Real Estate Advisory Drone (READ): system for autonomous indoor space appraisals, based on Deep Learning and Visual Inertial Odometry

A Quattrini1*, A Mascheroni1*, A Vandone1*, M Coluzzi1*, A Barazzetti2, F Cecconi2 and T Leidi1.*

1Department of Innovative Technologies, University of Applied Sciences of Southern Switzerland SUPSI, Viganello 6962 CH
2QBT sagl, Business Integrator, Chiasso 6830 CH
*Authors contributed equally

Corresponding author: andrea.quattrini@supsi.ch

Keywords: autonomous navigation, mobile platform, deep learning, automated indoor appraisals, visual inertial odometry

List of Acronyms: API (Application Programming Interface), SDK (Software Development Kit), LiDAR (Laser Detection and Ranging), FaaS (Function as a Service), GCP (Google Cloud Platform), SLAM (Simultaneous Localization and Mapping), IMU (Inertial Measurement Unit), CNN (Convolutional Neural Networks), Grad-CAM (Gradient-Weighted Class Activation Mapping)

Abstract. The present paper describes the development of a mobile platform as a support of the real estate appraisal procedure. Currently, the estate evaluation is performed by an expert that manually collects data, performs measurements, and grabs pictures of the inspected unit to finally evaluate its commercial value. The READ project aims at automatizing this process by developing a solution based on a mobile unit (drone or tablet) able to navigate the indoor environment and record data, which will be later processed on the cloud. To accomplish all these tasks, the platform is equipped with cameras, a LiDAR sensor, and a data process unit, with the goal of 1) understanding its motion and localization; 2) reconstructing a 3D map of the inspected space; 3) performing image-based analyses applying AI algorithms enabling the identification of the indoor space (e.g. bedroom or kitchen), the counting and the classification of furniture objects, and the detection of building imperfections or frauds. Tests have been performed in different scenarios providing promising results, laying the foundations for bringing these technologies into a real operational context.

1. Introduction
1.1. Background
Real estate sales are becoming increasingly a business reason for individuals, companies and especially banks that almost always intervene to provide mortgages to help buyers. In Italy, for instance, up to 440'000 property evaluations are carried out annually. Each appraisal currently costs around 300 euros to banks or private clients – for both the Italian and the European market - and it consists in a manual collection, cataloguing and evaluation of information (photographs, measurements, cadastral maps) made by the appraiser on site. The expert assesses the properties based on traditional parameters such as square footage, number of rooms, the location, and the general condition of the property. These
parameters are usually retrieved by the expert who cannot rely on cadastral values that are often dated, approximated, and not updated to the following renovations. Technology has placed more and more measuring and reporting instruments in the hands of the appraisers even if they must be handled with skill. In fact, the accuracy and the time taken for the surveys still depend on the competence of the expert. Banks, instead, require reliable, repeatable, and certified procedures for the estate evaluation so that the financial transaction could be safe and unchallengeable. READ project aims at changing radically the way a real estate appraisal inspection is performed by developing a mobile platform (e.g., a drone) to autonomously navigates the unit collecting and processing the required information in an automatic way (Figure 1).

Figure 1. READ automatic appraisal workflow process

1.2. State of the art for drone indoor inspection

Nowadays, the idea of using drones to perform real estate appraisals is new and there are no products already on the market ready for this scope. Different solutions for building and plant inspection are already available on the market, but most of them have been designed to work in large outdoor environments or they are able to navigate only large and empty spaces due to their bulky dimensions. Various solutions have been proposed in the literature in the past years for micro to nano drones [1-3] but unfortunately, the results remain limited to lab experiments. Nowadays, small drones are on the market that already support predefined autopilot flying modes (e.g., Points of Interest, Waypoint Navigation, Cinematic Mode, Follow Me), and are already equipped with different proximity sensors and cameras. Some of them offer APIs or SDKs to interact with the sensing and the navigation control systems. Therefore, the idea is to exploit these available devices and boost their capability, leveraging these software tools to implement an autonomous house navigation and collection of data. It is worth mentioning that in this paper the focus has been addressed to the activities that are related mainly to the integration of sensors and implementation of the algorithms for mapping, room and object classification and cloud communication. Therefore sometimes, for simplicity, the mobile platform shown in these tests is composed of a tablet properly sensorized moved around by an operator. However, the drone platform is equipped with the very same sensor unit, carrying out the same data processing, while performing autonomously the environment navigation task.

This paper is divided into four paragraphs, including the ‘Introduction.’ Section 2 describes the methodology proposed by the authors to accomplish all the appraisal tasks. Section 3 shows tests and the results obtained so far, and finally, Section 4 reports the conclusions and planned future works.

2. Methods

The proposed inspection solution should be able to accomplish numbers of different tasks:

- indoor environment navigation,
- reconstruction of a 3D map (the final output is a 3D mesh or a 3D point cloud),
- detection of frauds, defects, and system elements (e.g., heaters, plugs, switches),
- identification of the type of room (e.g., bedroom, kitchen) and objects located in the environment (e.g., chairs, tables, sofas),
- interaction with external software tools for unit evaluation and report generation.
To reach all the above-mentioned objectives, the READ mobile platform has been equipped with several sensors and a proper software infrastructure has been developed to assure the data collection, storage and elaboration as described in Section 2.1.

2.1. System architecture

The architecture was designed with a state of art cloud-native and fully serverless approach. Using a serverless architecture allows the system to be scaled in a fully automated way. The overall architecture is divided into two main modules:

- **Acquisition module**
  - This module is embedded into the device (i.e. drone or tablet) on board the device and consists of all the sensors (specifically a LiDAR and a tracking camera) that are used to collect the necessary data. The reconstruction of the 3D map is performed onboard because the reconstructed environment is used for obstacle avoidance in autonomous navigation.

- **Deep Learning module**
  - This module is hosted in the cloud. Exploiting Convolutional Neural Network algorithms, the content collected by the device is analysed to identify indoor spaces and objects.

Communication between the various elements involved in the system is based on an event-driven architecture. The cloud vendor provides the required infrastructure when the related action is triggered. The prototype was deployed on Google Cloud Platform (GCP) but, it can be moved to any other cloud vendor with minimal effort.

![Figure 2. Serverless architecture.](image)

Figure 2 shows the architecture of the first part of the processing pipeline involving room classification. Once uploaded, the content is stored in scalable and resilient storage. The API is implemented using FaaS components. An API Gateway service is used to ensure authentication and authorization features. All processing requests are enqueued and processed asynchronously to avoid system overloading. The deep learning processes run into containerized environments executed by serverless compute engines. The output of the processing is stored in the cloud and can be accessed at any time via API.

3. Tests and results

3.1. Indoor navigation and fraud detection tasks

3.1.1. SLAM

In computational geometry and robotics, SLAM (Simultaneous Localization and Mapping) is the problem of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it. The localization subtask consists in estimating the trajectory of the
mobile platform usually by tracking visual features extracted from images coming from pre-calibrated onboard camera systems. When visual features are not available (e.g. walls without textures) an Inertial Measurement Unit (IMU) can be exploited [4]. The mapping subtask, on the other hand, considers this estimated trajectory to merge local point clouds in a common reference system to generate a global environmental map. Optimization and refinement steps usually follow.

In the present work, the mobile platform is equipped with a stereoscopic system Realsense T265 comprising two fisheye cameras and an integrated IMU. This device is responsible for the localization task. An Intel Realsense L515 is used as a Lidar, for the final mapping of the environment.

Tests have been performed in SUPSI labs to evaluate the accuracy of the trajectory estimation. The trajectory provided by the T265 has been compared with the one computed using an Optitrack Mocap (motion capture system, Figure 3 left) composed of 6 cameras visualizing a room of approximately 20 square meters and tracking the position of reflective markers - fixed onto the READ platform (Figure 3 right) - with sub-millimeter accuracy, thus considered as ground truth.

![Figure 3. On the left, SUPSI lab equipped with a Mocap system, on the right the mobile platform equipped with the RealSense L515 and Realsense T265.](image)

As both trajectories can be specified in arbitrary coordinate frames, they first need to be aligned. This can be achieved, for example, using the method of Horn [5] which finds the rigid-body transformation $S$ corresponding to the least-squares solution that maps the estimated trajectory $P_{1:n}$ onto the ground truth trajectory $Q_{1:n}$. Given this transformation, the absolute trajectory error (ATE) at time step $i$ can be computed as:

$$F_i = Q_i^{-1} \ast S \ast P_i$$

As evaluation metrics the root mean squared error over all time indices of the translational components has been considered as defined in [6]:

$$RMSE(F_{1:n}) = \left( \frac{1}{n} \sum_{i=1}^{n} ||trans(F_i)||^2 \right)^{\frac{1}{2}}$$

Various paths have been tested for example motion on a plane or 3D closed loop path (Figure 4). Navigation has been performed smoothly at a constant speed of approx. 0.5 m/s. Pose estimations have been asked at 30 Hz rate.

In all the cases, the deviation of the positions estimated with the T265 to the ground truth are in the range of 0.05 m. This value is in accordance with the results found in literature [10] and can be considered acceptable for the project application. Moreover, it is worth to notice that this position estimation represents the first step in merging partial map acquisitions, but further optimization and map refinement are performed based on the geometries reconstructed (e.g. room corners or pieces of planar wall surfaces must match).
3.1.2. 3D reconstruction and mapping
The mapping task is addressed exploiting the LiDAR. The READ platform is equipped with a RealSense L515 Time of Flight (TOF) sensor that performs 3D measurements of the surroundings. The output is the so-called point cloud, thus a dataset containing the spatial position of thousands of measured points. The device includes also an RGB and an IR camera used to grab images provided as input for the indoor space recognition task described in Section 3.2.

Figure 4. Deviation of the trajectory estimated by the READ platform (coloured line) in comparison to the one provided by the Mocap system (black line).

Grabbing several clouds and knowing the relative motion between them thanks to the T265 device, it is possible to reconstruct the global map of the environment. Tests have been performed in several scenarios as shown in Figure 5.

Figure 5. 3D Indoor reconstructed environment: bedroom on the left acquisition noise on the right.

Performances of the L515 sensor has been evaluated: the acquisition noise estimated fitting the measured points representing a planar wall with a plane and computing the mean deviation of each point to this reference geometry turned out to be approx. 1.5 cm (see Figure 5 right). Once again, this is an acceptable measuring error in the real estate appraisal context.

3.2. Indoor space recognition task

3.2.1. Deep Learning approach
A Deep Learning approach based on Convolutional Neural Network (CNN) has been applied to tackle the indoor recognition problem. 91154 human-shot and labelled images were included in the analysis. The images are split in 5 unbalanced classes: bathroom, big bedroom, small bedroom, kitchen, living room (Figure 6). A state-of-the-art CNN model (DenseNet121 [7]) has been exploited using the so-
called fine-tuning transfer learning strategy. The model was trained for a total of 60 epochs as no major improvement in the accuracy and f1 score metrics was observed afterwards, and the images were divided in batches of 128.

![Figure 6. Samples from dataset.](image)

The Adam optimiser was used to minimise the categorical cross-entropy loss function. The learning rate was set to $3 \times 10^{-4}$ for the first 15 epochs of features extraction, then it was divided by 10 to reduce the risk of over-fitting when unfreezing more layers of the network (fine tuning approach). Also, to overcome the unbalance nature of the dataset, different weights have been considered for different classes in the computation of the loss function, thus less-sampled classes weighted more than over-sampled ones. Regularisation techniques were employed to reduce the over-fitting problem and therefore increase the generalisation capability of the model. Particularly, a dropout of 0.3 was applied on the train set between the last global average layer and the classification layer, and the following image augmentation techniques were applied on the input images during training: rotation, zoom, translation, flip, brightness.

3.2.2. **Real-time prediction: switch of context problem**

As mentioned above, the train dataset used to train the CNN model included human-taken pictures. These images were shot with consciousness such that all the objects of the room could fit in a single global picture. However, the goal of the READ system is to equip this technology on a mobile platform that continuously grabs pictures. This means that, in a real-time scenario, each picture will not cover the full area of an indoor space but only a portion, and some of them could be totally meaningless for the deep learning model for the purpose of classifying the room - e.g. ceilings, part of walls, etc.- (Figure 7).

![Figure 7. Example of meaningless images](image)

This issue led to an unexpected behaviour of the model during the real-time experiment phase, resulting in wrong predictions on the meaningless input images. Initially, to investigate the behaviour of the model and to exclude possible mode-related issues, a Grad-CAM [8] technique has been applied to highlight the pixels (by means of an heatmap) of an image that mainly influence the model decision when classifying (Figure 8). The Grad-CAM results led to the conclusion that the problem was not laying in the model itself, but rather it was more related to the images given as input. To tackle the switch-of-context problem, the Monte Carlo Dropout approach was applied. This method consists in activating the dropout also during the prediction phase (normally the dropout is applied only during the training) and performing multiple predictions over the same input image to classify. At each iteration, different neurons are “turned off” randomly leading to 2 possible scenarios: if the image is totally uncorrelated with the training dataset, the model assign to it a different class at each run demonstrating a high prediction uncertainty.
On the left image, the bed activated the decision of the model: the model is looking for the right feature; on the right, a meaning-less image, where the model focuses on a random area.

On the contrary, if the image presents known patterns, the prediction variability is low, suggesting a high classification confidence. Thanks to this method, it is therefore possible to recognise those meaningless images that could show up when using the model in real-time scenarios, avoiding misleading predictions of the model.

### 3.3. Indoor Space Recognition results

The classifier results have been evaluated in terms of average f1 score. This parameter takes into consideration the imbalanced nature of the dataset, weighting each class based on the number of observations they have and is computed as in the formula below:

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

Table 1 shows precision, recall and f1-score for each class for the dataset analysed. The average f1 score across all classes is 0.73.

![Table 1. Metrics for each class compute at epoch 60.](image)

These results can be considered acceptable for our purposes when compared with a similar study found in literature by Othman [9]. This author’s classifier reaches an accuracy level of 88%, slightly higher than the proposed one, but different aspects must be taken into consideration. First, no distinction between bedroom and small bedroom classes are reported while it is clearly noticeable from the confusion matrix shown in Figure 9 that most of the mistakes made by the proposed classifier are between these two classes due to their natural ambiguity. Merging them together would increase the classifier performance by around 11%, reaching an f1 value of 0.84. Moreover, differently from them, our input dataset was not cleaned from some miss-leading images (e.g. empty rooms). As reported in their results, when running their model on the full dataset (not cleaned) they loose around 5% of accuracy. If we hypothesize the same improvement for our dataset, and the improvement gained from merging the two bedrooms classes, the final performance would be comparable. In addition, compared with [9], the Monte Carlo strategy was applied to overcome the issues that could come up during real-
time predictions, enabling the model to discard the meaningless images, a limitation of the previous study.

4. Conclusions

In this work we presented a solution for real estate appraisal based on a mobile platform able to collect and evaluate data of the environment inspected in an automatic way. SLAM algorithms run in the acquisition module designed for READ and equipped with 2 cameras and an IMU. An error of about 0.05 m in the computation of the mobile agent motion navigating the indoor environment has been estimated. This result was obtained through the comparison of the READ trajectory with a ground truth one obtained by a sub-millimetre accurate Mocap system. The trajectory is then refined in a second level of the process, during the creation of the global map, which has been evaluated in terms of reconstruction accuracy at both qualitative and quantitative levels. The global map has been retrieved using the 3D measurements provided by a LiDAR device comprised into the READ mobile solution and an average acquisition noise of 1.5 cm has been estimated. This value is acceptable for the use of these techniques in a real estate context.

The indoor space recognition task can correctly identify indoor environments with an average f1-score of 0.73, representing performances comparable to the state of the art. Correct space predictions are obtained for bathroom and kitchen classes; incorrect predictions might occur while classifying single bedroom and bedroom images due to their intrinsic ambiguity. Thanks to the application of the Grad-CAM and Monte Carlo dropout strategies, we managed to tackle and overcome the switch-of-context problem during real-time experiments, being able to identify meaningless images fed into the model.

As future development, to increase even more the READ solution performances we will consider excluding from the train dataset images of empty rooms, that were misleading for the model.

Acknowledgments

The research in this paper has been partially funded by Innosuisse Application #: 46479.1 IP-ICT.

References

[1] B. B. Moshe, N. Shvalb, J. Baadani, I. Nagar, and H. Levy, "Indoor positioning and navigation for micro UAV drones — Work in progress," in 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, 2012, pp. 1-5.

[2] L. Dowling et al., Accurate indoor mapping using an autonomous unmanned aerial vehicle (UAV), 2018.

[3] A. A. Ravankar, A. Ravankar, Y. Kobayashi, and T. Emaru, "Autonomous Mapping and Exploration with Unmanned Aerial Vehicles Using Low Cost Sensors," Proceedings, vol. 4, no. 1, 2019.

[4] G. J. I. C. o. R. Huang and Automation, "Visual-Inertial Navigation: A Concise Review," pp. 9572-9582, 2019.

[5] B. K. P. Horn, H. M. Hilden, and S. Negahdaripour, "Closed-form solution of absolute orientation using orthonormal matrices," Journal of the Optical Society of America A, vol. 5, no. 7, pp. 1127-1135, 1988/07/01 1988.

[6] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 573-580.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," presented at the CVPR, 2017.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626.

[9] K. Othman and A. Rad, "An Indoor Room Classification System for Social Robots via Integration of CNN and ECOC," Applied Sciences, vol. 9, p. 470, 01/30 2019.