This is a preprint version of the article: Woller, D., Kozák, V., and Kulich, M. (2021). **Robust Visual Teach and Repeat Navigation for Unmanned Aerial Vehicles**. *European Conference on Mobile Robots 2021*,

Robust Visual Teach and Repeat Navigation for Unmanned Aerial Vehicles

Viktor Kozák[1,2], Tomáš Pivoňka[1,2], Pavlos Avgoustinakis[1], Lukáš Majer[1,2], Miroslav Kulich[1], Libor Přeučil[1] and Luis G. Camara[3]

***Abstract*— Vision-based navigation is one of the leading tasks in mobile robotics. It, however, introduces additional challenges in long-term autonomy due to its reliance on stable visual features. As such, visual navigation methods are often sensitive to appearance changes and unreliable in environments with low feature density. We present a teach-and-repeat navigation system for unmanned aerial vehicles (UAVs) equipped with a low-end camera. We use a novel visual place recognition methodology based on high-level CNN features to localize a robot on a previously traversed trajectory and to directly calculate heading corrections for navigation. The developed navigation method is fully vision-based and independent of other sensory information, making it universal and easily transferable. The system has been experimentally verified and evaluated with respect to a state-of-the-art ORB2-TaR navigation system. It showed comparable results in terms of its precision and robustness to environmental changes. In addition, the system was able to safely navigate in environments with low feature density and to reliably solve the wake-up robot problem.**

I. Introduction

The use of Unmanned Aerial Vehicles (UAVs) has gained popularity in various fields. With the increasing availability of (especially quadrotor-based) UAV systems, their use has become highly cost-effective for deployment in various industrial applications in logistics, surveillance, or inspection. However, the widespread use of UAVs in the industry is still limited due to numerous navigation issues. In order to achieve full autonomy, the challenging task of localization and navigation without the use of external systems has to be solved, since many application scenarios require operation in potentially confined and GPS-denied environments.

Due to the characteristics of aerial vehicles, traditional odometry-based navigation methods such as dead reckoning suffer from a high level of unreliability and may result in drifts caused by cumulative measurement errors. Simultaneous Localization And Mapping (SLAM) [1]–[3] is a widely adopted approach to GPS-denied navigation, providing the robot position in a global map of the environment. Such approaches are generally more reliable but may entail an additional payload due to the use of exteroceptive sensors and higher demand on memory and computational cost. Additionally, such approaches often perform loop closure to achieve global map consistency, which is very costly in large-scale scenarios.

Teach and Repeat (T&R) techniques [4], [5] have become vastly popular, as they provide an effective solution for autonomous navigation along previously traversed trajectories. As the name implies, T&R consists of two parts. The desired route is traversed during the *teach* phase, creating the reference trajectory and storing relevant information. During the *repeat* phase, the robot navigates autonomously along the learned path. Visual Teach and Repeat (VT&R) [6]–[9] techniques are especially relevant for the use with Micro Aerial Vehicles (MAVs), which are constrained by their carrying capacity and are often limited to simple monocular cameras.

In this work, we present a highly robust appearance-based visual teach & repeat navigation system for UAVs, based on a novel Semantic and Spatial Matching Visual Place Recognition (SSM-VPR) methodology [10]–[12]. The system proved to be invariant to large viewpoint and superficial appearance changes. It also achieved good results in feature-scarce environments. The presented VT&R methodology is fully vision-based and thus easily transferable to different robots. We extend the system to allow 3D navigation and we demonstrate its performance and potential experimentally on a commercial DJI Ryze Tello drone. We also perform a direct comparison with a VT&R method [8] based on the current state-of-the-art ORB-SLAM2 methodology [13].

The rest of this paper is organized as follows: Section II introduces related work in VT&R. In Section III, we present the methodology behind the proposed VT&R system. Section IV describes the experimental setup and results, whereas Section V is devoted to conclusions and future work.

II. Related work

With recent advances in visual odometry pipelines [14], [15], mapping, and navigation frameworks [13], [16], visual navigation has become a leading task in mobile robotics [17]. However, it introduces additional challenges in long-term autonomy as the environment may exhibit significant appearance changes caused by illumination or weather variations [18], [19].

Most visual navigation methods can be classified into two categories: pose-based approaches and appearance-based approaches [20].

Pose-based (or quantitative) approaches [1], [3], [8], [16] directly reconstruct the position of the vehicle, detected landmarks, and the desired route with respect to a global

[1]Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, Jugoslávských partyzánů 1580/3, 160 00 Praha 6, Czech Republic (e-mail: viktor.kozak@cvut.cz)
[2]Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo náměstí 13, 121 35 Praha 2, Czech Republic
[3]Inria Grenoble Rhône-Alpes, 655 Avenue de l'Europe, 38330 Montbonnot-Saint-Martin, France

coordinate frame, creating a virtual model of the environment. The precise mapping introduces a high computational cost. Furthermore, pose-based approaches require calibrated cameras to compute 3D positions of detected landmarks as well as a sufficient quantity of reliable visual features [6].

Appearance-based (or qualitative) approaches [6], [7] are used to estimate the current segment of operation based on the similarity of the current scene view and previously stored scenes, and directly infer the desired heading corrections. These approaches are generally easier to implement and show considerable benefits in computational costs since there is no necessity for a global map of the environment or exact pose estimation.

Pose-based VT&R approaches demonstrated their ability to repeatedly travel learned paths and even improve on the previously traversed trajectories [21]. SLAM-based methods proposed in [1] and [2] use a downward-looking camera to navigate using visual features in combination with data from inertial sensors. A forward-looking fisheye camera and a low-cost IMU were used in [3] to construct a 3D mesh map and navigate through the environment. A complex VT&R system based on the Tango [22] visual-inertial mapping framework was presented in [16], where the feature-based localization map can be generated both by a teleoperated drone or by a non-expert human operator using a hand-held tablet with a camera. The system also provides a user-friendly interface that allows autonomous UAV operation in collaboration with the operator. These approaches require a sufficient number of robust image features to achieve reliable performance in various and often changing environments. As such, the selection of suitable visual features is crucial and many works focus on this specific detail [18], [19]. For example, in [23] the selection of features suitable for various illumination and weather conditions was performed on multikilometer trajectories during repeated traversals using the Bag-of-Words model [24].

On the other hand, appearance-based methods are often robust to outliers and landmark deficiency situations, delivering satisfactory performance even in feature-scarce environments. An extreme example was presented in [25], where the simplicity of the proposed VT&R method enabled the drone to traverse individual path segments even when sensing only one landmark at a time. Although the number of necessary features varies depending on the exact use, the importance of the selection of suitable and robust features is still a crucial part of many systems. Different methods have been introduced over the years, resulting in VT&R systems based on SURF [6], SIFT [2], ORB [8], and the combination of various other feature detection methods [5].

Recently, neural networks have shown excellent performance in various computer vision applications and several feature extraction methods similar to the ones mentioned above were introduced [26], [27]. However, to our knowledge, they are yet to be used for VT&R. In contrast to conventional handcrafted features, the CNN-based VT&R system presented in [9] uses semantic object detections as high-level image descriptors. The work shows that objects

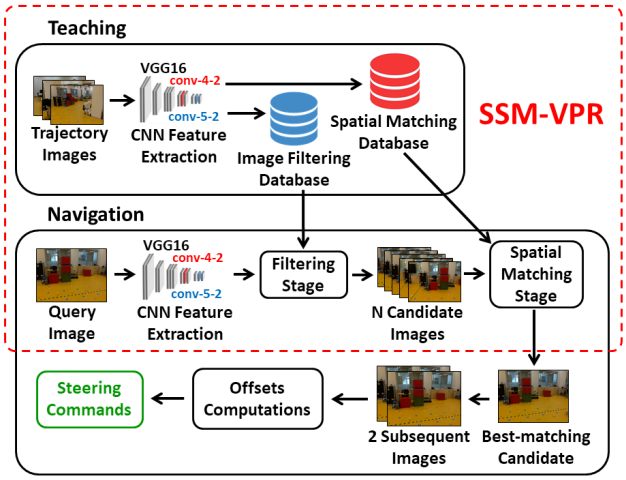


Fig. 1: Flowchart of the visual teach-and-repeat navigation system based-on SSM-VPR.

are detected with high viewpoint and lighting invariance that compares favorably with traditional feature-based methods. The downside of this approach is the dependence on distinct ambient objects and the need to pre-train the detection network for these specific objects.

The system proposed in this work is based on a novel Visual Place Recognition (VPR) methodology presented by us in [12]. The method does not directly work with individual features or predefined object descriptors, but rather groups CNN features from the latest layers of a pre-trained VGG16 CNN [28], generating high-level semantic descriptors. Spatial matching of these descriptors provides a very reliable measure of similarity between images, which is highly robust to variations in viewpoint and appearance. The output can also be directly used to infer desirable heading corrections for appearance-based navigation. In this aspect, the presented approach is similar to our previous VT&R system for Unmanned Ground Vehicles (UGVs) proposed in [7]. The proposed system is fully vision-based without reliance on odometry and modified to provide corrections in the vertical axis.

III. Teach-and-Repeat Methodology

A. System Overview

The structure of the proposed Semantic and Spatial Matching based Visual Teach and Repeat (SSM-VTaR) system is depicted in Fig. 1. During the *teach* phase, the UAV is guided along the required trajectory, and scene images are captured. The images are processed by a CNN, returning visual features that are subsequently stored in a database. During the flight in the repeat phase, visual features are extracted from the current image by the CNN and compared with the reference database created in the *teach* phase. In this fashion, the best match is retrieved and its vertical and horizontal offsets with respect to the current image are computed. Finally, the UAV velocities are adjusted according to the computed offsets.

B. Visual Place Recognition system

The presented navigation system is based on the output of the SSM-VPR system previously introduced by us in [10]–[12]. This method has been successfully employed on the teach-and-repeat task for a UGV with a monocular camera [7] and achieved high accuracy. The SSM-VPR pipeline depicted in Fig. 1 consists of two stages. First, image descriptors are extracted by a pre-trained VGG16 CNN [28]. In the filtering stage, a fast preselection is performed and the method returns a list of N best matching candidate images from the database. In the spatial matching stage, the retrieved candidates are exhaustively compared to the query through a geometrical consistency check. Apart from the similarity of corresponding features, the system evaluates the consistency of their mutual positions in the query and candidate images. The candidate with the highest score is chosen as the best matching image.

The latency of the system in the filtering stage depends on the size of the database, and it directly influences the control rate and the precision of the navigation. In order to reduce the latency in the SSM-VTaR system, we introduce a confidence threshold. If the VPR system recognizes an image with high confidence, it is assumed that the next query image is located in its neighborhood, and the filtering stage is skipped. The candidates for the spatial matching stage are selected directly from the neighborhood of the last best match as its previous and subsequent images in the ratio of 1 to 4.

C. Visual Servoing

The navigation is based exclusively on the horizontal and vertical offsets between the currently observed image and its matched images from the reference database. These are calculated by a modified method used in the spatial matching stage of the VPR system. This step generates a histogram of offsets from displacements of $i \times j$ corresponding feature segments between images, and the bin with the most votes is returned (Figure 2).

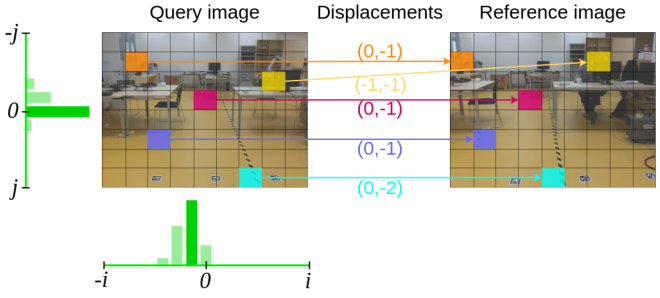


Fig. 2: Offset histogram voting visualization.

Retrieved horizontal and vertical offsets are directly transformed into steering commands. During navigation, the UAV moves with a constant forward velocity v_f, whereas yaw (v_h) and altitude (v_v) velocities are continuously adjusted. The system controls the velocities independently using two proportional regulators with parameters α and β, as described by Eq. 1.

$$\begin{aligned} v_h &= \alpha \cdot v_f \cdot o_h \,, \\ v_v &= \beta \cdot v_f \cdot o_v \end{aligned} \tag{1}$$

Instead of the offset for the retrieved best-matching image i, the horizontal (o_h) and vertical (o_v) offsets are calculated from the averaged offsets of the subsequent two images $i+1$ and $i+2$. This is used to predict the desired immediate movement and to provide adequate action values during sharp turns. The approach also allows us to achieve fully vision-based navigation by removing the dependence on odometry from the previously introduced system [7].

We have implemented a simple recovery procedure for situations when the UAV finds itself in an unknown or significantly changed environment. Such situations are referred to as the kidnapped or wake-up robot problem, where the robot moves (or is moved) to an arbitrary place, which is not necessarily on the taught trajectory. Undesirable movement in such scenarios can lead to failure or collisions. To prevent this situation, if the system recognizes a retrieved image with the recognition score under a specified threshold, the UAV is immobilized. It then starts to rotate and adjusts its altitude on the spot until the system retrieves an image with high confidence again. To guarantee a robust behavior, we use a 5-moving average for the confidence estimate.

IV. Experiments and Results

This chapter describes the experimental setup and presents the results of real-world experiments. A feasibility study on the rotational invariance of the SSM-VPR method is performed in subsection IV-A. Subsection IV-B describes the experimental setup and system optimization. Subsection IV-C presents the accuracy of the developed VT&R system in comparison with a state-of-the-art ORB2-TaR system, and subsection IV-D expands on this comparison by verifying the behavior of both systems in challenging environments. Lastly, subsection IV-E shows experiments on displacement and perturbation robustness of both systems in the wake-up robot scenario.

We use a commercial DJI Ryze Tello drone, equipped with a built-in camera with a resolution of 960x720 pixels. All operations and control of the UAV were done remotely on a personal computer equipped with an Nvidia GeForce 2080Ti GPU, Intel Core i7-7700 CPU @ 3.60 GHz processor, and 64 GB RAM. The navigation of the drone is based on a non-official DJITelloPy [29] Python interface. The experiments were performed under a Vicon motion capture system, which provides ground truth positions with submillimeter precision on a 10x10 meters area.

A. Rotational invariance

Preliminary experiments on the rotational invariance of the SSM-VPR system were performed to verify the feasibility of the proposed navigation system. Due to the characteristics of the spatial matching used in SSM-VPR, we expect a direct relation between the rotation around the camera axis and the image matching capability of the system, which could pose

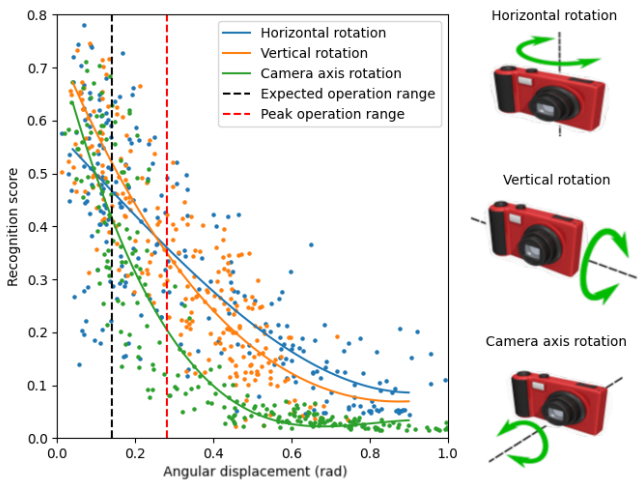


Fig. 3: Recognition score in relation to rotations around individual axes.

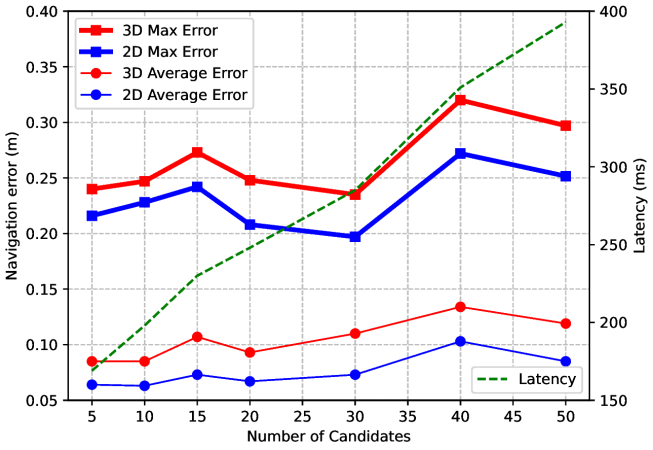


Fig. 4: Navigation error with respect to the number of candidates for the spatial matching stage in SSM-VPR. System latency is depicted with the scale on the right vertical axis.

a problem for aerial vehicles (see [12] for implementation details).

The rotational invariance was tested on a custom dataset that we created using a hand-held camera and the Vicon motion capture system. The reference trajectory was created by traversing a simple path with the camera aimed in the direction of motion. Subsequently, the path was repeated three times, each time adding random rotations around one of the camera axes: the vertical, horizontal, and the camera axis. Figure 3 shows SSM-VPR recognition scores between images from the reference trajectory and images from the three datasets with individual axis deviations.

As expected, the recognition score decreases with rotational deviations, and the decrease is especially sharp in the case of rotation in the camera axis. However, during our experiments, the Tello drone operated within the range of 8 degrees ($[-0.14, 0.14]$ rad) with occasional peaks up to 16 degrees (0.28 rad). Therefore, we consider the robustness of the image matching method as sufficient. Arguably, the sensitivity of the SSM-VPR system to camera axis rotation is one of its drawbacks and could pose problems on more dynamic systems. This deficiency can be compensated, for instance, by combining the system with an IMU.

B. Experiment setup and system optimization

1) Trajectory representation and memory requirements: SSM-VPR uses an image resolution of 224x224 pixels for the filtering stage and 416x416 pixels for the spatial matching stage. During the *teach* phase, the UAV is manually guided along the desired trajectory at a constant speed and images of the scene are captured with a set frequency. We have empirically determined the optimal distance between the captured images to be approximately 0.15 meters, which results in the reference images being captured every 750 ms. The experiments were performed on indoor trajectories ranging from 20 to 65 meters in length, represented by reference databases having 60 to 220 images per trajectory. The memory requirements for these databases vary between 76 MB and 310 MB, respectively. The density of captured images can be reduced in open-space environments. It can also be reduced at specific path segments if we consider using an informed system (i.e., specified by the operator).

2) Number of candidates and system's latency: The exhaustive geometrical comparison performed in the spatial matching stage of SSM-VPR plays a significant role in the computation time. The latency of the system is an important parameter for UAV navigation, as it is required to frequently adjust the motion of the drone. A higher number of candidates used to compare with the current image may increase the retrieval accuracy. However, it introduces additional latency. This might cause the navigation system to underperform.

We have conducted experiments to assess the interplay between the number of candidates for the spatial matching stage and the latency of the system. Figure 4 shows the latency of the system and the divergence from the learned path averaged over 5 laps for 5 to 50 candidates. The robustness of the system increases with the number of candidates. However, rising the number of candidates results in higher computational cost, which negatively affects the latency. The selection of 30 candidates shows optimal performance and results in the system control loop running at approx. 285 ms. This setting was used in the experiments described below.

3) ORB2-TaR - Reference System for Comparison: The system accuracy and properties were compared with ORB2-TaR teach-and-repeat system [8] based on state-of-the-art SLAM system ORB-SLAM2 [13]. The original ORB2-TaR system serves for UGV navigation; therefore, the following modifications were introduced for the UAV application. The modified system contains an additional proportional regulator for altitude control. The altitude difference is computed by subtraction of the currently estimated height and the stored altitude of the closest point on the reference trajectory. Instead of considering the current distance from the trajectory only, the new bearing controller directly combines a 2D lat-

eral displacement with a difference in orientation to compute a correction. The system works at full camera resolution and the control command refresh rate was set to 10Hz.

C. System accuracy (qualitative comparison)

The navigation accuracy was evaluated on three different trajectories. Trajectory **A** was created by guiding the UAV along a circular path with constant altitude, while trajectory **B** was created likewise but with varying altitude. Trajectory **C** was created by guiding the UAV along a more complex path with varying altitudes and sharp turns. Visualizations of individual trajectories are presented in Figure 5.

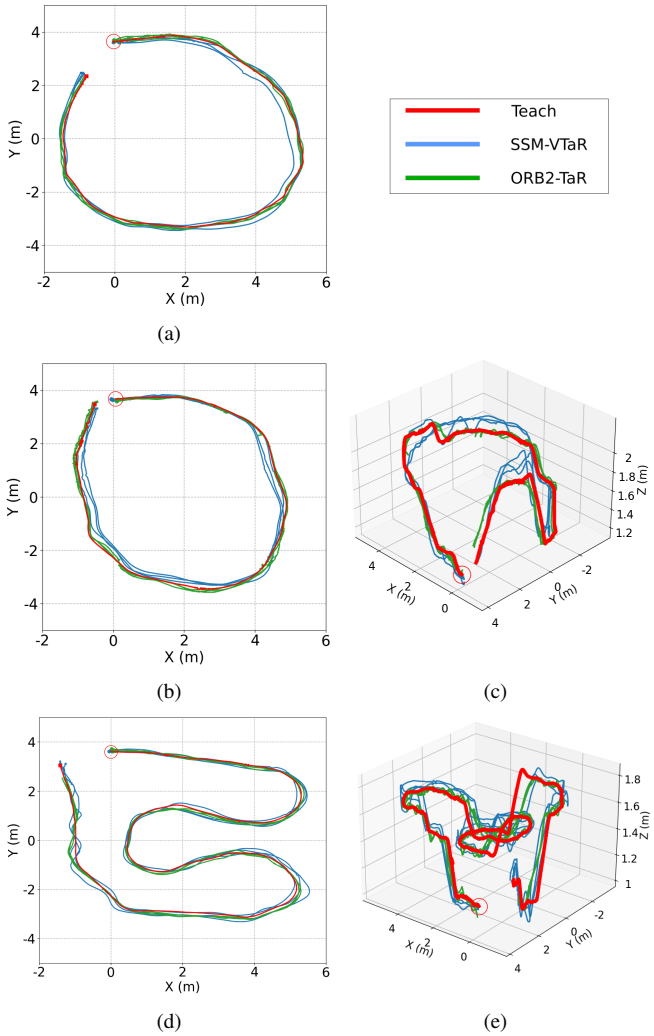


Fig. 5: Results of the SSM-VTaR and ORB2-TaR systems on experimental trajectories. Ground truth positions were taken from the Vicon system. (**a**) Trajectory **A**. (**b**), (**c**) 2D and 3D projections of trajectory **B**. (**d**), (**e**) 2D and 3D projections of trajectory **C**.

The results of experiments on all trajectories are shown in Table I. Three experimental rounds were performed on each trajectory and both systems were able to successfully follow the reference path on all runs. The ORB2-TaR system achieved better results on all trajectories, however, the results are on a comparable level. Since UAVs are highly dynamic systems, we justify the difference by the significantly lower latency of the ORB2-TaR system.

TABLE I: Trajectory errors averaged over 3 experimental rounds for each trajectory.

Trajectory	System	Max. 2D (m)	Mean 2D (m)	Max. 3D (m)	Mean 3D (m)
A	SSM-VTaR	0.238	0.093	0.255	0.115
	ORB2-TaR	0.140	0.051	**0.154**	0.056
B	SSM-VTaR	0.317	0.112	0.339	0.154
	ORB2-TaR	0.153	0.042	**0.251**	0.072
C	SSM-VTaR	0.263	0.082	0.341	0.116
	ORB2-TaR	0.198	0.063	**0.273**	0.082

D. Experiments in challenging environments

Changing and dynamic environments as well as scenes with low feature density pose a difficult challenge to most visual-based navigation methods. In this section, we test the robustness of the developed SSM-VTaR system to such conditions.

1) Changing and dynamic environments: The robustness to changing environments was tested by rearranging the equipment and furniture around the reference path. In addition, the stability of the system in dynamic environments was also tested by introducing several passersby in the experimental area and by moving various objects during the traversal (Figure 6).

(a) (b) (c)

Fig. 6: Example camera views of **a)** the teach phase environment setup, **b)** the changed environment setup, **c)** the dynamic environment setup.

The experiments were performed using the reference trajectory **C** introduced in subsection IV-C and the qualitative results are presented in Table II. Both systems finished the desired trajectory successfully at all attempts in both the changing and the dynamic environment. The ORB2-TaR system showed superior performance in both cases, with approximately half of the maximum trajectory error in comparison with our system.

TABLE II: Position errors in challenging environments.

System	Environment	Max. 3D (m)	Mean 3D (m)
changing	SSM-VTaR	0.680	0.217
	ORB2-TaR	**0.338**	0.094
dynamic	SSM-VTaR	0.606	0.253
	ORB2-TaR	**0.303**	0.102

2) Environments with unreliable features and low feature density: The type and quantity of visual features can be a deciding factor in visual navigation. We performed experiments on a path leading through corridors with repeating structures and matte glass walls, and several doorways with a very low feature density. Since most of the 65 meters long path is out of the scope of the Vicon system, no ground truth positions are provided for these experiments. The path is shown in Figure 7 together with numbered example views from the drone camera. View number 1 represents the starting position with a vast quantity of reliable visual features. Numbers 2 and 5 are example views with a very low feature density. Environments with repetitive structures and unreliable features on glass surfaces are shown in views number 3, 4, and 6.

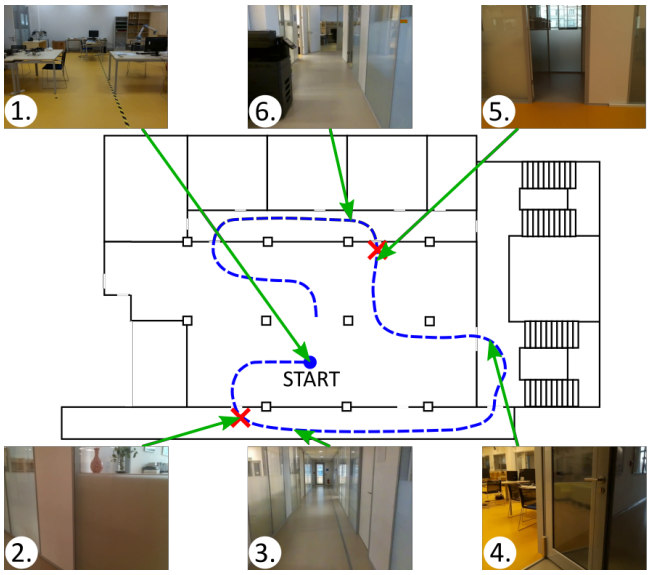
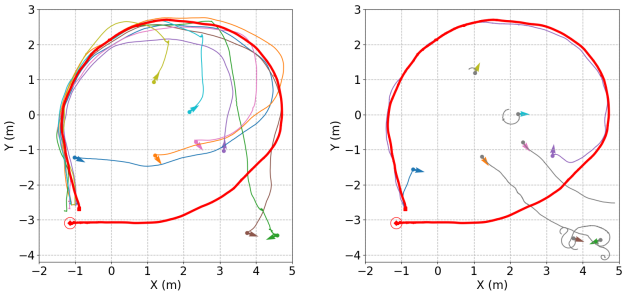


Fig. 7: Reference path with example views from the drone camera. Places where ORB-SLAM2 was unable to reliably map the environment are marked with red crosses.

The reference trajectory was created by manually guiding the UAV through the path, after which SSM-VTaR was able to follow it without a problem. However, due to the lack of suitable visual features in several areas along the path, ORB-SLAM2 was initially unable to create a reliable representation of the environment. The desired flight path had to be adjusted by changing the flight altitude or shifting the view angle to create a valid map. The operation space was reduced in comparison to SSM-VTaR and a significant effort had to be made to deploy the system.

E. Wake-up robot problem

To test the robustness of the developed system, we have simulated the wake-up robot problem. For this purpose, we conducted experiments with different UAV starting locations and headings, as shown in Figure 8a. In most cases, the recognition score was below the allowed navigation threshold (< 0.06) and the recovery procedure had to be initiated. The UAV rotated and adjusted its altitude on the spot until it recognized a segment from the reference database with sufficient confidence. Recognition scores for individual trajectories from Figure 8a are presented in Figure 9. The scores are aligned to the finishing point at $t = 0$ seconds and the confidence threshold used to skip the filtering stage in SSM-VPR (introduced in Section III-B) is also marked in the figure. The UAV managed to successfully return to the reference path and finish the trajectory in each case.



(a) (b)

Fig. 8: UAV trajectories in the wake-up robot scenario. Arrows indicate initial positions of the UAV. The thick red line represents the reference trajectory. **a)** The SSM-VTaR system, **b)** the ORB2-TaR system.

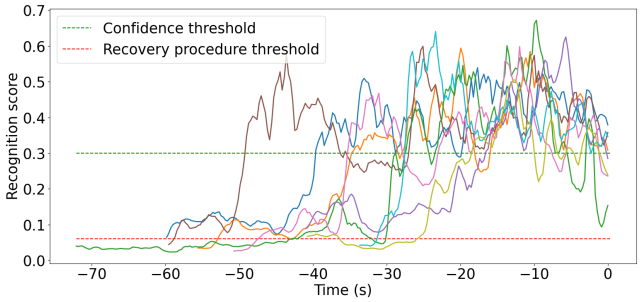


Fig. 9: Recognition scores of SSM-VPR for the wake-up robot experiments (5-moving average).

As can be seen in Figure 8b, the ORB2-TaR system showed much higher sensitivity to the starting position. Although we used the same recovery procedure, the ORB-SLAM2 system was unable to localize itself after take-off in most cases during the wake-up robot scenario. This can be attributed to the fact that most conventional feature-based navigation methods are highly sensitive to viewpoint changes and SLAM-based systems require a larger quantity of stable features. We have identified the system's capabilities in follow-up experiments and ORB-SLAM2 was able to reliably localize itself when the starting position of the drone was within a distance of 0.8 meters from the reference trajectory. However, with increasing distance, the reliability dropped rapidly. The robustness can be increased by mapping the full environment. However, this would once again increase the effort in creating the trajectory model, and it would fundamentally change the teaching procedure.

V. Conclusions

In this paper, we proposed a robust appearance-based visual teach-and-repeat navigation system (SSM-VTaR) for unmanned aerial vehicles. The presented solution is fully vision-based and uses a low-end uncalibrated monocular camera. It improves on our previous work [7] by removing the reliability on odometry and adding altitude control for use with aerial vehicles, making the method suitable for 3D navigation. We have also identified rotational invariance limits of the original SSM-VPR method [12].

The system's performance was compared with a state-of-the-art ORB2-TaR system [8] under a precise external motion capture system. While the SLAM-based ORB2-TaR achieved higher overall precision, SSM-VTaR showed comparable performance and both systems were able to navigate in environments with scene changes and dynamic objects. SSM-VTaR was also able to reliably navigate in environments with low feature density and during perturbations such as viewpoint changes or different initial positions. However, ORB2-TaR performed poorly in such scenarios. This shows the advantages of our system.

SSM-VTaR proved to be a reliable stand-alone VT&R navigation system. It is also fully vision-based, making it platform-independent and easily transferable to different problems. In future work, we intend to test the system in large-scale outdoor environments and its potential use in different applications, such as loop closure or as a separate measure for traversing path segments unsuitable for traditional methods.

Acknowledgment

The work has been supported by the European Regional Development Fund under the project Robotics for Industry 4.0 (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000470) and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS21/185/OHK3/3T/37.

References

[1] J. Surber, L. Teixeira, and M. Chli, "Robust visual-inertial localization with weak gps priors for repetitive uav flights," 05 2017, pp. 6300–6306.

[2] C.-L. Wang, T.-M. Wang, J.-H. Liang, Y.-C. Zhang, and Y. Zhou, "Bearing-only visual SLAM for small unmanned aerial vehicles in gps-denied environments," *International Journal of Automation and Computing*, vol. 10, pp. 387–396, 10 2013.

[3] Y. Lin, F. Gao, T. Qin, W. Gao, T. Liu, W. Wu, Z. Yang, and S. Shen, "Autonomous aerial navigation using monocular visual-inertial fusion: Lin et al." *Journal of Field Robotics*, vol. 35, 07 2017.

[4] M. Nitsche, F. Pessacg, and J. Civera, "Visual-inertial teach repeat for aerial robot navigation," in *2019 European Conference on Mobile Robots (ECMR)*, Sep. 2019, pp. 1–6.

[5] T. Krajník, F. Majer, L. Halodová, and T. Vintr, "Navigation without localisation: reliable teach and repeat based on the convergence theorem," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1657–1664.

[6] T. Nguyen, G. Mann, R. Gosine, and A. Vardy, "Appearance-based visual-teach-and-repeat navigation technique for micro aerial vehicle," *Journal of Intelligent Robotic Systems*, vol. 84, 12 2016.

[7] L. G. Camara, T. Pivoňka, M. Jílek, C. Gäbert, K. Košnar, and L. Přeučil, "Accurate and robust teach and repeat navigation by visual place recognition: A cnn approach," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst*, 2020.

[8] T. Pivoňka and L. Přeučil, "ORB-SLAM2 based teach-and-repeat system," in *Modelling and Simulation for Autonomous Systems*. Springer International Publishing, 2021.

[9] A. G. Toudeshki, F. Shamshirdar, and R. Vaughan, "Robust UAV visual teach and repeat using only sparse semantic object features," in *2018 15th Conference on Computer and Robot Vision (CRV)*, May 2018, pp. 182–189.

[10] L. G. Camara and L. Přeučil, "Spatio-semantic convnet-based visual place recognition," in *2019 European Conference on Mobile Robots*. IEEE, 2019, pp. 1–8.

[11] L. G. Camara, C. Gäbert, and L. Přeučil, "Highly robust visual place recognition through spatial matching of CNN features," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[12] L. G. Camara and L. Přeučil, "Visual place recognition by spatial matching of high-level CNN features," *Robotics and Autonomous Systems*, vol. 133, p. 103625, 2020.

[13] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[14] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. PP, 12 2016.

[15] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," 09 2015, pp. 298–304.

[16] M. Fehr, T. Schneider, and R. Siegwart, "Visual-inertial teach and repeat powered by google tango," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 1–9.

[17] Y. Lu, X. Zhucun, G.-S. Xia, and L. Zhang, "A survey on vision-based uav navigation," *Geo-spatial Information Science*, pp. 1–12, 01 2018.

[18] L. Kunze, N. Hawes, T. Duckett, M. Hanheide, and T. Krajník, "Artificial intelligence for long-term robot autonomy: A survey," *IEEE Robotics and Automation Letters*, pp. 1–1, 07 2018.

[19] T. Krajník, P. De Cristóforis, K. Kusumam, P. Neubert, and T. Duckett, "Image features for visual teach-and-repeat navigation in changing environments," *Robotics and Autonomous Systems*, vol. 88, 11 2016.

[20] A. Vardy, "Using feature scale change for robot localization along a route," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2010, pp. 4830–4835.

[21] F. Gao, L. Wang, B. Zhou, X. Zhou, J. Pan, and S. Shen, "Teach-repeat-replan: A complete and robust system for aggressive flight in complex environments," *IEEE Transactions on Robotics*, pp. 1–20, 05 2020.

[22] E. Marder-Eppstein, "Project tango," in *ACM SIGGRAPH 2016 Real-Time Live!*, 2016, pp. 25–25.

[23] M. Paton, K. MacTavish, M. Warren, and T. D. Barfoot, "Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 1918–1925.

[24] K. MacTavish, M. Paton, and T. D. Barfoot, "Visual triage: A bag-of-words experience selector for long-term visual route following," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 2065–2072.

[25] T. Krajník, M. Nitsche, S. Pedre, L. Přeučil, and M. Mejail, "A simple visual navigation system for an UAV," 03 2012.

[26] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," 06 2018, pp. 337–33 712.

[27] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable CNN for joint description and detection of local features," 06 2019, pp. 8084–8093.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[29] D. F. Escoté and J. Löw, "DJI tello drone python interface." https://github.com/damiafuentes/DJITelloPy, 2018.