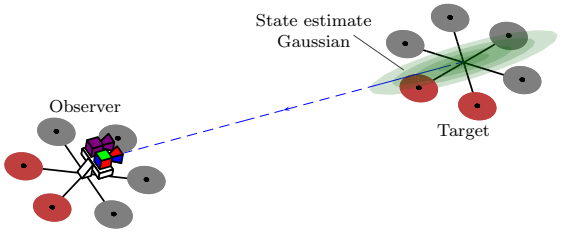
The color camera footage is recorded by default at $3\,\mathrm{Hz}$ to avoid producing numerous frames of high similarity, but it can be increased (*e.g.*, for applications that use inter-frame tracking). The subsequent processing steps are performed onboard of the observer in real time, so that the raw datasets are available immediately after the flight. They can also be done offline after deployment, if the observer does not possess sufficient computational resources, or if the operator wishes to adjust the settings of the consecutive processing steps, the outputs of which are all stored.



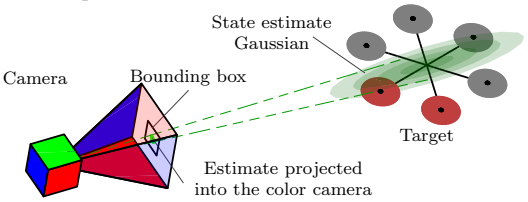
(a) Relative pose estimation from *UVDAR* input (see [25])



(b) Measurement covariance estimated with UT on the input



(c) Filtered pose estimate is transformed into the color camera frame



(d) The pose estimate is projected into the color camera using UT

Fig. 4: Consecutive phases of dataset generation - automatic pose estimation and reprojection

B. Detection and localization

As described in [25], the annotated markers detected on a target MAV can be used to retrieve an estimate of its relative pose. In the proposed system, a new position estimation approach needed to be designed. Since the camera producing the dataset views is not synchronized, and is even potentially delayed w.r.t. the *UVDAR* sensor, it is necessary to use a tracking mechanism that allows for the retrieval of relative pose estimates for the instants when images are produced. A linear Kalman filter, which additionally provides robustness to the target being temporarily lost from view due to occlusion or exiting the field of view of the sensor, is used as a core mechanism in the proposed system. To be input into the Kalman filter, the relative measurements must include covariances, approximating the measurement noise or a measure of the reliability of a given measurement (Fig. 4b). The precision of the *UVDAR* measurement depends primarily on the ratio between the resolution of the UV camera, the field of view of its lens and the distance of the target, due to perspective foreshortening.

The precision of a measurement in the form of a Kalman filter-compliant multivariate Gaussian distribution is difficult to express analytically, due to the highly non-linear dependence between the image positions of the markers and the 3D pose of the object carrying them. Unscented Transform (UT) [27] is used to translate the known precision of detection of the markers in the *UVDAR* image into the approximate covariance of the 3D pose of the MAV itself. The input to the transform is a vector containing the image positions of the markers belonging to a given MAV, mean values of the variables that express the ambiguities in some cases of the detection, and also an error covariance matrix for all of these variables. The values expressing the ambiguities are set to their mean expected values, *e.g.*, the angle by which the target MAV is rotated away from perpendicular alignment if only two markers are seen. The input covariance matrix thus expresses the error in the image positions stemming from pixel size and the image processing involved as well as the approximate ranges in the known ambiguities. For example, the input vectors for observing two and three adjacent markers on a hexarotor, as described in [25], are

$$\mathbf{x_2} = \left[x_1,\ y_1,\ {}^1\!/_{f_1},\ x_2,\ y_2,\ {}^1\!/_{f_2},\ \delta{=}0,\ \alpha{=}0,\ \varphi{=}0 \right]^T \quad (1)$$

$$\mathbf{x_3} = \left[x_1,\ y_1,\ {}^1\!/_{f_1},\ x_2,\ y_2,\ {}^1\!/_{f_2},\ x_3,\ y_3,\ {}^1\!/_{f_3},\ \beta{=}0 \right]^T . \quad (2)$$

Here, x_i,y_i and f_i are the measured image coordinates of each marker i and the measured frequency of its blinking, to account for the possibility of incorrect matching with a known template. Symbol δ refers to the angle by which two observed markers differ from the pose where their connecting line is perpendicular to the line of sight of the observer, while α represents the ambiguity in orientation when observing two markers of the same ID on a MAV with six markers of only two different marker IDs. Element φ represents the unknown amount of tilt that the target has w.r.t. the line of sight of the observer. Symbol β in three-marker observation, where the above ambiguities do not apply, resolves an observation of an ID sequence that does not fit the known marker layout, by introducing a wide additive orientation component in that case. The input measurement covariances entering into the UT for each situation are

$$\mathbf{P_2} = \mathrm{diag}\left(\left[{e_x}^2, {e_x}^2, {e_p}^2, {e_x}^2, {e_x}^2, {e_p}^2, {e_\alpha}^2, {e_\delta}^2, {e_\varphi}^2 \right]^T \right) \quad (3)$$

$$\mathbf{P_3} = \mathrm{diag}\left(\left[{e_x}^2, {e_x}^2, {e_p}^2, {e_x}^2, {e_x}^2, {e_p}^2, {e_x}^2, {e_x}^2, {e_p}^2, {e_\beta}^2 \right]^T \right) \quad (4)$$

where $\mathrm{diag}(\mathbf{x})$ represents a diagonal matrix with elements of $\mathbf{x}$ on the main diagonal, e_x is the mean image position error in pixels, e_p is the mean error of the blinking period measurement, and the rest of the variables represent the mean ranges of the associated ambiguities. We set the values on the basis of geometrical assumptions, and we refined them empirically to the following values: $e_x = 2, e_p = 0.2/f_c, e_\alpha = {}^\pi\!/_{20}, e_\delta = {}^\pi\!/_3, e_\varphi = {}^\pi\!/_{18}, e_\beta = {}^{2\pi}\!/_3$ where f_c is the current UV camera frame-rate.

The input measurement vectors $\mathbf{x_2}$, resp. $\mathbf{x_3}$, together with the appropriate input covariance, are used with the UT to

produce a set of sigma-points, representative of the spread of the values of the vectors. These sigma-points are each converted into a relative target pose estimate, as described in detail in [25], where some are affected by the ambiguity elements being non-zero. These output poses are combined into a single weighted average, and their spread is used to approximate the error covariance of the final pose estimate. For other marker layouts, we progress equivalently. Note that measuring a blinking frequency that is close to two expected values will increase the error covariance, since for certain sigma-points some of the observed markers can be matched with different body markers than others in the 3D pose calculation. The output covariances are invariably strongly elongated in the direction from the detector to the target, showing the characteristic property of visual localization that the distance estimate is significantly less precise than the bearing of the observed object. Since the detection of only two markers in the image contains more ambiguity than with three, the covariances are larger for the former.

If only a single marker is detected (due to an occlusion or of large distance of the target) no distance information is retrieved, except for the known detection range that provides the upper distance limit. The relative position of the marker can therefore be anywhere along its corresponding optical line, up to the maximum detection distance. The markers lie on the extreme points of the target, and the marker that is currently detected may, from the perspective of the observer, lie on the silhouette of the target. The center or the target MAV can therefore reasonably be expected to lie inside a cylinder, the longitudinal axis of which points towards the detected marker, with radius equal to the maximum distance of the markers from the target MAV center. For use in a Kalman filter, this cylindrical subspace is approximated by an elongated Gaussian. While this specific case of measurement is less informative than with multiple markers, it still proves useful with a Kalman filter if a better prior estimate initiated the filter with a distance estimate. In that case the change from the previous bearing can preserve reasonably precise tracking. Without applying this new information the process noise of the filter would expand the state covariance beyond useful size, in addition to the mean value not following the changing pose of the target at all. Furthermore, for the purposes of image annotation, the lacking distance information is admissible, since the *UVDAR* sensor and the color camera - both essentially bearing sensors - are close enough to each other for the reprojection of the covariance.

C. Data post-processing

The estimates of the relative poses of the target MAVs in the frame of the *UVDAR* camera are input into a linear Kalman filter, since the estimates are expressed in Cartesian 3D coordinates (Fig. 4c). Since the relative pose between the *UVDAR* camera and the color camera is fixed and known, the transformation into the camera frame from the external frame will negate the effects of inevitable errors in the absolute observer pose estimate. This is possible because both the transformation from the *UVDAR* frame to an external frame

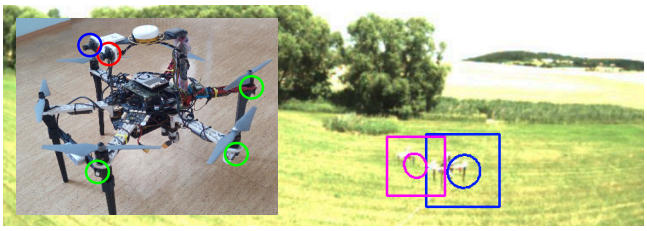


Fig. 5: The MAV platform used in our experiments, here equipped both with a *UVDAR* camera (red) and with a color camera (blue). The markers (green) do not need to be applied to the observer unit, and here they merely demonstrate their layout on the target. In the background, two target MAVs overlap, but are still both detected since some markers of each are seen by *UVDAR*.

and the transformation from an external frame to a color camera frame are burdened by the same error in opposite directions. Correction step of the Kalman filter is applied when a new UVDAR measurement is available. A state of the filter at the time of the measurement is predicted using the latest available filter state, and then the measurement is used to correct it. To obtain a state estimate at the current time, another prediction step is applied to the corrected state. This approach enables avoiding incorrect estimates due to camera delays.

The annotations provided by the proposed system with the camera footage take the form of bounding boxes enclosing the target MAVs in the image, in addition to the estimated range of target distances. The last step of converting the relative pose estimates into these bounding boxes is reprojection, which is done by applying the UT to the current relative position estimates, transformed into the color camera frame (Fig. 4c), in order to obtain a projection of the position estimates, including the covariances, into the image of the color camera (Fig. 4d). This 2D covariance is then converted into a rotated ellipse, by selecting a boundary probability level, *e.g.*, 2σ, but since computer vision ML systems typically [9]–[11] work with rectangular areas, an axis-aligned rectangular bounding box is derived from the ellipse, and is then further expanded.

During creating of MIDGARD, over 85% of the images, containing MAVs in range of the UVDAR system, were automatically labeled with sufficient precision to be included in the final dataset. This outcome is representative of the capabilities of the proposed system, as the values are similar in all recorded scenarios. The pictures where no UAV was detected (most of the incorrectly labelled pictures) were automatically deleted and some remaining outliers were removed using a GUI software that allows the user to discard ranges of annotated views that they consider unsuitable for a specific application.

IV. MIDGARD DATASET

A. Platforms

The experimental platforms used for generating the MIDGARD dataset are based on the *DJI F550* frames, equipped with *Pixhawk* flight controllers and *Intel NUC*

computers. In theory, for our method of dataset generation the computers are needed only for the observer units, while both the targets and the observers can even be piloted manually if no other option presents itself. For MAVs that are highly visually dissimilar to the model above, new footage has to be made. Since MIDGARD is an open-ended dataset, footage of other models will be provided in the future.

B. Environments

The dataset provided was gathered in indoor and outdoor environments. The outdoor parts of the dataset were obtained in various locations, including forest, meadow, fields and urban facilities. These showcase various backgrounds that can be found outdoors, including trees and fields as well as houses and repetitive man-made structures (see table I). The indoor footage includes complicated indoor backgrounds and lighting conditions. The trajectories used for the observer and target MAVs were designed in such a way that the dataset they produce will contain a number of significant challenging situations, in addition to normal views with MAVs against a full variety of backdrops in the given area at a full operational range of distances from the observer. These situations include temporary loss of line-of-sight by a target MAV leaving the field of view of the observer and UAV eclipsing each other in footage with two targets.

C. Examples of main locations

1) Countryside: The footage gathered in the countryside involves observation of two target MAVs, presented against various backdrops - fields, hills, deciduous trees, a distant village and a coniferous forest. Notably, in this footage the two targets eclipse each other in the view (Fig. 5). This is intentional, and was ensured by applying the Model predictive control (MPC) tracker with predefined trajectories.

2) Semi-urban landscape: This footage includes backgrounds of wide buildings, distant hills, vehicles and covered stands. The footage contains one MAV as a target.

3) Classical interiors: Currently contains footages from a vestibule in our departments with ornate stairwells, arched windows and stuccoed ceilings and a historical church under reconstruction, obtained under our DRONUMENT project[1].

These represent complex backgrounds that can be encountered inside and outside of historical buildings. Both footages have one MAV as a target. Notably, due to the low lighting in the vestibule, the exposure levels of the color camera were high enough for the markers to be visible, which was addressed in the church footage by applying low-pass optical filter onto the color camera.

4) Modern interiors: This footage represents a modern, utilitarian architectural background representative of what can often be seen inside and outside modern buildings, with one MAV as the target. The first setting here is a transitory room built inside former courtyard of a building belonging to the faculty of mechanical engineering of the Czech Technical University. The room has glass walls, framed by steel beams, and an uncovered concrete entrance. The second location was an industrial warehouse, where we captured both purely internal footage, as well as footage of transition into the exterior, showcasing the effects of radical change in lighting.

[1]mrs.felk.cvut.cz/research/historical-monuments-documentation

Background	Lighting	FoV	Frames
Fields, hills	Direct sunlight	180°	780
Fields, hills	Direct sunlight	96°	554
Coniferous forest	Direct sunlight	180°	763
Coniferous forest	Direct sunlight	96°	769
Semi-urban	Direct sunlight	96°	475
Stands	Direct sunlight	96°	586
Modern architecture	Strong indirect natural light	96°	534
Historical stairwell	Low light through windows	96°	319
Church interior	Very low mixed light	96°	984
Church exterior	Overcast, late evening sky	96°	697
Warehouse interior	Low fluorescent lightbulbs	96°	564
Warehouse exit	Changes halfway	96°	272
Appartment buildings	Overcast sky	96°	300

TABLE I: Primary characteristics of the current dataset

5) Future additions: Our team is actively involved in projects involving flights inside industrial and historical buildings. This will be leveraged to keep the MIDGARD dataset gradually expanding with footage obtained with the proposed system from flights in these environments.

V. Conclusion

In this paper, we have proposed a new method for fast, automatic generation of datasets for training ML methods for visual relative localization of MAVs by other MAVs. The method uses our specialized system incorporating the *UVDAR* system for relative localization, and makes it possible to develop training datasets on the fly, specifically tailored for the needs of an ML application, by deploying desired models of MAVs into arbitrary operational environments. The software used for the processing is based on Robot Operating System (ROS), and is provided on-line. On demand, we can also provide the hardware of the *UVDAR* system.

As an additional contribution, we provide a large dataset called MIDGARD, which was generated using the proposed method. We believe that this dataset will promote the development of ML approaches for practical field deployment of multi-robotic flight systems, and also of MAV systems intended for interaction with other MAVs. The annotated dataset comprises color camera images with MAVs in various environments, with their positions and bounding boxes provided. The dataset, together with a brief video demonstration of the proposed system, is available at `mrs.felk.cvut.cz/midgard`. We are enthusiastic about the possibility of opening cooperation with peers who can use our proposed system, which can potentially yield footage of other unusual MAVs or normally inaccessible flight locations.

References

[1] M. Vrba, D. Heřt, and M. Saska, "Onboard marker-less detection and localization of non-cooperating drones for their safe interception by an autonomous aerial system," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3402–3409, Oct 2019.

[2] M. Vrba and M. Saska, "Onboard marker-less MAV detection and localization using neural networks," *IEEE Robotics and Automation Letters*, 2020, in review.
[3] J. Li, D. H. Ye *et al.*, "Multi-target detection and tracking from a single camera in Unmanned Aerial Vehicles (UAVs)," in *IROS*, 2016.
[4] A. Rozantsev, V. Lepetit, and P. Fua, "Flying objects detection from a single moving camera," in *CVPR*, 2015.
[5] K. R. Sapkota, S. Roelofsen *et al.*, "Vision-based Unmanned Aerial Vehicle detection and tracking for sense and avoid systems," in *IROS*, 2016.
[6] R. Opromolla, G. Fasano, and D. Accardo, "A vision-based approach to UAV detection and tracking in cooperative applications," *Sensors*, vol. 18, no. 10, 2018.
[7] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "A study on detecting drones using deep convolutional neural networks," in *IEEE AVSS*, 2017.
[8] A. Schumann, L. Sommer *et al.*, "Deep cross-domain flying object classification for robust UAV detection," in *IEEE AVSS*, 2017.
[9] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *CVPR*, 2017.
[10] W. Liu, D. Anguelov *et al.*, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015.
[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017.
[12] Q. Ali, N. Gageik, and S. Montenegro, "A review on distributed control of cooperating mini uavs," *International Journal of Artificial Intelligence & Applications*, vol. 5, pp. 1–13, 07 2014.
[13] D. A. Mercado, R. Castro, and R. Lozano, "Quadrotors flight formation control using a leader-follower approach," in *ECC*, July 2013.
[14] T. Chen, Q. Gao, and M. Guo, "An improved multiple uavs cooperative flight algorithm based on leader follower strategy," in *CCDSC*, 2018.
[15] S. van der Helm, M. Coppola, K. N. McGuire, and G. C. H. E. de Croon, "On-board range-based relative localization for micro air vehicles in indoor leader–follower flight," *Autonomous Robots*, 2019.
[16] M. Saska, T. Báča *et al.*, "System for deployment of groups of unmanned micro aerial vehicles in gps-denied environments using onboard visual relative localization," *Autonomous Robots*, vol. 41, no. 4, pp. 919–944, 2017.
[17] M. Saska, "Mav-swarms: Unmanned aerial vehicles stabilized along a given path using onboard relative localization," in *ICUAS*, 2015.
[18] A. Censi, J. Strubel *et al.*, "Low-latency localization by active led markers tracking using a dynamic vision sensor," in *IROS*, 2013.
[19] A. G. *et al.*, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361.
[20] J. Deng, W. Dong *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
[21] P. Zhu, L. Wen *et al.*, "Vision meets drones: A challenge," *CoRR*, vol. abs/1804.07437, 2018.
[22] T. Krajník, M. Nitsche *et al.*, "A practical multirobot localization system," *Journal of Intelligent & Robotic Systems*, vol. 76, no. 3-4, pp. 539–562, 2014.
[23] V. Walter, M. Saska, and A. Franchi, "Fast mutual relative localization of uavs using ultraviolet led markers," in *ICUAS*, 2018.
[24] V. Walter, N.Staub, M. Saska, and A. Franchi, "Mutual localization of uavs based on blinking ultraviolet markers and 3d time-position hough transform," in *(CASE 2018)*, 2018.
[25] V. Walter, N. Staub, A. Franchi, and M. Saska, "Uvdar system for visual relative localization with application to leader–follower formations of multirotor uavs," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2637–2644, July 2019.
[26] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A flexible technique for accurate omnidirectional camera calibration and structure from motion," in *ICVS*, 2006.
[27] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *IEEE*, vol. 92, no. 3, pp. 401–422, March 2004.