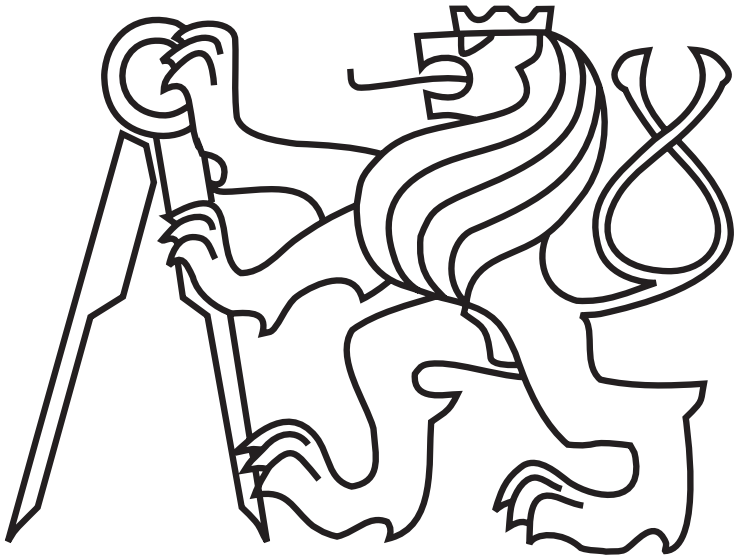
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Electrical Engineering

MASTER THESIS



Jan Blaha

Unsupervised Learning of Semantic Landmarks for Visual Navigation over Extended Periods of Time

Department of Computer Science

Thesis supervisor: **George Broughton, MSc.**

May, 2022

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Blaha Jan** Personal ID number: **478158**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Open Informatics**

Specialisation: **Data Science**

II. Master’s thesis details

Master’s thesis title in English:

Unsupervised learning of semantic landmarks for visual navigation over extended periods of time

Master’s thesis title in Czech:

Učení sémantických orientačních bodů bez učitele pro dlouhodobou vizuální navigaci robotů

Guidelines:

1) Read related work on visual teach-and-repeat navigation [1] and the proof-of-concept paper of an unsupervised method for learning landmarks for visual navigation [2].
2) Implement the methods of [2], so that they can be properly tested in visual navigation using real-world robotic datasets.
3) Propose and implement alternative methods for producing annotations, for the training of landmark detecting ANNs.
4) Propose a method for evaluating the performance of the designed methods in the task of long-term visual navigation.
5) Perform a comparison of proposed and current methods using real-world robotic datasets.
6) Discuss and assess the advantages of the deployment of such methods in autonomous robotic operations.

Bibliography / sources:

[1] Krajník, Tomáš, Jan Faigl, Vojtěch Vonásek, Karel Košnar, Miroslav Kulich, and Libor Přeučil. "Simple yet stable bearing only navigation." Journal of Field Robotics 27, no. 5 (2010): 511-533.
[2] Peconková, Veronika, George Broughton, and Tomáš Krajnık. "Unsupervised Learning of Landmarks for Visual Navigation in Changing Environments."
[3] Wang, Dominic Zeng, Ingmar Posner, and Paul Newman. "What could move? Finding cars, pedestrians and bicyclists in 3D laser data." In 2012 IEEE International Conference on Robotics and Automation, pp. 4038-4044. IEEE, 2012.
[4] Paz, Lina Maria, Pedro Piniés, and Paul Newman. "A variational approach to online road and path segmentation with monocular vision." In 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 1633-1639. IEEE, 2015.

Name and workplace of master’s thesis supervisor:

George Broughton, MSc. Department of Computer Science FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **02.02.2022** Deadline for master's thesis submission: **20.05.2022**

Assignment valid until: **30.09.2023**

George Broughton, MSc. Supervisor’s signature	Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

. _____ Date of assignment receipt

_____ Student's signature

Author statement

I declare that the presented work was developed independently and that I have listed all sources of information used in accordance with the Methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, date ....

Signature

Ackowledgment

I would like to thank my supervisor George Broughton for all his guidance and pushing. My gratitude also belongs to the rest of the Chronorobotics laboratory, who supported me both emotionally and with expert advice and did who not cut me slack in following the ways of science until exhaustion.

Abstract

Representations used by mobile autonomous systems for visual navigation have trouble effectively dealing with changes which inevitably happen with time and are detrimental to their performance. The goal of this thesis is to propose and evaluate methods for the unsupervised learning of semantic landmarks from a long-term deployment of the visual teach-and-repeat navigation system. Their main assumed benefit is the ability to enable such navigation over extended periods of time where classical methods based on local image features fail. Deployment of autonomous robotic systems is heavily dependent on large annotated datasets because of neural networks in perception. These, however, fail to satisfy the requirements for representativeness, which makes even partial automation of creating these datasets very desired. In the particular task of visual navigation, learning more abstract and high-level representations of the environment is also interesting as it brings the algorithmic solution closer to human cognition. The work proposes three methods for automatic semantic landmarks discovery, each based on a different principle specifically designed to make use of the data from long-term visual teach-and-repeat navigation. The proposed methods are evaluated on their ability to align images necessary for their integration into the chosen navigational system. The results indicate that they are more robust albeit less precise than the classical methods based on local image features. The integration is then also done and tested in a real-world experiment with a robot repeating a path recorded two weeks previously.

Keywords: teach-and-repeat navigation, long-term autonomy, unsupervised learning, autolabeling, chronorobotics

Abstrakt

Reprezentace používané mobilními autonomními systémy pro vizuální navigaci trpí změnami v prostředí, které nevyhnutelně v čase nastávají a zhoršují jejich výkon. Cílem této práce je navrhnout metody pro učení sémantických orientačních bodů bez učitele pomocí dlouhodobého nasazení navigačního systému pro opakovanou navigaci. Dále je zhodnocen jejich schopnost umožnit dlouhodobou navigaci, protože klasické metody založené na lokálních obrazových vlastnostech mohou v takových případech selhávat. Nasazování autonomních robotických systémů je silně závislé na velkém množství anotovaných dat pro neuronové sítě z oblasti počítačového vidění. Dostupná data však nesplňují požadavky na reprezentativnost, a proto je i částečná automatizace tvorby těchto datových sad velmi žádoucí schopnost. Přímo pro úlohu vizuální navigace je také zajímavé učení abstraktnějších reprezentací prostředí, která jsou bližší lidskému uvažování. Práce navrhuje tři metody pro automatickou identifikaci sémantických orientačních bodů, z nichž každá je založena na jiném principu, specificky navrženém pro využití dat z opakované vizuální navigace. Navržené metody jsou hodnoceny z hlediska jejich schopnosti horizontálně zarovnávat snímky, což je nutné pro jejich integraci do zvoleného navigačního systému. Výsledky naznačují, že jsou v tyto metody robustnější, i když méně přesné než klasické přístupy založené na lokálních obrazových příznacích. Samotná integrace byla otestována v reálném experimentu s robotem opakujícím cestu dva týdny po jejím naučení.

Klíčová slova: opakovaná navigace, dlouhodobá autonomie, učení bez učitele, automatické anotace, chronorobotika

Contents

List of Figures

1 Introduction

Countless determined hands and minds of researchers in the field of robotics have for a long time worked very hard to realize the idea of robotic autonomy. All the advances in new sensors, the recent revolution in techniques for robotic perception and others have brought us extremely close to its realization, but why is it that the robots are still mostly tied to confined space of industry sites? One of the main problems of integration of autonomous robots into the lives of ordinary people is the principal difference in reasoning, knowledge extraction and other competencies between the human and the robot. These make tasks that are very easy for humans hard for robots, which can cause frustration to the people having to interact with such autonomous systems. One particular case is the task of visual navigation.

For a human, it is hard to imagine the technical problems tied to traversing an environment based on visual sensors. Moreover, once a human is guided through a specific path between two places, they are able to repeat the path easily even under heavy changes like the fall of leaves during the autumn or construction work alongside the route. If some parts of the route were blocked, the human would just go around, but for a robot, this can be rather complicated problem to solve.

Robots simply navigate differently. They are often equipped with an external localisation system like the GPS, which allows them to repeat some path by simply checking their exact position with the reference one and correcting this along the path. If then these systems need to enter a building, they are completely lost because such external localisation systems are rendered unusable. One of the approaches not relying on external localisation are the methods collectively referred to as Simultaneous Localization And Mapping (SLAM). These build complex maps of the environment where the robot operates, in which the robot localises itself even without the use of external infrastructure. Most currently existing systems, however, need very precise and high-end equipment to be able to do so, being it 3D-LiDAR scanners or industrial grade 4k, stereo or depth cameras. These can be very expensive and, in general, try to solve the problems the robots have by adding more data into more and more detailed maps, while at the same time the human might be able to explain the route—communicate their internal map of the path—in few sentences and based on a simple pair of cameras only.

The efforts to mimic the visual-only navigation through a known path with a simple, even monochromatic camera have been very successful in the last ten years showing that no external infrastructure like Earth-orbiting satellites or expensive sensors like 3D-LiDAR scanners are needed. As the ability of these systems has been tested for longer and longer periods of time, it became apparent that the internal representations of the path they used are not good enough because with the passing time between the first and repeating navigation, the ability declined rapidly. The representations were not robust enough because they were based on local information embedded in the images taken by the camera and did not consider any high-level information about the environment.

With the new technologies based on deep learning, we are now able to equip the robots with perception modules able to identify and understand the image and scene at

a much higher level. This brings the option to use higher-level features of the environment as a basis for the repeating navigation and to create internal map representations at a semantic level—much closer to those of humans.

Unfortunately, even these methods are not without their own problems. The enormous need for large datasets that have previously been hand-annotated does present a big hurdle in deploying such models in real applications. The creation of such datasets is very expensive and human-labour intensive, so it is not possible to create a completely new dataset for every particular deployment environment and scenario. And humans are still able to learn the significant and important landmarks from the actual navigation itself—they do not need to distinguish a stationary white building from a moving white van to figure out that one is a useful landmark and the other is not.

This particular ability to learn the effective landmarks simply from performing the navigation is the main topic of this thesis. If robots were to posses it, they would have a robust navigational method that works under various conditions and a dense representation of their paths, which they could easily communicate. With some level of effectiveness these representations could even be communicated to humans.

The problem to be solved is formulated as proposing methods for automatically generating annotations for unsupervised learning of semantic landmarks by a neural network and evaluating them. To fulfil the proposal, the author conducted an in-depth review of relevant literature, proposed and implemented three methods—one adopted and two original methods—designed an evaluation to compare the proposed methods in the task of long-term visual navigation and performed said evaluation including contribution to the dataset collection, complete processing of the data and a real-world robotic experiment. At the end of the work is a discussion of the prospects of the proposed methods and their potential.

1.1 Organisation of the Thesis

The thesis is organized into six sections in most parts corresponding with the proposal guidelines. The first Section is this Introduction.

The second Section, titled “Related Work”, provides an overview of the literature on four topics important for this thesis. The first topic is the teach-and-repeat navigation with particular focus on the variant based on cameras—the visual teach-and-repeat navigation (VT&R). The second one concerns itself with the problem of the long-term adaptation of the maps in VT&R to account for the changes in the environment. Then the topic of temporal modelling is reiterated as its methods are further used in this work. The last covered topic is the literature on the task of unsupervised object discovery, which is not a much-studied topic but corresponds strongly to the topic of this thesis.

The third Section, titled “Methods for Mask Generation”, proposes the methods for automatic generation of annotations for the unsupervised learning of semantic landmarks that are the main concern of this work. The section presents one method presented in a proof-of-concept two-page extended abstract paper [1] and two original methods based

on two different kinds of information present in the kind of data created during the long-term deployment of the VT&R navigational system. Alongside the specification of the methods, the section also contains implementation details necessary for reproducing the work.

The fourth Section, titled "Experiments", contains the design of the evaluation scheme. As the evaluation is supposed to reflect the use of presented methods for long-term VT&R navigation, additional methods had to be proposed to make the neural networks trained using the methods from the third Section integrable into such a system. The Section also presents the dataset used for the evaluation and the details on its collection. And finally, two experiments are designed to compare the proposed methods, one based on the collected dataset and one real-world robotic experiment.

The fifth Section "Results" presents all the findings. First, the methods are evaluated qualitatively based on what they were able to learn. Then the outcomes of the two experiments are presented and discussed together with statistical evaluation.

The last sixth Section "Conclusion" closes the thesis. An overview of the work done is given with its main outcomes. The findings and the possibility of deploying such methods in autonomous robotic operations are discussed in the context of their capabilities. The Section also contains the directions of future work and hypotheses to be tested.

2 Related Work

This section presents the context of this work in the current state of the research on related topics. First, the concept of teach-and-repeat navigation is discussed with particular focus on its visual variant. Then, as the existing standard implementations used in this thesis exhibit problems when navigating over long periods of time, and one of the goals of this work is to make these systems less likely to fail due to this, the approaches for fixing these problems are presented. Next, as they are used in existing adaptive visual teach-and-repeat systems and are a large part of this work, temporal models are discussed. The section closes with the topic of methods for object discovery from visual data.

2.1 Teach and Repeat Navigation

Teach-and-repeat navigation (T&RN) is an extensively studied concept in robotics, originally especially popular with industrial robotic manipulators, which could be adapted to a new task by technicians at the site by simply guiding the manipulator through its operation. In the field of mobile robotics, T&RN refers to robot movement in the environment, which is once performed under some supervision and then repeated autonomously.

This particular formulation of navigation task is special because it does not require a precise global map—it is sufficient to build a locally consistent topometric map [2, 3]. A particular kind of map building methods can also provide global localisation during the building of the map—Simultaneous Localisation And Mapping (SLAM) [4]—but since the global consistency of the map is not required, this is not necessary for T&RN. In fact, no explicit localisation against the map is needed [5], though it can be beneficial and is often used as in [6] for along-path corrections.

There are solutions to this problem labelled as map-less, as they employ implicit representation of the navigational algorithm using deep neural networks like [7, 8]. Some have been shown to perform well in outdoor environments, but no deep study of their reliability was conducted, and therefore deployment of such methods in practical robotic applications is still far, given the inherent inexplicability of neural networks.

Efforts have been made to build T&RN systems using various sensors, including LiDARS [9, 10] that performed well in indoor experiments. Cameras turned out to be more suitable for this task as they can be used both indoors and outdoors without such strict requirements on the environment—structures need to always be in the range of LiDAR sensors if any matching should be done. Using cameras for T&RN falls under visual navigation, defining the visual teach-and-repeat navigation task. Of course, different kind of cameras can be used—monocular [11], stereo [12] or omnidirectional [13], monochromatic [14] or RGB [6]. As this work builds on the existing BearNav system, it is limited to the use of the monocular camera.

Individual approaches are also characterisable by whether they use robots odometry or not. Systems not using odometry need to provide the whole navigation, while those that do use it only need to correct for errors occurring due to problems like wheel slipping.

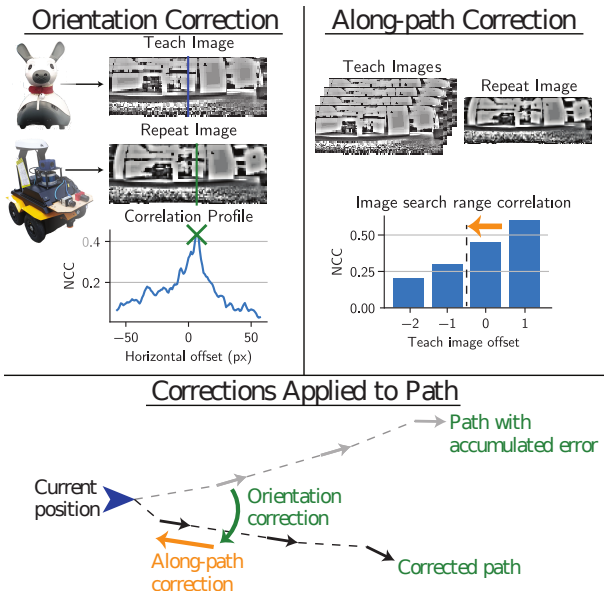


Figure 1: This Figure shows an example schema of a visual navigation teach-and-repeat system. This particular system corrects the odometry information both in the heading and along its path. Courtesy of [6]

.

The authors of [11] developed VT&RN for navigation in corridors without odometry. Their system navigates only by computing the steering angle from the template, matching the centre of the current image to the current position on the map. For a map, they use a sequence of images captured during learning and localise on it by incrementing the current position when the correlation of the current position becomes lower than the next one.

In [14] the system is in the teaching phase recording steering commands to the robot from the teacher and also a sequence of images along its path. During operation, it replays stored commands and corrects for errors in orientation only by computing correction steering angle from matching the image features in the current view to the ones in the map. Localisation in the map is done using the information about travelled distance from robots odometry. It was shown that only by correcting the heading, the trajectory of the robot converges in time to the one performed during training, given it is a closed curve [5].

Finally, in [6] authors extend on orientation correcting with along-path correction computing it by searching for the best-matching image in the map in the neighbourhood of the currently estimated position. This particular system is as an example shown in Figure 1.

An important distinction is also the method used for computing the steering command, i.e. the image alignment. Some have used the alignment by maximising correlation [11, 6]

described above, some derived alignment of the whole images from matching of individual images features, and there have been efforts to perform the matching directly using CNNs. In [15, 16] the used features were based on whole lines. In [17] the authors used SIFT [18] features in combination with KLT point tracking. [14] used very simple histogram voting based on pair-wise mapping of SURF [19] features in both images. Other image features like BRIEF [20] or later ORB [21] have also been used. Authors of [22] show that just by incorporating the information about time, it is possible to achieve better accuracy in image alignment.

2.2 Adaptation

Alignment based on rather low-level image features is precise but prone to very fast map degradation over time as the appearance of the scene changes. This effect has been extensively studied with respect to standard image features [23, 24] and navigation using them [25, 26]. There are currently two approaches to solving this problem—one is to cleverly update the map during the operation [27, 28, 29, 30], the other is to use higher-level features that are unlikely to change and detectable under diverse conditions. In [31] the authors successfully try to use higher-level features based on the output of convolutional masks of a pre-trained CNN. An alternative to feature matching then presents the use of CNN on the images directly—the authors of [32] used a higher level output of a pre-trained CNN on which they then applied the Discrete Fourier Transform for matching, the authors [33] trained Siamese networks to predict the alignment between two images which was then further developed [34]. They all show that such an approach brings significant improvement in robustness. Once a robust but maybe not precise alignment is reached it can also be further improved by image features for a finer alignment [35].

As in other domains of artificial intelligence, even the field of robotics is lately undergoing a transformation caused by the success of deep artificial neural networks in high-dimensional data processing [36]. Unlike in other applications, though, roboticists have to be warier as allowing a physical machine to be operated by an algorithm of low explainability with potentially erratic behaviour can lead to property damage or even injury. That is why deep learning (DL) models are usually deployed as modules of larger systems performing a rather specific task, most likely of perception, where the prevalent kind of models are convolutional neural networks (CNN), e.g. YOLO architecture for object detection [37] or Mask-RCNN for semantic segmentation [38].

DL has been successfully applied in many applications trained in a supervised manner [39]. Due to quite large feature spaces of typical DL applications, e.g. images, useful models often end up being large as well. This, in turn, creates extreme requirements for training dataset—one of the first large datasets for object detection "ImageNet" [40] contains millions of annotated images, another popular Microsoft COCO dataset [41] contains 328k images with 2.5 million labelled instances. When a new application for DL models arises, it is first necessary to collect immense amounts of data and hand annotate them, which is why unsupervised training of DL models has such potential.

There are multiple ways to achieve unsupervised learning of deep models like transfer

learning [42] or GAN networks [43]. Often, the network is taught on a proxy task in a supervised manner, which allows it to learn low level convolutional or other features that are applicable in other contexts and then it is used in other tasks where it needs lower amounts of data to adapt [42]. Such data can sometimes be generated automatically as in [44, 45, 46], either algorithmically or by cleverly withholding parts of the whole dataset to use as training labels.

2.3 Temporal Modelling

Being it the work on improving the image alignment accuracy [22] or the work on the adaptation of the maps over time [27] it has been shown that robotic systems operating over a long time benefit from explicitly taking into account the temporal information. As this is an effect present also in other robotic tasks like localisation, there have already for some time been efforts to study ways to model temporal phenomena for robots. The robotics subfield with such focus is called the chronorobotics [47]. This section first provides an overview of the efforts in this direction until the time of writing this thesis and then focuses on the evolution of two particular methods used later in this work.

Mapping has been a key capability of robots operating over the long term, and one of the problems such systems have to deal with is the uncertainty of their measurements as well as the changes in the environment itself. The first approaches chose to neglect the changes in the environment and only deal with the former [48], which was highly enabled by the development of probabilistic methods based, among others, on the Bayes theorem and other probabilistic techniques. While that allows the robots to operate in a static environment like industrial halls, the long-term operation in environments that change has to deal with those changes as well [49].

There are different strategies for dealing with environmental change—it is possible to suppress them, filter them and learn from them. Suppressing the changes is mostly tied to the visual appearances, which are highly affected by the changes in illumination typical for the day-night cycle [50, 51], but choosing structure over appearances for navigation [52] can also be understood here. The filtering approach consists of many techniques like filtering of features in feature-based maps by estimating their chance of persistence [53], learning visual representations that isolate features invariant to the changes [54] or learning visual feature descriptors that are robust to the change [55, 56] which was further enabled by the success of neural networks [57]. Finally, the learning from the changes strategy recognizes that changes are inherent to the environments where autonomous mobile robots are set to operate and even more heavily when humans are involved [58]. There are works trying to learn various representations of the environment based on the conditioning of the changes like [59] others choose to embed the description of the changes into the maps themselves, allowing, among other things, for predictive queries over the map [60, 61] with existing efforts employing generative networks to make such predictions over visual appearances [62].

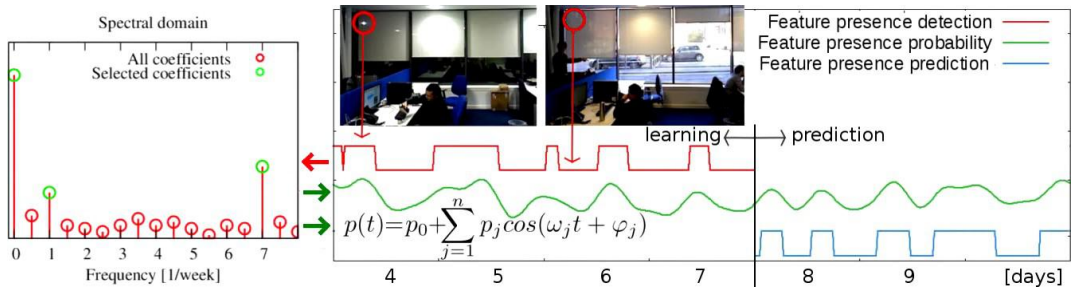


Figure 2: One of the approaches to dealing with changes in the environment is to learn from them. This has been shown to be beneficial in many tasks like visual localisation. Courtesy of [47].

2.3.1 FreMEn

Particular methods stem from the efforts to build universal maps of periodic temporal phenomena. The first efforts to do this were trying to use spectral decomposition based on the Fast Fourier Transform (FFT) to model the evolution of binary states in a robot's environment, including grid occupancy maps [63]. The idea of using spectral representations was then tested on various robotic tasks like localisation [64], navigation [65] or activity recognition [66]. To overcome limitations imposed by the FFT algorithm like the need for regular observations of modelled phenomena which can happen in different parts of the environment forcing the robot to often revisit all places of observation, and the inability of model updating, a new method was derived called the Frequency Map Enhancement (FreMEn) [22]. The FreMEn method was used in many scenarios, much like its predecessor based on FFT. It was used, for example, to enhance occupancy maps [67], predict traversability of edges of a topological map [68] or improve task allocation based on predictions of human movement [69].

Following up on the success of FreMEn, another model was proposed called the AAM [70]. The authors also show its qualities for semantic clustering of locations on a topological map. Similarity based on spectral analysis of time series is an idea of interest to this work as it is at the heart of one of the methods for unsupervised object mask generation.

2.3.2 Other Methods

Apart from FreMEn method, there are others that deal with modelling periodic spatio-temporal phenomena, for example, STeF-map [71]. These are often developed for a particular application but still provide a hint for the development of a more general method. Many works introduce the knowledge about the periodicities into the models in form of a priori knowledge and only use a small number of periodicities like predicting the demand for ambulances [72, 73], predicting street crime [74, 75] or modelling the spreading of disease [76]. Others seek to extract the periodical nature in its entirety from the

data itself, like efforts for approximations of periodic kernels for Gaussian process-based methods [77], but aside from already mentioned work on FreMEn there are not many.

One noteworthy method building on FreMEn is the Warped Hypertime method [78] and its derivatives [79] which is based on a warping projection of spatio-temporal data onto a surface of a hypertorus where the non-temporal data is then processed as depicted in Figure 3. This allows the model to be built from temporally sparse observations. It was applied in the robotic [80, 81] and other [82, 83] tasks and found to be equal or better than FreMEn [79].

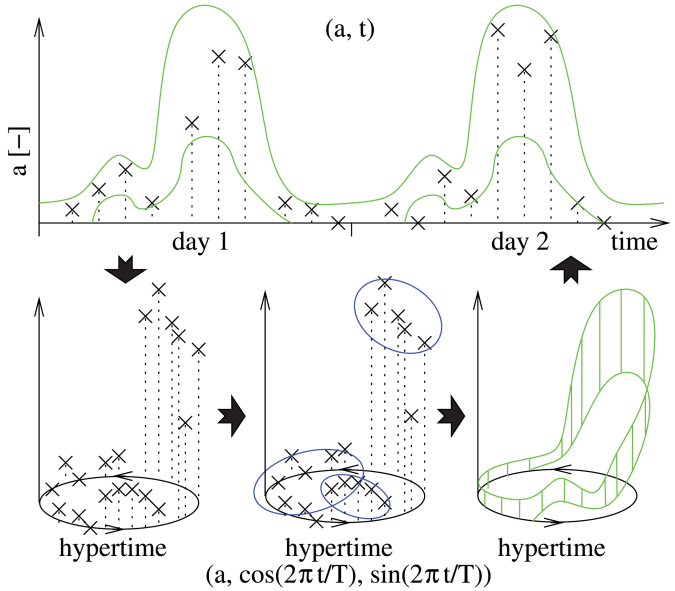


Figure 3: The process of warping the temporal dimension of spatio-temporal data to allow for periodic model creation from sparse observation by an autonomous system. Courtesy of [78].

2.4 Unsupervised Object Discovery

This last section tries to capture various interesting works tied to the topic of unsupervised object discovery. While that is a research topic of its own, the application in this thesis has its specifics mostly tied to the robotics domain, so most of the general work is not directly applicable. Although it still gives an interesting perspective, more focus is put on efforts directly in the robotics domain as these better capture the context of the task.

The approaches to object discovery use various kinds of data, which heavily determines the principles they exploit. If the data is visual, the main difference is if the methods only use one image or full videos. One image only methods sometimes try to solve the salient object detection task [84], and even active learning can be employed for them [85]. However, for the learning from a video, the possibilities are numerous. The spatio-temporal

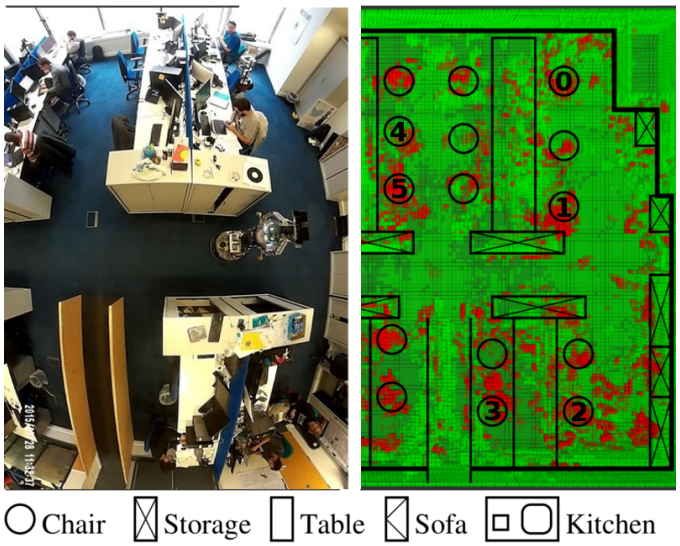


Figure 4: A visualisation of a spatio-temporal occupancy model and the environment it is supposed to capture. The red color indicates daily periodicity of the data. Courtesy of [99].

context provided by the video data gives much more information to the differentiation of the objects, the knowledge to be exploited mostly comes from the motion [86, 87, 88, 89], with techniques usually using some kind of optical flow method. Still, other approaches can be found, like the segmentation of the dynamic textures [90]. In robotics, another popular kind of data is the RGB-D maps which, when created dense, can combine the information about appearance with geometry [91, 92, 93] or maps made from the LiDAR scanners [94, 95].

Of particular interest to this work are efforts in object discovery based on temporal information gathered over longer periods of time. The temporal aspect of the autonomous operations does not just bring more data over time, but it gives the robot an entirely new source of information—the changes in the environment. These changes can, analogous to the previous section, be also used to discover objects as the sets of sensor inputs exhibiting similar change characteristics. The authors of [96] explored this idea in the long-term autonomy scenario for unsupervised learning of objects and further developed it into a system which is able to perform this for a long time with no human interaction [97] and show that temporal patterns of the changes are also usefull [96]. The semantic relevance of temporal properties has been explored, for example, in the work [98] which is set to classify spatial areas based on temporal properties (temporal grids). There are works supporting that temporal patterns of the changes are also useful, like [96]. In the results of [99] the authors also show that sectors of an office which exhibit daily periodicity are relatively compact and spatially tied to places of human work, which also suggests the meaningfulness of efforts to segment data using temporal characteristics. Their figure is included as Figure 4.

3 Methods for Mask Generation

This section presents in detail the individual methods proposed by this work as parts of the whole pipeline for VT&R navigation. The overall idea is that an autonomous vehicle uses the VT&R navigation system to travel the same path over and over for some time. The system could eventually use the data it collects to investigate landmarks specific to the particular environment and therefore learn better representations to enable the navigation further. Because the current methods based on image features are known only to be able to repeat the given trajectory for some limited time due to changes in the environment and their inherent frailness, this approach could prove to be a viable solution to enable the navigation to happen for really extended periods of time.

First, proposed methods for automatic generation of masks for training neural networks are presented as the main focus of the thesis. Then the training of the neural networks in the given application is discussed. Finally, albeit not the main focus of the work, it was also necessary to design methods to provide alignment information based on the output of the trained models on pairs of images to be able to test them using the selected visual navigation system BearNav. The image matching approaches are therefore also discussed.

3.1 Mask Generation

The main goal of this thesis is to propose a set of methods for the automatic generation of annotations for a landmark detecting artificial neural network. This process has to be unsupervised to allow its deployment in long-term autonomy navigation scenarios, and for the same purpose, it has to be able to process the data produced in such an application. As the automatic generation of masks for training the landmark detecting neural network necessarily encompasses the establishment of what the landmarks in the data are or could be, and the network used performs the task of instance segmentation, the terms "generation of annotations", "mask generation" and "landmark detection" will be used interchangeably where no confusion can arise.

The first question to answer is, what is a good landmark for visual navigation based on which the navigation is possible even over extended periods of time. The current methods rely heavily on standard local image features like SIFT or BRIEF, but these have been shown to drop fast in their matching ability with the changes in the environment as their local nature does not allow for much robustness. Hence, the landmarks necessarily have to be higher level than local pixel values based features. From experience with human navigation in a known environment, it follows that the landmarks useful for such navigation often correspond to prominent structures, such as buildings, their parts, trees or hills in nature. The ideal landmark is distinguishable under various conditions and carries some semantic information in the given environment. While larger objects do often fall into this category, one does not want to pay attention to those that move a lot, only to those that are stationary, so in general, not all objects are good semantic landmarks.

Based on the previously described requirements, two approaches to generate masks—ultimately providing the ability to detect useful semantic landmarks—alternative to [1] were proposed to exploit the specific data generated during the operation of the BearNav VT&R system. The system operates by repeating the same control commands as it received during the teaching phase and correcting its heading by aligning the current view with the images stored regularly in a given time interval during the teaching. This generates specific maps where the same place from a very similar viewpoint is visited as many times as the system repeats the traversal with the number of "places" determined by the interval parameter of the system. The standard interval is one meter of along-the-path travelled distance.

The reference approach dubbed Autodidact uses pairs of images of the same place from different traversals, i.e. different times, to establish dominant colour changes. The second proposed method—the first alternative—uses pairs of images that are consecutive in the traversal, i.e. from the same time but slightly different places, to establish the motion patterns in the image. The third proposed approach—the second alternative—tries to detect temporally stable and interesting regions of the images based on long-term observations of the same place, i.e. all available times. All these methods aim at detecting regions of the images that show some similar characteristics in the sense relevant to navigation. The Autodidact approach targets extracting structure from appearance changes, the first alternative targets extracting structure from camera motion and the second targets extracting landmarks from temporal stability.

All the presented implementation details and parameter settings have been developed and tailored to data used in this thesis which are further discussed in detail in Section 4.3 and contain images of the dimensions 640 x 320 pixels which for the actual mask generation were rectified using the actual camera calibration and distortion parameters.

3.1.1 Autodidact

The first method for generating annotations, presented in [1], is the Autodidact which extracts the masks from pairs of images taken at different times.

From the historical runs of the navigational method, the collected data present a large set of image pairs in which the same place is viewed up to a relatively small change of viewpoint, mostly in terms of a horizontal shift. The Autodidact method assumes that the change in the viewpoint is minimal, as the images have been aligned by the navigational system over the course of traversal so far. This assumption is quite strong and possibly could be weakened by prior prealigning the images by image features or whatever method was used for the navigation so far, but this improvement is not part of the original paper. Because of the nature of the navigational task—it takes some time to complete the whole traversal—the two images come from different times, and the appearance of the scene is therefore necessarily a subject to change. This change comes from various sources depending on the time difference between the two images analysed, but the change sought by the method is simply the change in appearance.

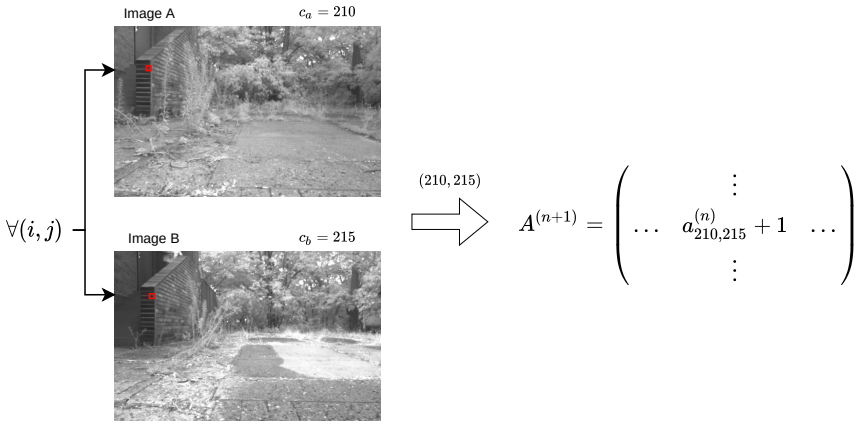


Figure 5: A visualisation to the process of construction of the accumulator matrix and its contents.

As stated previously, the method works under the assumption that the images are shot from the same spot only at different times. This allows for the reasoning that one can associate individual pixels in one image with the respective pixels in the other image, and by analysing this association, segment regions of interest—the landmarks. The method, therefore, calculates statistics over the changes in colour between associated pixels and searches for dominant changes across the whole image. The pixels of respective colours in both images are then masked. With the use of morphological operations, the masks are refined and split into connected components. This results in the set of masks for the given pair of images.

Formally, the method has five steps.

1. First, the colour mapping quantitative statistics are computed. The process starts with two grayscale images $\mathbf{P}$ and $\mathbf{Q}$ understood as matrices which are pixel-wise overlaid, and the relative frequencies of the generated colour mapping are computed. This frequency information can be organised into *the accumulator*—a matrix $\mathbf{A} = \left(a_{ij}\right)_{i,j=0}^{C}$, where C is the total number of colors in the picture and

$$a_{ij} = \#\{(k,l);\ \mathbf{P}_{k,l} = i \wedge \mathbf{Q}_{k,l} = j \text{ with } k = 0, ..., H \text{ and } l = 0, ..., W\}, \quad (1)$$

   where H and W are the height and width of the images, respectively. A more intuitive understanding of the construction of the accumulator matrix is shown in the Figure 5

2. The accumulator matrix is then interpreted as a grayscale image, which is first blurred using a gaussian kernel and then normalized to a standard 8-bit image. In the resulting image, blobs are searched for using a simple blob detection algorithm based on thresholding the image on various levels and filtering the resulting blobs by certain characteristics like colour, area or circularity. The original paper presenting

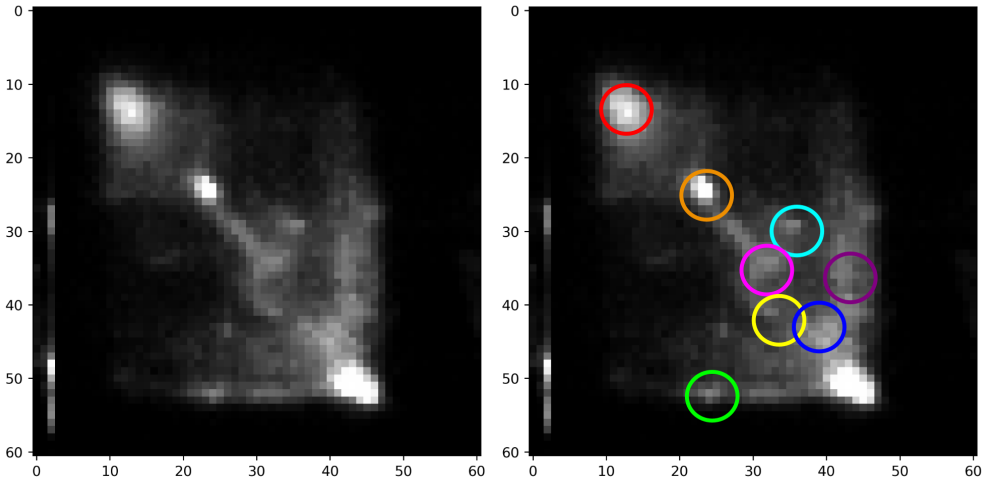


Figure 6: An example of the image interpretation of the accumulator matrix and the result of blob detection in the images. The particular example was generated from images quantized to 64 colors.

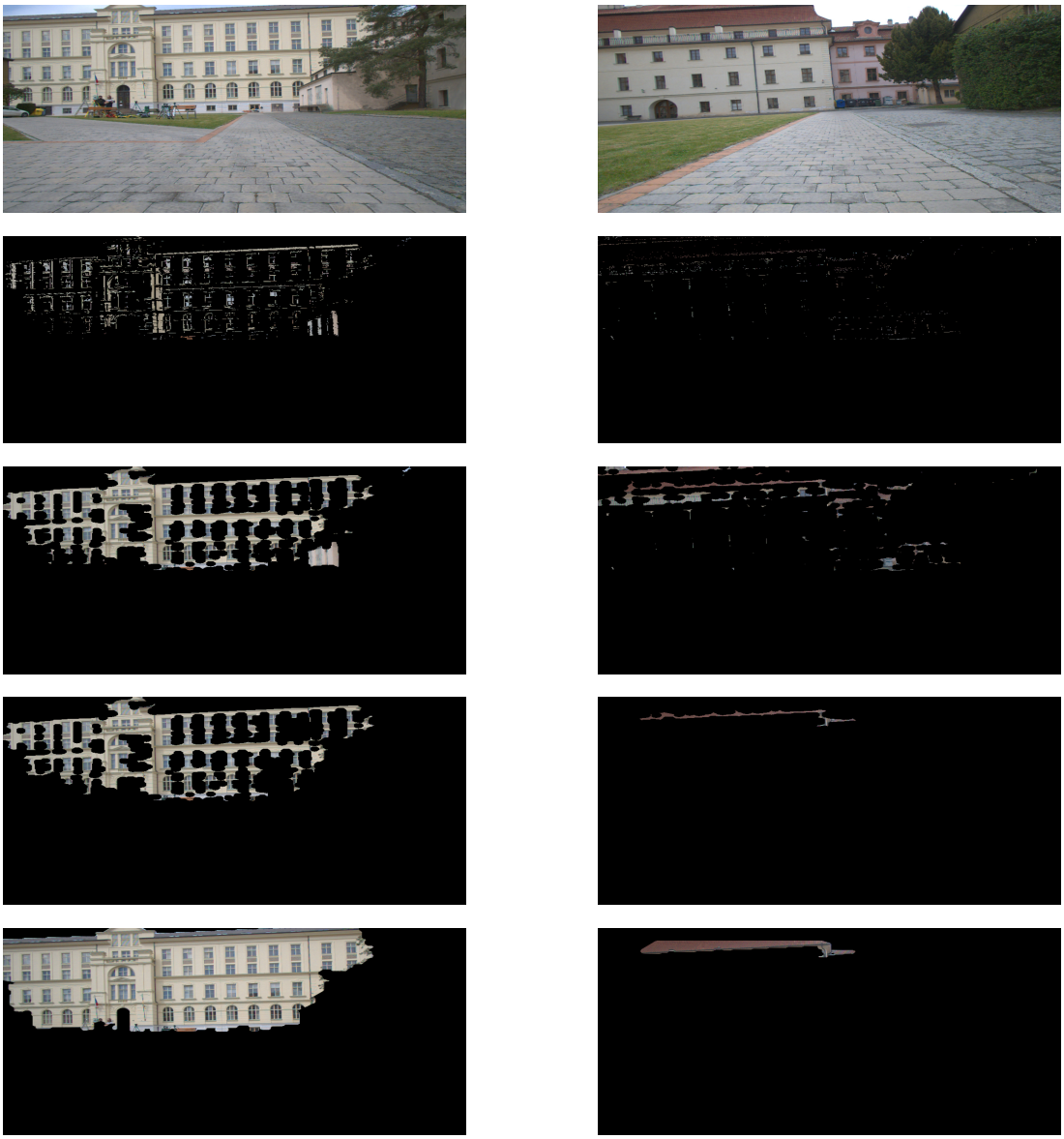
Autodidact references a particular implementation of this algorithm in a popular computer vision library OpenCV [100]; this work follows.

A visualisation of the resulting accumulator matrix image can be seen in Figure 5 with the blobs detected by the detection algorithm shown as well.

3. The blobs extracted in the previous step represent the prominent colour shifts between the two captures of the scene. These now have to be translated back to the original image. Because of the way the accumulator matrix is constructed, this is straightforward—the blob defines a set of positions in the accumulator matrix, which give pairs of colours at the corresponding pixels in the two images, so these just have to be masked. This happens for all the blobs, which generates a set of binary masks for each image.

4. In the next step the binary masks are refined. Due to the pixel-wise analysis and the fact that in the accumulator matrix, values of pixels from different regions of the image are all projected in the same position by having a common colour, the masks that result from the backprojection are not particularly compact. To improve them, standard image processing morphological operations dilate and erode are applied.

5. Finally, to extract individual landmark masks from the binary masks for the whole image refined in Step 4, the connected components algorithm is used to extract connected regions. As the morphological operations make the connected regions more compact, the result of the connected components splitting the whole mask tends to output realistic individual masks that have intuitive properties—they are dense, compact, localised and have some common colour structure.



(a) (b)

Figure 7: Two examples from different places of how the process of mask extraction happens using the Autodidact method. On the left, the extracted landmark is the whole building, and on the right, the corner of the roof with a characteristic gutter drain. The sequence of images from the top represents the original image (one from the pair), the mask extracted directly from the autodidact method through backprojection of one of the colour blobs from the accumulator matrix, the mask after morphological refinement, one of the connected components and finally the mask, as it is further refined by the GrabCut algorithm. The whole process happens in binary masks. The overlay over the original image is added for visualization purposes only.

Unfortunately, the original paper is not very descriptive of the details, and the presented proof-of-concept experiments seem to involve some manual interaction with the method and limited dataset of a natural environment. From experimentation with the method, the following improvements had to be introduced or specified to reach the wanted performance of the method.

- The first significant change in the implementation done for this work is that to reach stable performance of the method, the colour space was quantized to a lower number of colours. Namely, two versions were studied, one with 32 and one with 64 colours instead of the full 256 colour spectrum of 8-bit images.
- Because, ultimately, the masks are supposed to be used for navigation and come from the camera on a robot, the lower half of the image was cut before searching for the masks. The reasoning behind this decision is that the camera is positioned relatively low and on the robot. The lower half of the image captures the ground, which does not present very good landmarks for navigation. Even in the BearNav navigation system, features from the lower part of the image are discarded.

  If this was not applied, the method would produce masks, for example, for parts of the sidewalk. While this is a nice result, due to the structures on sidewalks being repetitive and the method picking it up differently each time, the network trained on these would not produce good navigational landmarks.
- In the Step 2 of the method, one percent of the highest values in the accumulator matrix image was clipped before the gaussian blurring for the stability of the blob detections.
- The blob detection itself is an algorithm with a lot of parameters that depend heavily on the rest of the settings, like the number of colours necessarily affecting the size of the blobs detected. Out of the set of parameters to the OpenCV, the library function used in this work sets the maximal threshold value to 250 and filters resulting detections only by inertia, i.e. the mathematical measure of resistance to the rotation around the principal axis, and sets the minimal inertia ratio parameter to 0.05.
- The refinement of the masks in the Step 4 using morphological operations misses important information about the setting of parameters. The size of the structural element obviously depends on the size of the images and the qualitative properties of the masks, which in turn depend on the size of the colour blobs and of the colour quantization. Therefore this has to be tuned to the rest of the algorithm. This work for the particular data used in the experiment and further explained in Section 4.3 used a circular element with a radius of 6,4.
- After refinement of the masks and their splitting by the connected components algorithm in Step 5 the method, based again on the overall setting of parameters, produces a relatively large number of masks. Although these masks mostly correspond to interesting regions and, due to their number, are not really suited for the

training of the neural network directly. Extra filtering is therefore applied to only keep those individual masks that satisfy the size requirement for the area being larger than 1500 pixels.

- A completely new extension to the method was introduced as a sixth step to get a more consistent result over different pairs of images to enable better learning of landmarks by the network. It lies in the application of the GrabCut algorithm [101] to extract the foreground of the masked region from the background. This algorithm iteratively estimates the foreground and background colour models and then gives a refined mask based on what the foreground is estimated to be.

All the parameter values presented are based on preliminary qualitative experimentation with the method as it was implemented according to the original paper.

Finally, some examples of the created masks at different stages of refinement are presented in Figure 7.

3.1.2 Motion-Based Object Discovery

The second proposed method—alternative to Autodidact—for generating the masks of landmarks is based on the analysis of the motion of the camera with the robot through the environment. Data-wise it operates on pairs of images taken during one traversal on two consecutive places, i.e. for a video-like sequence.

If the robot operates using a VT&R system for long enough, based on the work presented in Section 2, the robot should be able to learn about the structure of the environment and even perform the object discovery task from the video captured during the operation. However, in the actual deployments, the processing of raw video feed can prove to be overly demanding on computational resources and moving it from the robot to a cloud solution is currently not possible due to the bandwidth limitations. While the introduction of 5G networks may lift the bandwidth limitations, the approach presented in this work tries to show that for the particular applications of long-term VT&R navigation, a sufficient amount of data lies in the regular sampling of the camera feed done by the BearNav system.

This method works by comparing the images from consecutive places, which are likely to contain most of the same scene only with the change of the camera position. Under this assumption, it is possible to employ algorithms for computing the dense optical flow, i.e. the field of vectors of displacement for the pixels in one image to best match the view in the second one. After this information is obtained, it is converted into polar coordinates, giving for each pixel the direction and magnitude, which are normalized and quantized into a lower number of possible values. Direction and magnitude information is turned into an HSV colour space image and segmented by unique colour values resulting in a set of binary masks. The relative movement captures the dimensionality of the environment thanks to the various distances of the objects to the robot. This splits the image based on

the relative movement of the objects projected into parts of the image. The binary masks are then filtered and refined in a process similar to the one of the Autodidact method.

The method has formally five steps.

1. The first step of the method is to compute the dense optical flow for the two input images. This is done using the Farneback dense optical flow algorithm [102] for which both images are first converted to grayscale. This algorithm iteratively computes the displacement field using a prior estimation and polynomial expansion of the neighbourhoods of the pixels. The particular implementation used was the one in the OpenCV library, with the parameters set as follows: the pyramidal scale to 0.5, the number of levels in the pyramid to 6, the size of the averaging window to 20 pixels, and at each level the algorithm performed five iterations, used 10-pixel neighbourhood for local polynomial expansion and sigma of 1.7 for the Gaussian smoothing. The numbers were determined experimentally and not tested rigorously. The specifics compared to the standard applications to video sequences are that the movement between the images can be substantial, and fast blurred motion is not really interesting for landmark detection.

2. The displacement field from the optical flow algorithm is then converted to polar coordinates—the angle and distance from the origin, in terms of the motion, this gives the direction and magnitude. The values are then normalized over the whole image to the standard 8-bit grayscale image interval $[0, 255]$ and then quantized to 10 distinct values.

3. Using the direction and magnitude of the motion, construct a new motion image. A particular choice of method does not matter as this is merely a different interpretation of the motion field given by the optical flow algorithm. As with the Autodidact method, the lower half of the image should now be removed to prevent focus on areas not suited for navigation.

4. Segment the motion image simply by colour. As the quantization has greatly reduced the number of colours present, this produces better binary masks, which can be spread over the whole image because the positions of colour pairs were lost in the accumulator. Therefore the masks do not have to be continuous, so they have to be further divided using the algorithm for finding connected components, and only those of sufficient size are kept.

5. In the last step, the masks produced in the previous one are refined using the GrabCut algorithm for the same reason as with the Autodidact method.

To summarize, this method produces masks that correspond to regions of the image that exhibit similar motion behaviour and are therefore believed to belong to the same structure in the physical world. To illustrate how the method behaves, the Figures 8 and 9 contain an overview of how the individual steps look on two selected views. In Figure 8 the situation is particularly challenging, as it happens in the time when the robot was

turning fast, so the part of the scene visible in the images changes significantly. Even in this situation, the method was able to produce masks.

3.1.3 Temporal Properties Based Object Discovery

The third proposed method—an alternative to the Autodidact method—is based on the temporal properties of the image features observed during a long-term operation. Unlike the two previous methods, where the long-term aspect only matters in the sense of more data collection and a limited environment, this method aims to explicitly observe and model changes in the environment over the long-term and reason about them.

When navigating over a limited time span, the current VT&R navigation systems perform well with the use of local image features for the alignment of images. The problem with image features is that only some of the detected features belong to some interesting semantic landmarks, and their detection is not robust to appearance change. The main idea of this method for generating masks is that if one could filter those features that are semantically interesting and train a neural network to detect their surroundings, one could get a robust detection algorithm for the underlying landmark.

As presented in Section 2, the research on the long-term autonomy of mobile robots over the last decade has shown that the assumption of a static world is one of the main hurdles of their deployments as they lose critical competencies with time. This led to the development of techniques to battle this problem. These were later adopted for various robotic tasks, including visual navigation. The experience with these methods led to the conclusion that systems that are able to understand the structure of time improve their ability to navigate under various conditions as they are, for example, able to distinguish between landmarks visible during the day and the ones visible during the night. The method for generating landmark masks based on temporal information is inspired by these ideas, as it analyses the image features for those that are stable and exhibit temporal properties that suggest usefulness to navigation.

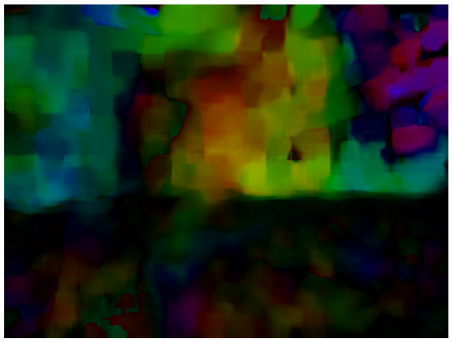
Specifically, the method analyses the time series mapping the time of place observation to the visibility of a specific landmark. Only those landmarks that are visible often enough are kept, and a descriptive temporal model based on spectral decomposition is trained on them. Then the landmarks are filtered according to the periodicities they exhibit to only retain those influenced by long enough processes relative to the length of the observation period. Once the interesting landmarks are separated, a square mask is drawn around them in each image where they are visible. As in the previous methods, a complete mask for each image is then split by the connected components algorithm and refined using the GrabCut algorithm.

Formally, the method has five steps.

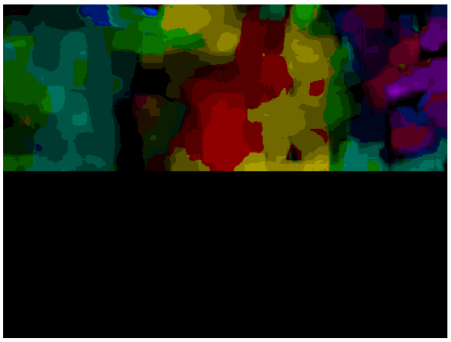
1. The first step is constructing the time series from a series of image observations of the same place. This process could be done in many ways. To select one, this work refers to the papers [22] and [78] where the authors applied the temporal modelling

(a) the original image 0

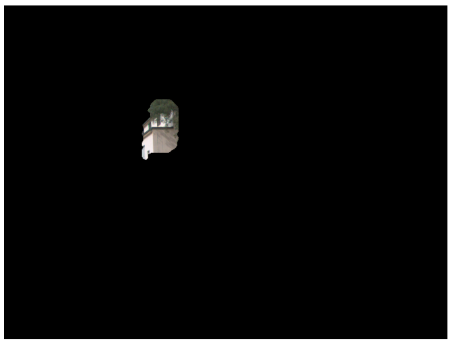
(b) the original image 1



(c) full estimated motion image



(d) cut and quantized motion image

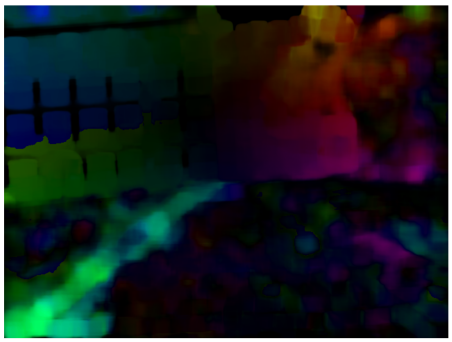


(e) example of the resulting masks

Figure 8: An example of the application of the motion-based mask generation approach. The first row depicts two consecutive images captured by the navigational system. The second shows the estimated motion image where the direction has been mapped to the hue and magnitude to the value component in the HSV colour space in a full-colour resolution cut and quantized version. The last row shows an example mask that was generated—the corner of the small technical building.

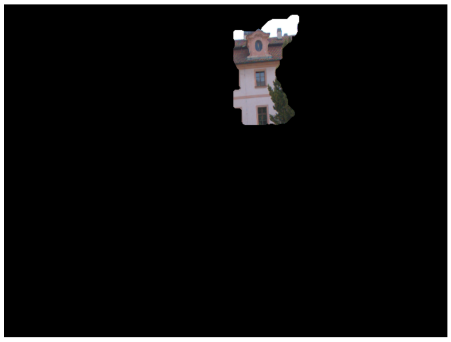
(a) the original image 0

(b) the original image 1



(c) full estimated motion image

(d) cut and quantized motion image



(e) example of the resulting masks

Figure 9: This figure analogously to the Figure 8 shows the application of the motion-based mask detection approach but on a different input images this time with simple forward motion.

techniques to the task of topological localisation for which they needed to construct the same kind of time series of feature visibility from the image data. This process works by feature matching in the series of images associating the features from the first image to the subsequent ones. The particular implementation of the process was the one provided by the authors of [78] as it contains a significant amount of implementation details.

The result of this step is the set of time series

$$v_i = \big(v_i(t)\big)_{t \in T} \quad \forall i \in I, \tag{2}$$

where $v_i(t)$ is the binary visibility value associated with feature i at time t, $T = t_i$, $i = 1, ..., n$ is the set of times at which the observations were made, represented in linear time as the number of seconds in the UNIX epoch, and I is the index set for the detected image features.

2. The generated time series are now filtered so that only those where the feature is visible in at least 10% of the observations are left, i.e. the index set I is restricted to $I' = \{i | i \in I \wedge avg_{t \in T}(v_i(t)) > 0.1\}$.

   The FreMEn temporal model of order two is then applied. The model of order o takes as input the time series as defined above and a set of candidate frequencies, in the literature defined as angular velocities $\omega_k \in \{2\pi i / L\}_{i=1}^{\lfloor L/S \rfloor + 1}$, with the L and S constants setting the longest and shortest considered periodicity respectivelly. Computing the complex coefficients of the discrete Fourier series for the visibility series v_i is done by the means of the following equations:

$$\begin{aligned} \gamma_0 &= avg_{t \in T}\big(v_i(t)\big), \\ \gamma_k &= avg_{t \in T}\big((v_i(t) - \gamma_0)\, e^{jt\omega_k}\big) \ \forall k \end{aligned} \tag{3}$$

   and selecting these (γ_k, ω_k) that have the o highest values of $abs(\gamma_k)$. The selected pairs are the converted into amplitudes $\alpha_k = abs(\gamma_k)$, phases $\phi_k = angle(\gamma_k)$ and periodicities $p_k = 2\pi/\omega_k$. While the method also defines a predictive function predicting the state visible if for time $\gamma_0 + 2\sum_{i=1}^{o} \alpha_i \cos(t\omega_i - \phi_i) > 0.5$ and invisible otherwise, this work only considers the learned parameters.

   Implementation by the author of this thesis done during the work as a part of the publicly available Python library Chronolib [103] was used.

3. With the parameters of the FreMEn model for all the image feature time series, these are further filtered, i.e. the index set I' is restricted to I'', so that $I'' = \{i | i \in I' \wedge max\big(per(\text{FreMEn}_2(v_i))\big) \in [12600, 43200]\}$, where the per function gives the set of periodicities and the filtering allows only those series where the maximal periodicity is between 3.5 and 12 hours. These numbers are more or less arbitrary but only leave the features where the affecting process is long enough and at the same time is not a day-night process of 24 hours. The upper limit is the highest attainable value below 24 hours.

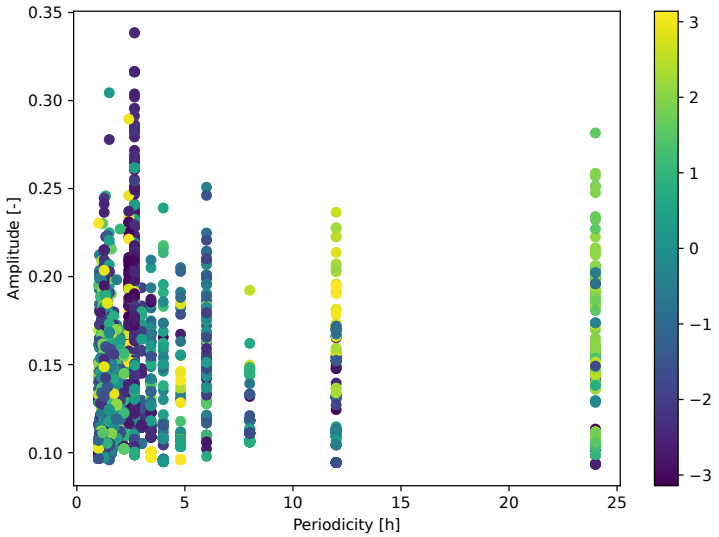


Figure 10: The relationship between learned periodicities and amplitudes of the FreMEn models trained on one example place in the used dataset. Colour is used to show the associated value of the phase shift in radians.

4. For every image feature left in the index set I'', a square binary mask of size 40 x 40 pixels is drawn for the images where it is detected. These masks are grouped for each image in the set, i.e. the one place.

5. The mask for one particular image from the previous step is split using the connected components algorithm and refined using the GrabCut algorithm as in the other methods.

The original idea included clever clustering of the parameters of the methods to split the landmarks also into distinctive categories allowing this method to also produce some additional labels for the mask. While this idea is supported by the literature, its implementation proved to be quite difficult. The clustering of the parameters themselves is not that complicated, and a preliminary analysis of the learned model on data presented later on suggests that it makes sense to do it. In Figure 10 the relationship between learned amplitudes and associated periodicities is shown, and Figure 11 presents the visualisation of all the learned parameters of individual models, where the parameters do form distinctive groups. The reason why this was not further studied for the generation of the masks is that upon closer inspection, some of the groups of features did have a similar semantic category, like a specific part of the balcony railing, but for most of them, it was not clear, and they did not exhibit any spatial coherence in the image that would allow for any segmentation.

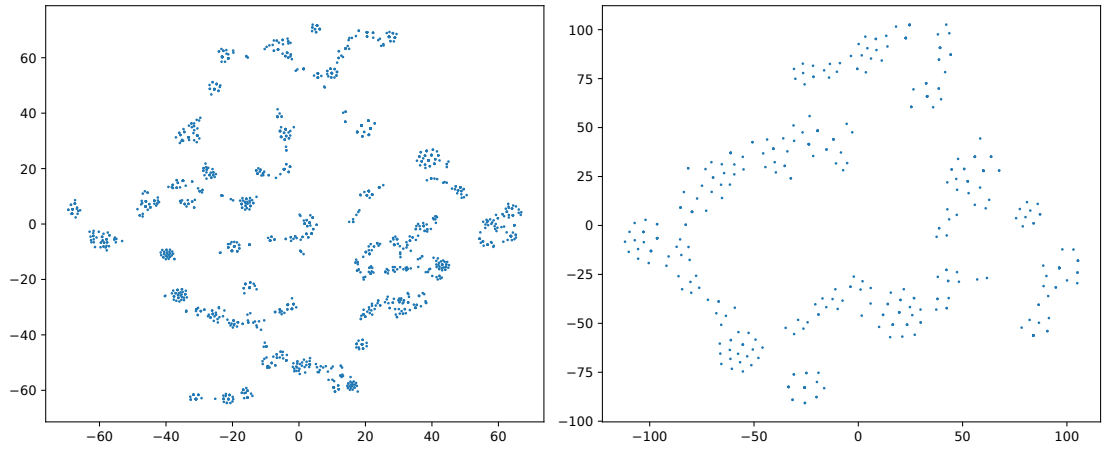


Figure 11: This figure shows the t-SNE visualization of 7-dimensional vectors of the model parameters in 2-dimensional plane. Because of mapping by the non-linear dimentionality reduction the values are unitless. On the left, all models for a particular place are shown, on the right they are shown after filtering by the maximal periodicity used in the method. The results indicate, that the models do actually form distinctive groups and therefore their clustering could bring useful information.

3.2 Summary of the Proposed Methods

To summarize the Section 3, the methods for generating the masks to be evaluated in the rest of the thesis are the Autodidact, the motion-based method referred to as "OptFlow", and the temporal properties-based method referred to as "Temporal". These methods were designed for the particular data generated by a long-term deployment of the BearNav VT&R system and represent three distinct approaches to the task. The first one has been adopted from a proof-of-concept paper and supplemented by a lot of practical notes and implementation details not covered by the original paper. The second and third have been developed completely from scratch for this thesis based on ideas present in the literature and tailored for the particular data.

4 Experiments

This section explains in detail the evaluation scheme designed to test the methods presented in Section 3 in the task of long-term visual navigation. That first involves training the neural networks from the generated annotations and their integration into the BearNav navigational system. Then a real-world robotic dataset used for the development and evaluation of the methods is described. Finally, two experiments are designed, one based on the metrics used in the literature on the topic and the other assessing the ability of a real-world robotic platform to navigate using the proposed methods.

Because the presented methods are, in general, supposed to automatically produce annotations for an artificial neural network that should learn to detect the given landmarks, the evaluation of the methods really only makes sense looking at the quality of the neural network models that were trained using these annotations. Furthermore, the methods are supposed to be evaluated in the task of long-term visual navigation, and because the methods were designed specifically for the BearNav navigation system, the evaluation requires their integration into this system. Namely, from the set of detections by the network, one has to be able to align two images well enough to eventually get the steering command for the robot. Because this has been thoroughly studied in the literature, the metrics and the evaluation scheme are adopted from relevant papers. Finally, of the most important metrics for any robotic task is whether a robot using the method can actually perform the task it is supposed to perform, so an experiment with a real robot is included as well.

4.1 Network Training

Following the original paper of the Autodidact method [1], the neural network used in this work is the Mask R-CNN architecture for instance segmentation [38]. The reason for choosing an instance segmentation network instead of an object detection one is that such a network actually combines object detection with extra mask segmentation. The masks produced by the methods have various sizes and shapes, and these often do not correspond well to standard rectangular bounding boxes, so having the information on a more detailed level should help the network learn and identify the relevant landmarks.

Because the dataset is naturally small and does not present much diversity for proper training of a deep convolutional neural network, the training scheme selected is the transfer learning, where the significant part of the network is pre-trained on a large but different dataset and then finetuned with a completely new head part. A popular choice for the large proxy tasks for transfer learning in object detection is the Common Objects in Context dataset [41], which was also used in the original Mask R-CNN paper. The particular implementation and model pre-trained on the COCO dataset used comes from the open-source deep-learning library PyTorch `torchvision` module [104].

In order to increase the robustness of the model to noise, which particularly during the night is from the camera very high, the images were for the network further downsized by

a factor of half, and gaussian blurring was applied in operation referred to in computer vision as transforming the image one level in the standard Gaussian image pyramid.

For the training, the dataset dedicated for the training of the networks was randomly split in the ratio of 9:1 into a training and validation set of images and corresponding binary masks representing the individuals to be detected. The network is trained using a very complicated loss based on the IoU metrics for bounding box detection and corresponding segmentation loss, and the validation loss used is the bounding box detection accuracy as designed for the instance segmentation challenges associated with the COCO dataset. Unfortunately, these have very little correspondence with the real target ability of the network—the detection of semantic landmarks useful for navigation—so in this work, these are only used to assess the level of training of the model.

4.2 Image Alignment

As mentioned in the introduction to this section, the methods to be tested for long-term visual navigation have to be integrated into the BearNav navigation system. For this, they need to be extended by some matching module that is able to horizontally align two images during the operation of the vehicle. One of these images comes from the original mapping phase of the T&R navigational scheme, the other from the current situation so that their alignment can be used to produce a steering command for the vehicle.

This turned out to be a hard problem of its own, making the comparison of the presented methods for mask generation quite difficult. The method for aligning the images presented in the original Autodidact paper [1] unfortunately does not contain enough information to reproduce it but is said to be based on the positions of the detected objects. Several methods for aligning two images based on the output of the network, therefore, had to be designed.

4.2.1 Feature-Based Matching

In the original BearNav system, the alignment of the two images is done using local image features. A technique used regularly for determining transformation between two images in the field of computer vision relies on computing a set of matches between points in the two images. In the particular case of BearNav, no complicated transformation is sought, and the only important quantity is the values of horizontal displacement.

The key points of interest for computing the local features are determined based on edges, corners and other structural primitives, and for them, a descriptor vector is computed as a specific function of their neighbourhoods. Using the nearest neighbour algorithm, the matches are determined between the descriptors from one and the other image, which are then further filtered based either on the cross-checking rule that the matching must be symmetric or by the ratio rule by D. Lowe, where the ratio of the distance to the nearest neighbour to the distance to the second closest neighbour must be less than a given value chosen typically around 0.8 based on the original paper [105].

Because of the local nature of the feature descriptors, even the filtering of not so good matches will not provide one with robust and reliable correspondences. In the literature, the solution to this is to determine the final one-value alignment for the two images not by, for example, averaging the individual displacements between the matched features, but by a histogram voting scheme, i.e. taking the modus of the sample with respect to some division of the containing interval. This works well in general when one can establish enough correspondences between the images, as statistically, a large number of not very reliable matches allows one to filter out the noise effectively. For details on this technique, one can refer, for example, to [35].

For this work, this algorithm in its original form stands as a baseline to compare to, as it has been shown to effectively perform the visual navigation over limited time spans as its effectiveness drops significantly with time. To also adopt this approach with the neural networks trained for landmark detection, an approach is tested where the feature matching algorithm is only allowed to detect features in the regions identified by the networks as landmarks.

Specifically, the feature used is the SIFT algorithm [105] with the contrast threshold parameter set to 0 as that should lead to the maximum number of detections. A larger number of features should then lead to better matching, and compared to the time requirements of the neural networks, the overhead of generating so many features is negligible, as shown later.

4.2.2 Position of Detections

The first proposed alternative to matching based on local image features is matching the detections based on the position of the detections in the image. Under the assumption that the method would detect the same landmarks in both images, one could build a matching scheme based on the minimal cost assignment problem. The simplest choice to determine the cost of matching two detections then is their euclidean distance. Obviously, the assumption that the detections will be the same is very strong. Moreover, treating the detections independently and solving the assignment problem completely neglect the structure of the environment. To deal with this, an additional constraint has to be introduced, which is the limit of vertical displacement to allow for two detections to be matched. The limit set in this work is 30 pixels which corresponds approximately to 12% of the height of the images used. Because the robot operates only in a locally planar environment, this covers the displacement caused by shaking of the camera and slight violations of the same viewpoint distance travelled assumption of the BearNav system.

To turn the matching into alignment of two images, one simply takes the differences in a horizontal position, but because the number of detections is generally not high enough, the same histogram voting scheme as used with the image features cannot be used. The first choice may seem to be to take the mean value, but as this is known to be heavily influenced by the outliers, the median was chosen instead.

4.2.3 Color Descriptors

The second proposed scheme for alignment is based on creating custom descriptors to be associated with the detected landmarks and to reduce the task back to feature matching as done in the standard way. The process then is to take the mask of the detection, overlay it over the original image and compute a function on the masked region producing a fixed-size vector descriptor.

Particularly two different descriptors based on colours were tested. The first one would simply give the channel-wise mean, i.e. produce the mean RGB colour of the segment. The second considered the image to be grayscale and computed a 50-bins histogram of its values over the interval $[0, 255]$ of colour values in a standard 8-bit grayscale image.

4.2.4 Convolutions

The third proposed matching algorithm is based on convolving the images. Convolution is generally known to align the images reasonably well—it has been used for VT&R in this way—but fails when big changes happen in the environment for example movement of dynamic objects can throw it off by introducing previously nonexistent edges. The idea here is that the network should be able to only detect a small subset of landmarks over the whole image, so it would filter the changes that cause the convolution to fail.

Again two variants have been designed. The first one only convolves the binary masks produced by overlaying the detections in one image with the same combined mask from the other image. The second further uses these combined masks to segment the original images in grayscale and tries to convolve these restricted images.

Unlike the matching between individual detections, this approach provides the alignment information directly as the shift maximizes the convolution value.

4.2.5 Template Matching

The last tested matching scheme is quite similar to the previous one—it also employs convolution but not over the whole images but detection-wise. The idea is based on template matching, where the detections by the network in the image from the current view are understood as templates, and their position is estimated in the image from the map—the teaching phase.

This approach produces again a set of displacement values that have to be consolidated into one alignment value. For this, the same scheme as with the matching based on the colour descriptors is used, where these are aggregated by taking the median value.

4.2.6 Implementation Details

During the development and experiments, few implementation details arose as important.

First, the networks output a lot of detections, and these have to be filtered to some reasonable set. Because the detections also have an associated value of confidence score, the standard approach is to select some confidence threshold which, based on preliminary testing, was set to a standard 0.5 level.

Sometimes the network also flagged most of the scenes as a plausible detection. While the reason for this can lie in such masks being present in the annotations, such detections are not very suited for aligning the images and, therefore, navigation. Because of this, another rule was introduced for filtering the detections, which is that detections which span more than one-third of the image are discarded.

4.3 Dataset

One of the main pillars of the proper testing and evaluation of the methods is having a proper dataset on which the methods can be run to produce annotations, the networks can be trained, and finally, the evaluation of the image alignment and the real-world experiment can be conducted.

The dataset used in this work comes from an experiment conducted by the Chronorobotic laboratory of the Artificial Intelligence Center at the Faculty of Electrical Engineering of the Czech Technical University in Prague with the participation of the author of this thesis, who is currently a member of said group. The experiment was done in order to test the ability of methods presented in [34] to navigate continuously over long periods of time under various lighting conditions by often letting the method repeat the same path in a structured urban exterior environment. This experiment happened in the courtyard of the Charles Square campus of the said faculty in Prague—top overview of the site is shown in Figure 12—started on the Saturday 30th of April at approximately 4 am in the morning and lasted until approximately 8 am on the Monday 2nd of May with a clear sky weather conditions. During this time, a robotic platform autonomously repeated the given trajectory at least once every hour with about two exceptions due to technical reasons and a particular focus on the time of transition from day to night and vice versa. As the method successfully navigated under the day-night changes, this experiment produced a lot of data from VT&R navigation, which is optimal for this work. In terms of the data contained, this dataset is not very complicated as it comes from a structured urban environment where the navigational structures are prominent.

The robotic platform used for the experiment and running the whole data collection was the Husky A200 platform mounted with an extra sensoric and computational tower. The computational resources present include one Jetson AGX running an ARM CPU with CUDA enabled GPU and 32GB of shared RAM and Intel NUC10i7 with i7 10th Gen x86 CPU with 64GB of RAM, both contained 1TB SSD for data recording. Both computers are interconnected via 1Gb networking and a 300Mb Wi-Fi allowing connection wireless connection of external hardware, which includes a command post with a laptop, geodetic total station and RTK-GPS base station. Sensor wise is a platform equipped with an MTI-30 IMU running at 100Hz, RTK-GPS running at 5Hz, Ouster OS0 3D lidar with 128 channels running at 10Hz and 4 Basler AcePro colour Ethernet cameras, one facing each

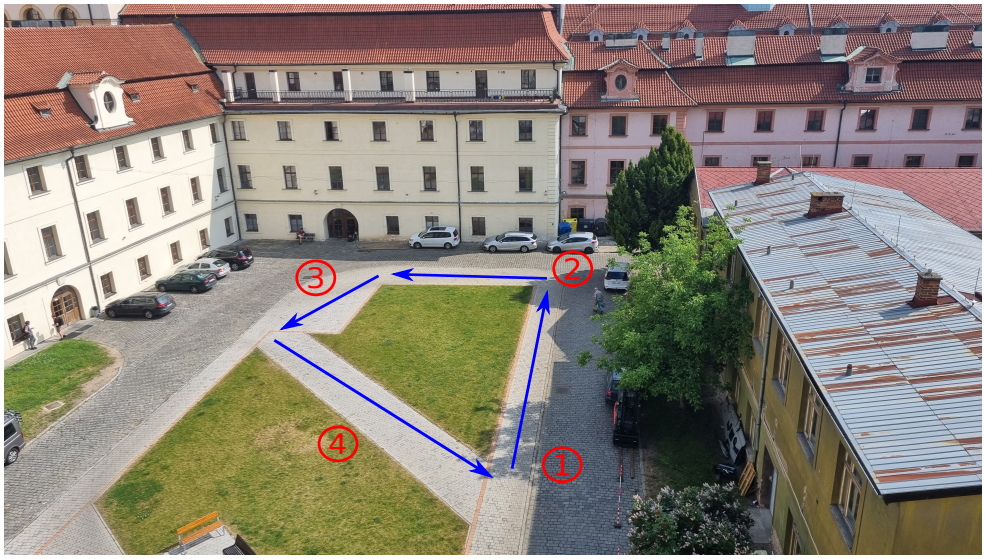


Figure 12: The view of the faculty courtyard at the Charles Square campus where the data collection and real robotic experiment took place. The path the robot was traversing is shown in blue with several special places marked. Place (1) corresponds to the starting position of the robot. Place (2) is the first turn, which depending on the error in the rotation of the initial pose, causes the first significant error in odometry, place (3) is another such turn, and place (4) denotes the position of the total station measuring the ground truth.

side with the front one running at 30Hz for navigation and the rest at 1Hz for additional and debugging data. The teaching was done using an Xbox controller, by which the robot was manually guided through the desired trajectory. The robotic platform used can be seen in Figure 13.

For collecting the ground truth on the robot's position, its tower was fitted with a geodetic 360° crystal to be tracked by a Laica TS-16 total station which gives the position of the crystal in a predefined 3D metric coordinate system with submillimeter precision. Small reflexive markers were placed on the surrounding buildings so that even after removing the total station and its reinstallation, the exact coordinate system can be easily recovered, so the experiment can be continued or repeated later. The also tested RTK-GPS was not used in the end because of the positioning of the experiment in the courtyard between buildings.

From the data produced, 24 traversals were selected at random, distributed over the whole duration of the data collection, and the stored video feed was processed to produce an image per every meter of the path based on the robot's internal odometry. This sample was taken so that it is representative of the whole dataset but reasonably big with respect to constraints in processing capabilities, given one traversal amounts to about 25-30GB of data and considering that its contents are very uniform and the network's training

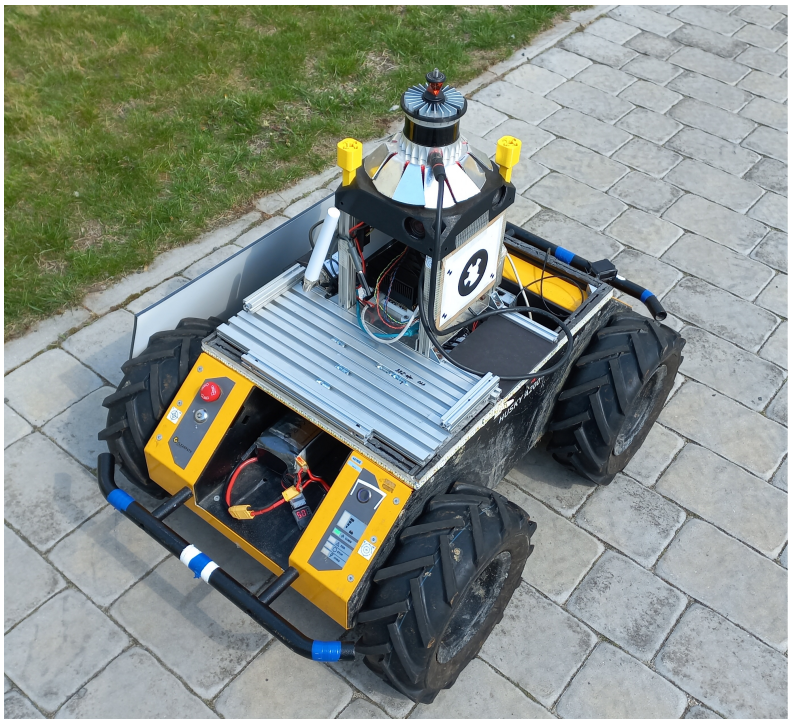


Figure 13: The robotic platform used for collecting the data and performing a real-world experiment with the navigating methods.

might be degraded if too much same data was used. This produced the set of traversals from 60 to 80 places each based on the context of the recording. To reach a uniform dataset, all were cut to 60 places for uniformity.

The complete dataset of 60 places by 24 views was split into a part for training and a part for evaluation by splitting the dataset temporarily to correspond to the way such methods would be used in real deployments. The particular division line was set to 5 am on Monday, to especially retract, for the image alignment evaluation, the data that were shot after the sunrise as the map against which the comparison is made was shot at 4 am during the night conditions and aligning two images under different conditions is harder that under the same conditions. To summarize, for the evaluation, three traversals were retracted and hand-annotated for their horizontal displacement relative to the corresponding map images by identifying 5 to 10 key points that match on both images and then averaging their displacements. Because of some faulty hand annotation, this process resulted in 174 evaluation image pairs.

4.4 Image Alignment Experiment

For the evaluation of the image alignment methods for their use in VT&R navigation, this work follows the approach used in the literature on BearNav and its derivates, pre-

sented in [27]. The method allows for qualitative and statistical evaluation of the quality of the horizontal image registration.

For each image pair i in the evaluation dataset and method m, an error e_i^m is calculated as the absolute difference between the estimated alignment a_i^m and the ground truth alignment a_i^{gt}, giving $e_i^m = \mid a_i^m - a_i^{gt} \mid$. This results in a sequence of error $\boldsymbol{e}^m$ for a particular method m. The cumulative distribution function is then estimated from the errors to evaluate the methods' performance qualitatively. Formally, this is the estimation of a probability, that a method m will have the error smaller than some threshold t, i.e. $F^m(t) = P(e^m \leq t)$. These distribution functions allow for a better understanding of how well the method performs compared to a single value metric, as well as reasoning on how to set the threshold for catastrophic failure to align the two images. In this work, this threshold for catastrophic failure is set to 70 pixels, corresponding to almost one-quarter of the aligned image.

On the sets of errors, the standard statistical methods can be used. The literature often chooses tests like t-test [27] or Wilcoxon test [34] in their pair variants and applied to all pairs of methods, but because these tests fail to control the family-wise error, they are not suited for comparing larger sets of methods. If all 29 combinations of methods in this work should be compared, the Šidak correction [106]—which dictates to set the new level of significance to $\alpha_{SID} = 1-(1-\alpha)^{\frac{1}{n}}$ where α is the original significance and n is the number of tests—would require to perform all the pairwise tests at the ridiculous level of $\alpha_{SID} = 6.3 \times 10^{-5}$. For this reason, in this work, the number of compared methods is first reduced significantly by their qualitative evaluation. As the ultimate evaluation should compare the methods for generating masks and not the methods for aligning images, one representative is selected for each mask generating method plus the baseline method based on SIFT features. Of course, most of the statistical tests also assume the independence of tested samples as the most basic condition, which is an assumption that could be challenged on almost any robotics dataset.

The procedure for selecting the representatives was not based on statistical testing but rather on the assessment of the qualities of individual methods. The literature compares the methods, as stated previously, by looking for the minimal value of the expected error. This choice is reasonable especially given the almost perfect performance of the methods on used datasets, but one can argue that the worst-case approach, which maximizes the probability of correct registration at the threshold of the critical failure, i.e. minimizes the percentage of never correctly registered images, is also of interest. This value is easily readable from the graphs of the error distribution functions F^m. In this work, the representatives were selected based on their worst-case performance, but as this was usually very similar between the alignment methods as a second ordering between the best was done by maximizing the area under the curve of the F^m which corresponds to minimizing the average error.

To evaluate the results of the alignment experiment also statistically, two tests are performed.

First, each of the selected representatives is tested whether or not its output for the

alignment is meaningful with respect to the ground truth. This test is performed by simply fitting a linear regression function on the estimated alignment as a function of the ground truth alignment and then performing the F-statistics to determine if the slope is statistically significantly not equal to zero (the model is better with the slope parameter than without it). To beware of violating the assumption of normality of the residuals too much, the outlier values—the ones corresponding to the critical failure of absolute error of more than 70 pixels—were removed from the test. Because of the way the representatives are selected, this amounts to about 5% or less samples. The assumption of heteroscedasticity or its violation does not affect too much the main goal of determining the positive relationship between the estimate and the true value as long as the samples are symmetrically distributed around the fitted line. However, no conclusions on the actual form or parameter values of this relationship can be drawn without proper testing of these assumptions. Because there are five hypotheses, their significance level of standard 5% is corrected by the aforementioned Šidak correction to $\alpha_{SID} = 0.0102$. If also the assumptions would need to be rigorously tested, the number of hypotheses would grow, and the correction would have to be stronger.

The second test, applied on methods passing the first one, the non-parametric version of ANOVA—the Kruskal-Wallis test—is used to determine whether or not there is a difference between the errors $\boldsymbol{e}^m$ of individual methods.

As the methods are supposed to run on a real robot, their runtime is also measured and evaluated as these have to be real-time able to correct the heading of the robot often enough relative to its speed.

4.5 Real-World Experiment

Based on the qualitative evaluation of the methods using the graphs of distribution functions of error, the representatives for every mask generating scheme with an image alignment strategy were evaluated through an experiment with a real robot. The representatives and a baseline method based on local image features were deployed on the same robotic platform described in Section 4.3, and this was then made to repeat the same path as was used during the data collection.

This real experiment was conducted on the 16th of May, approximately two weeks after the initial map recording and in the evening around 7 pm and with light rain to again create the hardest conditions for the methods, given that the map was recorded during the cloudless nighttime.

5 Results

In this section, all the results and comparisons are presented and commented on. First, the qualitative evaluation is performed to show how and what the neural networks were able to learn from the automatically generated masks. This is to understand how the methods proposed in Section 3 actually work and assess their actual performance compared to expectations. Then the results of the image alignment experiments are shown with a commentary, and the four representatives are selected for further testing. The representatives are statistically tested to whether their alignment does bring any information on the true alignment between the images and whether there is an actual difference between their performance. Finally, a field robotic experiment is conducted to test the ability of the methods to perform in the actual navigation scenario.

5.1 Qualitative Evaluation

To qualitatively evaluate how the methods work and what kinds of detections they lead to, this section looks at the outputs of the trained models in various situations.

(a) (b)

(c) (d)

Figure 14: A set of examples of the inference using the Mask R-CNN neural network trained using annotations automatically generated by the Autodidact method version with 32 colours discovering the landmarks from appearance change between two images.

The Figures 14, 15, 16 and 17 show several examples each of the inference of the individual methods on the pairs of two images to be aligned. The pairs were taken from the evaluation dataset for the image alignment. Therefore the one on the left always contains an image from the map, and the other data the method never saw. For each method, the inference is shown from several different places to get a representative sample,

(a) (b)

(c) (d)

Figure 15: A set of examples of the inference using the Mask R-CNN neural network trained using annotations automatically generated by the Autodidact method version with 64 colours.

as all the results would not fit the thesis. To also show how the detections are consistent, the views are duplicated to show a place that is close and possibly at a different time but where the image shows a very similar scene. For the visualisation, all the detections are overlaid over the original image, and the result of this is shown.

Figure 14 shows the behaviour of the Autodidact method quantised to 32 colours. The results, at first sight, do not suggest that the method is able to detect localised semantic landmarks. Nevertheless, under closer inspection, this is caused by overlaying the masks, as the Autodidact method simply produces larger masks that generally cover larger portions of the image, providing more segmentation into individuals than a set of localised detections. The method generally produces rather large masks expectably of surfaces of a similar colour. An example of one isolated mask generated by the method can be seen back in Section 3 in Figure 7.

The Figure 15 which shows the behaviour of the same Autodidact method version only quantising the colour space into 64 colours, exhibits a very similar kind of behaviour, but it, of course, provides slightly different detections. It seems to have higher resolution and produce smaller landmarks.

Because of the high coverage of the underlying images with the detected masks, the figures, unfortunately, do not allow us to really compare the detections made by networks trained by Autodidact between day and night.

The detections by the network trained on the masks generated from camera motion are shown in Figure 16. Clearly, this method has a bias for detecting—and therefore very

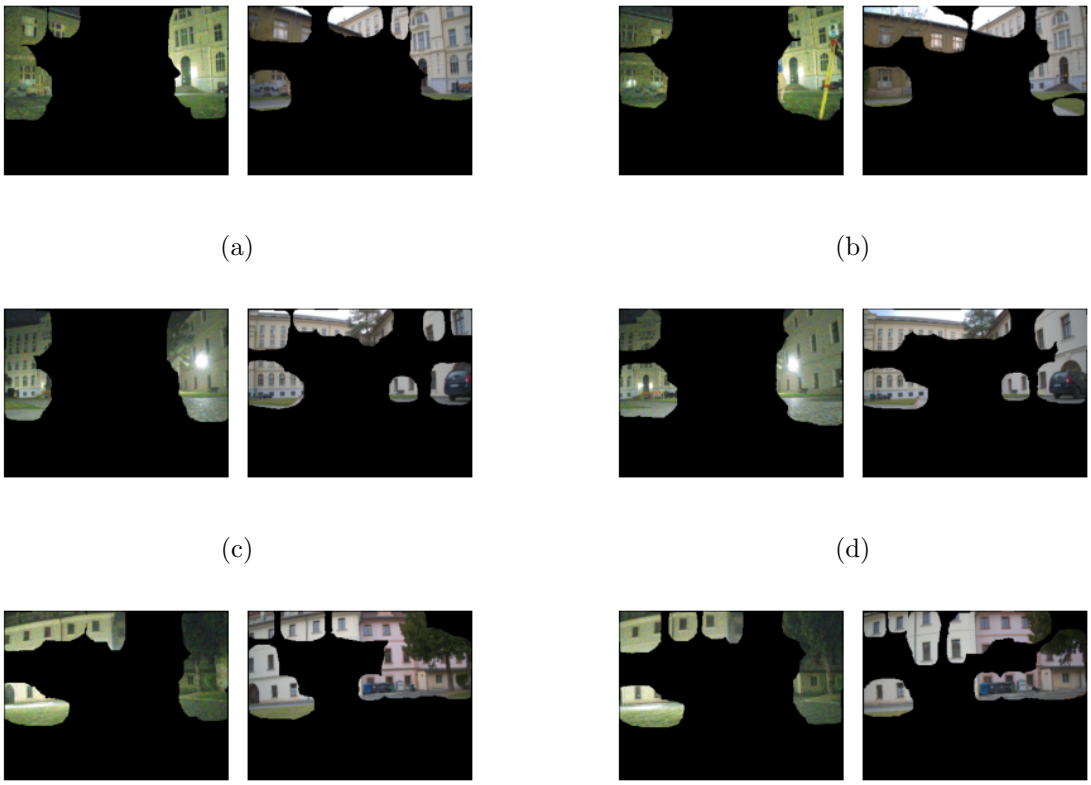
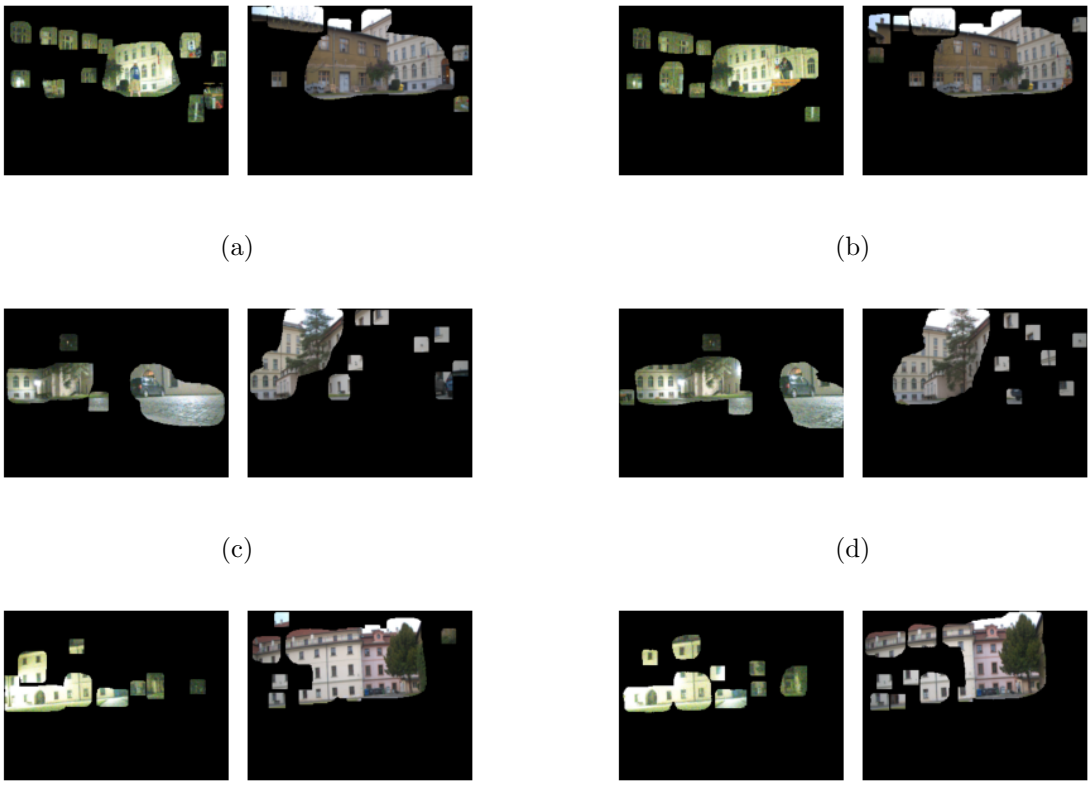


Figure 16: A set of examples of the inference using the Mask R-CNN neural network trained using annotations automatically generated by the OptFlow method discovering the landmarks from the camera motion.

likely also discovering—the landmarks closer to the edges of the picture, which corresponds to the areas where larger values of the movement-caused displacements happen. The detections on similar scenes are highly consistent, which is definitely a quality the method should have, and moreover, the detections from the day and from the night images are also consistent, which suggests the robustness of this method to the day-night appearance changes. This result is to be expected as the movement detected in the image should depend on the structure, not appearance, and the structure is not affected by the changes in lighting or other similar natural processes.

The last method presented is the method based on the analysis of long-term temporal properties of the images features in the environment. The examples of the detected masks can be seen in Figure 17. This method produces the most localised and semantically differentiated detections that seem to have the potential for well establishing the bearing direction. The obvious problem of this method is that the landmarks learned for the



(a) (b) (c) (d) (e) (f)

Figure 17: A set of examples of the inference using the Mask R-CNN neural network trained using annotations automatically generated by the Temporal method discovering landmarks based on stable image features with interesting temporal properties.

day and for the night barely correspond in the pair to be aligned. That means that the features, even when filtered to those with periodicities lower than a day, do not melange enough to provide some minimal amount of correspondences between day and night views.

5.2 Alignment Experiments

In the alignment experiments, the goal was to compare the ability of the methods to align images which is the necessary requirement to compute the steering command during the actual navigation using the BearNav system.

Figure 18 presents the estimated cumulative distribution functions for all the methods that came into existence by combining the networks trained by the mask generating methods with the methods for the image alignment. The methods are split so that the image alignment methods are compared for individual mask generating methods, and the

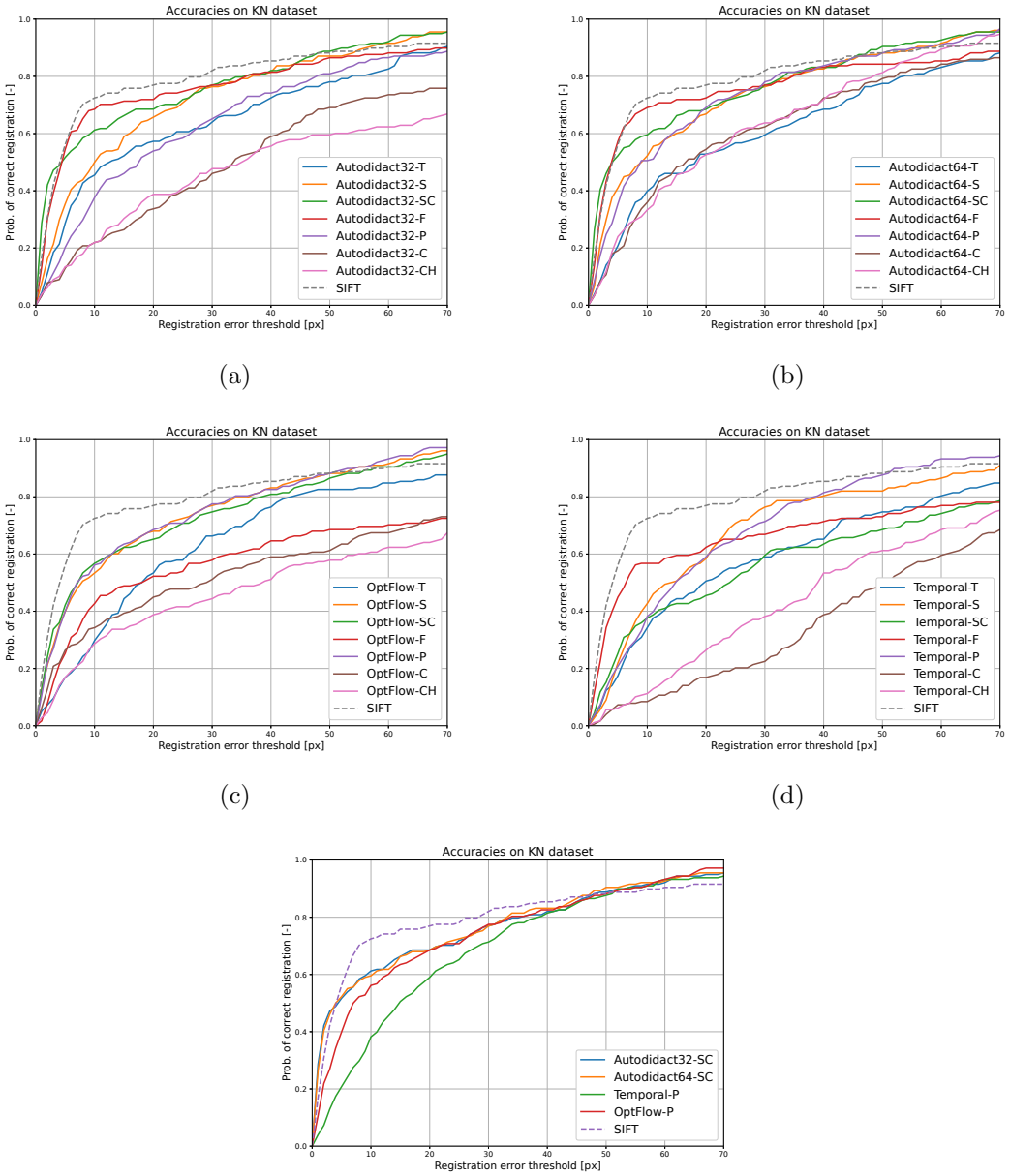


Figure 18: The resulting estimated cumulative distributions function F^m for different variants of the methods. In the top four figures, every image alignment method is compared for the individual mask generating method and the baseline alignment based on the SIFT features (distinguished by a dashed line). These figures are the base for selecting the representative for further experiments. In the bottom figure, the selected representatives are compared with each other. The labelling of the image alignment methods goes as follows: template matching (T), convolution (S), convolution with colour (SC), image features (F), position-based matching (P), mean colour descriptor (C), colour histogram descriptor (CH).

Method	Estimated slope	P-value	P-value $< \alpha_{SID}$
OptFlow	1.276	0.005	True
Temporal	0.293	0.007	True
Autodidact32	1.021	0.000	True
Autodidact64	1.046	0.000	True
SIFT	0.751	0.000	True

Table 1: Results of linear regression fitting.

representatives can therefore be selected. The selection was done by the scheme discussed in detail in Section 4.4, where the main criterion was to select the method with the lowest rate of catastrophic failures—the value of CDF to be highest at the error threshold of 70 pixels—and in case multiple methods have this value similar then select the one with the highest probability of correct registration in the left part of the graph in the change point around the threshold of 10 pixels. The selected matching schemes were—the convolution with colour information for the two Autodidact methods and matching by the minimal assignment problem over the positions for the OptFlow and Temporal mask generating methods. The figure compares the CDFs of these representatives is also shown.

The image alignment methods affect the performance of the methods greatly, and in general, the methods based on colour descriptors do not perform very well. Particularly, looking closer at the matching using SIFT image features over the regions determined by the landmark detections, the methods that perform better are the Autodidact ones that, as shown previously, detect large landmarks covering a significant part of the image. Because these still underperform the baseline method matching features across the whole image, the conclusion is that these do not benefit from preselection using the network and comparing them will only lead to selecting less useful landmarks.

The selected representatives are then compared statistically. First, the testing of whether the methods even bring any information based on the linear regression described in Section 4.4 showed that all methods exhibit a positive dependence of the estimated error on the actual ground truth alignment. The results of this test are summarised in Table 1.

The absolute errors of the individual methods were then compared using the Kruskal-Wallis test, comparing the distributions of the errors. The test allowed to reject the hypothesis at the level of significance of 0.05 with a p-value of 9.49×10^{-10}. The errors compared are shown in Figure 19, where the difference between the methods is clear at first sight. The results indicate that the methods based on the semantic landmarks are generally more robust than the SIFT features based baseline and have fewer outliers which comes at the cost of higher variance in the errors interpretable as lower precision. The mean values of the errors are shown in Table 2, and one can see that even with twice the variance compared to the baseline, the Autodidact and OptFlow methods have lower average absolute error.

Regarding the temporal requirements of the methods, the measurements during their

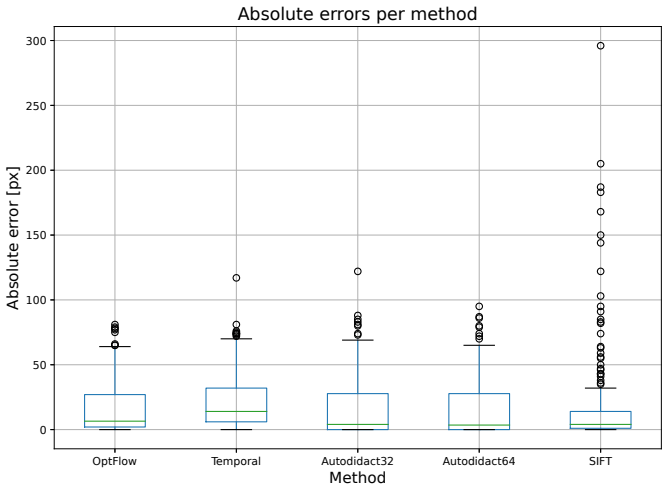


Figure 19: The distribution of the absolute error in alignment between individual methods. While the methods based on semantic landmarks have higher variance than expected, they exhibit a significantly lower amount of outliers and, therefore, can be believed to be more robust.

Method	Average error [px]
Autodidact64	**16.174**
Autodidact32	16.669
OptFlow	17.466
SIFT	19.730
Temporal	21.944

Table 2: The average error of individual methods, sorted from the best.

evaluation show that the most demanding part of the process is the evaluation of the neural network, and the image matching by any method is then negligible. The evaluation was done on a PC equipped with an AMD Ryzen 9 3950X 16-Core Processor and an NVIDIA GeForce RTX 3080 Ti dedicated GPU. This, as it later turned out, might have heavily biased the timing results as the bottleneck of the methods is the model evaluation, and the GPU used was very strong. The actual results are in Table 3 where all methods prove to be able to run at no less than 4 FPS, which is a real-time performance satisfactory for the BearNav system. The baseline not using any neural network was expectably more than twice as fast.

Method	Sec. per sample	Runtime FPS
OptFlow-P	0.246	4.067
Temporal-P	0.240	4.162
Autodidact32-SC	0.246	4.071
Autodidact64-SC	0.288	3.476
SIFT	0.103	**9.687**

Table 3: Time requirements of individual methods.

5.3 Experiment With a Real Robot

To truly evaluate the performance of the methods, it is important to test them in the actual application where they are supposed to function. For this work, the four representatives selected based on the previous results were together with the baseline method deployed on the same platform used for data collection and made to repeat the exact same path.

The results of this experiment were not conclusive, but three out of four methods were successfully deployed on a real robotic platform. Only one of the methods got lost completely—the OptFlow—suggesting some systemic error causing the method to lead the robot off of its path. The rest of the methods were able to finish the lap with some variance to the final position, but that can be expected given they are less precise in the image alignment task and would have to run longer to show whether they converge to the target path. The baseline method SIFT was the only method able to follow the map very closely all the way. The visualisation of the trajectories in the common coordinate frame given by the total station is shown in Figure 20.

When analysing the problems that might have had a negative effect on the proposed methods, two key observations were made. The total station was set up further from the sidewalk where the robot was driving. In the examples in Section 5.1 one can see that the total station that was completely stationary during the data collection has been identified as a landmark; therefore, its displacement might have affected the ability to align images in the last segment of the path. The processing power on the robotic platform equipped with an NVIDIA Jetson AGX board might have been insufficient to run the methods in their given state as the time of aligning one pair of images on the robot rose up to about 1.2 seconds which cannot be labelled as real-time and is completely unsatisfactory for robotic navigation at the speeds the robot was driving. For comparison, the SIFT-based method was still able to run at a rate close to 10 FPS which even strengthens the case for a weak GPU.

The problem with processing power could be solved in three ways. One is optimising the network for deployment on a device with limited resources with techniques like compiling the network for inference only and targeting the specific platform, pruning and other techniques. These would, at the loss of general library implementation, turn the model into a highly optimised engine. Changing the control scheme from control

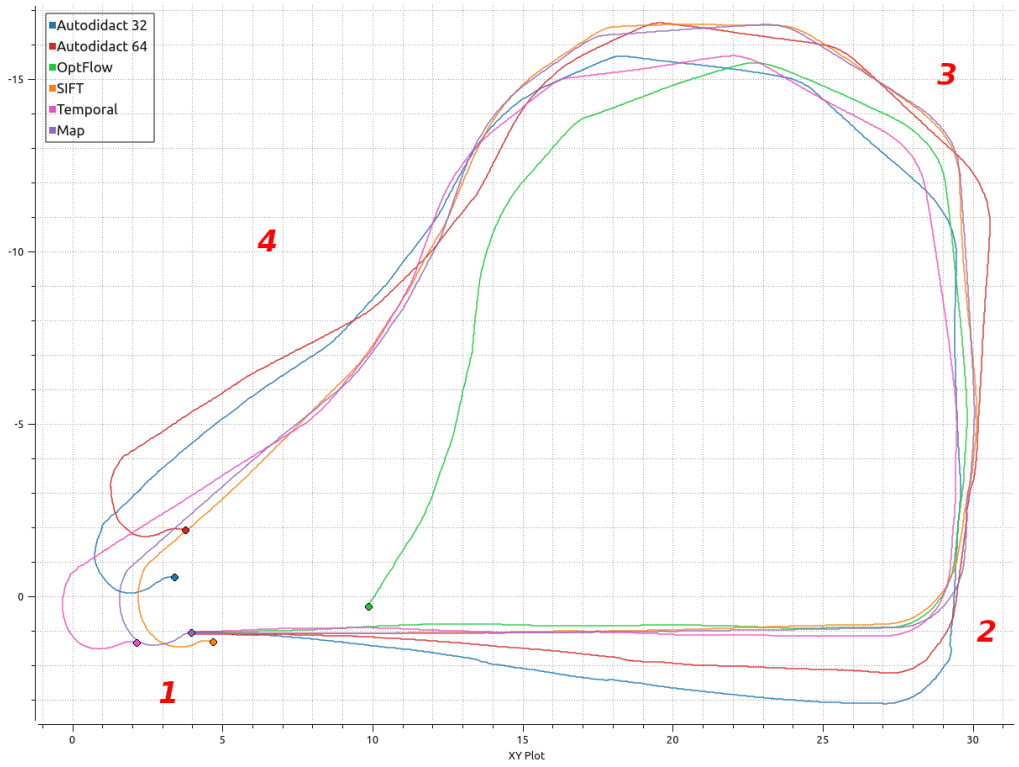


Figure 20: The 2-dimensional plot of planar trajectories performed by the evaluated methods in the effort to repeat the traversal performed during the teaching phase as recorded by the total station in the common coordinate frame, the grid size is 1 meter. The "Map" curve represents the target map trajectory, and all the others correspond to individual methods. A special method was added extra, which is the Odometry method introduced as a control sample. The numbered locations correspond to the spatial layout presented in the Figure 12 and the highlighted points are the final positions where the robot ended.

based on estimated turning speed but based on the error in heading. This would allow running a two-level controller with different frequencies eliminating the problem of only low-frequency control. The last and the simplest would be to run the robot at a significantly slower speed, which would make the task even harder as the internal odometry of the platform used should exhibit higher error during slower movement. None of these two solutions was tested for the temporal limitations as the first one is highly specialised, and the particular implementation of BearNav used does not yet allow for repeating the traversal at a different speed than was used for the teaching.

6 Conclusion

This thesis set out to propose and test methods that should, in the long run, allow autonomous systems to learn about their environment and discover useful semantic landmarks without any need for human input. Such semantic landmarks are much closer to the human path representations than current methods based on image features or black-box neural network algorithms, which could, among others, make the communication of the paths between the man and the machine easier.

6.1 Summary of the Findings

Based on the review of the literature, the thesis proposed three methods for the automatic creation of annotations for unsupervised learning of semantic landmarks over extended periods of time. One was adopted from a proof-of-concept paper [1] by which the whole task was inspired, and two were original, designed to each using a different kind of information hidden in the collected data. All methods were implemented and tested, and implementation details have also been presented in the text so that the work is reproducible.

The evaluation of the methods was a complex process that proved to be more demanding than designing and implementing the methods. To allow for integration of the methods into the BearNav navigational system, extra work was put into developing methods able to align images based on the learned and detected landmarks. The methods were evaluated qualitatively to assess their qualities and identify patterns in their functionality.

Based on the standard way of comparing and inspecting image alignment methods, the methods were evaluated. The resulting estimated CDFs of the absolute error were used to select a representative image alignment method for each mask, generating one which created a set of methods which were further tested in detail. Then to evaluate the methods, a complex set of experiments was designed and conducted. The experiments have shown the methods to be useful in the image alignment task by testing the dependence between the estimated displacement and the actual one and that the methods differ significantly, with the average absolute error of the alignment being between 16 and 22 pixels for all methods. Temporal requirements of the methods were also tested, which in the evaluation system amounted to about 4 FPS for the proposed methods and almost 10 FPS for the baseline method.

Finally, all five tested methods were integrated into a particular BearNav implementation, and a real-robotic experiment was conducted. The results of this experiment were optimistic as only one of the methods failed to complete the navigation. In comparison to the baseline method, their final position was not so precise, but that can be expected from less precise methods, and the experiment does not allow for a proper comparison. The problems identified that might have affected the results were the moving of the total station to a slightly new location or—which is more problematic—by the GPU performance drop compared to the testing environment.

To summarize the results of this work, the methods for automatic annotation generation from long-term VT&R operation were proposed and shown to work with interesting properties tied to the principle of their function. Their performance back in the visual navigation, however, was not established. In the dataset evaluation, the methods performed comparably and have shown to be more robust than traditional SIFT matching with better average performance but higher variance.

6.2 Future Work

Probably the main question for future work is whether the methods can actually be used for visual navigation. This includes investigating the image alignment based on the detection techniques or performance optimization on the computers with limited resources as are present on the robotic platforms.

Furthermore, the tests should be conducted on more complicated datasets with more heterogeneous structures or higher variance in the distance from the objects. The methods presented in this paper—especially Autodidact—might have trouble with an unstructured natural environment, but the highly structured urban environment used in this work might be too simple, so a semi-structured environment should be tested.

If the ability to navigate is established, then another important question to answer is the relative performance of the methods and their testing. The BearNav navigational system navigating in a loop is proven to converge to the correct trajectory, so a quantity to measure would, for example, be the speed of convergence. Because the presented methods are believed to be more robust, these tests should be conducted under dramatic changes of the environment. Based on an idea presented in [35] it should also be possible to fuse a more robust but less precise technique with more precise but less robust matching based on image features and achieve overall better performance.

6.3 Discussion of Prospects

The technologies studied in this work have been shown to perform well in the task of automatic mask generation of prominent structures from image data. This result is very strong for the long-term autonomy efforts because these techniques could be used for learning relatively general representations of the environment that promise to be robust and explainable with a simple camera.

The clear advantage of such an approach is that the autonomous mobile robot can adapt its models to a particular environment of operation without the need for additional manually prepared data. The ability to adapt is known to be important from the biological sciences. Moreover, creating datasets automatically without human supervision is a step toward lifelong learning systems able to update, refine and improve their perception models within their deployment.

While the field of perception has evolved considerably in the last years with higher and higher penetration of deep learning technologies, neural networks are always only as

good as the data used to train them. This quality obviously means their volume, which is crucial with the growing complexity and size of neural network architectures, but for robotic applications, even more importantly in their representativeness of the problem for which the learned models are supposed to be deployed. The so popular large computer vision datasets like the COCO dataset simply do not contain a representative sample of real-world conditions as, for example, adversary weather conditions are rarely included, which is caused by the human tendency to seek shelter instead of sacrificing comfort at the altar of science. Thus, each new project is now facing the problem of creating proper and specific datasets, which—when combined with requirements on their volumes and quality—is an investment that can prevent independent and small research teams from entering the field. This then harms the development of innovative industry nowadays, mostly revolving around the start-up companies and gives large behemoth corporations a monopoly on the data. Any methods that automate all or at least a significant portion of the dataset creation process in robotics are bound to have an immense impact on the democratization of robotic research.

Finally, regarding the specific topic of this thesis and the ability to navigate using the methods proposed, the results were simply inconclusive. Based on the qualitative evaluation, the author—and hopefully the reader as well—tends to the opinion that the methods have huge potential and present a very interesting option in the search for autonomous dataset creation. Unfortunately, because the real-world experiment of the navigation is the ultimate test of this ability and it was not favourable to the proposed methods, more investigation is needed. Assuming that the landmark detecting networks learned to detect good landmarks for navigation—note that, for example, the Temporal method was shown to be inconsistent between the day and night—the problem can either lie with the image alignment methods, which were not the main subject of this work so they might not have been studied enough or in the insufficient computational power of the robotic platform to achieve necessary real-time performance. Even if the latter was the reason, the problem might not have a simple solution, so the ability to navigate using these methods remains an open question.

References

[1] Veronika Pečonková, George Broughton, and Tomáš Krajnık. Unsupervised learning of landmarks for visual navigation in changing environments.

[2] Eric T Baumgartner and Steven B Skaar. An autonomous vision-based mobile robot. *IEEE Transactions on Automatic Control*, 39(3):493–502, 1994.

[3] Rodney Brooks. Visual map making for a mobile robot. In *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, volume 2, pages 824–829. IEEE, 1985.

[4] MWM Gamini Dissanayake, Paul Newman, Steve Clark, Hugh F Durrant-Whyte, and Michael Csorba. A solution to the simultaneous localization and map building (slam) problem. *IEEE Transactions on robotics and automation*, 17(3):229–241, 2001.

[5] Tomáš Krajník, Filip Majer, Lucie Halodová, and Tomáš 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)*, pages 1657–1664. IEEE, 2018.

[6] Dominic Dall'Osto, Tobias Fischer, and Michael Milford. Fast and robust bio-inspired teach and repeat navigation. *arXiv preprint arXiv:2010.11326*, 2020.

[7] Tristan Swedish and Ramesh Raskar. Deep visual teach and repeat on path networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1533–1542, 2018.

[8] Ashish Kumar, Saurabh Gupta, David Fouhey, Sergey Levine, and Jitendra Malik. Visual memory for robust path following. *Advances in neural information processing systems*, 31, 2018.

[9] Joshua Marshall, Timothy Barfoot, and Johan Larsson. Autonomous underground tramming for center-articulated vehicles. *Journal of Field Robotics*, 25(6-7):400–421, 2008.

[10] Paul Newman, John Leonard, Juan D Tardós, and José Neira. Explore and return: Experimental validation of real-time concurrent mapping and localization. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, volume 2, pages 1802–1809. IEEE, 2002.

[11] Yoshio Matsumoto, Masayuki Inaba, and Hirochika Inoue. Visual navigation using view-sequenced route representation. In *Proceedings of IEEE International conference on Robotics and Automation*, volume 1, pages 83–88. IEEE, 1996.

[12] Stephen D Jones, Claus Andresen, and James L Crowley. Appearance based process for visual navigation. In *Proceedings of the 1997 IEEE/RSJ International*

Conference on Intelligent Robot and Systems. Innovative Robotics for Real-World Applications. IROS'97, volume 2, pages 551–557. IEEE, 1997.

[13] Alan M Zhang and Lindsay Kleeman. Robust appearance based visual route following for navigation in large-scale outdoor environments. *The International Journal of Robotics Research*, 28(3):331–356, 2009.

[14] Tomáš Krajník, Jan Faigl, Vojtěch Vonásek, Karel Košnar, Miroslav Kulich, and Libor Přeučil. Simple yet stable bearing-only navigation. *Journal of Field Robotics*, 27(5):511–533, 2010.

[15] Lixin Tang and Shin'ichi Yuta. Vision based navigation for mobile robots in indoor environment by teaching and playing-back scheme. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, volume 3, pages 3072–3077. IEEE, 2001.

[16] T Geodeme, Tinne Tuytelaars, G Vanacker, M Nuttin, and L Van Gool. Omnidirectional sparse visual path following with occlusion-robust feature tracking. In *OMNIVIS Workshop, ICCV*. Citeseer, 2005.

[17] Zhichao Chen and Stanley T Birchfield. Qualitative vision-based mobile robot navigation. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 2686–2692. IEEE, 2006.

[18] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. IEEE, 1999.

[19] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[20] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer, 2010.

[21] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. IEEE, 2011.

[22] Tomáš Krajník, Jaime P Fentanes, Joao M Santos, and Tom Duckett. Fremen: Frequency map enhancement for long-term mobile robot autonomy in changing environments. *IEEE Transactions on Robotics*, 33(4):964–977, 2017.

[23] Christoffer Valgren and Achim J Lilienthal. Sift, surf & seasons: Appearance-based long-term localization in outdoor environments. *Robotics and Autonomous Systems*, 58(2):149–156, 2010.

[24] Tomáš Krajník, Pablo Cristóforis, Keerthy Kusumam, Peer Neubert, and Tom Duckett. Image features for visual teach-and-repeat navigation in changing environments. *Robotics and Autonomous Systems*, 88:127–141, 2017.

[25] Michael Paton, François Pomerleau, and Timothy D Barfoot. In the dead of winter: Challenging vision-based path following in extreme conditions. In *Field and Service Robotics*, pages 563–576. Springer, 2016.

[26] Winston Churchill, Chi Hay Tong, Corina Gurău, Ingmar Posner, and Paul Newman. Know your limits: Embedding localiser performance models in teach and repeat maps. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4238–4244. IEEE, 2015.

[27] Lucie Halodová, Eliška Dvořáková, Filip Majer, Jiří Ulrich, Tomáš Vintr, Keerthy Kusumam, and Tomáš Krajník. Adaptive image processing methods for outdoor autonomous vehicles. In *International Conference on Modelling and Simulation for Autonomous Systems*, pages 456–476. Springer, 2018.

[28] Winston Churchill and Paul Newman. Practice makes perfect? managing and leveraging visual experiences for lifelong navigation. In *2012 IEEE International Conference on Robotics and Automation*, pages 4525–4532. IEEE, 2012.

[29] Peter Biber, Tom Duckett, et al. Dynamic maps for long-term operation of mobile service robots. In *Robotics: science and systems*, pages 17–24. Citeseer, 2005.

[30] Feras Dayoub and Tom Duckett. An adaptive appearance-based map for long-term topological localization of mobile robots. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3364–3369. IEEE, 2008.

[31] Luis G Camara, Tomáš Pivoňka, Martin Jílek, Carl Gäbert, Karel Košnar, and Libor Přeučil. Accurate and robust teach and repeat navigation by visual place recognition: A cnn approach. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6018–6024. IEEE, 2020.

[32] George Broughton, Pavel Linder, Tomáš Rouček, Tomáš Vintr, and Tomáš Krajník. Robust image alignment for outdoor teach-and-repeat navigation. In *2021 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE.

[33] Zdeněk Rozsypálek, George Broughton, Pavel Linder, Tomáš Rouček, Keerthy Kusumam, and Tomáš Krajník. Semi-supervised learning for image alignment in teach and repeat navigation. In *Proceedings of the 37th Annual ACM Symposium on Applied Computing*, 2022.

[34] Zdeněk Rozsypálek, George Broughton, Pavel Linder, Tomáš Rouček, Jan Blaha, Leonard Mentzl, Keerthy Kusumam, and Tomáš Krajník. Contrastive learning for image registration in visual teach and repeat navigation. *Sensors*, 22(8):2975, 2022.

[35] Tomáš Rouček, Arash Sadeghi Amjadi, Zdeněk Rozsypálek, George Broughton, Jan Blaha, Keerthy Kusumam, and Tomáš Krajník. Self-supervised robust feature matching pipeline for teach and repeat navigation. *Sensors*, 22(8):2836, 2022.

[36] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.

[37] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[38] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[42] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[43] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[44] Xiao Ke, Jiawei Zou, and Yuzhen Niu. End-to-end automatic image annotation based on deep cnn and multi-label data augmentation. *IEEE Transactions on Multimedia*, 21(8):2093–2106, 2019.

[45] Stefan Hoermann, Martin Bach, and Klaus Dietmayer. Dynamic occupancy grid prediction for urban autonomous driving: A deep learning approach with fully automatic labeling. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2056–2063. IEEE, 2018.

[46] Karsten Behrendt and Jonas Witt. Deep learning lane marker segmentation from automatically generated labels. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 777–782. IEEE, 2017.

[47] Tomáš Krajník, Tomáš Vintr, George Broughton, Filip Majer, Tomáš Rouček, Jiří Ulrich, Jan Blaha, Veronika Pěčonková, and Martin Rektoris. Chronorobotics: Representing the structure of time for service robots. In *Proceedings of the 2020 4th International Symposium on Computer Science and Intelligent Control*, pages 1–8, 2020.

[48] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016.

[49] David Austin, Luke Fletcher, and Alexander Zelinsky. Mobile robotics in the long term-exploring the fourth dimension. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, volume 2, pages 613–618. IEEE, 2001.

[50] Graham D Finlayson and Steven D Hordley. Color constancy at a pixel. *JOSA A*, 18(2):253–264, 2001.

[51] Will Maddern, Alex Stewart, Colin McManus, Ben Upcroft, Winston Churchill, and Paul Newman. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China*, volume 2, page 5, 2014.

[52] Abel Gawel, Titus Cieslewski, Renaud Dubé, Mike Bosse, Roland Siegwart, and Juan Nieto. Structure-based vision-laser matching. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 182–188. IEEE, 2016.

[53] David M Rosen, Julian Mason, and John J Leonard. Towards lifelong feature-based mapping in semi-static environments. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1063–1070. IEEE, 2016.

[54] Stephanie Lowry, Gordon Wyeth, and Michael Milford. Unsupervised online learning of condition-invariant images for place recognition. In *Proc. Australasian Conf. on Robot. and Automation*. Citeseer, 2014.

[55] Nicholas Carlevaris-Bianco and Ryan M Eustice. Learning visual feature descriptors for dynamic lighting conditions. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2769–2776. IEEE, 2014.

[56] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016.

[57] Peer Neubert and Peter Protzel. Local region detector+ cnn based landmarks for practical place recognition in changing environments. In *2015 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2015.

[58] Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, 3(4):4023–4030, 2018.

[59] Winston Churchill and Paul Newman. Experience-based navigation for long-term localisation. *The International Journal of Robotics Research*, 32(14):1645–1661, 2013.

[60] Stephanie M Lowry, Michael J Milford, and Gordon F Wyeth. Transforming morning to afternoon using linear regression techniques. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 3950–3955. IEEE, 2014.

[61] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Predicting the change–a step towards life-long operation in everyday environments. *Robotics Challenges and Vision (RCV2013)*, page 17, 2014.

[62] Horia Porav, Will Maddern, and Paul Newman. Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1011–1018. IEEE, 2018.

[63] Tomáš Krajník, Jaime Pulido Fentanes, Grzegorz Cielniak, Christian Dondrup, and Tom Duckett. Spectral analysis for long-term robotic mapping. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3706–3711. IEEE, 2014.

[64] Tomáš Krajník, Jaime P Fentanes, Oscar M Mozos, Tom Duckett, Johan Ekekrantz, and Marc Hanheide. Long-term topological localisation for service robots in dynamic environments using spectral maps. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4537–4542. IEEE, 2014.

[65] Tomáš Krajník, Miroslav Kulich, Lenka Mudrová, Rares Ambrus, and Tom Duckett. Where's waldo at time t? using spatio-temporal models for mobile robot search. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2140–2146. IEEE, 2015.

[66] Claudio Coppola, Tomáš Krajník, Tom Duckett, Nicola Bellotto, et al. Learning temporal context for activity recognition. In *ECAI*, pages 107–115, 2016.

[67] Tomáš Krajník, Jaime Pulido Fentanes, Marc Hanheide, and Tom Duckett. Persistent localization and life-long mapping in changing environments using the frequency map enhancement. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4558–4563. IEEE, 2016.

[68] Jaime Pulido Fentanes, Bruno Lacerda, Tomáš Krajník, Nick Hawes, and Marc Hanheide. Now or later? predicting and maximising success of navigation actions from long-term experience. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1112–1117. IEEE, 2015.

[69] Filip Surma, Tomasz Piotr Kucner, and Masoumeh Mansouri. Multiple robots avoid humans to get the jobs done: An approach to human-aware task allocation. In *2021 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2021.

[70] Ferdian Jovan, Jeremy Wyatt, Nick Hawes, and Tomáš Krajník. A poisson-spectral model for modelling temporal patterns in human data observed by a robot. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4013–4018. IEEE, 2016.

[71] Sergi Molina, Grzegorz Cielniak, and Tom Duckett. Go with the flow: Exploration and mapping of pedestrian flow patterns from partial observations. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9725–9731. IEEE, 2019.

[72] Zhengyi Zhou and David S Matteson. Predicting ambulance demand: A spatio-temporal kernel approach. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2297–2303, 2015.

[73] Andrea Gilardi, Riccardo Borgoni, and Jorge Mateu. A spatio-temporal model for events on road networks: an application to ambulance interventions in milan. *Preface XIX 1 Plenary Sessions*, page 702, 2021.

[74] Sergio Hernán Garrido Mejía et al. Predicting crime in bogota using kernel warping. 2018.

[75] Juan S Moreno Pabón, Mateo Dulce Rubio, Yor Castaño, Alvaro J Riascos, and Paula Rodríguez Díaz. A manifold learning data enrichment methodology for homicide prediction. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pages 1–4. IEEE, 2020.

[76] Ransalu Senanayake, Simon O'Callaghan, and Fabio Ramos. Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[77] Anthony Tompkins and Fabio Ramos. Fourier feature approximations for periodic kernels in time-series modelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[78] Tomáš Krajník, Tomáš Vintr, Sergi Molina, Jaime Pulido Fentanes, Grzegorz Cielniak, Oscar Martinez Mozos, George Broughton, and Tom Duckett. Warped hypertime representations for long-term autonomy of mobile robots. *IEEE Robotics and Automation Letters*, 4(4):3310–3317, 2019.

[79] Tomáš Vintr, Zhi Yan, Kerem Eyisoy, Filip Kubiš, Jan Blaha, Jiří Ulrich, Chittaranjan S Swaminathan, Sergi Molina, Tomasz P Kucner, Martin Magnusson, et al. Natural criteria for comparison of pedestrian flow forecasting models. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11197–11204. IEEE, 2020.

[80] Tomáš Vintr, Zhi Yan, Tom Duckett, and Tomáš Krajník. Spatio-temporal representation for long-term anticipation of human presence in service robotics. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2620–2626. IEEE, 2019.

[81] Tomáš Vintr, Sergi Molina, Ransalu Senanayake, George Broughton, Zhi Yan, Jiří Ulrich, Tomasz Piotr Kucner, Chittaranjan Srinivas Swaminathan, Filip Majer, Mária Stachová, et al. Time-varying pedestrian flow models for service robots. In *2019 European Conference on Mobile Robots (ECMR)*, pages 1–7. IEEE, 2019.

[82] Tomáš Vintr, Kerem Eyisoy, Vanda Vintrová, Zhi Yan, Yassine Ruichek, and Tomáš Krajník. Spatiotemporal models of human activity for robotic patrolling. In *International Conference on Modelling and Simulation for Autonomous Systems*, pages 54–64. Springer, 2018.

[83] Martin Rektoris. Anomaly detection in periodic stochastic phenomena. B.S. thesis, Czech Technical University in Prague, 2021.

[84] Dominik Maximilián Ramík, Christophe Sabourin, Ramon Moreno, and Kurosh Madani. A machine learning based intelligent vision system for autonomous object detection and recognition. *Applied intelligence*, 40(2):358–375, 2014.

[85] Victoria Florence, Jason J Corso, and Brent Griffin. Robot-supervised learning for object segmentation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1343–1349. IEEE, 2020.

[86] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[87] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9994–10003, 2019.

[88] Tristram Southey and James J Little. Object discovery through motion, appearance and shape. In *AAAI Workshop on Cognitive Robotics*, page 9, 2006.

[89] Lina Maria Paz, Pedro Piniés, and Paul Newman. A variational approach to online road and path segmentation with monocular vision. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1633–1639. IEEE, 2015.

[90] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):909–926, 2008.

[91] Ross Finman, Thomas Whelan, Michael Kaess, and John J Leonard. Toward lifelong object segmentation from change detection in dense rgb-d maps. In *2013 European Conference on Mobile Robots*, pages 178–185. IEEE, 2013.

[92] Shuran Song, Linguang Zhang, and Jianxiong Xiao. Robot in a room: Toward perfect object recognition in closed environments. *CoRR, abs/1507.02703*, 2015.

[93] Yoshikatsu Nakajima, Keisuke Tateno, Federico Tombari, and Hideo Saito. Fast and accurate semantic mapping through geometric-based incremental segmentation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 385–392. IEEE, 2018.

[94] Nils Bore, Rares Ambrus, Patric Jensfelt, and John Folkesson. Efficient retrieval of arbitrary objects from long-term robot observations. *Robotics and Autonomous Systems*, 91:139–150, 2017.

[95] Dominic Zeng Wang, Ingmar Posner, and Paul Newman. What could move? finding cars, pedestrians and bicyclists in 3d laser data. In *2012 IEEE International Conference on Robotics and Automation*, pages 4038–4044. IEEE, 2012.

[96] Rares Ambrus, Johan Ekekrantz, John Folkesson, and Patric Jensfelt. Unsupervised learning of spatial-temporal models of objects in a long-term autonomy scenario. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5678–5685. IEEE, 2015.

[97] Thomas Fäulhammer, Rareş Ambruş, Chris Burbridge, Michael Zillich, John Folkesson, Nick Hawes, Patric Jensfelt, and Markus Vincze. Autonomous learning of object models on a mobile robot. *IEEE Robotics and Automation Letters*, 2(1):26–33, 2016.

[98] Daniel Arbuckle, Andrew Howard, and Maja Mataric. Temporal occupancy grids: a method for classifying the spatio-temporal properties of the environment. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 409–414. IEEE, 2002.

[99] Joao Machado Santos, Tomáš Krajník, Jaime Pulido Fentanes, and Tom Duckett. Lifelong information-driven exploration to complete and refine 4-d spatio-temporal maps. *IEEE Robotics and Automation Letters*, 1(2):684–691, 2016.

[100] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[101] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[102] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[103] Jan Blaha, Tomáš Vintr, and Jiři Ulrich. *Chronolib: The Chronorobotic library.* Chronorobotic Laboratory, CTU, 2022. [Available at `https://chronolib.readthedocs.io`].

[104] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[105] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[106] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

Appendix

List of attachments

This appendix lists the contents of the digital attachment by directory, including a short description of each one of them.

Directory name	Description
data	Structure and sample of data used by the rest of the software
experiments	Source codes for the experiments including the image alignment methods, neural networks training and results analysis
results	Results from which the figures for this thesis were generated
src	Source codes of the methods proposed in this thesis

Table 4: Attachment content