ANA RITA SILVA GASPAR

CLOSE-RANGE LOCALISATION FOR INSPECTION OF UNDERWATER STRUCTURES

SUPERVISOR: PROF. DOCTOR ANÍBAL CASTILHO DE COIMBRA MATOS

PH.D. THESIS

Doctoral Program in Electrical and Computer Engineering
2024, December

Ana Rita Silva Gaspar
Close-Range Localisation for Inspection of Underwater Structures
Ph.D. Thesis
2024, December.

E-MAIL:
argaspar@inesctec.pt

ABSTRACT

In the underwater world, there are some areas that require regular inspection, such as harbour infrastructure, which includes berths, quay walls, adjacent piles, ladders or bollards, to detect damage and corrosion or to clean and monitor the hulls and structures of large ships. These routine tasks need to be automated, which is why autonomous underwater vehicles are increasingly being used. To fulfil these tasks, it is crucial that these vehicles navigate in a feasible way to be successful: they must therefore have an accurate pose estimation over time and thus also recognise a previously visited place, as this is an essential aspect to compensate for accumulated pose deviations. There are several traditional sensor solutions that are typically used for positioning in this environment. However, due to their limited perception, they are not accurate and are not suitable for the close-range use required for these missions. In contrast, geophysical navigation uses information about the external environment based on magnetic, optical or acoustic sensors, i.e. their characteristics are used for localisation and thus serve as a reference for navigation. Vision-based systems can be deployed at a distance of less than 3m and offer a wealth of information, high resolution and simplicity, making them the most attractive solution for environmental sensing in robust short-range missions. It is therefore crucial that vehicles are able to select very robust features. However, underwater scenarios typically have a higher level of distortion and noise caused by factors such as wind, currents, suspended particles or physical factors related to vehicle control, so the appearance of the environment may change over time. In addition, harbour infrastructures are shallow waters, which means that they are strongly influenced by the effect of light. On the other hand, they are also usually characterised by poor visibility, i.e. turbidity due to suspended matter or even debris, and there are some areas with little or no texture. These problems lack robust features, making it difficult to perceive the scenario. However, in order to perform accurate inspection missions, it is crucial that vehicles understand their surroundings regardless of the environmental conditions.

Therefore, hybrid methods for pose estimation and loop closure based on unsupervised learning are proposed. The idea is to fully exploit the detailed information from cameras and recognise the situations where this sensor cannot provide useful data. In such situations, images from a forward-looking sonar are used. These sensors do not suffer from clouding effects and have recently proven to be a promising alternative to optical detection in difficult visibility conditions.

Inspection tasks require close-range navigation, which reduces the number of sensors that can be used due to limited perception. Despite technological developments such as automation that reduce costs, the collection of data on the marine environment is still a very costly, time-consuming and technically demanding activity. There is a gap in the available data. The data available online is still limited to certain situations and tests. They therefore do not meet the requirements or are not suitable for experiments in the intended context, i.e. there is no pose information, no ground truth of loops or synchronised data, specific trajectories without revisited areas or strong variations in the depth and speed of the vehicle. However, validation of the algorithms is crucial before they are applied in a real-world context. Therefore, a harbour scenario is created in the Stonefish simulator for initial validation and pre-testing. By selecting perception solutions depending on the context, a setup for real data acquisition is proposed to validate the effectiveness of the algorithms. Based on a preliminary evaluation of the influence of environmental conditions on the visual place recognition using raw images, it is demonstrated that these sensors are sensitive to some visibility conditions, achieving a recall of 69.57% and 35.29% in very low and low turbidity scenarios, respectively. In light and bright scenes, the recognition performance was also poor, reaching a maximum recall of 62.50%. The application of an enhancement technique has shown that this pre-processing step is effective for an adequate perception of the environment by increasing the inliers between the images by about 60%. In the best case, the algorithm achieves a recall of 75% or 66.77% for very low turbidiy and light scenes. Nevertheless, visual similarity detection is still sensitive to some appearances of the scene. Therefore, a boundary characterisation of the visual images depending on the environmental conditions is performed to understand to what extent cameras can solve perception problems in harbour scenarios. In the best case, the algorithm achieves a recall of 71.43% and 53.85% for scenarios with high and very high turbidity, respectively. It is therefore obvious that the performance in recognising similar places decreases with the degradation of the scene, which is dangerous in a navigation context. Therefore, acoustic place recognition is also proposed to understand whether the robots can fulfil their inspection tasks regardless of the challenges of the environment. To this end, knowledge of visual data is applied to acoustic images to assess whether it is effective in recognising loops, and then to exploit its potential in conditions where the cameras no longer provide such distinguishable information. In the best case, this approach only fails in 2 situations where a loop has closed, achieving a recall of 87.5%. This shows that the forward looking sonar supports loop detection in poor visibility conditions.

Based on these assumptions, a hybrid odometry method is proposed. To gain an initial understanding of the behaviour and limitations of cameras for accurate positioning, a monocular, feature-based odometry for clear water was first proposed. In this case, the enhancement technique proved to be efficient/necessary to improve the perception of the environment even under these visibility conditions. Despite some orientation deviations, the algorithm approaches the ground truth in 60% of the samples, achieving an overall average error of about 20°. Next, the limits of visibility were therefore validated for operation with cameras to estimate the vehicle position between consecutive images under difficult visibility conditions. In clear waters with very low turbidity, the algorithm is able to estimate the intended motion, achieving an overall average error of about 29°. In considerable turbidity, the perspective changes are less well perceived by the algorithm, resulting in an average total error of 49°, i.e. an increase of 20°. In scenes with high turbidity, the optical-based odometry achieves an average total error of approximately 58°, i.e. an increase of approximately 30° compared to the performance in clear water. Thus, as water turbidity increases, pose estimation performance also decreases, making the detection and tracking of robust and distinguishable features more difficult. Therefore, feature-based odometry based on acoustic images was also proposed to analyse the behaviour of pose estimation and to understand whether the robot can navigate with this data, even if it is weakly structured, regardless of the conditions encountered. The acoustic odometry achieved an average total error of about 51°. It has thus been shown to be able to provide pose information close to that of the optical method under considerable turbidity conditions, characterised by an error reduction of about 7° compared to the results obtained at high turbidity.

These methods therefore open up new horizons by improving the quality of positioning, as current solutions do not allow autonomous navigation in the vicinity of structures and cannot overcome the challenges posed by the appearance of the scene.

Keywords: Appearance-based Localisation, Autonomous Underwater Vehicles, Inspection, Machine Learning, Poor visibility, Place Recognition.

RESUMO

No mundo subaquático existem algumas áreas que requerem inspeção regular, como as infraestruturas portuárias, que incluem cais, paredes de cais, estacas de suporte, escadas ou meios de amarração, para deteção de danos e corrosão ou para limpeza e monitorização de cascos e estruturas de grandes navios. Para automatizar estas tarefas, os veículos autónomos subaquáticos têm sido cada vez mais utilizados. Para realizar estas tarefas com sucesso, estes veículos devem navegar de uma forma viável: ter uma estimação precisa de posição ao longo do tempo e reconhecer um local previamente visitado, sendo um aspeto essencial para compensar desvios de posição acumulados. Existem várias soluções de sensores tradicionais que são tipicamente usadas para posicionamento neste ambiente. No entanto, devido a limitações de foro percetual, não são precisas nem adequadas para utilização a curta distância necessária para estas missões. Em contrapartida, a navegação geofísica com base em sensores magnéticos, óticos ou acústicos utiliza informações sobre o ambiente externo, ou seja, são utilizadas características do ambiente para localização e que servem de referência para navegação. Os sistemas baseados em visão operam a uma distância inferior a 3m e oferecem grande quantidade de informação, alta resolução e simplicidade, tornando-os a solução mais atraente para a perceção do ambiente para missões robustas de curto alcance. Desta forma, os veículos devem ser capazes de selecionar características muito robustas. No entanto, os cenários subaquáticos são caracterizados por apresentarem um nível elevado de distorção e ruído causado por fatores como vento, correntes, partículas em suspensão ou mesmo fatores físicos relacionados com o controlo do veículo, pelo que a aparência do ambiente pode mudar ao longo do tempo. Além disso, as infraestruturas portuárias são ambientes de águas rasas, o que significa que são fortemente influenciadas pela exposição do sol. Por outro lado, estes cenários são também ser caracterizados por pouca visibilidade, ou seja, turbidez causada por partículas em suspensão ou mesmo lixo, e existem algumas áreas com pouca ou nenhuma textura. Estes problemas levam à carência de características robustas, o que dificulta a perceção do ambiente. Ainda assim, para realizar missões de inspeção precisas, é crucial que os veículos compreendam o seu meio envolvente, independentemente das condições ambientais.

Desta forma, são propostos métodos híbridos para estimação de posição e reconhecimento de locais, i.e. *loop closure* baseado em técnicas de aprendizagem não supervisionada. A ideia é tirar o máximo partido da informação detalhada das câmaras e reconhecer as situações em

que este sensor não consegue fornecer dados úteis. Nessas situações, são utilizadas imagens de um *forward looking sonar*. Estes sensores não sofrem com os efeitos de turbidez, sendo uma alternativa promissora à solução ótica em condições de visibilidade mais severas.

As tarefas de inspeção exigem navegação de curto alcance o que reduz o leque de sensores que podem ser usados. Apesar dos desenvolvimentos tecnológicos, como a automação, que reduzem custos, a recolha de dados sobre o ambiente marinho ainda é uma atividade muito dispendiosa, demorada e tecnicamente exigente. Portanto, ainda existe uma lacuna nos dados disponíveis. Por outro lado, aqueles que estão disponíveis *online* acabam por ser ainda restritos/limitados a situações e testes específicos, não atendendo aos requisitos nem são adequados ao contexto pretendido: não fornecem informações de posição, informação das imagens que representam o mesmo local ou dados sincronizados; apresentaram trajetórias especificas, sem locais revisitados ou com grandes variações de profundidade e velocidade do veículo. No entanto, é crucial a validação dos algoritmos antes de serem aplicados num contexto real. Neste sentido, com base no simulador *Stonefish*, é desenvolvido um cenário portuário para validação inicial e testes preliminares. Com base na seleção de soluções de perceção de acordo com o contexto, é proposto um sistema de aquisição de dados reais para validar a eficiência dos algoritmos. Com base numa avaliação preliminar da influência das condições do ambiente no reconhecimento visual de locais utilizando as imagens adquiridas, i.e. dados *raw*, é demonstrado que estes sensores são sensíveis a algumas condições de visibilidade, alcançando um *recall* de 69.57% e 35.29% em cenários com condições de turbidez muito baixas e baixas. Em cenas claras e brilhantes, o desempenho de reconhecimento ficou também aquém, atingindo um *recall* máximo de 62.50%. A aplicação de uma técnica de melhoramento de imagem mostrou que esta etapa é eficaz para uma perceção adequada do ambiente, aumentando os *inliers* entre as imagens em cerca de 60%. No melhor caso, o algoritmo atinge um *recall* de 75% ou 66.77% para cenas com alguma turbidez e incidência de luz. No entanto, a detecção de similaridade visual é ainda sensível para algumas aparências da cena. Neste contexto, é realizada uma caracterização dos limites das imagens visuais em função das condições ambientais para compreender até que ponto as câmaras podem resolver problemas de perceção em cenários portuários. No melhor caso, o algoritmo atinge um *recall* de 71.43% e 53.85% para cenários de turbidez alta e muito alta, respetivamente. É portanto óbvio que o desempenho no reconhecimento de locais semelhantes diminui com a degradação da cena, o que é perigoso num contexto de navegação. Portanto, o reconhecimento acústico de locais é proposto para ser possível compreender se os robôs podem cumprir as suas tarefas de inspeção independentemente dos desafios do ambiente.

Os conhecimentos e pressupostos dos dados visuais são aplicados a imagens acústicas para avaliar se estes dados são eficazes no reconhecimento de locais e, em seguida, para explorar o seu potencial em condições onde as câmaras já não fornecem informação distinguível. Esta abordagem falha apenas em 2 situações de *loop*, alcançando um recall de 87.5%. Isto evidencia que este sensor acústico suporta o reconhecimento de locais em condições de baixa visibilidade.

Com base nestes pressupostos, é proposto um método de odometria híbrido. Para uma análise inicial do comportamento e das limitações das câmaras para um posicionamento preciso, o método de odometria visual com base em *features* é proposto para águas claras. Neste caso, a técnica de melhoramento mostrou-se eficiente para melhorar a percepção do ambiente mesmo nestas condições de visibilidade. Apesar de alguns desvios de orientação, o algoritmo aproxima-se do *ground truth* em 60% do trajeto e atinge um erro médio total de cerca de 20^{o}. Desta forma, foram validados os limites de visibilidade para operação com câmaras para estimar a posição do veículo entre imagens consecutivas em condições de visibilidade restrita. Em águas límpidas com turbidez muito baixa, o algoritmo é capaz de estimar o movimento pretendido, atingindo um erro médio total de cerca de 29^{o}. Em turbidez considerável, as mudanças de perspectiva são menos percetivas, resultando num erro total médio de 49^{o}, ou seja, um aumento de 20^{o}. Em cenários com alta turbidez, a odometria baseada em imagens óticas atinge um erro total médio de aproximadamente 58^{o}, ou seja, um aumento de cerca de 30^{o} em comparação ao desempenho em águas claras. Assim, a *performance* do algoritmo de estimação de posicionamento também diminui com o aumento da turbidez da água, tornando mais difícil a detecção e seguimento de *features* robustas e distinguíveis. Neste sentido, é proposto um método de odometria baseada em *features* com recurso a imagens acústicas para analisar o comportamento do algoritmo de posicionamente e entender se o robô consegue navegar com esses dados, mesmo sendo fracamente estruturados, independente das condições encontradas. A odometria acústica alcança um erro total médio de cerca de 51^{o}. Demonstrou ser capaz de fornecer informações de posição similares às do método óptico sob condições de turbidez considerável e é caracterizado por uma redução de erro de cerca de 7^{o} em comparação com os resultados em alta turbidez.

Desta forma, estes métodos abrem novos horizontes ao melhorar a qualidade do posicionamento, uma vez que as soluções atuais não permitem a navegação autónoma perto de estruturas e não são capazes de lidar com os desafios da aparência da cena.

Palavras chave: Aprendizagem computacional, Fraca Visibilidade, Inspeção, Localização baseada em Aparência, Reconhecimento de Locais, Veículos Autónomos Subaquáticos.

AGRADECIMENTOS

Esta tese de Doutoramento representa não só o fim de um ciclo, mas também o alcançar de um grande objetivo pessoal: desafiante mas, acima de tudo, gratificante por vivenciar a cada dia a saída da zona de conforto. Ainda assim, tudo se tornou possível e mais leve graças ao apoio de quem percorreu este caminho ao meu lado.

Agradecer desde já, ao meu orientador, Prof. Dr. Aníbal Matos, por toda a dedicação, empenho, motivação e conhecimento que sempre me transmitiu. Por todas as palavras de incentivo, sem esquecer todas as oportunidades proporcionadas ao longo destes anos e, principalmente, pela constante troca de ideias, por me incutir liberdade de pensamento e autonomia. Sem dúvida que foram fatores determinantes para o fecho deste capítulo, e que contribuiram para o meu crescimento profissional e pessoal! Ao Prof. Dr. Nuno Cruz, enquanto Professor e Coordenador do CRAS, por todas as oportunidades proporciondas ao longos dos últimos três anos e por todos os conhecimentos e sentido crítico: todas as trocas de ideias, foram sem dúvida uma mais-valia! Agradecer ainda ao Prof. Dr. João Pavão por ter contribuído para o início deste percurso e pelo incentivo a nunca desistir dos meus objetivos. Sem eu ainda saber, preparou-me para este Mundo da Ciência, ao dizer-me "nunca te esqueças que um mau resultado, também é um resultado".

Agradeço a todos os meus colegas do CRAS, principalmente ao Jorge Barbosa, por todo o apoio, disponibilidade e ajuda na componente experimental e nas campanhas de aquisição de dados.

Não tenho palavras suficientes para agradecer aos meus pais e ao meu irmão pelo apoio incondicional, pela força e por me incentivarem a acreditar e lutar sempre pelos meus sonhos.

Aos meus avós maternos, por estarem sempre comigo, pelo amor, mimos e aconchego. Avô, acabei! É para ti! ⋆

À minha Tia Céu, por querer sempre o meu sucesso e por ser o meu "112"!

Aos meu sogros, por toda a preocupação, apoio e palavras de conforto.

Ao meu marido, Bruno, por estar incondicionalmente ao meu lado, por acreditar sempre em mim, por ser o meu maior apoio, pela preocupação e carinho. Obrigada pela força e motivação nos momentos menos bons que atravessei ao longo deste percurso. Por toda a paciência: foste sem dúvida o meu suporte, o meu porto-seguro. Sem

ti, nada disto teria sido possível! Ao Yoshi e ao Locky, por serem os meus fíeis companheiros e darem leveza aos meus dias!

À Xana, por estar comigo em todos os momentos, pela garra e ânimo, mesmo quando as próprias forças também estavam em baixo... Por lutarmos sempre lado a lado pela concretização deste objetivo! Percorrer este caminho contigo foi, sem dúvida, muito bonito! Somos uma grande equipa! Ao Jacob, pelas palavras de incentivo e apoio.

À Marta, por mesmo "longe", estar sempre "perto"! Obrigada pela força, apoio e constante preocupação ao longo deste percurso!

À Ana Luísa, à Kathy Lee e à Tânia, por continuarem sempre ao meu lado!

À Rita, ao Renato e ao Du, pelo apoio e por todos momentos que tornaram mais coloridas as fases menos boas!

Obrigada a todos por, de forma tão particular, contribuirem para o culminar desta etapa!

Rita Gaspar

OFFICIAL ACKNOWLEDGMENTS

Ana Rita Silva Gaspar would like to thank the Portuguese Government through the FCT - Fundação para a Ciência e Tecnologia - and FSE - Fundo Social Europeu through the Norte 2020 - Programa Operacional Regional do Norte - for its doctoral scholarship SFRH/BD/146460/2019, without which this research would not have been possible.

In addition, Ana Gaspar would also like to thank the Centre for Robotics and Autonomous Systems (CRAS) of INESC TEC for being the host institution and also supporting this research.

"A smooth sea never made a skilled sailor."

FRANKLIN D. ROOSEVELT

CONTENTS

1 INTRODUCTION 1
1.1 Context . . . 1
1.2 Motivation . . . 1
1.3 Objectives . . . 2
1.4 Problem Statement . . . 3
1.5 Contributions . . . 4
1.6 Thesis Overview . . . 6

I FUNDAMENTALS

2 BACKGROUND 11
2.1 Underwater Perception . . . 11
2.2 Pose Estimation . . . 19
2.3 Place Recognition . . . 23
3 STATE OF THE ART 33
3.1 Close-range Pose Estimation . . . 33
3.2 Visual Place Recognition . . . 39
3.3 Underwater Imaging Datasets . . . 44

II LOCALISATION AND LOOP CLOSURE

4 APPEARANCE-BASED PLACE RECOGNITION 51
4.1 Introduction . . . 51
4.2 Binary Bag-of-Words . . . 53
4.3 Experimental Results . . . 57
4.4 Summary . . . 63
5 TOOLS AND ROBOTIC PLATFORMS 65
5.1 Introduction . . . 65
5.2 Stonefish Simulator . . . 66
5.2.1 Scenario Setup . . . 66
5.2.2 Loop Closure Decision . . . 69
5.2.3 Datasets . . . 71
5.3 Acquisition System . . . 75
5.3.1 Sensor Setup . . . 76
5.3.2 Datasets . . . 84
6 VISUAL PLACE RECOGNITION 87
6.1 Introduction . . . 87
6.2 Impact Assessment of Environment Conditions . . . 88
6.2.1 Methodology . . . 88
6.2.2 Experimental Results . . . 90
6.2.3 Discussion . . . 98
6.3 Impact Assessment of CLAHE on Environment Conditions 99
6.3.1 Methodology . . . 100

6.3.2 Experimental Results . . . 101
6.3.3 Discussion . . . 115
6.4 Summary . . . 117
7 HYBRID PLACE RECOGNITION 119
7.1 Introduction . . . 119
7.2 Limit Characterisation of Visual Images on Environment Conditions . . . 120
7.2.1 Methodology . . . 121
7.2.2 Experimental Results . . . 121
7.2.3 Discussion . . . 128
7.3 Behaviour Assessment of Acoustic Images . . . 129
7.3.1 Methodology . . . 132
7.3.2 Experimental Results . . . 134
7.3.3 Discussion . . . 144
7.4 Summary . . . 147
8 HYBRID POSE ESTIMATION 149
8.1 Introduction . . . 149
8.2 Visual Pose Estimation . . . 150
8.2.1 Methodology . . . 151
8.2.2 Experimental Results . . . 153
8.2.3 Discussion . . . 162
8.3 Acoustic Pose Estimation . . . 164
8.3.1 Methodology . . . 164
8.3.2 Experimental Results . . . 165
8.3.3 Discussion . . . 168
8.4 Summary . . . 169

III FINAL CONSIDERATIONS

9 CONCLUSION AND FUTURE WORK 175

IV APPENDIX

A PLACE RECOGNITION IN REAL ENVIRONMENTS 183

B POSE ESTIMATION IN REAL ENVIRONMENTS 187

LIST OF FIGURES

Figure 1 Overview of the SLAM process. 20
Figure 2 Block diagram of the most important steps of a VO system. 22
Figure 3 Overview of a visual PR system. 24
Figure 4 General CBIR framework. 25
Figure 5 Overview of indexing techniques. 28
Figure 6 Illustrative example of the images captured in the UWSIM dataset, based on different turbidity levels. 44
Figure 7 Illustrative example of the images captured in the Underwater Caves Sonar & Vision dataset. . 45
Figure 8 Illustrative example of the images captured in the AQUALOC dataset. 46
Figure 9 Illustrative example of the images recorded in the AFRL dataset: Cemetery (a), bus (b) and cave (c). 47
Figure 10 Illustrative example of the images recorded by Sound Metrics in inspection applications. 48
Figure 11 Illustrative example of the vocabulary tree process with direct and inverse indices. 55
Figure 12 Block diagram of the most important steps of the BoW-based PR approach. 56
Figure 13 Example scenario from KITTI 05 (a), St. Lucia (b) and New College (c) datasets. 57
Figure 14 Comparison of ORB and BRISK descriptors against rotation, scale, viewpoint and brightness invariants of scenes from KITTI 05 dataset. 58
Figure 15 Precision-recall curves for the KITTI 05 and St. Lucia datasets when varying the similarity threshold α. 60
Figure 16 BoW-based loop closure for the KTTI 05 (a), St. Lucia (b) and New College (c) datasets. 62
Figure 17 lllustrative example of the simulated harbour scenario recreated with the Stonefish simulator. . 67
Figure 18 Illustrative example of the recreated underwater harbour floors for the rendering process in the Stonefish simulator. 67
Figure 19 Illustrative example of the rendering process for creating the terrain (a) and the texture of a pile (b) in the Stonefish simulator. 67

Figure 20 Labelling of the underwater data depending on the levels of turbidity (a) and position of the sun (b) in the Stonefish simulator. 68
Figure 21 Schematic representation of the sensor's field of view process. 70
Figure 22 Illustrative example of the images captured in the PortSeafloor datasets. 72
Figure 23 Predefined trajectories of the AUV for PortSeafloorA (a) and PortSeafloorB (b) datasets. . . 72
Figure 24 HarbourArea dataset: Illustrative example of the captured images (a) and the predefined trajectory followed by the AUV (b). 73
Figure 25 HarbourMissionCAM dataset: Illustrative example of the captured images (a) and the predefined trajectory followed by the AUV (b). . . . 74
Figure 26 Predefined trajectory of the AUV for HarbourMissionFLS dataset. 75
Figure 27 Illustrative example of the images captured in the HarbourMissionFLS dataset. 75
Figure 28 Overview of the operating modules of robotic vehicles. 76
Figure 29 Illustration of the calibration process of the underwater camera using images of a chessboard. . 78
Figure 30 Schematic representation of the perception acquisition system. 82
Figure 31 Sensor models of the perception acquisition system. 83
Figure 32 Remote-controlled underwater vehicle: CRAB. . 83
Figure 33 Perception acquisition system mounted on CRAB. 84
Figure 34 Illustrative example of optical images taken by CRAB during an operation in the Port of Ponta Delgada, Azores. 85
Figure 35 Block diagram of the proposed VPR algorithm based on optical images. 89
Figure 36 Predefined trajectories followed by the AUV for PortSeafloorA (a) and PortSeafloorB (b) datasets. 91
Figure 37 Illustrative example of the effects of turbidity (1), the position of the sun (2) and the presence of suspended particles (3) on the appearance of the images. 91
Figure 38 Precision-recall curves in all visibility scenarios for the PortSeafloorB dataset when varying the similarity threshold α. 92
Figure 39 Appearance-based loop closure for the PortSeafloorA dataset in the clear water scenario. . . 93

Figure 40 Appearance-based loop closure for the PortSeafloorB dataset for clear (a), very low turbid (b) and low turbid (c) waters. 94
Figure 41 Appearance-based loop closure for the PortSeafloorB dataset for dark (a), light and bright (c) scenes. 97
Figure 42 Appearance-based loop closure for the PortSeafloorB for water with suspended particles. . . 98
Figure 43 Block diagram of the proposed VPR algorithm based on optical images with CLAHE enhancement technique. 100
Figure 44 Overview of the CLAHE-based enhancement process. 101
Figure 45 Predefined trajectories of the AUV for PortSeafloorB (a) and HarbourArea (b) datasets. . . . 102
Figure 46 Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied at very low turbidity (a) and light scenes (b) in the PortSeafloorB dataset. 106
Figure 47 Comparison of keypoint matches with and without CLAHE-based image enhancement for a scene with very low turbidity. First row: original images; second row: CLAHE-enhanced images. 109
Figure 48 Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied for very low turbidity (a) and light (b) scenes in the HarbourArea dataset. 111
Figure 49 Appearance-based loop closure for the HarbourArea dataset in a scenario with very low turbidity when varying CL. 112
Figure 50 Appearance-based loop closure for the HarbourArea dataset in the light scenes when varying CL. 114
Figure 51 Predefined trajectory of the AUV for the HarbourMissionCAM dataset. 122
Figure 52 Illustrative example of the effects of considerable (1) and high (2) turbidity on the appearance of images. 123
Figure 53 Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied in scenes with considerable (top), high (middle) and very high (bottom) turbidity in the HarbourMissionCAM dataset. 124

Figure 54 Appearance-based loop closure for the HarbourMissionCAM dataset in the considerable turbidity at NT = [6,4]. 125
Figure 55 Appearance-based loop closure for the HarbourMissionCAM dataset in the high turbidity waters at NT = [3,2] (a) or NT = [6,4] (b) and CL = 1. 127
Figure 56 Appearance-based loop closure for the HarbourMissionCAM dataset in waters with very high turbidity at NT = [3,2] and CL = 3. 128
Figure 57 Illustration of FLS operation: The sound energy is returned based on r and θ and viewed as a map of 3D points on the null plane. 131
Figure 58 Block diagram of the proposed VPR algorithm based on acoustic (FLS) images. 133
Figure 59 Illustration of the geometry model used for FLS operation. 133
Figure 60 Illustrative example of the effectively mapped area of each FLS image. 134
Figure 61 Predefined inspection route by the AUV for HarbourMissionFLS dataset. 135
Figure 62 Comparison of detected keypoints (shown as dots) and matches (shown as straight lines) in FLS images on scenes with (top) and without texture (bottom). 136
Figure 63 Appearance-based loop closure for the HarbourMissionFLS dataset at α = 0.3 with minFPnts = 12. 138
Figure 64 Appearance-based loop closure for the HarbourMissionFLS dataset at α = 0.3 and minFPnts = 10 (top) or minFPnts = 8 (bottom). 140
Figure 65 Block diagram of the proposed VPR algorithm based on acoustic (FLS) images with CLAHE enhancement technique. 141
Figure 66 Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied in scenes with (bottom) and without texture (top) in the HarbourMissionFLS dataset. 142
Figure 67 Appearance-based loop closure for the HarbourMissionFLS dataset based on α = 0.3 and minFPnts = 12 at CL = 1 and NT = [2,3]. 143
Figure 68 Block diagram of the proposed VO algorithm based on optical images with CLAHE enhancement tecnhique. 152

Figure 69 Samples of trajectories of the AUV for PortSeafloorA (a) and HarbourMissionCAM (b) datasets, used for optical odometry. 154
Figure 70 Appearance-based odometry of a straight line for the PortSeafloorA dataset in clear water at CL = 1 and NT = [3,2] at acquired frame rate (a) and at $\frac{1}{3}$ (b) or $\frac{1}{6}$ (c) of frame rate. 155
Figure 71 Appearance-based odometry of a straight sample (S1) for the PortSeafloorA dataset in clear water at CL = 1 and NT = [3,2]. 155
Figure 72 Appearance-based odometry of a curve sample (S2) for the PortSeafloorA dataset in clear water at CL = 1 and NT = [3,2]. 157
Figure 73 Appearance-based odometry of a curve sample (S2) for the PortSeafloorA dataset at very low turbidity at CL = 2 and NT = [6,4]. 158
Figure 74 Illustration of the scene of the HarbourMissionCAM dataset, recorded with a titled (a) and a planar (b) camera. 159
Figure 75 Appearance-based odometry of an inspection sample (S1) for the HarbourMissionCAM dataset in clear water at CL = 2 and NT = [3,2]. 160
Figure 76 Appearance-based odometry of an inspection sample (S1) for the HarbourMissionCAM dataset at considerable (a) and high (b) turbidity at CL = 2 or CL = 6 and NT = [3,2]. 161
Figure 77 Block diagram of the proposed VO algorithm based on acoustic (FLS) images with CLAHE enhancement technique. 165
Figure 78 Sample of the trajectory performed by AUV for HarbourMissionFLS dataset, used for acoustic odometry. 166
Figure 79 Appearance-based odometry of an inspection sample (S1) for the HarbourMissionFLS dataset at CL = 5 and NT = [2,3]. 167
Figure 80 Illustrative example of recorded scenes of the floating piers A (a) and B (b) in the harbour of Ponta Delgada, Azores. 183
Figure 81 Illustrative example of recorded scenes of a pile in the harbour of Ponta Delgada, Azores. 185
Figure 82 Illustrative example of recorded acoustic images of concrete blocks, recorded by SoundMetrics. . 185
Figure 83 Appearance-based odometry of an inspection sample (Foating pier A) for the PortAzoresCAM dataset at CL = 1 and NT = [4,3]. 187

Figure 84 Appearance-based odometry of an inspection sample (concrete blocks) for the SoundMetrics data for raw (a) and using CLAHE (b) data at CL = 1 and NT = [5,13]. 188

LIST OF TABLES

Table 1 Overview of the specifications of perception sensors for underwater applications. 18
Table 2 Overview of specifications from KITTI 05, St. Lucia and New College datasets. 58
Table 3 Performance of the BoW-based loop closure at 100% of precision for the KITTI 05 and the St. Lucia datasets when varying the similarity threshold α. 60
Table 4 Performance of the BoW-based loop closure for the KITTI 05 and St. Lucia datasets, where the number of keypoints extracted in the vocabulary varies. 60
Table 5 Performance of the BoW-based loop closure for the KITTI 05, St. Lucia and NewCollege datasets, for the extraction of 2000 features per image. 61
Table 6 Industrial and underwater commercial laser solutions. 77
Table 7 Commercial Forward-Looking Sonar solutions. . 80
Table 8 Performance of the VPR algorithm for the PortSeafloorB when varying the water type (wt), i.e. the turbidity level. 95
Table 9 Performance of the VPR algorithm for the PortSeafloorA when varying the position of the sun, i.e. the angle of incidence. 95
Table 10 Performance of the VPR algorithm for the PortSeafloorB when varying the position of the sun, i.e. the angle of incidence. 96
Table 11 FR-IQA and functional evaluation for very low turbidity scenes using different CL. 107
Table 12 FR-IQA and functional evaluation for light scenes using different CL. 108
Table 13 Functional evaluation between images taken at different times with very low turbidity and light scenes. 109
Table 14 NR-IQA for very low turbidity scenes using different CL. 110
Table 15 NR-IQA for light scenes using different CL. . . . 110

Table 16 Performance of the VPR algorithm for the HarbourArea dataset in the very low turbidity scenario with varying CL. 113
Table 17 Performance of the VPR algorithm for the HarbourArea dataset in the light scenario with varying CL. 114
Table 18 Performance of the VPR algorithm for the HarbourArea dataset in the waters with very low turbidity and light scenes at CL = 4. 115
Table 19 Performance of the VPR algorithm for the HarbourMissionCAM dataset at different NT and CL in the considerable turbidity waters. 125
Table 20 Performance of the VPR algorithm for the HarbourMissionCAM dataset at different NT and CL in the high turbidity waters. 126
Table 21 Performance of the VPR algorithm for the HarbourMissionCAM dataset at different NT and CL in the very high turbidity waters. 126
Table 22 Functional evaluation of grid and floor regions for clear waters based on FLS images. 135
Table 23 Functional evaluation of grid and floor regions for clear waters based on optical images. 136
Table 24 Functional evaluation based on inliers for floor areas in turbid waters based on acoustic and optical images. 137
Table 25 Functional evaluation based on inliers for floor areas in light scenes based on acoustic and optical images. 137
Table 26 Performance of the VPR algorithm for the HarborMissionFLS dataset when varying the similarity threshold α, with minFPnts = 12. 139
Table 27 Performance of the VPR algorithm for the HarbourMissionFLS dataset when varying the threshold for minimum correspondences minFPnts, at α = 0.3. 139
Table 28 Performance of the VPR algorithm for the HarbourMissionFLS dataset with varying CL and threshold for minimum correspondences minFPnts, at NT = [2,3]. 143
Table 29 Performance of the VPR algorithm for the HarbourMissionFLS dataset with varying CL and threshold for minimum correspondencesminFPnts, at NT = [4,6]. 144

Table 30 Performance of the VO algorithm for the HarbourMissionCAM dataset under different water visibility conditions. 162
Table 31 Performance of the VPR algorithm for the PortAzoresCAM dataset - floating pier A - with ?? at NT = [4,3]. 184
Table 32 Performance of the VPR algorithm for the PortAzoresCAM dataset - floating pier B - with ?? at NT = [4,3]. 184
Table 33 Performance of the VPR algorithm for the PortAzoresCAM dataset - pile - with ?? at NT = [4,3]. 184
Table 34 Performance of the VPR algorithm for the SoundMetrics dataset - concrete blocks - with ?? at NT = [5,13]. 185

ACRONYMS

AHE Adaptive Histogram Equalization.
AI Artificial Intelligence.
AUV Autonomous Underwater Vehicle.

BoW Bag-of-Words.
BRIEF Binary Robust Independent Elementary Features.
BRISK Binary Robust Invariant Scalable Keypoints.
BRISQUE Blind/Referenceless Image Spatial QUality Evaluator.

CBIR Content-based Image Retrieval.
CL Clip Limit.
CLAHE Contrast Limited Adaptive Histogram Equalization.
CNN Convolutional Neural Networks.
CRAB Customisable Remotely-operated Aquatic Bot.

DOE Diffractive Optical Elements.
DVL Doppler Velocity Log.

EKF Extended Kalman Filter.

FAST Features from Accelerated Segment Test.
FLS Forward Looking Sonar.
FN False Negative.
FOV Fied of View.
FP False Positive.
FR-IQA Full reference image quality assessment.
FV Fisher Vector.

GPS Global Positioning System.

HE Histogram Equalization.

IDF Inverse Document Frequency.
IMU Inertial Measurement Unit.
IR Image Retrieval.
IS Imaging Sonar.

KLT Kanade-Lucas-Tomasi.

LbSL Laser-based Structured Light.
LS Laser Stripe.

MBS Multi-Beam Sonar.
MSE Mean Squared Error.

MSIS Mechanically Scanned Imaging Sonar.

NIQE Natural Image Quality Evaluator.
NR-IQA No reference image quality assessment.
NT Number of Tiles.

ORB Oriented FAST and Rotated BRIEF.

P3P Perpective-from-3-Points.
PF Particle Filter.
PIQUE Perception-based Image QUality Evaluator.
PR Place Recognition.
PSNR Peak Sginal to Noise Ratio.

RANSAC RANdom SAmple Consensus.
RMSE Root Mean Squared Error.
ROI Region of Interest.
ROS Robot Operating System.
ROV Remotely Operated Vehicle.
RR-IQA Reduced reference image quality assessment.

SBL Short Base Line.
SBS Single-Beam Sonar.
SIFT Scale-Invariant Feature Transform.
SL Structured Light.
SLAM Simultaneous Localisation and Mapping.
SSIM Structure Similarity Index Method.
SSS Side-Scan Sonar.
SURF Speeded-Up Robust Feature.
SV Stereo Vision.
SVD Singular Value Decomposition.

TF Term Frequency.
TF-IDF Term Frequency - Inverse Document Frequency.
TP True Positive.
TVG Time Varying Gain.

USBL Ultra-Short Baseline.
UV Ultraviolet.
UWSim Underwater Simulator.

VIR Visual Image Retrieval.
VLAD Vector of Locally Aggregated Descriptors.
VO Visual Odometry.
VoS Velocity of Sound.
VPR Visual Place Recognition.

WAVES Advanced Vehicles for Exploration of the Seas.

1 INTRODUCTION

1.1 CONTEXT

In the still unknown underwater world, there are some areas that require regular inspection, such as the harbour infrastructure, which includes berths, quay walls, adjacent piles, ladders and bollards. Here it is important to detect damage and corrosion or to clean and monitor the hulls and structures of large ships. The inspection of such structures is normally carried out by divers or Remotely Operated Vehicle (ROV). However, diving operations are often dangerous and the use of ROV requires an operator via a cable, which can limit their use in these semi-structured scenarios. Therefore, routine tasks need to be automated and so Autonomous Underwater Vehicle (AUV) are increasingly being used. To perform these tasks, these vehicles must navigate in a practical manner and recognise a previously visited place (also known as loop closure) in order to navigate successfully, as this is an essential aspect of compensating for accumulated pose deviations. However, the nature of the underwater environment can present a number of challenges to accurately positioning the vehicle, and operating in close proximity to the desired object is still a challenge due to perceptual limitations. Typically, in these scenarios, there is a higher level of distortion and noise caused by factors such as wind, currents, airborne particles or physical factors related to vehicle control. Nevertheless, vision-based systems for robust close-range operations are the most attractive solution for environmental sensing. Therefore, it is crucial that vehicles are able to understand their surroundings and deal with all environmental conditions by selecting very robust features. In this area, there is an increasing need for more accessible and accurate technologies for short-range sensing.

1.2 MOTIVATION

Autonomous vehicles can help people in some dangerous situations, such as inspecting underwater structures. In order to fulfil these tasks, it is crucial that they know their location in a feasible way over a long period of time. These missions require close-range operations. Therefore, the vehicles must be equipped with a sensory system that enables them to perceive the environment, i.e. they must be able to recognise landmarks and measure relative position regardless of

environmental conditions. Vision-based systems are the most attractive solution for environmental detection as they can be deployed at a distance of less than 3m and offer rich information, high resolution and simplicity. In harbour infrastructures, however, there are several factors that affect the quality of the images captured: Firstly, they are shallow waters, which means that they are strongly influenced by light irradiation (orientation of the sun angle), and secondly, they are usually also characterised by poor visibility conditions (turbidity) caused by the presence of suspended matter or even rubbish. These problems make it difficult to perceive the scenario. It is therefore important to find solutions to take full advantage of the detailed information provided by the cameras, even at short distances. It is also necessary/interesting to understand whether it is possible/beneficial to use other sensors in situations where the cameras no longer provide useful information, namely in scenarios with poor visibility and in featureless scenarios, i.e. without robust features, so that the robots do not lose pose information that interferes with the inspection tasks. For accurate pose estimation, it is also crucial that the vehicles are able to recognise a visited place. The loop closure mechanism corrects the errors that have accumulated during navigation. Nevertheless, this is a difficult problem as there is no inherent rule to determine whether two images are similar and show the same place or not. When using visual sensors in underwater environments, the challenge is even greater as a large amount of information needs to be mapped and the appearance changes over time due to the dynamics typical of these scenarios. Therefore, it is crucial to develop methods that allow the vehicle to recognise similar places without prior information and continuously adapt to the environmental conditions encountered. Conventional approaches to recognising locations are usually based on a training set, which is limited and critical when the robot is operating in uncontrolled environments such as underwater.

1.3 OBJECTIVES

The main objective of this work is to develop techniques for reliable localisation at close range to enable underwater vehicles to perform inspection missions in unknown and unstable conditions. It is expected to have a significant impact on underwater robotics by improving the quality of motion estimation, as current solutions still do not allow autonomous navigation close to structures due to perception limitations, as well as the transferability of the principle to different unknown environments related to inspection tasks.

This research therefore focuses on three goals:

1. To improve the perception abilities of the AUV based on the appearance/features of the environment under strict visibility conditions;

2. To reinforce navigation behaviours so that the AUV can operate close to underwater structures and cope with changing visibility conditions;

3. To equip the AUV with the ability to recognise a place based on current observations of the vehicle;

1.4 PROBLEM STATEMENT

The proposed work thus focuses on the formulation of a multisensory localisation system based on visual information and applied to a semi-structured environment that allows the AUV to deal with the environmental conditions encountered.

To this end, the work deals in particular with the following topics:

- Characterise the inherent challenges of underwater environments and understand the perception needs and operational constraints for the development of a vision-based solution: visual and acoustic data;

- Develop a method for recognising similarities between optical images based on an Artificial Intelligence (AI) agent for place recognition;

- Develop a method for recognising similarities between acoustic images for place recognition based on unsupervised learning. Acoustic data is used based on visual assumptions;

- Formulate a re-observability evaluation method depending on the input data to consistently and efficiently identify the loop closure candidates provided by the AI agent;

- Develop a positioning approach to restore the trajectory between two defined points, based on both perception solutions: visual and acoustic;

- Data collection in real environments to ensure the performance of the proposed methods

1.5 CONTRIBUTIONS

Therefore, the most important achievements are:

- Development of a test setup for real data acquisition with a camera, a Forward Looking Sonar (FLS) and a positioning system. A laser is also included to create a structured light system (camera + projector) to treat areas with no or poor texture;
- Evaluation of the use of appearance-based techniques for place recognition in a challenging environment, i.e. dynamic scenarios with changing conditions. A comparison of some feature extractors is performed to find out which extractor is best suited in such environments;
- Evaluation of the impact of the inherent perception problems of harbour scenarios on visual place recognition;
- Evaluation of the impact of applying enhancement techniques to overcome the inherent perception problems of port scenarios for visual place recognition to maximise the effectiveness of visual sensors for close-range operations;
- Characterisation of visibility boundaries for precise operations with visual sensors;
- Evaluation of the behaviour of using acoustic images for place recognition to understand whether these sensors can support robotic operations based on image processing assumptions in semi-structured scenarios where the cameras already do not provide useful information;
- Development of a method to estimate pose in close-range operations using optical images based on a characterisation of visibility limits;
- Evaluation of the behaviour of using acoustic images to estimate pose in scenarios where the cameras already do not provide detailed information;

Publications

In order to pursue these tasks, the following publications contain some of the work carried out as part of this thesis:

- **Feature-Based Place Recognition Using Forward-Looking Sonar**

  Ana Rita Gaspar and Aníbal Matos.

  Journal of Marine Science and Engineering, 2023

(10.3390/jmse11112198)

The inherent conditions of harbour scenarios pose a major challenge for the accurate estimation of a vehicle's position and the subsequent recognition of similar images. Under these operating conditions, imaging sonars are a promising solution as the behaviour of FLS is the same regardless of the environmental conditions. Therefore, an acoustic approach for unsupervised detection of similar images based on FLS data is proposed to solve the perception problems in harbour facilities. The proposed method provides a good compromise in terms of discriminability and achieves 87.5% of recall. This work opens new horizons for the use of these sensors as a great help for underwater perception, namely to avoid navigation performance degradation in muddy conditions.

- **Visual Place Recognition for Harbour Infrastructure Inspection**

  Ana Rita Gaspar, Alexandra Nunes and Aníbal Matos.

  OCEANS 2023 - Limerick, 5-8 June

  (10.1109/OCEANSLimerick52467.2023.10244576)

  The harbour infrastructure has some structures that still need to be inspected regularly. Visual sensors have a major impact on this work at close range. There are some factors that degrade the quality of the captured images, but image pre-processing steps are increasingly used. Therefore, a visual similarity detection based on an enhancement technique is proposed to overcome the inherent perception problems in a harbour scenario under some poor visibility conditions typical of these scenarios. The place recognition approach provides a good balance in coping with turbidity and light incidence at a low computational cost and achieves a performance of about 80%.

- **Limit Characterization for Visual Place Recognition in Underwater Scenes**

  Ana Rita Gaspar, Alexandra Nunes and Aníbal Matos.

  ROBOT2022: Fifth Iberian Robotics Conference - Zaragoza, 22-24 November

  (10.1007/978-3-031-21065-5)

  The nature of this environment presents a number of challenges when it comes to achieving accurate vehicle position and therefore successful image similarity detection. Although there are some factors - water turbidity or light attenuation - that affect the quality of the images captured, visual sensors have shown

a strong impact on mission scenarios - close range operations. Therefore, this research investigated whether this data is capable of overcoming the above-mentioned underwater challenges on its own. The situations where the Bag-of-Words (BoW) approach closes the loop are correctly recognised, but it is sensitive to some severe conditions and at best achieves a recall of about 77%.

- **Evaluation of Bags of Binary Words for Place Recognition in Challenging Scenarios**

  Ana Rita Gaspar, Alexandra Nunes and Aníbal Matos.

  2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) - Santa Maria da Feira, 28-29 April

  (10.1109/ICARSC52212.2021.9429799)

  To fulfil autonomous tasks, robots in the real world must be able to navigate in dynamic and unknown spaces. They must recognise previously seen places in order to compensate for accumulated position deviations. This task requires effective identification of recovered landmarks and the use of binary descriptors is increasing. The visual BoW algorithm is one of the most widely used techniques for fast and robust detection of loop closures based on appearance. Therefore, this paper presents a behavioural evaluation of a conventional BoW scheme based on Oriented FAST and Rotated BRIEF (ORB) features for image similarity detection in challenging scenarios with scene changes, perceptual aliasing conditions or dynamic elements. The BoW technique shows a good balance to deal with such challenging conditions at low computational cost.

1.6 THESIS OVERVIEW

This document is divided into 3 parts and 8 chapters and is therefore organised as described below.

The first part, entitled **"Fundamentals"**, comprises chapter 2 and chapter 3, which provide an overview of autonomous underwater perception and navigation:

- **Chapter 2** provides a better understanding of the characterisation of the underwater environment with regard to the challenges for perception. It also describes the different approaches to localisation and place recognition.

- **Chapter 3** presents recent progress in the main areas of this research, i.e. an overview of the most important work developed in

the field of underwater position estimation and place recognition techniques. It also includes a summary of currently available datasets.

The next part, entitled **"Localisation and Loop Closure"**, comprises chapter 4 to chapter 8, in which the main developments of the work proposed in this thesis are presented:

- **Chapter** 4 presents a behavioural evaluation of a conventional image similarity detection technique to understand its effectiveness in challenging scenarios. The architecture of a visual place recognition system is briefly described and the place recognition approaches are presented.
- **Chapter** 5 gives an overview of the challenges of working in underwater environments and addresses the lack of available data to meet the requirements of the planned experiments. In this context, a simulation tool is presented that can be used to validate the proposed methods in advance. In addition, an acquisition system is described that can be used to collect real data to evaluate the effectiveness of the proposed approaches.
- **Chapter** 6 proposes an approach to visual place recognition for harbour scenarios. In this way, its behaviour is evaluated under different visiblity conditons, also using an enhancement technique to understand whether cameras can overcome inherent perception problems typical of these environments.
- **Chapter** 7 presents a hybrid approach to place recognition. Therefore, a boundary characterisation of the visibility conditions for place recognition is first performed to understand to what extent the cameras can overcome perception problems. Next, an acoustic place recognition approach is proposed to understand whether the acoustic images can be used as an aid in an operational inspection mission based on visual assumptions.
- **Chapter** 8 presents a hybrid approach for estimating the robot position. First, a visual odometry based on optical data is described and a boundary characterisation of the environmental conditions is presented to obtain an accurate estimation of the vehicle position. Then, a visual odometry based on FLS images is proposed for such conditions, i.e. when the camera no longer provides distinguishable information.

The final part, entitled **"Final Considerations"**, consists of the chapter 9, in which the main conclusions of the entire work are discussed and some ideas for future work are presented.

Part I

FUNDAMENTALS

2 BACKGROUND

Autonomous Underwater Vehicle are increasingly being used to carry out inspection tasks. It is crucial that these vehicles localise themselves accurately. Therefore, it is essential to equip these vehicles with the ability to collect high-quality information to understand their surroundings and perform close-range operations. There are a variety of methods for autonomous underwater perception and navigation, which is justified by the inherent complexity of the problem in these challenging environments.

Therefore, this chapter is organised as follows: Section 2.1 provides an overview of the characterisation of the underwater environment, particularly with respect to perception. Section 2.2 describes the inherent approaches to the localisation/navigation process by autonomous vehicles. Finally, section 2.3 presents techniques for closing loops, i.e. place recognition, which is a key aspect direct linked to the success of the navigation task.

2.1 UNDERWATER PERCEPTION

The underwater world is still unknown, and so the robotic vehicles can help humans in some dangerous situations. Therefore, autonomous navigation of these vehicles is a growing area of research and application as the need for underwater data for activities such as mine clearance, building inspection, environmental monitoring or geological exploration increases. In order to fulfil these tasks, it is crucial that the vehicles can navigate autonomously in a practicable manner. To do this, an AUV must first be able to answer the question "*Where am I?*", i.e. that you are aware of your exact position. However, the nature of the underwater environment poses a number of challenges to accurate vehicle location.

There are various sensor solutions that are typically used for positioning (and vehicle navigation) in this environment: Inertial methods have very good short-term accuracy, but their position and speed estimates are known to drift over time. Acoustic navigation techniques based on beacons can provide navigation aids, usually in the form of range and bearing signals. However, they have a low update rate and the need to deploy acoustic beacons or use a vessel as a backup is highly impractical and in some cases not even feasible. In addition, the lack of electromagnetic signal propagation prevents the use of long-

range beacon networks. The Global Positioning System (GPS) provides a global position (latitude and longitude) with an accuracy of centimetres, but has certain errors that are often referred to as measurement noise. This technology has a strong attenuation that affects the electromagnetic radiation in the underwater environment [1]. As it requires the use of carefully placed beacons, the range of the GPS technology is severely limited, so the range of the vehicle is restricted to the area between the beacons. Geophysical navigation uses information about the external environment, i.e. features are used for localisation that serve as a reference for navigation with magnetic, optical or acoustic sensors. The acoustic category can be subdivided into imaging and range sensors, with feature extraction for Imaging Sonar (IS) based on image processing, as with optical sensors. For robust close-range operations, due to perception limitations, underwater vehicles must therefore be equipped with a sensory system that enables them to perceive the environment, i.e. to be able to detect landmarks and measure the relative position between them and the vehicle independently of the structure to be inspected [2]. For these applications, sonar-based (acoustic) and vision-based systems are the most attractive solutions.

Sonar sensors are the most reliable and widely used sensors for underwater height-to-bottom measurements. There are four main types: Multi-Beam Sonar (MBS), Single-Beam Sonar (SBS), Side-Scan Sonar (SSS) and FLS. These systems can be towed or mounted on a ship or underwater vehicle to perform a scan as they have a long range, making them a viable technology for bathymetric data collection. The MBS ensonifies multiple locations on the seafloor with a single ping, at a higher resolution than conventional echosounders. This feature makes the MBS an almost optimal sonar sensor for bathymetric terrain navigation. However, this is an expensive solution. The SBS also emits a single beam and can perform a three-dimensional (3D) swath by rotating the head, but data retrieval is not as fast as with an MBS. The SSS analyses the time of flight and the acoustic backscatter of the returned signal. This sensor transmits a wide swath (port and starboard channel), which can be advantageous in shallow water scenarios or close-range operations where the footprint of other sensors would necessarily be small. After suitable pre-processing of the raw data, two-dimensional acoustic images of the seabed can be obtained by recording successive swaths. To do this, the sonar must be in motion, so the movement and attitude of the sonar must be taken into account when pre-processing the data. Therefore, SSS images cannot provide an accurate/detailed visual representation of the seabed as the intensity values depend on several factors such as the graze angle, terrain surface and local water properties. However, this solution has been used for AUV localisation and navigation for landmark detection and matching, especially for known local features. FLS

uses a similar principle to the previous one, but in this case only a single, forward-facing one. It recognises vertical features and can approach underwater structures at very low speed. Feature extraction can be performed online. Although the behaviour of sonar sensors is frequency-dependent, sound waves can be transmitted over long distances (in the order of kilometres) without significant attenuation. These systems are fast, work in poor visibility, provide measurements and have detection capabilities with a large amount of information. Nevertheless, the sonar sensors are dependent on inertial sensors, i.e. a small movement error can have a major impact on the subsequent position determination.

The visual systems are hardly considered for underwater applications because there are several factors that degrade the quality of the captured images (they tend to be blurred and/or poor in colour), such as:

- interfering noise, caused by suspended particles in the water - bubbles, small sand particles and small fish or plants that inhabit the aquatic ecosystem. These particles block the light and the cameras produce noisy images with distorted colours;
- refraction of light, which means that the objects in the image are distorted differently depending on the environment. Therefore, the positioning is not accurate; and
- attenuation of light, as the light intensity decreases with increasing distance to the objects. Therefore, the colours of the observed underwater objects look different from those perceived in the air.

Nevertheless, the cameras have shown a strong impact at short distances, as in these situations the visibility is usually good enough and the measurement accuracy achieved is higher than with other sensors, even if it is necessary to apply some image quality enhancement [3]. In addition, these systems provide rich/comprehensive information and high resolution with fast acquisition. Therefore, visual positioning is suitable for operational scenarios (at close range) such as docking operations, object detection, manipulation, seabed reconstruction, and underwater inspection and maintenance, where work locations change frequently and are limited to relatively small areas [4]. The image processing systems can be monocular or stereo. The monocular system consists of one camera and therefore only displays the environment in two dimensions. Consequently, this solution is not able to perform motion estimation without scale ambiguity, i.e. the scaling factor is unknown. In contrast, two or more cameras, i.e. Stereo Vision (SV) systems, can determine the metric scale based on the known baseline between both sensors (assuming the stereo system is calibrated). Stereoscopy is a triangulation method in which images of an object or scene

are captured from two or more angles. Image features can be recognised from the images from these cameras and compared between the left and right images to calculate 3D correspondences. In order to know the relative position of one camera in relation to the other, the camera system must be calibrated and the scale ambiguity resolved. Despite the advantages of stereo systems over monocular cameras, embedding a single camera is much more practical as classic camera housings can be used and problems with camera synchronisation are avoided. In addition, the development of monocular solutions is transferable to any type of underwater vehicle as long as it is equipped with a camera. Even if there is a projective ambiguity in monocular systems, it is possible to retrieve the scaling factor from an external source of information, i.e. from a complementary sensor that can measure a metric quantity. However, the methods based on image processing rely on the association of features and landmarks. To perceive a scene, one can consider local or global features: Local features refer to a pattern or a specific structure in an image, e.g. a dot, an edge or a small section of the image. They are usually associated with a section of the image that differs from its immediate surroundings in terms of texture, colour or intensity. Global features provide information that allows the viewer to quickly identify the essence of a scene, i.e. they are associated with specific types of scenes, such as the degree of naturalness, the degree of openness, the degree of roughness, the degree of expansiveness and the colour. However, the seabed and some underwater infrastructure such as dams or locks, which need to be inspected regularly to detect damage or corrosion, may not be structured, i.e. they may be featureless. The use of such techniques can then usually be limited to specific underwater regions.

Structured Light (SL) systems can therefore be used to help vehicles perceive their surroundings and navigate in front of backgrounds with little or no structure. These systems consist of a camera and a projector and are based on the triangulation principle between these two elements and the projected scene. This solution can be pattern-based, where the projector casts a known pattern onto the scene, i.e. usually a series of light planes, or laser-based, where a laser light is projected into the environment [5]. This system therefore projects patterns of light into the scene, giving an object the illusion of texture. This increases the number of identifiable features for each individual position in the image, minimising the difficulty of establishing correspondences and overcoming the problems of the camera on surfaces with few features. Subsequently, the information about how the patterns look after they have been distorted by the scene is used to finally recover the 3D geometry. Typically, SL systems make the basic assumption that each point of the scene receives more direct illumination from the light source than indirect illumination, usually

referred to as global illumination, and includes inter-reflection, sub-surface scattering and projector defocus. However, this assumption can be violated. Projector defocus is the main source of error for most structured light methods - the projected identifiable features may not be correctly recognised, leading to errors in the creation of correspondences. However, various methods have been developed to process scenes with global illumination or projector defocus effects. Nonetheless, it is the potential speed of data acquisition, the non-contact nature, the availability of the required hardware and the high measurement accuracy offered by modern structured light 3D scanning technologies that ensure they are utilised in many fields. There is promising evidence - accuracies in the order of *mm* when used within short distances [6] - but has not yet been extensively studied in this environment due to the nature of the underwater environment, which can lead to absorption and scattering.

A pattern-based structured light system (referred to here as PbSL) casts different patterns on the scene to create features. Most existing methods can be categorised into temporal and spatial patterns based on the number of projection patterns required. The spatial patterns, which require only a single shot, are used for dynamic scenes as the codes are not monochromatic. In addition, a single pattern enables very fast acquisition and simple optical elements with a regular frame rate (similar to cameras) are used, which enables the use of this system in a ROV and real time. In addition, these techniques are considered the most robust to distance limitations and motion artefacts. But usually the correspondences are sparse - only correspondences along the stripe edges or the grid corners - and no high resolution is achieved for reconstruction and generation models. Spatial methods therefore summarise all techniques in which the codeword of a specific position is extracted from the surrounding points. Intensity or colour variations are used to generate the codeword. The different strategies within this group can be categorised as follows: De Bruijn patterns and pseudo-random grid patterns. The De Bruijn sequence is one of the most commonly used spatial coding patterns and can be generated using a publicly available generator. It has the property of windowed uniqueness so that global optimisation can be applied to produce correspondences. It is usually a striped pattern. This method is easy to use and understand, but some types of surfaces do not correctly reflect the incoming colour spectra back to the camera. It is also prone to noise and does not achieve high resolution. Typically, a single pattern allows for very fast acquisition, so this system can be used in a ROV. This technique can also be used in combination with a video camera for close-range mapping [6]. The pattern is projected onto the scene and the camera records the image. Next, these methods usually require an image processing step to extract

the stripe edges in the captured image. Finally, a global optimisation method is applied to match the extracted stripe edges in the captured image and the projected pattern. Once a set of matches is obtained, triangulation can be applied to reconstruct the object surface. The differences between the various proposed methods usually consist of three aspects: Patterns, image processing techniques and global optimisation methods. In addition to fringe patterns, grid patterns are often used. Using pseudo-random sequences, a grid patterns can be generated in which the grid corners can be marked with dots or other primitives. Within this array, each sub-window of a certain size that slides over the entire array is uniquely coded. There are patterns consisting only of black and white colours, and colour-coded grid patterns are also commonly used, which are an extension of the De Bruijn striped pattern. By extending the coding strategy to the vertical and horizontal directions and some other modifications, a 2D grid indexing pattern can be designed. Depending on the application, the coding strategies for the two directions can be either the same or different. The accuracy of spatial coding methods can be affected by the robustness of the image processing step in detecting the stripe edges or the grid corners. In addition, these methods can be sensitive to the colour of the object, as most patterns are designed in RGB colour space: Fewer colours mean less sensitivity to noise. On the other hand, the colours must be chosen to improve the accuracy of the 3D image and limit the required acquisition time.

Instead, temporal methods, known as time multiplexing, are the earliest and most commonly used. As the name suggests, a series of patterns are projected onto the surface of an object, and each pixel is encoded by a sequence of illumination values by varying the projected pattern in time. The last category - spatial patterns - generates features within a single pattern using the neighbourhood information in the pattern. The temporal patterns achieve denser and more accurate results, even in moving scenarios. However, they cannot be used in dynamic scenes, as the projection of multiple patterns is required and the different captured images move, so error in the code can occur due to the movement. These systems are very sensitive to the non-linearities of the camera, which reduces the accuracy and sensitivity to details on the target surface. The temporal methods can be:

- Binary pattern: uses only two states (black and white) of light stripes in the scene, usually with light subdivision of the previous pattern until the software can no longer segment two consecutive stripes. These patterns are the most commonly used as they are easy to obtain with a projector and simple to process. In addition, this solution is robust, namely against noise and changes in the object surface;

- *N*-ary pattern: uses more than two intensity values (from 0 to 255) to encode a scene. Using many intensity values allows an increase in potential sensing speed as the number of required patterns can be reduced. However, this solution is sensitive to the focussing plane of the projector and the camera;

- *Phase-Shifting* pattern: was created to avoid the mentioned problems of the last pattern (N-ary) by using the grey scale. There are different types of patterns that can be used: triangular, trapezoidal and sinusoidal phase encodings. The triangular patterns are relatively simple and fast, but they are sensitive to the gamma non-linearity of the projector and the defocusing of the image. The trapezoidal pattern copes better with image noise, but can still be somewhat sensitive to image defocus. To solve this problem, sine patterns - also known as fringe patterns - can be used. These patterns also increase the accuracy of the method used, as the match between camera and projector can be achieved more precisely. The method is also robust against changes in the surface structure.

A Laser-based Structured Light (LbSL) system projects a laser light into the environment (fixed line). This system can overcome the limitation of light in the medium and provide sparse and fast three-dimensional information over a small area, e.g. for autonomous manipulation - it is attractive when precise measurements of reduced areas are required [2]. In addition, this type of structured light is an optimal solution for moving objects compared to pattern-based systems. The projector emits a known laser pattern whose shape, deformed by the target area, is recovered by a camera. Depending on where the projector is located in relation to the camera (e.g. on the baseline), a relationship between range and resolution can be established. There is clearly a trade-off between spatial resolution and range: the greater the distance to the target surface to be measured, the larger the size of a pixel on the target, which decreases the resolution. The light intensity of the laser is of critical importance. The brightness of the laser must be clear and sharp compared to the background light/scene, i.e. system projects working in depth where there is no ambient light and artificial light is controlled or reduced. Lasers are available in many colours, with specific wavelengths available for each colour. The colour of the laser affects both the visibility and the penetration of the laser into the water medium. In general, the shorter the wavelength, the better the laser penetrates; green lasers penetrate turbid water better than red lasers. However, green lasers require much more power to operate (10× the operating current for the same output power). Green lasers are also much more sensitive to temperature fluctuations. Only the most expensive models have a comparable operating temperature

range to ordinary red laser modules. The Laser Stripe (LS) is a subset of these systems where the pattern is also a fixed line, but the projector pans across the camera's field of view. This configuration requires a motorised element where the system always holds the camera and the laser does not move. These systems usually have a higher resolution than stereoscopy, but they are limited by absorption and scattering. However, there is already a solution that reduces the scattering effects, but the setup is more complex and costly [7].

Table 1 shows a comparison of some perception solutions according to their typical range and resolution. The most important strengths and weaknesses of the individual solutions are also summarised.

Table 1: Overview of the specifications of perception sensors for underwater applications.

Sensor	Range	Strength	Weakness
IS	150 m	- Water turbidity - Work at close range - Feature information	- Low frame rate - Low contrast - Unknown dimension
SSS	150 m	- Water turbidity - High frame rate	- Needs constant speed - Unknown dimension
SV	3 m	- Simple - High accuracy	- Low frame rate - Visibility conditions
PbSL	3 m	- High frame rate - Poor texture	- Illumination conditions - Occlusions and shadows - Moving objects
LbSL	3 m	- High frame rate - Poor texture - Complex lighting conditions	- Occlusions and shadows - Light colour and intensity

To fulfil inspection tasks, vehicles must therefore have a multisensory system to be able to perceive the environment under different conditions. However, the selection of a 3D sensor system for these applications is not trivial, because behind the characteristics of each perception sensor there are some fundamental aspects that should be taken into account, such as the volume of the payload, the weight and available power, the measurement time, the budget and the expected quality of the acquired data.

2.2 POSE ESTIMATION

Autonomous navigation of underwater vehicles, i.e. an accurate localisation is crucial to enable the vehicles to fulfil demanding tasks, namely in regions with poor visibility or in areas with obstacles or that are difficult to access, such as monitoring and inspecting infrastructure. These tasks require close-range navigation, which reduces the number of sensors that can be deployed due to perception limitations. Typically, these scenarios are more characterised by distortion and noise and can be featureless, making the perception of the environment (feature extraction) a challenge. Therefore, it is important to select very robust features, regardless of the type of structure under investigation and even in low light conditions or from different viewing angles. In short, the visual sensors are attractive because they allow the extraction of landmarks taking into account 2D and 3D information. In addition, these systems are a cost-effective solution, operate at a range of less than 3m and offer high resolution, rich information and simplicity. However, these passive sensors require textured scenes to achieve satisfactory results, as typically the final reconstructions lack parts that correspond to non-textured regions. In addition, the cameras are very sensitive to water turbidity, which can affect the adaptability of the sensor to the location problem. The structured light system can therefore operate at high resolution even at short distances (less than 3m). It has a slow acquisition time caused by the projection of different patterns, which can limit the use of these systems in real conditions. In addition, this solution is strongly influenced by the lighting and contrast conditions of dark objects. Shiny objects also pose a challenge for the structured light system, as the reflected light can mislead the pattern decoder. As this solution illuminates the operating scene, it is also affected by scattering, which reduces its range. However, as this sensor casts features onto the environment, it also works on objects without texture. Sonar-based acoustic systems are a costly solution, as the speed of sound propagation depends on water temperature and salinity, and their proper operation can depend on the type of materials [7]. However, these sensors are independent of water turbidity and there are also sonar sensors that are suitable for shorter distances. These solutions can perceive the environment, allowing that the vehicles have the capability to perform inspection tasks – close-range navigation. Even so, the success of the autonomous navigation includes three essential main tasks:

- **localisation**, that determines the current pose of the robot in an environment from sensory information;
- **mapping**, that integrates the surrounding observation into a single consistent map;

- **path planning**, responsible for determining the best route in the map to a vehicle navigates in the environment.

Mapping and localisation can be studied independently, although it has been recognised that accurate localisation in an environment requires a correct map and that, consequently, adequate and accurate localisation is necessary for the creation of a good map when the features and landmarks of the environment are added to the map. Currently, this problem is known as Simultaneous Localisation and Mapping (SLAM) [8]: a process in which a robot equipped with onboard sensors creates a map of the environment through feature extraction and simultaneously uses this map for its location. A SLAM system therefore consists of two main components: the front-end (low level) and the back-end (high level). The front-end module extracts significant features from the sensor data, i.e. it abstracts the sensor data into models suitable for estimation, and it associates each measurement with a specific landmark (e.g. a 3D point) in the environment (elements that are already included in the robot's map) - data association - a fundamental key aspect of SLAM. The back-end module, on the other hand, draws conclusions from the abstracted data generated by the front-end. At this level, SLAM is a mixture of geometry, graph theory, optimisation and probabilistic estimation. In addition, the back-end can provide feedback to the front-end as explained below in section 2.3. SLAM can be divided into feature-based methods, where the features are measured by the sensors, and view-based methods (also called pose-based methods), where the entire view of each pose is stored and processed.

Figure 1 illustrates an overview of the SLAM process.

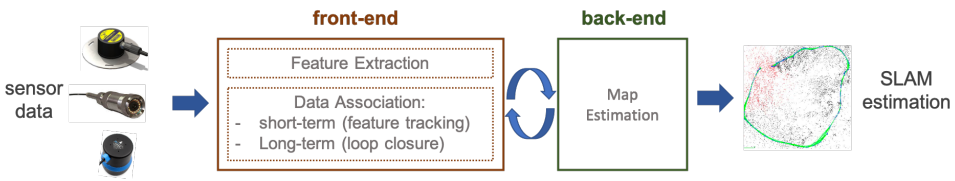

Figure 1: Overview of the SLAM process.

This technique is widely used in ground robotics, but there is great interest in the application of SLAM in the underwater world. In this environment, the most common SLAM approaches are based on the Extended Kalman Filter (EKF), the Particle Filter (PF) – FastSLAM – and Graph SLAM [9]. The limited visibility, the constantly changing conditions and the inability to use conventional sensors such as laser range finder make this environment a challenge and one of the most difficult environments. Several underwater SLAM algorithms using either optical or sonar sensors have played an important role in utilising this technique for close-range applications [10].

When using optical sensors - visual SLAM or vSLAM - the robot uses only the images from one or more cameras (monocular or stereo system, as mentioned in the previous section 2.1) as source information to determine its position while creating a representation of the surveillance area. This solution offers high resolution and update rate at relatively low cost, but is a challenge when operating in turbid waters. In addition, data association is more challenging with cameras due to the amount of data to be mapped and the high level of noise. With sonar sensors, on the other hand, the vehicle uses acoustic signals to simultaneously map the 3D positions of multiple sound sources and localise the observer within the scene map - aSLAM. These sensors can be used in environments with turbid water, as the sound is hardly attenuated in the water. In addition, imaging sonars can utilise three-dimensional features of the environment obtained after data processing, making them much less dependent on visibility than classic 2D features. However, their cost is much higher compared to cameras and their resolution and frame rate are lower. The underwater environment is dynamic and unknown, which increases the difficulties in data association due to many or very few salient features or sensor occlusions [7]. However, efficient data matching and correct data association are crucial for the success of navigation, as errors can quickly lead to incorrect maps.

A map can support other tasks or provide an intuitive visualisation for a human operator and allows to limit the error in state estimation as it is a set of distinguishable landmarks [11]. Nevertheless, localisation is the main task for autonomous vehicles to follow their path and correctly detect and avoid obstacles, namely when the vehicle needs to know its current position in order to perform a defined inspection task in a semi-structured scenario, as it is intended and the focus here. In this context, autonomous vehicles need to use data from motion sensors to estimate their position relative to their initial position, i.e. to estimate the change in robot position (translation and orientation) over time. This process is called odometry. Considering the effectiveness of visual sensors in underwater environments, Visual Odometry (VO) is one of the robust vehicle localisation techniques that has been an important research topic in recent decades, especially in underwater environments where localisation using image processing is even more complex than in air and land environments and modern open-source algorithms fail when the operating conditions become too harsh. VO is a localisation method that uses only an image stream captured by a monocular or stereo system, i.e. it provides an incremental online estimate of a vehicle's position by analysing the image sequences captured by a camera [12] [13]. Therefore, vSLAM differs from VO by maintaining a reusable global map that allows the detection of loop closures if it has already been remapped.

There are two categories for positioning with image-based odometry: geometry-based methods and deep learning-based methods.

The geometry-based methods can be further subdivided into feature-based approaches, appearance-based approaches (also called direct methods) or a hybrid of feature- and appearance-based approaches. Feature-based approaches extract the feature points of the image and track them in the image sequence. The goal of this feature point method is to minimise the error in the reprojection, which is usually achieved by calculating the difference between the coordinate values of the pixels. Direct methods consider the entire image and use the pixel intensity in the image sequence directly as visual input [14], minimise the photometric error and the function takes the form of a subtraction of the grey level values of the pixels. These methods are more suitable than the feature method when there are many repeating textures in the environment and there is a lack of corner points. However, the direct method is relatively computationally intensive and usually requires GPU-based computational acceleration. The feature-based approach is therefore particularly interesting in this context as it is suitable for textured scenarios, can deal with occlusions, is not dependent/sensitive to changes in the appearance of the captured images and the intensity of the pixel information they contain, and is fast. Provides sparse motion fields but potentially robust tracking and is flexible: outliers can be dealt with retrospectively.

A general flow of the VO process is summarised in Figure 2.

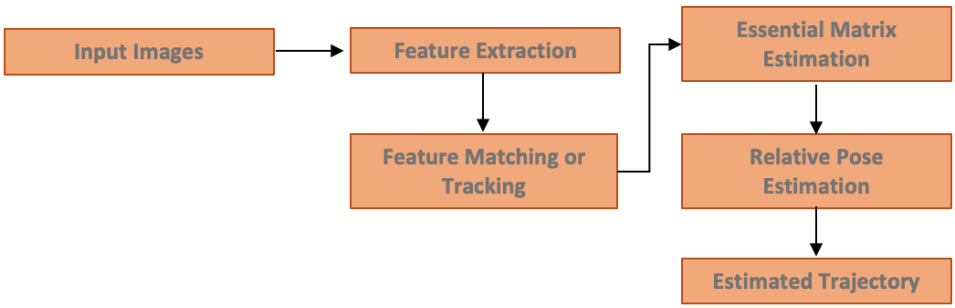

Figure 2: Block diagram of the most important steps of a VO system.

In short, for each new image, features are detected in a first step and compared with those of the previous image on the basis of feature matching or feature tracking. There are therefore two ways of estimating correspondences: Tracking for high temporal resolution and Matching for spatial resolution. In matching, the features are detected independently in all images and then matched based on some similarity metrics by comparing descriptors, which are described below in section 2.3. In feature tracking, the features are found in the first image and then tracked in the next images using a local search technique. Based on a performance evaluation of different combinations of detec-

tors and descriptors and the optical flow-based Kanade-Lucas-Tomasi (KLT) with Harris corner extraction on underwater images (two sets of 20 images), the tracking method was able to track the highest number of features and showed more robustness to the poor quality of underwater images [15]. This could be because the low texture of the images leads to the extraction of ambiguous descriptors that cannot be robustly matched, whereas the KLT method, which aims to minimise photometric cost, is less prone to these ambiguities. In addition, this method is very computationally efficient, as new features do not need to be extracted for each image. However, unlike descriptors, the KLT are only intended for tracking features between consecutive images. This can be a problem in a dynamic environment as good features can be lost due to short occlusions. A re-tracking mechanism, which has been considered in state-of-the-art approaches, can be used to make the KLT robust to such events.

Based on the matches obtained regardless of the technique used to estimate the correspondences, the relative motion is computed from the rotational and translational changes between the images. The camera pose is then computed by concatenating the relative motion with the previous pose.

2.3 PLACE RECOGNITION

Place Recognition (PR) places Image Retrieval (IR) in the practical context of physical agents operating in a physical world and is used to search an image database for images whose visual content is similar to a queried image. An important aspect for the successful localisation of robots is the creation of a map of their operational environment so that they can localise themselves. This task is usually referred to as loop closure and occurs when the robot returns to a point on its trajectory. This requires effective identification of the revisited landmarks to create a consistent map - a long front-end in the SLAM process, see Figure 1.

In this context, correct data association (consistent map) is required so that the robot can uniquely identify the landmarks that match the previously seen ones, which can be used to recognise the closing of the loop. Actually, autonomous robots need a PR technique because if it only moves based on odometry, it understands the environment as an "infinite corridor" and constantly explores new areas. Instead, the robot is informed by the loop that this "corridor" is constantly overlapping. This technique therefore enables the robot to understand the actual topology of the environment. As a rule, the loop closure is detected using cameras, lasers or sonars. When using visual sensors, the challenge is greater due to the amount of information that needs

to be associated, but cameras are indeed effective when it comes to close-range. In addition, acoustic image sensors are mainly used in the underwater environment as light propagation is limited and they do not suffer from the effects of turbidity.

Therefore, given an image of a place, "*how can a robot decide whether this image is a place it has already seen or not?*" Firstly, a PR system must have an internal representation – a map, i.e. a set of distinguishable landmarks in the environment that can be compared with the incoming data. Secondly, the system must report whether the current visual information originates from a place already contained in the map, and if so, from which one. Each PR system therefore consists of three important modules, which are represented by green blocks in Figure 3.

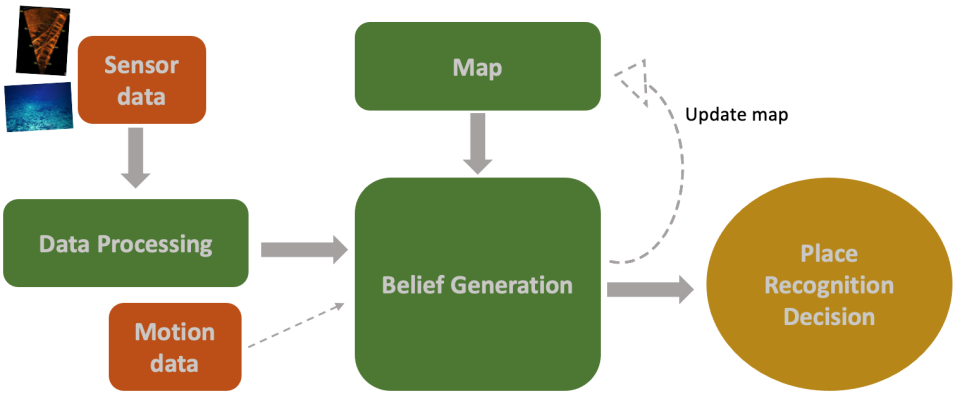

Figure 3: Overview of a visual PR system.

The processing module interprets the incoming data, the map contains a representation of the robot's environment, and the belief generation module compares the incoming sensor data with the map and decides whether the vehicle is in a known or new place. Finally, the map is updated accordingly.

Based on [16] there are the following categories for place recognition:

- feature-based techniques, also known as appearance-based techniques, which are successful for smaller environments (walkways or indoor corridors). However, these techniques lack the invariance of the environmental conditions, which can limit their robustness in long-term operation;
- lighting-invariant image representations that rely on a state-invariant approach and represent images in daylight in state-invariant forms or use non-visible spectra as in Ultraviolet (UV) images. These strategies offer reasonable performance but assume a single sunlight source and can therefore only be applied to daylight observations;
- trained-based algorithms, which are used to learn and predict changes in appearance under different conditions or to build a

database for each place, e.g. to learn a new visual experience when the appearance of a place changes. These approaches have inherent disadvantages, such as the need to visit a place multiple times or to train data, which can happen in an unsuitable environment.

An emerging trend is the application of Convolutional Neural Networks (CNN) as a robust feature extractor (which learns the visual features of the entire image), especially in dynamic environments [17]. These approaches are very powerful as pre-trained generic models can be customised to specific tasks (eliminating the need for vocabulary training). However, since they use global features, they are not viewpoint invariant. Furthermore, these methods are unchallengeable for large contexts (millions of training examples) [18]:

- sequence-based methods that incorporate temporal information to improve evidence accumulation. In general, these approaches are not viewpoint invariant;
- probabilistic methods based on graph representations. These methods use the entire image and rely on filters for hypothesis formulation. They are suitable for state-invariant applications with training-free models.

In a pure image retrieval model, the map simply consists of the stored images or image representations and is therefore based on appearance. Furthermore, appearance-based methods are one of the most commonly used cues for PR of mobile robotic platforms, as the information about the appearance of a scene is independent of the pose and error estimate. This makes these methods a potential solution that enables fast and robust loop closure detection.

Figure 4 illustrates Content-based Image Retrieval (CBIR).

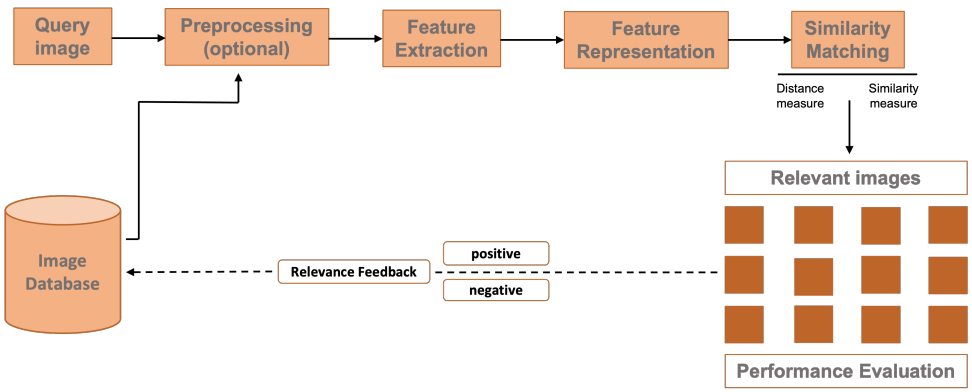

Figure 4: General CBIR framework.

Feature extraction is the first process in CBIR that aims to convert human perception into a numerical description that can be processed

by machines - the accuracy of the retrieved images is strongly influenced by the extracted features. In this context, appearance methods can be divided into three categories for data representation (feature aggregation): Image-to-Image, Map-to-Map and Image-to-Map. These approaches differ mainly in where the association data comes from – the metric or the image space: Image-to-image correspondence is sought between the most recent image and the previously viewed images and is often done at the pixel level using intensity information about each pixel; map-to-map correspondence is sought between features in the two maps, taking into account the appearance of the features and their relative geometric location; and image-to-map correspondence is sought between the most recent image and the features on the map [19].

The first matching approach, i.e image-to-image performs better than image-to-map and map-to-map methods and comprises three main approaches for image representation approaches [8]:

- Bag-of-Words (BoW), which typically groups the local descriptors using the K-means clustering technique and therefore requires a codebook of visual words. Each local descriptor is assigned to the nearest centroid and the representation is done as a histogram;
- Vector of Locally Aggregated Descriptors (VLAD), which summarises descriptors based on locality criteria in feature space. Like BoW, this approach requires a codebook of visual words using k-means. However, a smaller number of visual words is used, as the descriptor assigned to a cluster centre is calculated as the sum of the signed differences between the cluster centre and the nearest feature descriptor. The computational effort is therefore significantly lower than with the BoW approach. However, the VLAD representation is more demanding (KD-dim), as it aggregates the residual vectors element by element and the local space-time features are high-dimensional. Therefore, this representation leads to high time and storage costs, especially for large data sets;
- Fisher Vector (FV), i.e. the gradient of the probability of the sampling probability according to the parameters of this distribution (scaled with the inverse square root). This method requires only a few visual words compared to BoW, but is significantly worse in terms of representation (2KD-dim).

The BoW models are the most popular for place recognition. Although they are not pose invariant due to their neglect of geometric structure and may reduce the accuracy of place recognition, they have many advantages over other visual place recognition approaches. Firstly, due to the fact that they ignore the geometric distribution of

their input data (the local features), they are implicitly invariant to the camera pose from which the images were taken - a desirable property for any location recognition task, as the amount of change in viewpoint can have a significant impact on the difficulty of the place recognition problem and the suitability of a particular approach. They are also very efficient thanks to the use of an inverted index and potentially hierarchical structures, as will be explained later. Finally, thanks to their simple architecture, they can be easily integrated into other complex systems (e.g. SLAM systems).

These models represent local features in a fixed-length vetor. To do this, they use a visual vocabulary to represent an image as a single numerical vector, i.e. each image feature is assigned to a word. The vocabulary is created offline (the usual technique) or online [1] and histogram matching is used to find the similarity between the captured image and the images in the map [17]. First, a local feature extraction is performed: Detection of relevant image regions is crucial and in image classification, the use of multiscale features (which increase the amount of low-level information) has a favourable, strong influence on classification performance. Moreover, in low-contrast areas, edge-based feature extraction is not recommended as it is prone to noise. Next, a set of code words is needed – the so-called visual codebook or vocabulary. This phase can be carried out using different strategies. Typically, the vocabulary is learnt using an unsupervised clustering algorithm applied to local descriptors randomly selected from an image dataset. According to [18], the "naive" unsupervised learning methods such as K-means or random sampling are sufficient to learn vocabularies that achieve high performance on different image datasets. This representation can thus be used to train statistical learning models or to model visual maps by weighting features. This approach is not able to deal with strong illumination changes: visual words can no longer be assigned. Nevertheless, among appearance-based methods, BoW stands out [8] due to its simplicity and speed in finding similar images. This method provides results with less computational effort [20] and does not take into account the spatial distribution between the recognised features and the 3D geometric information that can be useful in making associations. These techniques can be arranged in hierarchical vocabulary trees, allowing efficient search in large datasets, and have shown reliable performance in detecting loop closures in a single session [11].

As databases can be very large, it is important to develop fast access and search techniques to simplify the search for similarities - this requires a data structure that stores data in an appropriately abstracted and compressed form - indexing. The search to compare

1 Offline vocabularies are created à priori from a training set, instead the online vocabulary is created incrementally during navigation.

the features between the query and the actual images is therefore based on the indexing technique used and the extracted features. A similar representation of the query helps with retrieval. The indexing phase is one of the most important in Visual Image Retrieval (VIR) as it is a powerful means for a quick comparison between the query and the object in the database. In the context of place recognition, the inverse index, which maps each index term to a document list in which it occurs, is most appropriate. If a query contains multiple terms and some of them may be more important, term weighting is used.

To facilitate the indexing and similarity of images, it is necessary to pack the visual features [21]. The indexing techniques can be divided into hash-based and non-hash-based, as shown in Figure 5.

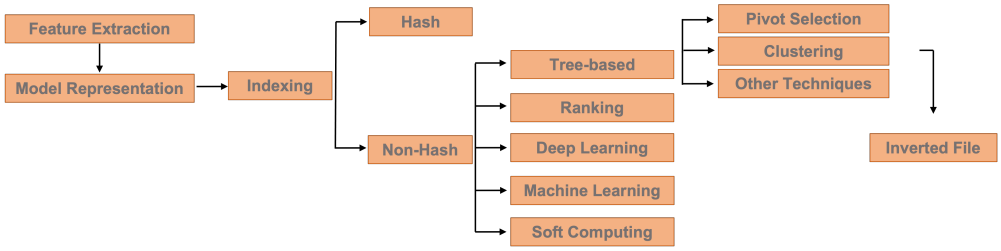

Figure 5: Overview of indexing techniques.

As the name suggests, non-hash-based indexing techniques are generally used to speed up access to the data. An index is basically a type of data structure that is used to quickly find and access data in a database table. Indexes can be easily developed or created by using one or more columns of the database table. Hash-based techniques, on the other hand, use hash functions with search keys as parameters to generate the address of a data record. The direct storage location of the data record on the hard drive is calculated without using an index structure. In simple terms, it is a process of converting a specific key into another value called a hash value or simply a hash. Hash-based data structures are used as an efficient way to store, grouping and searching data, especially for large amounts of data. However, in the context of place recognition, tree-based structures (non-hash-based indexing) are the most important. They are not suitable for large databases, but only for small databases. These structures are easy to understand, allow efficient insertion of data, i.e. ordered data and efficient searching.

There are two main strategies for clustering: hierarchical (or agglomerative) clustering and algorithms based on point assignments. Hierarchical clustering methods start by assigning a separate cluster to each data point. The clusters are then merged iteratively merge clusters depending on a selected closeness measure until a specific stop criterion is met. This criterion can be, for example, the desired number of resulting clusters. If the number is not known in advance,

the algorithm can be stopped if not enough clusters are generated. Otherwise, algorithms based on point assignments start with an initial set of clusters. They then assign each point in turn to the best matching cluster. These techniques usually assume that the desired number of clusters is known in advance, although there are also methods that can be used to infer this number (e.g. a simple trial-and-error method). One of the best known point assignment algorithms is called k-means (standard technique of BoW models). The hierarchical structure - tree - is intended to increase efficiency, as already mentioned, where each node represents a cluster and the complete set of terms is then defined by the leaves of the tree.

The comparison between the current image and the images in the database is based on a "similarity measure". This measure determines which images are considered particularly relevant to the queried image and should be returned from the database. It is therefore an important process that affects CBIR performance, as an inappropriate measure will result in fewer identical images, which reduces the accuracy of the CBIR system. The selection of the similarity measure is influenced by the structure of the constructed feature vector (type and dimensionality of the input data). Nevertheless, the choice of a suitable measure is a difficult task and still an open research area that needs to be further explored.

The similarity measure can be divided into two categories:

- distance measure, which normally measures the dissimilarity between two feature vectors. The smallest value therefore indicates that the images are most similar to the query image. Distance measures are divided into two main categories: bin-by-bin and cross-bin. The bin-by-bin distance measure compares the bins of two feature vectors. This category is widely used due to its ease of calculation and implementation, but is affected by scaling, quantisation, noise, shape changes and illumination changes. On the other hand, the cross-bin function takes into account the cross-bin relationship between non-corresponding bins in the feature vector. It is better than the bin-by-bin distance, but has a rather high computational cost.

  The Minkowski family distance belongs to the bin-by-bin distance category and is often used in CBIR as a dissimilarity measure because it is easy to implement and calculate. The distances L1 and L2 are among the most commonly used distance metrics in CBIR and other areas of image processing. L2 is known as Euclidean or Pythagorean distance and is invariant to orthogonal transformations. While L1 is known as Manhattan or City block or Taxi-cab distance. It is a variant of the rotation of the coordinate system, but robust against reflection and translation.

Square Chord is another distance metric used in image search and is unsuitable for feature vectors with negative values. L1 and L2 have a lower complexity than e.g. Square Cord, but are sensitive to noise.

- Similarity metric - Cosine distance, where the larger value indicates the most similar images to the query image.

In addition to data association for recognising loop closures, another problem of visual place recognition is the phenomenon of perceptual aliasing - in relation to the application area (environment). This term describes the effect that different locations in the environment can have a very similar appearance and be incorrectly recognised as the same. This problem occurs with repetitive features, e.g. corridors, similar architectural elements or areas with a few bushes.

A good place recognition method must not produce false positives and must minimise false negatives. Mainly because of this phenomenon, BoW methods are not considered a perfect solution for place recognition due to the vector quantisation achieved by clustering. Therefore, a verification step must be performed later to check whether the matching images are geometrically consistent, requiring feature correspondences and rejecting outliers. RANdom SAmple Consensus (RANSAC) is usually used for this purpose [22]: This algorithm can be used to derive a strong estimate of the transformation between image A and image B, which is used to find effective correspondences, and then the affine transformation is derived to map the inliers in image A to image B. This step is usually performed using the metrics precision[2] and recall[3] and their relationship (precision-recall curve). But the perceptual aliasing can severely affect the quality of the map. Therefore, methods have recently been developed that take into account the residual error from the optimisation, or others that try to detect outliers before optimisation. In addition, similarity can be calculated by direct feature matching (without vector quantisation). This approach shows a better recovery rate, but requires more computational effort, which makes it prohibitively expensive for larger maps [18]. To overcome this limitation, some approaches tend to merge different methods. Recently, [20] proposed to incorporate spatial co-occurrence information directly into the vocabulary itself and create an additional vocabulary consisting of word pairs (formed by using a spatial neighbourhood) - this approach provides better discrimination in loop closure detection. It uses a K-D tree to store and searching features, a Bayesian probabilistic framework for loop closure decisions, and RANSAC to confirm

2 The precision measures how correct the predictions were - its value should be closer to 100% to be useful in a SLAM approach.

3 The recall measures how many loop closures the detector missed - a lower value means that the detector missed many loop closures and cannot help with the SLAM problem.

loop closure for the best candidates. The K-D tree requires very little computational cost as it does not require distance computation, unlike the hierarchical K-means that uses distance computation for clustering. In experiments with some literature datasets, the proposed method showed higher recall performance compared to state-of-the-art methods without geometric information. The method needs to be integrated into a SLAM approach so that the camera motion can be taken into account. Furthermore, it is important to deal with the computational overhead as the map grows: a forgetting factor can be incorporated where older features are gradually forgotten. In this context, [11] also considers that these approaches must be used to determine what to remember and what to forget, or approaches that draw on multiple map representations of each place. With the latter mechanism, the mean path error decreases by up to 80% when applied in an indoor office. In addition, topological information is becoming increasingly important as incoming sensor data becomes less reliable and difficult to reconcile with previous observations. In the context of the application area, the environment also has an important influence on the expected type and quantity of dynamic objects.

Another challenge - and an active area of research - in closing loops is the recognition of places despite strong changes in environmental conditions caused by day-night cycles, changing weather or seasons. In the appearance change question, the most active problems are the character/intensity of change - strong appearance, frequency of occurrence - continuous state, and whether the state can change within a sequence. In contrast to a general image retrieval problem, the images for the visual place recognition problem are typically captured by a robot, a car or a person equipped with a camera travelling or walking along a path. Consequently, the images can be captured in databases and queries as spatio-temporal sequences. However, depending on the trajectories, velocities and frame rates, the spatio-temporal structure (and thus the sequences in the similarity and ground truth matrices) can vary considerably, making it difficult to recognise loop closures. Thus, although PR is an extensively studied problem, there are still many open research questions. The sheer number of properties that can change when setting up a PR experiment makes this area of research an important and challenging task.

In a dynamic environment, such as underwater, it is more difficult to localise locations and create maps, as the appearance of a place can change over time due to various causes, such as changes in lighting, weather or turbidity levels. In these environments, the challenge for loop closure detection is therefore twofold: firstly, the robot must recognise, reject or track changes - dynamic elements are part of the model. Secondly, it must model permanent or semi-permanent

changes and understand how and when to update the map accordingly in order to continue to faithfully represent the environment.

Loop Closure Detection

The performance of a place recognition algorithm is usually evaluated using precision and recall metrics/curves.

Therefore, the evaluation is done in comparison to the ground truth by counting the True Positive (TP), False Positive (FP) and False Negative (FN). TP defines the cases in which the algorithm successfully recognises the queried image as a known place. In contrast, FP are the cases in which the algorithm incorrectly recognises the queried image as a known place. The cases in which the place recognition algorithm does not incorrectly recognise the queried image as a known place are referred to as FN (unrecognised loops).

The metrics precision and recall are then calculated as follows:

$$\text{Precision (\%)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{1}$$

$$\text{Recall (\%)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100, \tag{2}$$

which means that precision represents the robustness of the algorithm to recognise a place without error, while recall determines the strength of the algorithm to recognise known places without loss.

To find a suitable combination of the two performance measures, the F1-score, i.e., the harmonic mean, is also determined:

$$\text{F1-score (\%)} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

where a higher value indicates the most suitable combination.

Even so, in the context of navigation, the maximum recall, which can be achieved with a precision of 100%, is of particular interest: A false positive will cause the algorithm to generate inconsistent maps and consequently unreliable pose estimates, which can lead to irreparable errors in many robotic applications. A high recall value also enables better motion estimation as it avoids incorrect path adjustments.

3 STATE OF THE ART

Autonomous navigation of underwater vehicles is crucial to enable the vehicles to fulfil demanding tasks, namely in regions with poor visibility or in areas that are difficult to access, such as monitoring and inspecting infrastructure. These tasks require close-range navigation, which reduces the number of sensors that can be used due to perception limitations. Therefore, some solutions have been proposed to overcome these challenges and improve the behaviour of AUV so that they can help people more and more in dangerous situations.

Therefore, this chapter gives an overview of the recent progress in the main areas of this research: close-range pose estimation in section 3.1 and place recognition in section 3.2.

3.1 CLOSE-RANGE POSE ESTIMATION

Underwater scenarios usually have a higher level of distortion and noise and can be featureless, making the perception of the environment, i.e. feature extraction, a challenge. Therefore, it is important to select very robust features, regardless of the type of structure under investigation and even in low light conditions or from different viewing angles. In short, the visual sensors are attractive because they allow the extraction of landmarks taking into account two-dimensional (2D) and 3D information. In addition, these systems are a cost-effective solution, operate at a range of less than 3m and offer high resolution and simplicity. However, these passive sensors require textured scenes to achieve satisfactory results. In addition, the cameras are very sensitive to water turbidity, which can affect the adaptability of the sensor to the localisation problem. The structured light system can operate at high resolution even at short distances (less than 3m). It has a slow acquisition time caused by the projection of different patterns, which can limit the use of these systems in real conditions. In addition, this solution is strongly influenced by the lighting and contrast conditions of dark objects. Shiny objects also pose a challenge for the structured light system, as the reflected light can mislead the pattern decoder. As this solution illuminates the operating scene, it is also affected by scattering, which reduces its range. However, as this sensor casts features onto the environment, it also works on objects without texture. Sonar-based acoustic systems are a costly solution. The propagation speed of the sound depends on the water temperature and salinity,

and correct operation may depend on the type of materials. However, these sensors are independent of water turbidity, and although they have a long range, there are sonar sensors that are also suitable for shorter distances.

Given the great need for inspection tasks and the effectiveness of visual approaches in underwater environments, visual odometry has been significantly improved in recent years, but it still lacks a certain robustness and reliability when navigating in complex underwater scenarios. Based on optical sensors, a mosaic-based visual localisation method for navigation of underwater vehicles near the seabed is proposed in [23]. This is a monocular approach that uses the Harris corner detector and KLT as a tracking method for matching with the following image. It estimates the robot's movements by computing homographies assuming purely planar scenes and movements with 4 degrees of freedom (x, y, z, yaw). This early work on navigation near the seabed shows the possibility of using image processing as the only positioning system for relatively large distances and longer periods of time. Nevertheless, complete coverage of the area of interest is not guaranteed, which is crucial to avoid area gaps and ensure sufficient overlap area between swaths for accurate motion estimation: Poor visibility conditions prevail in the underwater environment, which can affect feature extraction and thus navigation success. Considering the effectiveness of cameras in close-range operations and the need to have a suitable perception of the environment, i.e. to extract robust features, [24] proposes an approach to detect and select suitable landmarks to improve the matching phase and thus the pose estimation. A method is developed to detect Region of Interest (ROI) with salient SURF features that are used as landmarks to localise the vehicle when it is observed to provide repeatedly observed features to reduce the uncertainty caused by the vehicle drift. The method works satisfactorily as the motion estimation was significantly improved compared to dead reckoning alone. However, these tests were performed offline, so the method needs to be further improved to become a real-time solution. Considering the lack of robust features in the underwater context due to the unstructured and hazardous conditions of the environment, which consequently lead to complexity in determining the global position of the vehicle, [25] proposes an image processing-based approach that relies on robust recognition of the artificial markers in a semi-structured environment with poor visibility to guide the vehicle, highlighting the importance of real-time image enhancement for effective navigation. The method directly explores the use of perceptually uniform colour spaces, in particular the $l\alpha\beta$ colour space, which works satisfactorily when navigating in a pool under challenging conditions with a constant navigation tracking error of less than 25 cm, provided that more than one visual marker must be present in the robot's field

of view. However, this solution is limited to specific/controlled applications and the use of natural landmarks is foreseen and necessary. Therefore, a method for visual odometry adapted to the underwater context is proposed in [26]. It is based on the Harris detector and KLT for tracking to ensure fast pose estimation. An optimisation procedure is used to improve the trajectory errors. It was found that the use of five images represents the best compromise between accuracy and runtime and also represents the limit for real-time performance. This validation was performed underwater using different scenarios, e.g. long trajectories with loop closures or raster scans. On the other hand, dealing with errors in visual odometry is an important challenge to consider, especially in the context of underwater imagery, which is mainly due to poor image quality. In this context, a monocular odometry method for estimating the robot's ego-motion in real time was later proposed, which is robust to the usual challenging underwater conditions such as turbidity and dynamism (short occlusions) that remain complex for accurate vehicle positioning [15]. This approach utilises optical flow to track features and a re-tracking mechanism to recover temporarily lost features. This means that the tracking is not based on descriptor matching, but on the KLT method. The pose is then calculated using the Perpective-from-3-Points (P3P) formula within a RANSAC loop to remove inaccurate correspondences (outliers). The back-end relies on a bundle adjustment to ensure minimal drift and consistency of the reconstructed trajectory. Its performance was compared with some state-of-the-art approaches: ORB-SLAM, LSD-SLAM and SVO, in scenarios with different some turbidity levels and with or without fish. The SVO is not robust in dynamic scenarios (moving fish). The LSD-SLAM method does not provide a valid result for any sequence. The ORB-SLAM method loses accuracy in turbid conditions and cannot detect loops, i.e. the offline approach is not suitable for an underwater environment which does not offer many discriminating features. On the other hand, this proposed approach can indeed handle noisy images and calculate all trajectories. Nevertheless, it show that the optical flow is not as efficient as using descriptors in terms of robustness to occlusions. A maximum absolute trajectory error of 1.88% or 2.08% (without the re-tracking mechanism enabled) was achieved, based on averaging over five runs. The performance achieved still needs to be improved, in particular by sensor fusion to also determine the scaling factor or by adding a mechanism to close the loop. Recently, a modified version of LIBVISO2, i.e., a library to compute the 6-DOF motion of a moving mono/stereo camera, was proposed by [27] so that it can be used in underwater scenarios with different types of seagrass (Posidonia oceanica) or algae. It was mainly modified in the feature matching phase, comparing different combinations of feature detectors and descriptors. The data were collected in

campaigns along the southern coast of the Balearic Coast, where the seabed has long Posidonia oceanica leaves, pebbles, algae, low and textured sandy areas. SIFT obtains highest number of matches and inliers. It also offers the best compromise between the robustness of feature tracking and the cumulative drift of the trajectory, and significantly improves the performance of the original algorithm in the underwater domain. Nevertheless, these seabeds have areas with low texture and repetitive patterns, i.e. similar features, that make robust and stable feature matching difficult.

The first investigation of the potentialities and feasibility of an optical technique based on the projection of structured light patterns captured by an image processing system (stereo vision) in an underwater environment was proposed in 2011. Although it has been used for 3D reconstruction, it suggests that it will be a potential approach to help vehicles not lose their position in scenarios with little or no texture. In short, a series of n temporal binary patterns are proposed to establish a correspondence between two images [6]. Each binary code uniquely identifies each point of the object surface, even if the surface has no texture. The experiments were carried out in a water tank with three different reflective objects (shiny, opaque and dark) under different light and turbidity conditions. The data recorded at higher turbidity levels are affected by scattering: They produce many outliers within the volume of water illuminated by the projector. When the turbidity level changes (low to medium), the number of detected points in the dark object decreases sharply (from 1091 to 144 points), which could be due to the fact that the acquisition of dark objects is generally strongly influenced by the illumination and contrast conditions. The use of these multi-image methods (pattern projection and acquisition) requires a long detection time, which can be a relevant problem in practise. Nevertheless, reducing the number of projected patterns is a suitable solution that can be installed on a ROV or AUV, for example, to fulfil inspection tasks, albeit with some loss of accuracy. Later, [5] describes a structured light with grey sinusoidal stripes in four steps, i.e. a time-division multiplexing approach in an underwater environment, which shows a better behaviour than a stereo view. Next, a comparison between a laser-based structured light system (active) - projecting an uncoded pattern of 25 parallel lines - and a stereo vision device (passive) in a common underwater environment is presented to recognise the 3D structure of a studied area [7] [28]. For the experiments, both sensors were attached one after the other to a Cartesian robot in a swimming pool with clear water and four different objects were used. The stereo system provides a reconstruction consisting of 685872 points, while the structured light only provides 370261 points. The accuracy with which the sensors determine the dimensions of the objects was also evaluated. The LbSL

showed better accuracy in the z dimension and allows the capture of sharper and clearer details, but loses colour information. This system proves to be largely invariant at low to medium turbidity levels. At high turbidity, the number of detected points decreases until the camera can no longer recognise the objects from the scattered light. Correct detection of the laser depends on the contrast between the lines and the background and therefore may not work if the ambient light is high in a highly scattering medium. The stereo data is sufficient to restore the overall shape of the underwater environment if sufficient visibility and texture information is available. Later, a bathymetric SLAM is therefore proposed, which is obtained with structured light systems based on a single-line laser [29]. With this approach, a bathymetry of the seabed can be generated with a resolution in the millimetre range and is therefore also used in this case to reduce the relative uncertainties in vehicle localisation. The experiments, which are based on data collected during a research cruise in the Iheya North Field of the Okinawa Trough, show that the registration works well even with noisy data and generates self-consistent maps with high resolution. The system is able to compensate for misalignments by correcting localisation errors. In the same line of research, this SL approach is used to inspect an underwater structure: A laser scanner is integrated into a state-of-the-art pose-based EKF SLAM system to map a 3D underwater infrastructure in a water tank as used in the offshore industry [30]. Since there is not too much texture, the use of laser-camera systems for inspection is advocated. The idea is that the robot uses the map to correlate its sensor measurements with the map and determine its position. Experiments show that it is possible to limit the high navigation deviation. Nevertheless, the generated map is not completely consistent due to internal errors in the point clouds, which is mainly due to vehicle movement during a scan.

In order to avoid localisation errors, to which conventional and visual navigation methods are prone, sonar-based acoustic approaches have been investigated in the field of perception. Although sonar suffers from distortion and masking effects caused by acoustic shadows due to its physical properties, acoustic image sensors have been used for visual underwater navigation. Extracting information from sonar images therefore poses a particular challenge, namely the weak feature textures, acoustic shadows caused by objects blocking the passage of transmitted waves, high noise levels - mutual interference causing speckle noise, i.e. underwater engines from nearby surface vehicles or other acoustic sensors) [31]. Nevertheless, these sensors do not suffer from turbidity effects and have recently proven to be a promising alternative to optical sensing in such situations, as they limit the ability to extract features, i.e. discriminant features. Since acoustic images are of very low quality and may incompletely represent the target, image

processing methods commonly used for video images can also be applied to acoustic images. In this context, a method that utilises the fusion of Ultra-Short Baseline (USBL) measurements and sonar images to track underwater targets was proposed by [32]. The algorithm therefore relies on clustering contours and finding those that are most likely to belong to a target. First, basic image processing is performed, which includes a noise removal filter and binarisation of the image. The binarised image is used to determine the contours and perform clustering. Finally, the best candidate within the region of interest is searched. This approach was evaluated in a real environment in Croatia, where two data sets were performed: In scenario 1, the vehicle is moving while the target is static or drifting slowly; in scenario 2, the vehicle is static and the target is moving. Whenever no sonar measurements are available, the position estimate deviates from the true value. However, in both scenarios, sonar measurements were available for about 30% of the samplings, resulting in a mean position error of $0.5m$ and $1m$ in scenarios 1 and 2, respectively. Later, to compensate for the position deviations caused by traditional solutions during a mission, [33] proposes a SLAM method that uses a SSS. In contrast to cameras or FLS, objects in sonar images appear only once, it takes time to generate a side-scan image from individual pings (about 30 to 40 seconds to generate an image with 500 pings), and there are strong variations in the appearance of structures. To reduce complexity, only compact objects are of interest, as they usually produce distinct light-shadow structures over a wide range of distances and relative positions, which increases the reliability of follow-up observations. To ensure a high detection rate in real time, simple and fast descriptors were used and a noise reduction filter was applied. The method was evaluated using 3 missions of an AUV in a real scenario. The experiments show that the SLAM achieves a reduction in position deviations between 20%-80% (on average 38%) compared to the exclusive use of inertial navigation methods in combination with a GPS. Moreover, the performance mainly depends on the number of landmarks observed and the time intervals between observations. The concatenation of 500 pings leads to a delay in the observations. Therefore, if you only use this sensor as a positioning system, it is not possible to perform real-time navigation. However, it proves to be an attractive technique for correcting pose estimates obtained from other information. Recently, a real-time positioning method for AUV navigation using FLS has been proposed [34]. The approach is based on the detection of features in two consecutive FLS images. When the seabed has little texture information, a simple dynamic model is used. To evaluate the method, a deployment was carried out in La Spezia (Italy) in two different scenarios: one in an area with no features and the other in an area with high information content where two

underwater structures were located on the seabed. The results show that the navigation method has a mean error of about $1m$ compared to the ground truth solution. To try to achieve accurate positioning of the vehicle and understand whether sonars can assist the visual sensors, a Pose-Graph SLAM technique has recently been proposed that uses a pipe profiling sonar and visual sensors [35]. The sonar helps the visual sensor when it is difficult to find visual features, i.e. when visibility is poor. The proposed approach was tested with data sets recorded in three real-world environments. The experiments show that the sonar can be used in a less structured scenario, such as coral reef areas. In one cave, a video light was used for illumination, but it only illuminated part of the scene. This configuration made it possible to perceive the surroundings, but as the scene was not evenly lit, the visual features were sparse. The quality of the image of a flooded bus is quite poor due to the large number of particles. Nevertheless, the distance travelled is clearly visible and the gaps in the sonar data are visible, which correspond to the areas where the windows of the bus were located. So the sonar features were actually used to correct the pose estimation, which allowed the optimisation to converge and the error to remain very small - improving the mapping of the underwater structures.

Hybrid data acquisition is therefore of interest and an advance to enable the robot to perform controlled missions regardless of environmental conditions. To overcome the limitations of each individual sensor type, active perception systems combining different sensors of the same or different types are crucial, as underwater missions often require different speeds and distances to the seabed or to the infrastructures to be inspected.

3.2 VISUAL PLACE RECOGNITION

At its core, visual PR is an image retrieval problem that aims to search an image database for images whose visual content is similar to a query image. PR is an ill-defined problem where there is no inherently right or wrong answer to the question of whether two images show the same place or not. Nevertheless, correct loop closure detection must be used to correct accumulated deviations during navigation and mapping. Therefore, SLAM systems rely on a robust place detection method for a robot to successfully navigate autonomously, especially when detecting loop closures or estimating the position of candidates. In order to create a consistent map, effective identification of landmarks is required. In this sense, given the need to use viewpoint invariant feature extractors and descriptors, and given the high time cost of extracting SIFT/ Speeded-Up Robust Feature (SURF) features,

binary algorithms are increasingly used for place recognition [36]. These features have a very compact representation, which in turn requires less memory and computation time. Their proof has also been demonstrated in underwater environments by evaluating several feature detectors/extractors in different scenarios categorised by seabed features, object type, illumination, hue and turbidity, with orb excelling in detection and matching performance with the lowest computational time [37]. There are several methods for place detection, but the visual BoW algorithm is one of the most commonly used techniques for fast and robust loop closure detection based on appearance models. These models represent local features in a fixed-length vector and use a visual vocabulary where each image feature is associated with a word. They are an efficient approach thanks to the use of an inverted index - to enable fast queries - and potentially hierarchical structures.

In this context, in [38] the first vocabulary approach for recognising previously visited locations in an image sequence based on efficient, i.e. binary, features was proposed. It uses Features from Accelerated Segment Test (FAST) keypoints and Binary Robust Independent Elementary Features (BRIEF) descriptors and relies on the BoW model to perform image comparisons. This proposed method introduces a direct index to efficiently obtain correspondence points between two images and uses the RANSAC algorithm to perform the geometric check between the candidate matching scenes. The vocabulary is created offline and structured as a tree, i.e. a hierarchical BoW, where each word is weighted according to its relevance in the training. Using indoor, outdoor and dynamic mixed indoor and outdoor data without highly dynamic objects such as cars or people, it was shown that the BRIEF descriptor tolerates only small changes in scaling or rotation and shown a very good compromise between discriminative power and computational time: The time required for the query is very low (between 5 and 37ms), which means that this stage scales well with tens of thousands of images. Furthermore, the entire algorithm only requires less than 49ms to close the loop on a dataset with more than 19K images. However, in order to not limit the system to only rectilinear trajectories and loop events with a similar viewpoint, this work was later extended by using ORB features [39]. It achieved a higher recall than BRIEF in outdoor scenarios with larger viewpoint differences. Normally, the BoW model is predefined based on features from a training set, which is a constraint: The model is dependent on the environment and cannot handle changes. However, this is critical in cases where the robot operates in uncontrolled environments, such as the underwater scenario. Therefore, in [40] a visual vocabulary was proposed that is created when visual information becomes available during vehicle navigation. The approach relies on clustering criteria - it uses a modified version of agglomerative clustering - that takes

into account the entire distribution of clusters, which increases the efficiency of the vocabulary. This appearance-based incremental vocabulary outperforms offline vocabularies and, in outdoor scenarios with some dynamic elements, comes very close to the full-indexing vocabulary, which is also created offline but with images of the trajectory performed. Nevertheless, in a few cases false loop closures were detected, related to the highly repetitive patterns on the images: Grass, trees or soil. However, these first experiments in underwater environments show how difficult these scenarios are. This means that the performance of the method is slightly worse than in underwater scenarios due to the perceptual aliasing (the sequence is characterised by repetitive textures) and the decrease in image quality, i.e. less descriminative features. To deal with perceptual aliasing, an algorithm for validating the matched scenes and deleting outliers is crucial. On the other hand, there are also some approaches that avoid vector quantisation and use direct feature matching to calculate similarity. This alternative shows better recall performance, but is associated with higher computational costs [18]. To overcome this limitation of conventional BoW while maintaining its simplicity, the use of spatial co-occurrence of words directly in the vocabulary itself was later proposed [41]. It provides better discrimination in recognising loop closures and therefore showed higher recall performance under such conditions. Besides the difficult visibility, the lack of robust, stable and matching features in some areas is also a relevant challenge for performing similarity detection underwater. Therefore, a stereo SLAM was later proposed to detect loop closure situations in an underwater coastal area without limited visibility [42]. The proposed method improves loop detection by matching the features of clusters instead of the features of keyframes. The method was validated in an indoor water tank (marine scenario without 3D structure) and in a real outdoor region without restricted visibility. Compared to other state-of-the-art methods, this proposed approach recognises all situations of loop closure in both scenarios with less errors in motion estimation, computational time and system resources. ORB-SLAM did not detect any loop situation in the small mission in the outdoor environment, which could be due to the use of an offline vocabulary or the lack of robust features given the application in a shallow scenario with sandbanks and algae/seaweed. Due to the importance of building the codebook online later an incremental vocabulary approach based on a hierarchical scheme (BoW) and relying on the FLANN library was proposed [43]. This approach combines global and local binary descriptors and shows high performance, but is still vulnerable to dynamic elements. Next, an incremental BoW strategy has been proposed where the vocabulary is created using ORB features and is based on hashing techniques [44]. It provides a word deletion policy

that reduces the size of the visual vocabulary with little impact on performance and consequently requires less computation time. This approach provides good recall in indoor and outdoor environments with static and slightly dynamic objects.

The underwater environment is often affected by turbidity or illumination (in shallow or deep waters), which often makes it difficult for cameras to perform navigation and mapping tasks, as the range of optical devices is severely limited in very poor visibility. Such conditions reduce the ability to extract features on optical images, making it difficult to detect loop closures, and the vehicle may fail to detect some loops correctly or detect some erroneous loops, so that the trajectory is not adjusted or is adjusted incorrectly [2]. For this reason, acoustic image sensors have been used for vision-based underwater navigation [45]. Although these sensors suffer from distortion and masking effects due to their physical properties, they do not suffer from clouding effects as the images are based on the emitted and returned sound. Therefore, this category of sonar is seen as a promising solution for these difficult environmental conditions. FLS and SSS are the most commonly used sonars for environmental exploration. FLS is characterised by the fact that it provides a representation of the environment in front of the robot and allows overlapping images during movement. Due to the characteristics of FLS data, namely low signal-to-noise ratio, low/inhomogeneous resolution and weak feature textures, traditional feature-based registration methods are not yet used for acoustic images. Vilarnau [46] proposes a pairwise registration of FLS images for the mosaic pipeline based on a Fourier method, that can provide robustness to some artefacts commonly associated with acoustic imaging and noise for all image contents. For the inspection of ship hulls with FLS images, a machine learning method for the detection of loop closures based on saliency was used in 2018. To deal with the sparse distribution of sonar images, it is based on the evaluation of the potential information gain and the estimated saliency of the sonar image [47]. Later, a loop closure detector was proposed for a semi-structured underwater environment using only acoustic images captured by an FLS [48]. A topological relationship between objects in the scene is analysed based on a probabilistic Gaussian function. It is based on regions with higher acoustic intensity fluctuations, i.e. segments. Datasets recorded in a harbour area were used: ARACATI 2014 data have more features and ambiguity problems, i.e. the robot moves between two piers, and ARACATI 2017 data have fewer features but the sonar range is larger (30m to 50m), resulting in more and smaller objects per image. The method achieves at best a precision of 95.38% and a recall of 12.24%, which is dangerous for the navigation context as there are some incorrect trajectory adjustments and also many adjustments fail. It is therefore important to get a more accurate

description of the environment or consider using other sensors to fuse data.

But the performance of these sonars has greatly improved and the resolution of their images continues to increase, allowing the FLS to provide comprehensive underwater acoustic images. Deep learning is a powerful technique that is increasingly used in this field of research [49] [50]. However, it requires a large amount of data and a long training phase, which is a complex task in these semi-structured and unknown scenarios. On the other hand, traditional machine learning techniques focus on identifying previous observations, which makes the navigation principles very stationary and dependent on the representativeness of the collected observations. Although it is challenging to make this decision independently of the environment, i.e. through unsupervised learning, it is important that the vehicle continuously learns the environment during the inspection mission in order to adapt its behaviour accordingly. Therefore, developing efficient approaches to extract visual data from sonar images and understanding their performance remains crucial to ensure that the FLS data is appropriate for the vehicle's perception of the environment. In addition, these data can later be matched with the camera data as needed when environmental conditions change. Image matching is the first problem to solve, as it is the crucial step for pose estimation or place recognition. These algorithms can be based on feature point and regions, but given the FLS characteristics and real-time constraints of underwater operations, feature point matching is more appropriate. Given the need for viewpoint invariant feature descriptors, binary methods are increasingly being used for similarity detection. These features require less memory and computation time. Evidence for this was provided in underwater scenes categorised by seabed features, turbidity and illumination. Here, the ORB descriptor proved to be more effective for detection and matching with the least computational time [37]. Recently, its behaviour was also demonstrated for acoustic images using a performance comparison of different feature detectors, where ORB achieved the best overall performance [51]. Therefore, can be interesting to apply the knowledge of the visual to the acoustic data to evaluate whether it is effective in detecting loops at close range.

3.3 UNDERWATER IMAGING DATASETS

Underwater scenarios are unknown and it is therefore important to validate the various algorithms first. However, creating a dataset in a real underwater environment is an expensive and difficult task, as it is not easy to operate an underwater vehicle to record a planned trajectory. In addition, the working conditions are difficult and specialised personnel are needed.

Nevertheless, efforts have increased and some, albeit limited, data is available for the planned context. In terms of optical data, the following datasets are available:

- **UWSIM** [52]

  As the name suggests, this dataset consists of a collection of simulated datasets that were created with Underwater Simulator (UWSim) and are available for the preliminary analysis and execution of SLAM. These datasets contain multiple trajectories recorded by a robot equipped with two downward-facing cameras, with loop closure situations based on four turbidity levels common in underwater environments, as shown in Figure 6.

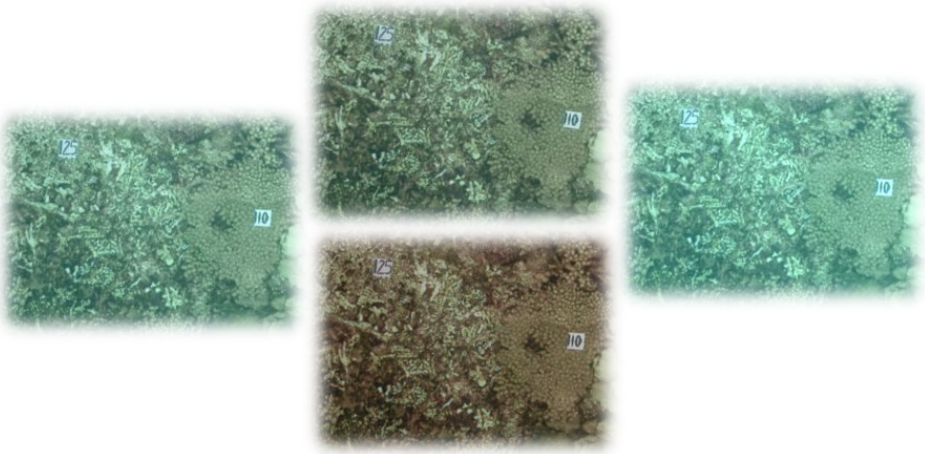

Figure 6: Illustrative example of the images captured in the UWSIM dataset, based on different turbidity levels.

  The seabed was created based on two mosaics of real coral: high texture, with repeating patterns. The information from Inertial Measurement Unit (IMU) and Doppler Velocity Log (DVL) is used to estimate the movement of the robot. The ground truth is based on the use of GPS data. The provided data is available as a Robot Operating System (ROS) bag file for each trajectory. However, the ground truth is not provided, i.e. the GPS topic is empty for all trajectories.

- **Underwater Caves Sonar & Vision** [53]

  Dataset collected with an AUV testbed in the unstructured environment of an underwater cave complex designed primarily for mapping difficult scenarios. The vehicle is equipped with two Mechanically Scanned Imaging Sonar (MSIS) for simultaneous mapping of the horizontal and vertical surfaces of the caves, one DVL, two IMU and one depth sensor. Added to this is a vertically mounted camera that images the seabed to verify ground truth at specific points[1], with the vehicle's architecture capturing still images from the video feed at a frequency of 4Hz.The investigation was carried out with the aim of demonstrating the feasibility of locating an AUV in this difficult environment without external navigational aids. To ensure the safety of the AUV and the acquisition of a complete data set in a limited time, the vehicle was guided through the caves by a diver. The trajectory of the vehicle was designed to contain multiple loop closure situations. The data from all sensors was stored asynchronously. The original ROS bag files are provided along with a derived version that combines images and human-readable text files for processing in other environments. Data provided contains ground truth information based on the standard ROS odometry message and calibration parameters.

  Figure 7 shows an example of the images taken in the underwater cave.

Figure 7: Illustrative example of the images captured in the Underwater Caves Sonar & Vision dataset.

1 Underwater Caves Sonar & Vision dataset: https://cirs.udg.edu/caves-dataset/

- **AQUALOC** [54]

  To pave the way for more in-depth studies on visual localisation techniques near the seafloor, this dataset comprises data recorded in a harbour and includes several sequences of synchronised measurements from a monocular camera (20Hz), an IMU and a pressure sensor[2]. The data is provided in a ROS bag format and as raw data (plain files). For the plain files, the IMU, pressure and compass measurements are provided in independent files with a timestamp for each measurement. The distorted and undistorted versions of the images are also provided as png images and the timestamp of each image is written to a separate file. A total of 7 sequences of varying difficulty were recorded and there are sometimes areas where the view becomes unusable: some backscatter and poorly visible areas. In addition, the trajectories performed by ROV have different depths during the movement and the speed can be very different, which affects the overlap between the images and the number of images of a given region. The ground truth is calculated using the colmap library: While the poses of each image are computed to create the 3D reconstructions, the estimated camera trajectories are extracted and are then available as ground truth.

  Figure 8 illustrates an example of the captured images.

Figure 8: Illustrative example of the images captured in the AQUALOC dataset.

- **AFRL** [55]

  Since most standard benchmark datasets only represent a single scenario with high visual quality, datasets in more challenging environments, such as underwater, have been collected over the years for close-range operations[3]. For visual-inertial-based state estimation algorithms, data is collected from an underwater sensor system, consisting of an IMU and a stereo camera, by a diver around a sunken bus and in an underwater cave. An AUV with the same sensor system collected data in an underwater structure, i.e. a fake cemetery cave.

2 AQUALOC dataset: https://www.lirmm.fr/aqualoc/
3 AFRL dataset: https://afrl.cse.sc.edu/afrl/resources/datasets/

Figure 9 shows an example of the images captured in each scenario. The datasets were recorded as ROS bag files and the intrinsic and extrinsic calibration parameters for both sensors are also provided. However, the ground truth is not provided.

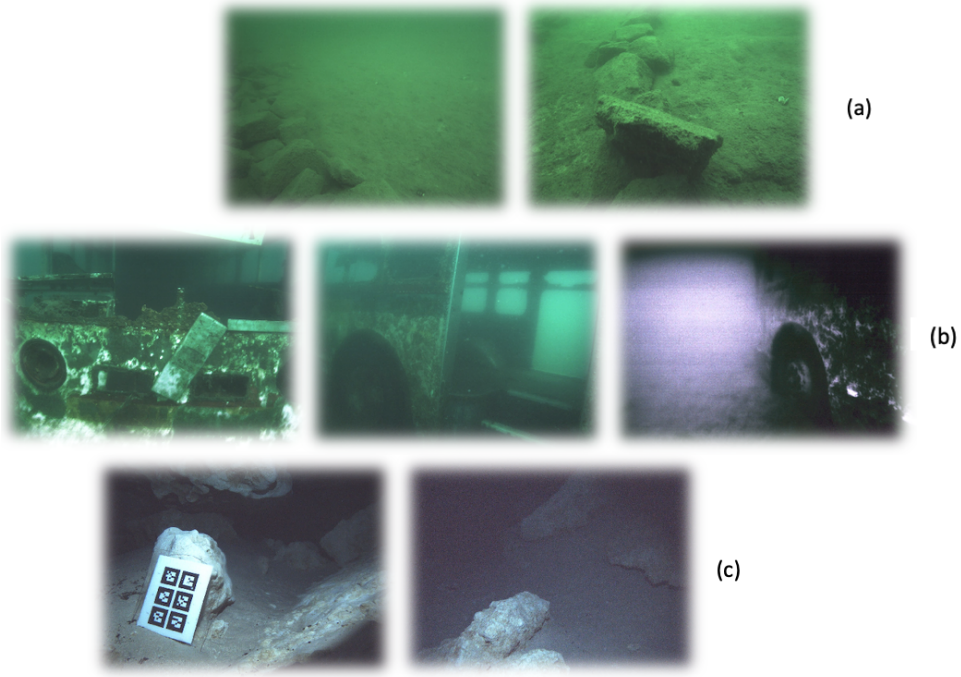

Figure 9: Illustrative example of the images recorded in the AFRL dataset: Cemetery (a), bus (b) and cave (c).

FLS do not suffer from clouding effects because the images are based on the emitted and returned sound and therefore their acoustic images can be used today to perceive the environment. However, as it is a new and promising solution for close-range operations, there is no available data for the intended context. Sound Metrics develops and manufactures imaging sonars (ARIS and DIDSON models) that provide the clearest and most detailed video images in their class. The products can be deployed from ships, ROV, AUV and divers. The company's research team conducts data acquisition campaigns to achieve optimal clarity for imaging sonar work that provides clarity and detail in dark, dirty waters for military, oil and gas, fishing, underwater construction and law enforcement applications[4].

They have provided us with some data that fits the intended context, i.e. data collected for inspection applications. The raw images and the corresponding programmes to read this data are provided.

4 Sound Metrics data: http://www.soundmetrics.com/Industries-Served. I would like to thank you in advance for providing the data and thus contributing to the development of this work.

Figure 10 shows an example of images captured by imaging acoustic sensors in the context of inspections in the fields of underwater construction, oil and gas or law enforcement.

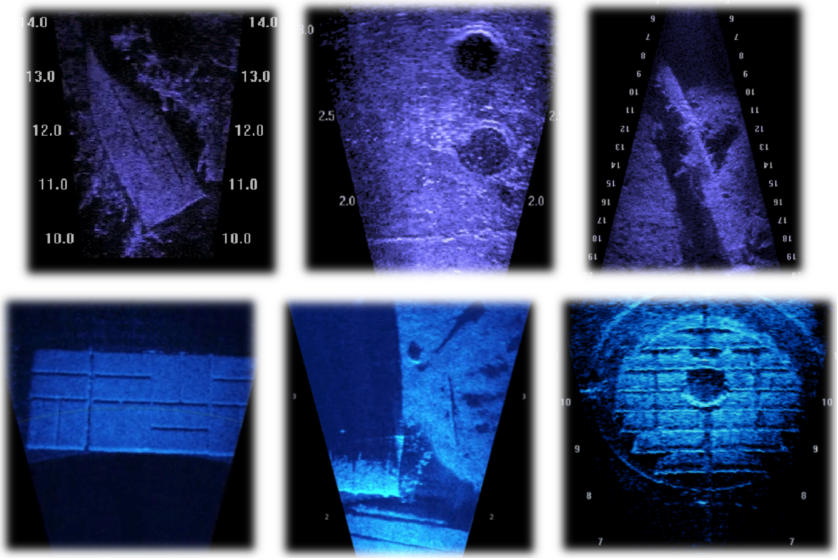

Figure 10: Illustrative example of the images recorded by Sound Metrics in inspection applications.

Part II

LOCALISATION AND LOOP CLOSURE

4 APPEARANCE-BASED PLACE RECOGNITION

To fulfil autonomous tasks, robots in the real world must be able to navigate in dynamic and unknown spaces. To do this, they must recognise previously seen places in order to compensate for accumulated position deviations. This requires effective identification of recovered landmarks to create a consistent map. Therefore, this chapter presents a behavioural evaluation of a conventional BoW method for image similarity in challenging scenarios.

The chapter is organised as follows: Section 4.1 gives an overview of the techniques for a place recognition system and a brief description of the proposed deployment environment. Section 4.2 presents the architecture of a visual place recognition system and describes a modified version of a BoW-based approach. Subsequently, section 4.3 describes the community datasets used to evaluate the recognition of previously seen places and reports on the different experiments performed. Finally, the main conclusions of this chapter are discussed in section 4.4.

4.1 INTRODUCTION

Autonomous robots operating in real-world environments need to map their operational environment to localise themselves (SLAM). Recognising a place that the vehicle has already visited is an essential aspect of compensating for accumulated position deviations and thus contributing to successful navigation. This task requires effective identification of previously seen landmarks to create a consistent map. In other words, there must be consistent correspondences of key points between images. Making this decision independent of the environment - unsupervised learning - is challenging but essential, as the appearance of a place can change over time, especially in uncontrolled environments such as underwater.

Given the need to use viewpoint-invariant feature extractors and descriptors, and given the high time cost of SIFT/SURF features, binary algorithms are increasingly used for place recognition [36]. These features have a very compact representation, which in turn requires less memory and less computing time. Among the various approaches for grouping descriptors based on appearance methods, the BoW algorithm is one of the most commonly used techniques for fast and robust loop closure detection. They are characterised by their simplicity and

speed when searching for similar images and provide results with less computational effort. Next, these features must be stored in a data structure that is suitably abstracted and compressed to simplify the search for similarities in the database - this is known as the indexing phase. As already described in section 2.3, indexing techniques can be divided into hash-based and non-hash-based methods. Non-hash-based indexing is generally used to speed up access to data. They are well suited for small databases, which leads us to believe that this is an appropriate approach for the intended context: inspection of a specific structure in harbour scenarios. For the context of place recognition, tree-based structures that are easy to understand, allow efficient data entry (ordered data) and efficient search are then particularly relevant. In such a context, BoW models can be arranged in hierarchical vocabulary trees, which enables efficient searching. Nevertheless, BoW approaches are susceptible to the phenomenon of perceptual aliasing, which is favoured by repetitive textures such as grass, trees, soil, sand or walls. On the other hand, some scenarios can also be very dynamic, making it difficult to recognise locations, as the appearance of a place can change over time due to changes in lighting or weather. This is particularly problematic for offline vocabulary-based approaches as the vocabulary may not be representative of the scenario. Considering the benefits of using binary features, it is therefore important to evaluate their effectiveness in extracting image features and performing scene matching. The SLAM technique is widely used in ground robotics and there are therefore datasets that contain the required data for the planned experiments, namely ground truth in the form of poses and loop closures. The preliminary evaluation of the behaviour of these loop closure detection techniques will therefore be conducted in outdoor scenarios, which in themselves may present similarly challenging conditions to underwater environments.

Therefore, a comparative analysis of some binary descriptors is presented, discussing the performance of the methods based on image transformations and computation time. Given the strengths of BoW techniques, a fundamental question also arises: *"are these models not really suitable for uncontrolled environments?"* Therefore, the behaviour of BoW approaches in challenging scenarios with scene changes, perceptual aliasing conditions or dynamic elements should be understood and evaluated. For each scenario, full-indexing vocabularies are created to model the operating environment and evaluate the performance in recognising previously seen places similar to online approaches.

4.2 BINARY BAG-OF-WORDS

Correct loop closure compensates for the accumulated position deviations, which is why this mechanism is crucial for successful autonomous navigation. In BoW models, a visual vocabulary is used to represent an image as a single numerical vector, i.e. each image feature is associated with a word. Typically, the vocabulary is learnt using an unsupervised clustering algorithm such as K-Means, which is applied to local descriptors from a training image dataset. A confidence measurement is then performed to check whether the current visual input data matches a specific place in the map representation of the robot world. The result is a list of loops. Due to the vector quantisation achieved by clustering, the obtained loops must finally be evaluated to check whether the match is correct or incorrect, to exclude outliers and consequently to determine the robustness of the system to identify similar places from a different viewpoint and with a different timestamp.

To understand the behaviour of this conventional approach in challenging outdoor scenarios, the C++ libraries DBoW2 [1] and DLoopDetector [2] are used to index images and convert them into a BoW representation and to detect loops in an image sequence, respectively. Therefore, a hierarchical tree is used to approximate the nearest neighbours in the image feature space and to create a visual vocabulary. It is used directly with ORB and BRIEF, but can work with any type of descriptor. The features and descriptors are first extracted for each image. Based on the features, the vocabulary tree is created with the agglomerative hierarchical K-Medians clustering using the K-Means++ seeding [56]. This is an algorithm for selecting the initial values (or "seeds"), i.e. an intelligent method for initialising the K-Means (or K-Medians) clustering algorithm. With this algorithm described in section 2.3, the approximation found in relation to the target function can be arbitrarily poor compared to the optimal clustering. K-Means++ remedies this shortcoming by defining a procedure for initialising the cluster centres before proceeding with the standard K-Means optimisation iterations. The initialisation of K-Means++ guarantees that the algorithm finds a solution that is O(log k) competitive with the optimal K-Means solution. The goal is to distribute the initial centroid by randomly assigning the first centroid and then selecting the remaining centroids based on the maximum square distance. The main difference between the K-Medians and K-Means algorithms lies in the way they calculate the centre of each cluster: The K-Means algorithm calculates the centroid of each cluster, which is the mean of all data points in that cluster. The K-Medians algorithm, on the other hand, calculates

1 https://github.com/dorian3d/DBoW2
2 https://github.com/dorian3d/DLoopDetector

the median of all data points in each cluster. The idea is to keep the centroids as far apart as possible.

This hierarchical scheme requires two user-defined parameters: the depth factor L, which defines the maximum level of the tree, and the branching factor K, which defines the maximum number of nodes (groups) for each level. Thus, based on the Manhattan distance, the clustering steps for each level are performed based on the clusters obtained in the previous level until the maximum depth L is reached. A previous study has shown that a higher number of clusters, i.e. K, does not improve the performance. Therefore, 10 clusters are used as this is the average performance obtained with this branching factor. As for the depth, there are no guarantees of its influence on the results obtained. Thus, L varies between 3 and 7, with $L = 6$ proving to be the best for a dataset with a greater variety of features and a high number of images. Given the intended scenarios, a 6-level vocabulary tree was thus created for all cases.

Finally, a tree with W leaves, i.e. words of the vocabulary, where each leaf node is a group containing a single data point. Each word is assigned a weight based on its relevance in the training corpus. To measure this relevance, DBoW2 allows the use of the algorithms Inverse Document Frequency (IDF) or Term Frequency - Inverse Document Frequency (TF-IDF). For the intended context, TF-IDF is used to determine the frequency of each word (TF term), in this case measured using a logarithmic scale, and to determine how often (or rarely) a word occurs in the corpus (IDF term). TF provides information about how often a term occurs in a document, and IDF provides information about the relative rarity of a term in the collection of documents. It is therefore a practical algorithm that uses the frequency of words to determine how relevant these words are to a particular document. When we multiply these values together, we get our final TF-IDF score. This reduces the weight of words that occur very frequently and are therefore less distinctive.

To enable quick queries and comparisons later on, a direct and an inverted file are kept in addition to the bag-of-words. In the inverted index, the weighting of each image and a list of the images in which it occurs are saved for each word. For each image, the direct index stores the nodes at the l level that are ancestors of the words occurring in the image, as well as the list of local features associated with each node.

Figure 11 shows an illustration of the process of creating the visual vocabulary.

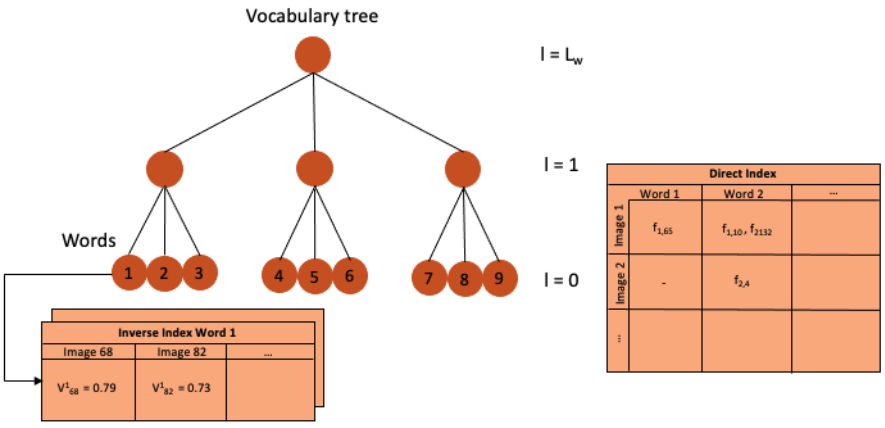

Figure 11: Illustrative example of the vocabulary tree process with direct and inverse indices.

Next, place detection uses the BoW database[3] to determine the candidate and final loops, as described below. Temporal and geometric constraints between each pair of candidate loop images are taken into account to achieve more consistent results in recognising similar places. The implementation includes versions that work with SURF-64 and BRIEF descriptors, but can also work with any other type of descriptor. Thus, for each current image, features are extracted and converted into a bag-of-words vector v_t, with the binary descriptors of the features traversing the tree from the root to the leaves by selecting the intermediate nodes at each level that minimise the Hamming distance. The database is then searched for v_t. The result is a list of matching candidates $< v_t, v_{t1} >$, $< v_t, v_{t2} >$, ..., based on the weighting of the words, i.e. using the inverted file, and their scores s (L_1-score), whose value is in [0..1]:

$$s(v_t, v_{tj}) = 1 - \frac{1}{2}\left|\frac{v_t}{|v_t|} - \frac{v_{tj}}{|v_{tj}|}\right| \tag{4}$$

The range of these values depends on the query image and the distribution of the words it contains. These values are then normalised with the best value we expected for the vector v_t in this sequence, i.e. approximated by $s(v_t, v_{t-\triangle t})$ to obtain the normalised similarity threshold score, η:

$$\eta(v_t, v_{tj}) = \frac{s(v_t, v_{tj})}{s(v_t, v_{t-\triangle t})} \ , \tag{5}$$

where $v_{t-\triangle t}$ is the bag-of-words vector of the previous image.

To avoid a low score leading to a falsely high score, the images that do not reach a minimum $v_{t-\triangle t}$ are skipped. This minimum score

3 images of previously visited places.

represents a compromise between the number of images that can be used to recognise loops and the correctness of the resulting score n. This means that only matches whose score exceeds a minimum threshold α are taken into account. This parameter can be predefined by the user, whereby the use of a small value prevents valid images from being sorted out. Nevertheless, a balance must be found, because a higher similarity threshold means that the algorithm is all the more demanding to deliver loop candidates, as the distance between the vectors of the two images must be smaller. To prevent images that are close to each other in time from competing with each other when the database is queried, they are grouped into islands and treated as a single match. Therefore, only the island with the highest score is selected as a matching group. If an island passes the temporal constraint - more than a user-defined number of k consistent islands ($k = 3$ by default) - only the match (database image) with the higher normalised score is considered as a candidate to close the loop. This candidate must then be accepted by the geometric test in order to be considered a loop. For this purpose, RANSAC is used to find a fundamental matrix between the matching images that is supported by at least 12 correspondences. However, this parameter can be defined by the user depending on the images entered. To calculate these matches, the direct index uses the vocabulary and compares only between the features associated with the same nodes on a user-defined level number (level l) in the vocabulary tree. This condition speeds up the calculation of correspondences. The parameter l is defined in advance and results in a compromise between the number of correspondences between It and It and the time required for this. By default, it is set to 2 (DI_2) as the features in the first levels of the vocabulary tree are usually more distinct and separated, resulting in a good balance between recall value and execution time. Figure 12 shows an overview of the process.

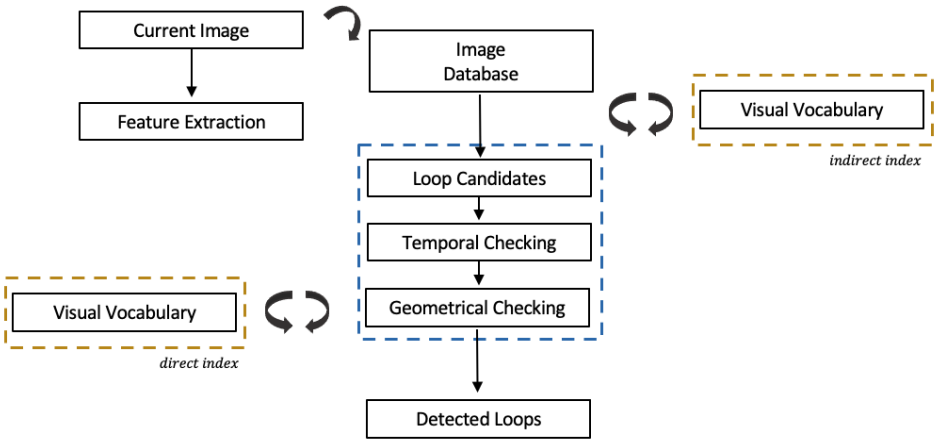

Figure 12: Block diagram of the most important steps of the BoW-based PR approach.

When using image processing for navigation tasks, it is crucial to extract reliable image features, both for key point matching and for subsequent vehicle position estimation and loop closure detection. To achieve this, there are multiple detectors/descriptors, but the binary category is increasingly used in the intended context. Based on the invariance assumed by each descriptor, ORB and BRISK are among the best scale-invariant, rotation-invariant and affine-invariant feature detectors [57] [36]. Therefore, each approach was modified accordingly to work with these binary algorithms. All experiments are based on the extraction of up to 1500 features per image and were performed on a workstation with an AMD Ryzen 7 2700X @ 3.7Ghz with 16GB RAM and NVIDIA RTX 3060.

4.3 EXPERIMENTAL RESULTS

Therefore, the ability of conventional BoW techniques to recognise previously seen places under different conditions is evaluated using the following outdoor community datasets with scene changes, dynamic objects and perceptual aliasing conditions: The KITTI odometry benchmark[4], which consists of 22 sequences in urban and highway environments [58]. Sequence 05 was used, which describes an urban scenario and has various loop closures; part of the St. Lucia dataset[5]. It contains building sites, bright scenes, dark scenes and multi-lane roads [59]; and the New College dataset[6] [60], in which images of a robot are captured as it moves through the environment, which includes buildings, open spaces and vegetated areas. The interest in this dataset lies in the high perceptual aliasing conditions.

Figure 13 shows an example scene from each dataset. The details of the individual datasets are summarised in Table 2.

Figure 13: Example scenario from KITTI 05 (a), St. Lucia (b) and New College (c) datasets.

4 KITTI dataset: http://www.cvlibs.net/datasets/kitti/eval_odometry.php
5 St. Lucia dataset: https://wiki.qut.edu.au/display/raq/UQStLucia
6 New College dataset: https://www.robots.ox.ac.uk/~mobile/IJRR_2008_Dataset/-data.html

Table 2: Overview of specifications from KITTI 05, St. Lucia and New College datasets.

Dataset	No. Images	Size (px)	Rate (Hz)	Length (km)
New College	281	1280 x 960	20	0.42
KITTTI 05	2761	1226 x 370	10	2.2
St. Lucia	2901	1024 x 518	10	2.7

A. Descriptors Effectiveness

An evaluation of the operational abilities of ORB and BRISK is then carried out under some difficult conditions. For this experiment, vocabularies were created with both extractors and the ability to recognise similarities between images was tested based on the DLoopDetector process, as described above in section 4.2. Figure 14 shows the scenarios selected from the KITTI 05 dataset to evaluate the behaviour of the feature descriptors in detecting similar places in some challenging situations, namely scale invariance, changes in viewpoint (including rotation and orientation) and illumination variations.

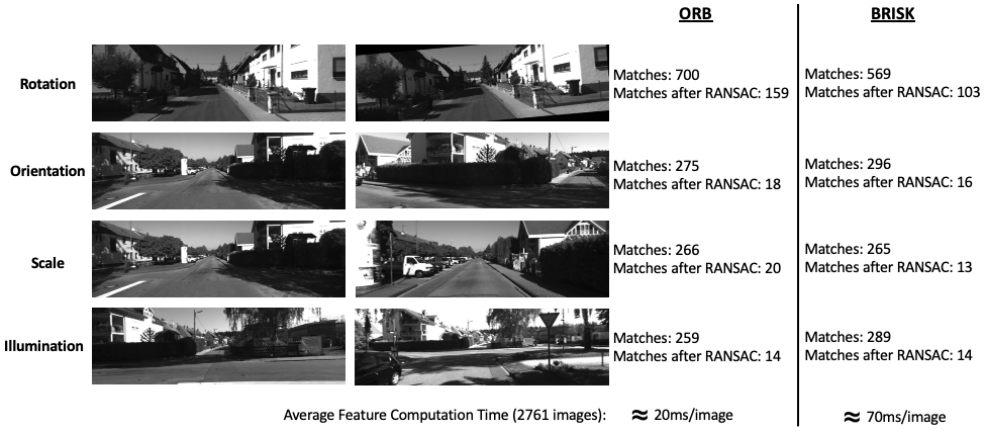

Figure 14: Comparison of ORB and BRISK descriptors against rotation, scale, viewpoint and brightness invariants of scenes from KITTI 05 dataset.

In both cases it is the same place, but it was recorded under different conditions and with different time stamps. Both descriptors have been shown to be able to recognise places as known, demonstrating their robustness to these scene changes. BRISK identifies more matching points between images in most cases. However, after removing outliers - finding a fundamental matrix with RANSAC - fewer matches are found. This means that the ORB matches are effective and therefore less sensitive to changes in the scene, especially with respect to orientation and illumination variations. Moreover, both descriptors are computationally efficient, but BRISK is slightly more expensive: about three times more expensive than ORB.

B. Loop Closure Detection

Next, the performance of the BoW scheme was evaluated based on the ORB descriptor for recognising previously seen places in image sequences. The performance of the place recognition algorithm is then evaluated using precision-recall metrics and the F1-score to find an appropriate combination of the two performance measures. In the context of navigation, however, maximum recall, which can be achieved with a precision of 100%, is of particular interest. Therefore, the evaluation in comparison to ground truth is carried out by counting TP, FP and FN.

Therefore, a ground truth for the loop closure was generated for each dataset. Each dataset provides either a pose file and a matrix or a list of images that are considered to be the same place, i.e. a loop closure. To generate a ground truth for loops based on the distance of camera positions, 4m, 15m and 4.5m are used as thresholds for placing two images at the same location for KITTI 05, St. Lucia and New College, respectively, based on the ground truth information provided for each dataset. In addition, only the loops where the distance to ground truth is greater than the frequency F are considered as final loops to prevent multiple loops from being detected in one second. Finally, the situations in which the camera motion is stopped were discarded. In all experiments, full-indexing vocabularies are used instead of using vocabularies with a training set as suggested by the DLoopDetector library and the usual in BoW techniques. In this way, the operating environment is modelled and the performance in detecting previously seen places is evaluated similarly to online approaches.

The precision-recall curves for KITTI 05 and St. Lucia when varying the threshold similarity parameter α between 0.1 and 0.45 and extracting 1000 keypoints per image are shown in Figure 15. For the sake of clarity, the best results at 100% of precision (Pr) are shown in Table 3.

A high recall rate (Re) without false positives is achieved in both datasets. For KITTI 05, a low similarity value means that false loops are recognised (FP), and the higher the similarity value, the better the algorithm is able to recognise loops without losing any. For St. Lucia, the best performance is achieved at the cost of simplifying the conditions required to recognise a place as known (low similarity threshold, $\alpha = 0.15$). Nevertheless, compared to KITTI 05, more detections are lost overall, with the maximum recall at $\alpha = 0.45$ being 45.90%. If the standard similarity is required for both scenarios ($\alpha = 0.3$), the algorithm is not robust enough to recognise places in St. Lucia without error, meaning that FP are recognised and do not achieve 100% precision. The keypoints extracted for each image therefore represent a simple neighbourhood well. Conversely, 1000 features are far from sufficient to model sophisticated scenes in which the algorithm can recognise places.

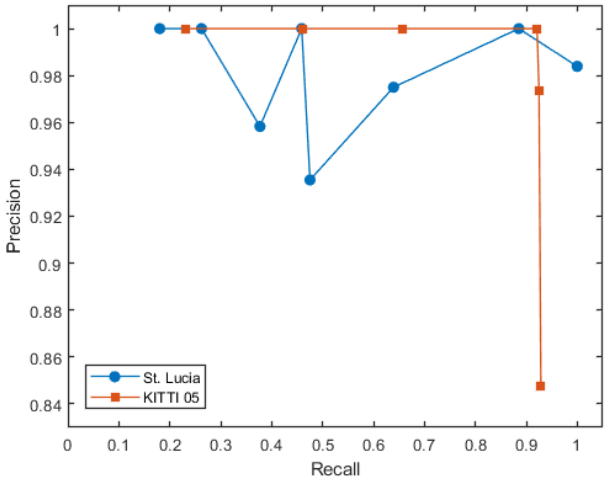

Figure 15: Precision-recall curves for the KITTI 05 and St. Lucia datasets when varying the similarity threshold α.

Table 3: Performance of the BoW-based loop closure at 100% of precision for the KITTI 05 and the St. Lucia datasets when varying the similarity threshold α.

Dataset	Pr (%)	Re (%)
KITTTI 05	100	92.11
St. Lucia	100	88.52

Table 4 shows the performance achieved in the KITTI 05 and St. Lucia datasets by varying the number of keypoints used to create the visual vocabularies (Voc). Specifically, vocabularies with 1000 and 1500 features are created. For image similarity detection, 2000 keypoints per image are extracted and the standard similarity threshold (α = 0.3) is considered for both datasets.

Table 4: Performance of the BoW-based loop closure for the KITTI 05 and St. Lucia datasets, where the number of keypoints extracted in the vocabulary varies.

Dataset	Voc	Pr (%)	Re (%)	F1-score (%)
KITTI 05	1000	100	86.44	92.96
	1500	100	65.79	79.37
St. Lucia	1000	87.88	47.54	61.70
	1500	100	29.51	45.57

Overall, the algorithm performs better in the KITTI 05 environment when the same number of features is used to create the codebook. Using 1000 features in this scenario achieves high recall at 100% of precision, while the algorithm in St. Lucia is not robust enough (87.88%

of precision) to recognise known places without loss and produces many false positives. If you use 1500 features instead, the algorithm in the St Lucia scenario achieves an precision of 100%, but loses more loop closures. This means that a more representative vocabulary contributes to the robustness of the algorithm, but the required similarity is high to recognise all loop closures in a difficult scenario like St. Lucia. The mixture of precision and recognition is higher when 1000 features per image are extracted. However, since this leads to false recognition (FP), it is not the better option in a navigation context as it may lead to inappropriate pose estimation. In KITTI 05, the robustness of the algorithm is not lost, but the recall decreases: since it is a simple scenario, a larger number of features implies too detailed a distinction of the environment, which, in combination with a high similarity threshold, causes the algorithm to lose the strength to recognise known places.

The best results at 100% precision by varying the similarity parameter α and the average calculation time (ms) obtained in the KITTI 05, St Lucia and New College datasets are shown in Table 5. For reasons of clarity, 2000 keypoints per image are used for all methods.

Table 5: Performance of the BoW-based loop closure for the KITTI 05, St. Lucia and NewCollege datasets, for the extraction of 2000 features per image.

Dataset	Pr (%)	Re (%)	Pr(0.25)	Re(0.25)	Time (ms)
KITTI 05	100	90	100	82.21	39.3
St. Lucia	100	68.85	100	50.82	32.8
New College	100	90	94.74	90	47.6

As you can see, more image features are required to distinguish the environment and improve performance in the St Lucia dataset, as this environment is very dynamic and has scene changes. Nevertheless, this result is achieved with a low similarity threshold and still about 1/3 of the loop closures are not recognised. The used sequence of the New College dataset consists to a large extent of walls and green areas (trees, grass and bushes) to evaluate the behaviour of the algorithm with many keypoints but different repetitive patterns. The proposed approach is able to deal with strong perceptual aliasing conditions as the algorithm successfully recognises the queried image as a known place (100% of precision) and identifies almost all loops (90% of recall). However, this best result is achieved at the expense of a high similarity threshold used compared to the other two scenarios. In terms of calculation time, the algorithm proves to be efficient. The biggest difference occurs in the feature calculation phase for the new university dataset, where the required maximum keypoints are extracted in all images.

Table 5 also shows the results obtained by setting a threshold for equal similarity in order to better analyse the impact of setting this parameter to a fixed value in terms of performance. For this purpose, the datasets were processed for $\alpha = 0.25$. As can be observed, the recall in the KITTI 05 and St. Lucia datasets decreases slightly when the similarity threshold is increased (more sensitivity). In contrast, the algorithm in the New College dataset incorrectly recognises situations with a loop closure. Since this scenario has similar features, the required similarity is not sufficient to distinguish it.

Following these results, Figure 16 shows the loops detected by DLoopDetector based on ORB features with a precision of 100% for each dataset. Note that most of the existing loops were detected, especially in the KITTI 05 and New College datasets.

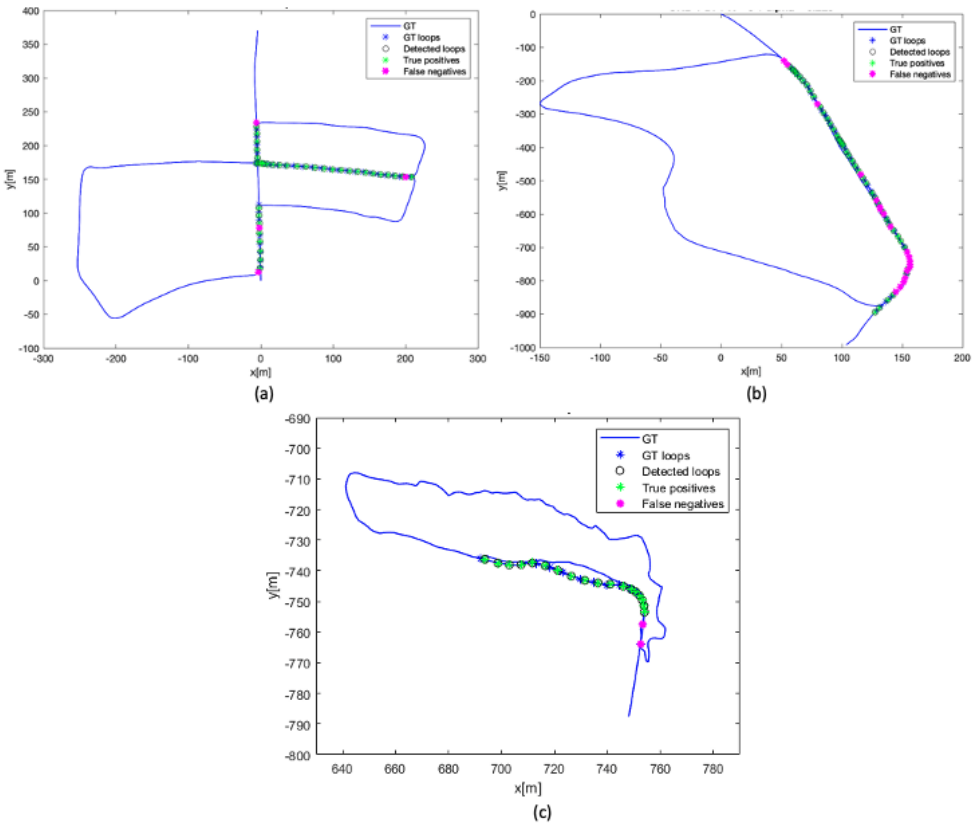

Figure 16: BoW-based loop closure for the KTTI 05 (a), St. Lucia (b) and New College (c) datasets.

4.4 SUMMARY

This behavioural evaluation of a traditional binary feature-based learning technique for uncontrolled environments with scene changes, perceptual aliasing conditions or dynamic elements thus shows that place recognition in KITTI 05 requires a small number of keypoints and a low similarity threshold due to its simplicity in order not to jeopardise recall. In contrast, for St. Lucia, a high number of features is required to represent the environment and achieve a recall of about 70% at 100% precision without high similarity thresholds. Finally, for New College, due to similar features, a high number of features combined with a high similarity is required to ensure distinctiveness for place recognition. Thus, it was proven that a more representative visual vocabulary contributes to the robustness of the loop closure algorithm the more challenging an environment is. The results obtained show that the BoW techniques are able to cope with challenging conditions, albeit at the expense of the number of features extracted (in some cases 2000 features per image). Nevertheless, the results have shown that the required computational effort is reduced and that this approach can be used in a real application with an appropriate hardware selection.

Since the outdoor environment is also unknown and can be just as dynamic as underwater, it is assumed that this technique can also cope with the challenges of underwater scenes, namely in the harbour environment. Therefore, its behaviour in such scenarios is evaluated and VPR method based on binary BoW is proposed later in chapter 6.

5 TOOLS AND ROBOTIC PLATFORMS

Work in the harbours is still carried out by divers or Remotely Operated Vehicle (ROV). However, diving operations are often dangerous and the use of ROV requires an operator over a cable, which is limited due to the semi-structured nature of these scenarios. Therefore, there is a lack of available data that fulfils the requirements for the planned experiments.

This chapter is therefore organised as follows: Section 5.1 gives an overview of the challenges of working in underwater environments and describes the commonly available tools that help us to obtain data from these scenarios. Section 5.2 introduces the Stonefish simulator and describes the creation scenario and the perception datasets collected for the intended context. Section 5.3 describes the acquisition system developed for real operation. It also describes the collected data.

5.1 INTRODUCTION

The underwater scenarios are still unknown and dynamic, i.e. they are characterised as harsh environments: Environmental conditions are uncertain and may change over time, e.g. due to currents, waves, light attenuation, high noise levels, etc. In addition, inspection tasks require close-range navigation, which reduces the number of sensors that can be deployed due to perception limitations. Despite technological developments such as automation that reduce costs, the collection of data on the marine environment is still a very costly, time-consuming and technically demanding activity. Therefore, there is a gap in the available data. Those that are available online do not meet the requirements or are not suitable for experiments in the intended context, i.e. there is no pose information, no ground truth of loops, different depth and speed values of the vehicle or no synchronised data. Nevertheless, it is important to validate the algorithms before they are applied in a real context. In this regard, environmental simulations are a powerful tool for initial validation and pre-testing, as they represent an important method of analysis that is easy to verify, communicate and understand in order to solve real-world problems safely and efficiently. They allow the variation of environmental or vehicle parameters, sensor configurations, different operating scenarios, etc., which is a good way to test the algorithm under different conditions and challenges. Of course,

real data is crucial to validate the efficiency, i.e. the strength and robustness of the algorithms, but for this, the perception solutions must be selected according to the context. This requires defining detailed scenarios that take into account important aspects for the operation, namely the volume of the payload, the weight and available power, the measurement time, the budget or the expected quality of the acquired data. This chapter therefore presents a simulator tool that was used to generate some data for the preliminary validation of the algorithms. A developed perception acquisition system is also described, which is used to generate various data to understand the strength and resilience of the algorithm under real-world conditions.

5.2 STONEFISH SIMULATOR

To facilitate the variation of environmental parameters and sensor configurations, and considering the lack of online data to fulfil the requirements in this context, the Stonefish simulator [61] was used to develop some datasets for the preliminary evaluation of the planned proposal. Stonefish is an open-source C++ library that enables the creation of realistic simulations of marine robots. Particular attention is paid to the representation of a realistic underwater environment, taking into account the effects of light absorption and scattering, instead of using only blue fog as is often the case with other simulators. It is possible to change the position of the sun in the sky, simulate optical effects in the water and also take into account the effects of suspended particles. In addition, it is possible to create special scenarios, including one in which so-called "static bodies" remain fixed at the origin of the world for the entire duration of the simulation; they are typically used for collision and sensor simulations. Static bodies include a simple plane, simple solids (obstacles), meshes and terrain.

5.2.1 Scenario Setup

Therefore, a harbour scenario is recreated to simulate inspection processes in harbour infrastructures that replicate various structures: Quay walls, floating piers and piles, i.e. a berth area. The floor is also simulated and consists of some objects commonly found in harbour facilities, such as rubbish, anchors and metal grids.

Figure 17 shows a wide view of the scenario of the simulated harbour facilities. To create the real appearance of the structures and the seabed, a graphical material called "looks" must be created, which defines how the objects are rendered. All looks are parameterised by reflectance (colour), roughness, metallicity and reflection factor (0 for no reflection to 1 for mirror) to obtain a real simulation of the

echoes returned by the sonar. All structures as well as the ground are considered as rock (material) with a certain roughness and thus without metallicity factor and without reflections.

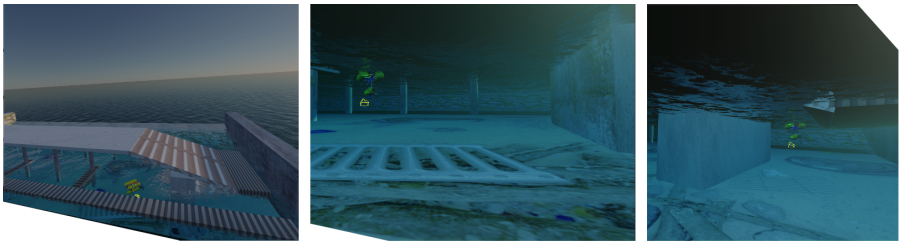

Figure 17: Illustrative example of the simulated harbour scenario recreated with the Stonefish simulator.

To add texture to a material, albedo, height and normal (or bump) maps are created based on original images to represent the appearance, depth or texture of the scenes. In detail, the appearance is based on an albedo map, which removes shadows and brightness from the original image. For the illusion of depth, the height map is used, on which the more detailed objects are displayed. Finally, the details of the scenario - the texture - are represented by the normal map, also known as the bump map. The correct setting of the individual maps is crucial for successful rendering, especially the thickness value used for the normal map. Figure 18 shows some floors that were created to simulate the rendering process. The entire rendering process described is shown in Figure 19.

Figure 18: Illustrative example of the recreated underwater harbour floors for the rendering process in the Stonefish simulator.

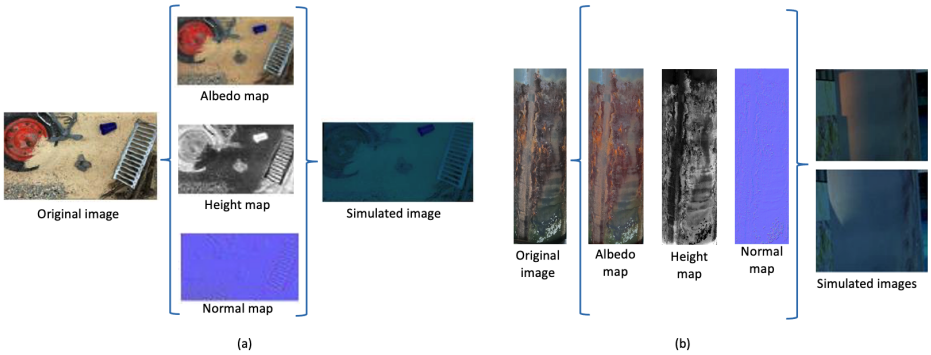

Figure 19: Illustrative example of the rendering process for creating the terrain (a) and the texture of a pile (b) in the Stonefish simulator.

.

The visibility conditions caused by turbidity or sun orientation can change dynamically during the simulation. They are called "WaterType" and "SunPosition" respectively. To implement the optical effects that occur in marine waters, the absorption and scattering coefficients for three wavelengths corresponding to the red, green and blue channels of the rendering pipeline are calculated based on Jerlov measurements covering a wide range of coastal waters. Water quality, i.e. the water type, is defined by a single parameter ranging from 0.0 to 1.0, with the lower limit corresponding to Jerlov water type I and the upper limit to Jerlov water type 9C. There are 10 Jerlov water types that characterise the oceans as clear to turbid: I, IA, IB, II, III, 1C, 3C, 5C, 7C, 9C. The first 5 types (I to III) characterise the ocean waters and the remaining 5 types (1C to 9C) coastal waters. This classification is based on *Kd* measurements (attenuation coefficient from 310 to 700 nm) in the first 100 m water depth [62]. Taking this scale into account, labelling is therefore used to classify each scenario accordingly based on the visibility conditions used (6 for water quality and 3 for sun path), as shown in Figure 20.

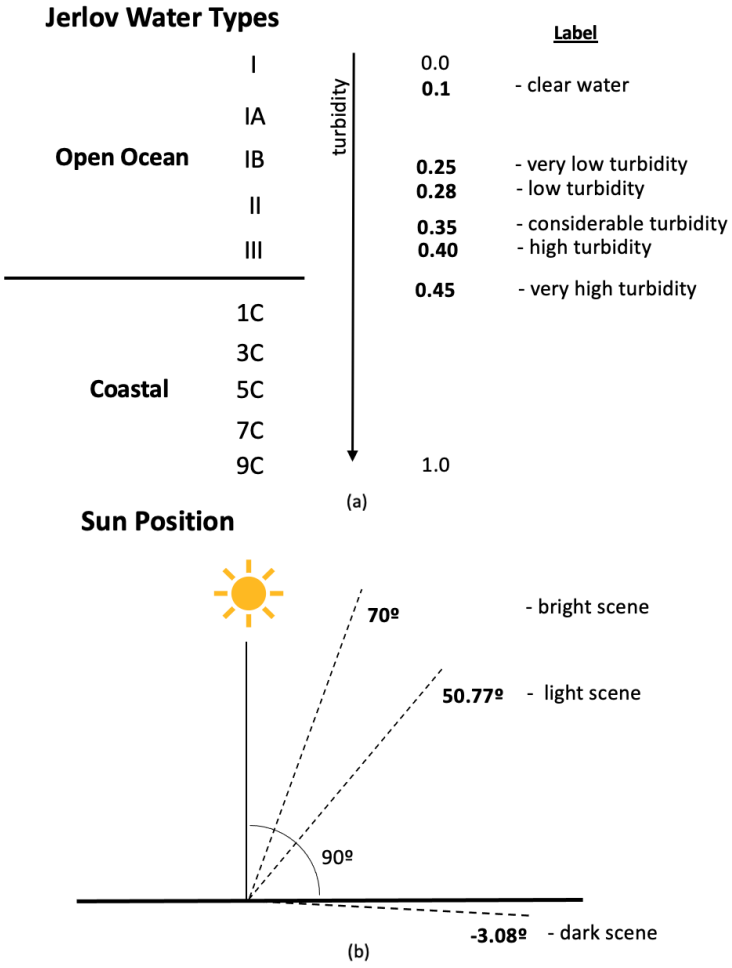

Figure 20: Labelling of the underwater data depending on the levels of turbidity (a) and position of the sun (b) in the Stonefish simulator.

To collect data, the simulated AUV – Girona 500 – was configured to autonomously follow predefined trajectories between different waypoints. Considering the propulsion system of Girona 500 — five thrusters — a state machine was designed for the AUV to perform an appropriate motion according to the required steering motion — e.g. straight ahead or change of direction, with a certain force for smooth navigation to collect reliable data, but also sufficient to respond to the difficulties of the underwater environment (waves, currents, wind, etc.) and even the payload of the vehicle, depending on the scenario as described below. In all experiments, the vehicle was set to navigate at a fixed distance above the bottom - a fixed z-value.

Therefore, according to the planned proposal, a camera, an FLS and an odometry sensor were used to provide position data as ground truth. The position data includes the position, orientation, linear and angular velocities of the vehicle. The standard resolution value is used for each optical image, i.e. each image consists of 300x200 pixels. The resolution of the acoustic images depends on the configured parameters, which have been set to mimic a TriTech Gemini 720ik Multibeam Imaging Sonar, as explained below in section 5.3. Once the simulation is complete, the captured images - optical and acoustic - are saved in the appropriate folders and a file with time stamps for each image is also saved in a text file. The position information (x, y and z) and the respective timestamp are also automatically saved in a text file. The corresponding data of the individual missions are synchronised so that each image has a corresponding timestamp.

5.2.2 Loop Closure Decision

Place recognition is a fuzzy problem, and when using image-based methods there is still no inherently right or wrong answer to the question of whether two images show the same place or not [63]. To solve this problem, ground truth information is usually created to add additional constraints, e.g. based on the maximum allowed spatial distance of the camera positions or the difference between the image indices.

In order to evaluate the behaviour of the place detection mechanism, a ground truth of the loop closures was therefore created for each scenario. To this end, based on the positioning information, i.e. based on the pose file created for each trajectory as a ground truth, a threshold is used to consider two images as the same place based on the effectively mapped area. More specifically, it is assumed that the images must overlap by 30%, which corresponds to a maximum distance of 70%. Two images therefore represent the same location if the difference between their poses (Euclidean distance) is less than 70% of their image dimension (width or height), depending on the

vehicle movement. The width of each image is calculated based on the sensor configuration, i.e. the depth and angle of view, as shown in Figure 21 by the schematic process of the image sensor's field of view.

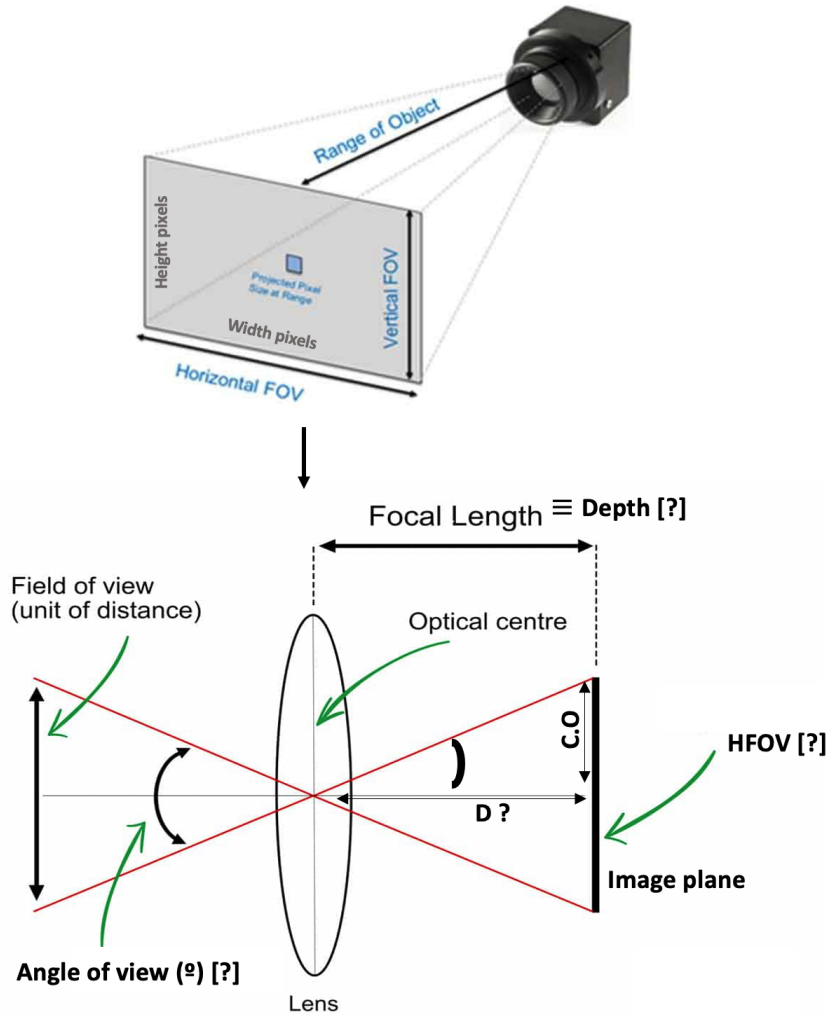

Figure 21: Schematic representation of the sensor's field of view process.

The height is determined from this value and the resolution of the image:

$$\text{VFOV (m)} = \frac{\text{Height pixels} \times \text{HFOV}}{\text{Width pixels}} \tag{6}$$

A mechanism that takes into account the frequency, F, and the speed of the vehicle, v, has also been added to prevent multiple loops from being recognised within 1 second (s): only the loops where the difference between their frame number and the frame number of the previously recognised loop is greater than F x v are considered as final loops.

5.2.3 Datasets

To evaluate the different perception algorithms, the simulator is then used to create multiple underwater datasets that consider different scenarios/paths:

- PortSeafloorA: The vehicle travels a simple route that ends at the starting position, i.e. a situation in which a loop closes with the same viewpoint. The camera has been configured to point downwards with a 44.2° field of view to capture images of the ground. Each image consists of 300×200 pixels, which corresponds to an area of the seabed of 2.7m^2 (camera footprint of 2.04×1.36m). The vehicle was set down at a height of about 2.5m - fixed z-value - above the seabed to simulate a real inspection mission.

  Given the intended context, different parameter configurations were considered to assess the impact of changes in scene appearance on visual place recognition. Therefore, three water types (wt) and sun angles were tested to cover a wide range of coastal waters and to simulate the effects of different depths (shallow or deep water). Suspended particles were also included in the scenario to simulate some dinamism. 7 different trajectories were created, comprising about 1700 images. For the cases with suspended particles, clear water (wt = 0.1) and low turbidity (wt = 0.28), 8 real loops were considered. For the scenario with very low turbidity (wt = 0.25) and for dark (sun angle = -3.08°) and bright scenes (sun angle = 70°), 7 true loops are recognised. For the remaining case, i.e. a light scene (sun angle = 50.77°), there are 6 true loops.

- PortSeafloorB: The vehicle performs a trajectory in the form of an "8", i.e. there are two different situations in which the vehicle revisits the area with a different viewpoint. As with the previously described dataset, the camera was pointed downwards and configured with a 44.2° field of view to capture images of the ground. Each image consists of 300×200 pixels, which corresponds to a seabed area of 2.7m^2. The vehicle was set down at a height of approximately 2.5m - fixed z-value - above the seabed. The same three water types and sun angles of the previous dataset were tested and suspended particles were also included in the scenario. 7 different trajectories were created in this way, comprising about 2500 images. For the case with suspended particles, 18 real loops were then considered. For the scenarios with clear water, very low turbidity, dark and bright scenes, 16 real loops are recognised. The trajectory for the light

scene has 17 real loop closures. In the remaining case, i.e. the low turbidity scenario, there are 15 real loops.

Both data were generated with the Stonefish simulator version 1.1.0. Both sensors, i.e. the camera and the odometry sensor, were configured with approximately the same acquisition rate (30Hz). Figure 22 shows examples of images from the PortSeafloor datasets. Figure 23 shows the two complete trajectories and highlights the respective loops.

Figure 22: Illustrative example of the images captured in the PortSeafloor datasets.

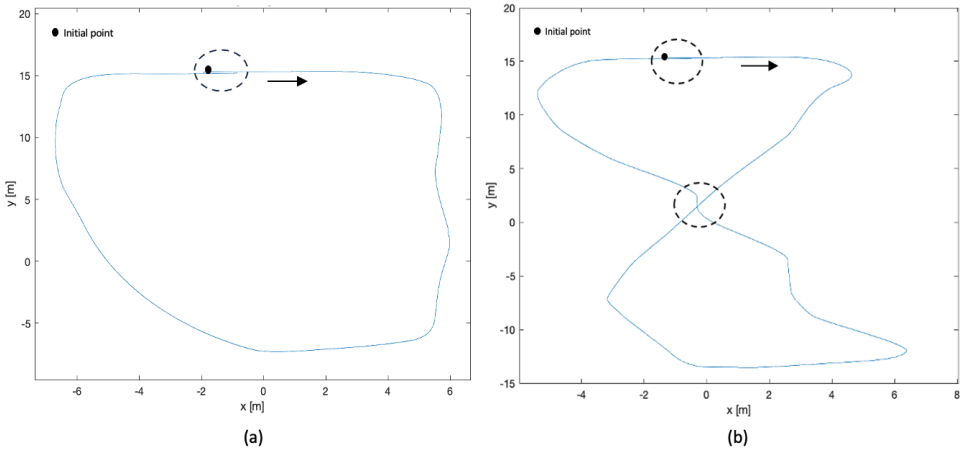

Figure 23: Predefined trajectories of the AUV for PortSeafloorA (a) and PortSeafloorB (b) datasets.

- HarbourArea: The AUV performs a simple trajectory around the harbour structures, as shown in Figure 24 (a), and follows a loop with the same viewpoint, see Figure 24 (b). The camera was pointed at the structures with a viewing angle of 35°. Each image consists of 300×200 pixels: you get a footprint of 1.58×1.05m (an area of 1.7m^2). The vehicle moves at a distance of about 2-3m from the structures to simulate an inspection mission. Both sensors, i.e. the camera and the odometry sensor, were configured with an acquisition rate of about 30Hz. Strict visibility restrictions were taken into account: Cases with very low turbidity and light scenes were tested, with each scenario comprising 1440 (20 true loops) and 1457 images (21 true loops), respectively. This data was generated using the Stonefish simulator version 1.1.0.

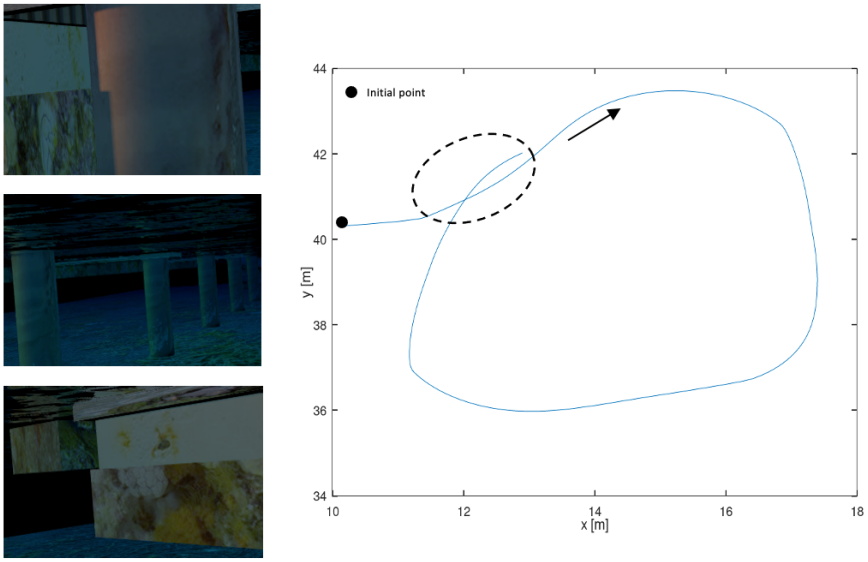

Figure 24: HarbourArea dataset: Illustrative example of the captured images (a) and the predefined trajectory followed by the AUV (b).

- HarbourMissionCAM: The vehicle follows a trajectory close to the harbour structures (the AUV moves about 2–3m away from the structures), i.e. it simulates an inspection around a concrete wall, at a fixed height, z, of 2m above the seabed. It is therefore a closed loop in which the robot maintains the viewpoint. The camera was tilted by 35° and pointed downwards. This sensor was also configured with a 30° field of view to capture images of the ground near structures. As each image consists of 300×200 pixels, this results in a camera footprint of 1.07×0.71m, which means that a seabed area of 0.75m^2 is covered. The camera and odometry sensors were set to a sampling rate of approximately 10Hz and 5Hz. Three water types were considered: considerable, high and very high turbidity scenarios, with trajectories comprising 857, 840 and 805 images respectively. For considerable and high turbidity there are 14 true loop closures and for the very high turbidity scenario 13 true loops are generated.

  In Figure 25 (a) and (b) you can see an illustrative example of the captured optical images and the trajectory performed by the AUV to simulate the inspection of a concrete wall, where this loop closure is particularly emphasised.

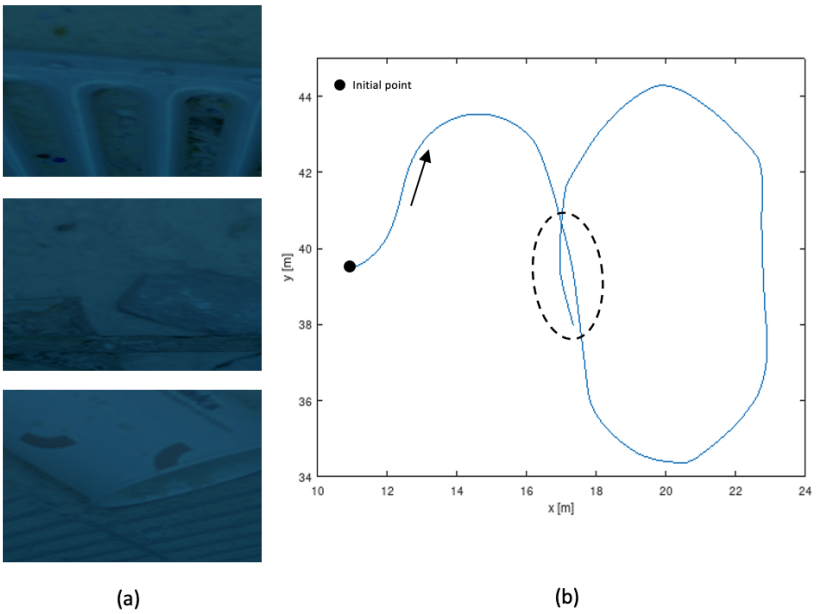

Figure 25: HarbourMissionCAM dataset: Illustrative example of the captured images (a) and the predefined trajectory followed by the AUV (b).

- HarbourMissionFLS: Like the previously described data, Girona 500 performs a trajectory close to the harbour structures (the AUV moves about 2–3m away from the structures), i.e. it simulates an inspection around a concrete wall, at a fixed height z of 2m above the seabed, as shown in Figure 26, which in this case is based on the FLS for data collection.

  It is therefore a closed loop in which the robot maintains the point of view, with 16 real loops. The FLS is a top-down sensor, and since it is intended to provide data about the ground near the structures under inspection, its design and configuration have been adapted accordingly. It has been configured for a range of up to 3m (maximum measured range) at a constant standard height of 2m above the seabed. To create suitable imaging conditions, the sonar was tilted by 35° and had a horizontal FOV of 40°. In addition, the sonar measurements have 512 beams and 750 bins (range resolution of the sonar image) to mimic the Gemini 720ik sonar that will be used later for real tests. Thus, each sonar image consists of 514×720 pixels. Each image covers a width of about 2m and a height of 3m, i.e. it represents an area of about 6 m^2. Both sensors, i.e. the short-time sensor and the odometry sensor, were set with an acquisition rate of about 7Hz. A total of 1067 images were acquired during the trajectory of the mission. Both HarbourMission data were created with the Stonefish simulator version 1.3.0. Figure 27 shows a clear example of the recorded acoustic images.

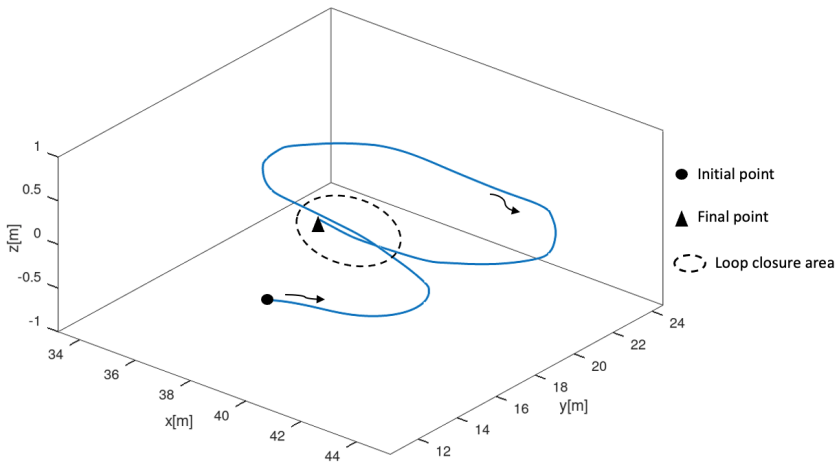

Figure 26: Predefined trajectory of the AUV for HarbourMissionFLS dataset.

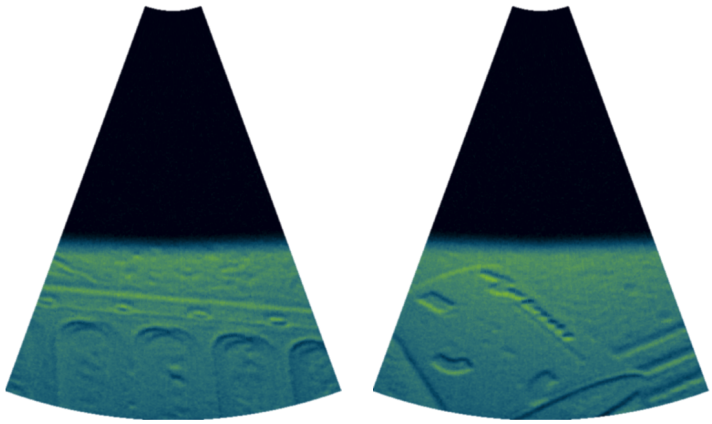

Figure 27: Illustrative example of the images captured in the HarbourMissionFLS dataset.

5.3 ACQUISITION SYSTEM

The simulated data reflect the underwater conditions and thus contribute to an initial validation and a more systematic parameterisation of the various algorithms. But of course they do not yet fully reflect the conditions in the real environment. Therefore, a perception acquisition system that records real data in harbour scenarios with uncontrolled conditions has been developed to evaluate the algorithms that make these features valuable for the development of situational awareness systems for autonomous robot navigation. An autonomous robotic vehicle uses a combination of modules based on various advanced technologies, as illustrated in Figure 28.

The idea is to act on the sensing and perception module to understand the environment and then enable an improvement in navigation concepts. To this end, based on the intended context, the proposed system includes a structured-light system, i.e. a camera and a laser projector, and an imaging sonar, namely a forward-looking sonar.

Each sensor model was initially selected based on an overview of various modern industrial or underwater applications described in the following section 5.3.1.

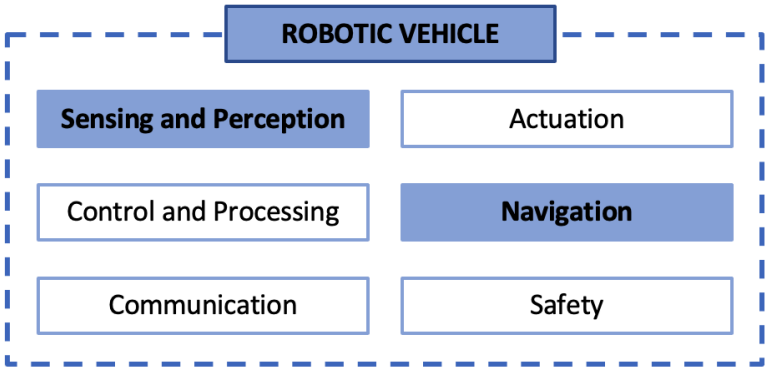

Figure 28: Overview of the operating modules of robotic vehicles.

5.3.1 Sensor Setup

The structured lighting systems - camera + projector - are able to capture closer distances with high resolution, regardless of the texture or colour of the scanned object: by projecting lighting patterns onto the scene, this system creates the illusion of texture on an object and increases the number of identifiable features for each individual position in the image, which can help vehicles navigate in places without texture, which is often the case in the underwater world. The speed of data capture, the non-contact nature, the availability of the required hardware and the high measurement accuracy that modern 3D scanning technologies offer ensure that they are used in many different areas. The development of a robust and practical structured light system therefore remains an active area of research and is still an open problem in computer vision, especially in the underwater environment, as absorption and scattering can occur in this scenario: Most of the existing methods for structured light have been proposed for land-based camera systems and cannot be applied to an underwater system due to the need for a waterproof housing, which in turn can cause refraction. As mentioned in section 2.1, laser-based structured light systems can overcome the limitations of light in the medium - they are insensitive to light conditions - and provide sparse and fast three-dimensional information over a small area, e.g. with the aim of autonomous manipulation. Compared to pattern-based systems, this system is an optimal solution for moving objects. Therefore, several commercial laser-based solutions for the projection of a structured light system were analysed.

Table 6 shows an overview of each solution, whether for industrial or underwater applications.

Table 6: Industrial and underwater commercial laser solutions.

	Industrial			Underwater		
	Multi Projector	*ODSXP30*	*S2000*	*Micro SeaLaser*	*LMNI Subsea*	*HELIOS4*
Group	Osela	Stemmer Imaging	A.G.O	DeepSea	Savante	NovaSub
Type	(multi) dot line	(multi) line grid cross half shpere	dot line		dot line grid	dot line cross
Wave Length (nm)	375, 450-488, 520, 670-810	470, 530, 625, 850 WHI	405, 532, 630-680	635, 532	450, 520, 532, 635	520
Power (V)	10-25	24	12-24	7-30	18-30	15-36
Current (mA)	200	1500	200	47	250	360
Weight (g)	750	410	-	40	275	230
Size (cm)	18x5.5	10x5.0	-	12x2.9	22x3.1	11x2.6

The biggest difference between the two solutions is that the underwater lasers project a simple pattern with a single line, dot or cross. The solution that can project a grid pattern - developed by *LMNI subsea* by Savante - is not yet available. As the intended context requires the creation and increase of the number of identifiable features, the choice had to fall on one of the industrial solutions.

For this reason, Osela's solution - *Multi Projector* - was chosen, also with regard to the required consumption. This projector is based on a combination of diffractive and refractive optics. The Diffractive Optical Elements (DOE) split the beam - the number of lines or dots is preset at the factory - at a precise angular distance while maintaining the characteristics (dimension, divergence and polarisation) of the input laser. In this case, a line projector was used. Lasers are available in many colours, with specific wavelengths available for each colour. As already mentioned, the colour affects both the visibility and the penetration of the laser into the water medium: The shorter the wavelength, the better the laser penetrates. Green lasers therefore penetrate cloudy water better than red lasers.

To evaluate the feasibility of this solution for the intended application - supporting vehicle tracking and place recognition in environments with little or no texture - and its behaviour in underwater environments, a green 3-line laser (520nm) was used[1]. In addition, the most suitable lens type for the working range of the intended application (up to about 3 metres) was chosen in order to achieve greater

1 Laser: https://www.osela.com/multi-line/

homogeneity of the lines in the distance range considered and also a lower thickness, thus improving the energy density (and of course it will be easier to detect the pattern with the camera). In industrial applications, the working depth is irrelevant. In this case, the housing must be designed to allow operation to a depth of approximately 300m: an aluminium housing that allows heat dissipation and a flat acrylic end cap.

As it has proven to be efficient in previous underwater experiments and allows a high acquisition rate (30 fps at full resolution) as well as the configuration of parameters according to the conditions, an Allied Vision Mako G-125B/C GigE camera was chosen as the visual solution[2] with a 6mm lens properly encapsulated in an aluminium housing that allows heat dissipation to capture optical images and work as a structured system when conditions require the capture of images with laser information.

Camera calibration

Some pinhole cameras lead to significant distortions of the images: radial and tangential distortions are the two main types. In short, radial distortions make straight lines appear curved, and tangential distortions occur because the imaging lens is not perfectly parallel to the image plane and therefore some areas in the image appear closer than expected.

The calibration process was therefore carried out in the CRAS laboratory pool using images of a chessboard (9x5 internal corners).

Figure 29 shows an example of captured chessboard images.

Figure 29: Illustration of the calibration process of the underwater camera using images of a chessboard.

In short, we need to find five parameters, the so-called distortion coefficients, which are given by:

$$\text{Distortion Coefficients} = (k_1\, k_2\, p_1\, p_2\, k_3) \tag{7}$$

2 Camera:https://www.spectratech.gr/Web/Allied-Vision/pdf/Mako-G-125B-PoE.pdf

In addition, the geometric properties of the camera such as distortion, focus, rotation and translation must be described so that intrinsic and extrinsic parameters can be estimated. Intrinsic parameters are specific to a camera: they include information such as the focal length (f_x, f_y) and the optical centres (c_x, c_y). You can use these parameters to create a camera matrix that allows you to remove the distortion caused by the lenses of a specific camera. The camera matrix is unique for a specific camera, so once it has been calculated, it can be reused for other images taken with the same camera. It is expressed as a 3x3 matrix:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

Extrinsic parameters correspond to rotation and translation vectors that translate the coordinates of a 3D point into a coordinate system.

So, based on the OpenCV (C++) library, the process is to find the corners of the chessboard in several images, which means that at least 10 test patterns with different perspectives are needed, refine the found corners, and then use the object points and image points to calibrate the camera. Subsequently, all the calibration values mentioned are determined, which are later used for image processing algorithms.

To obtain data when visibility conditions no longer permit the use of visual information, a forward-looking sonar is integrated into this perception acquisition system. As previously described, FLS are becoming a strong alternative for environments with limited visibility as they can provide high quality acoustic images with a frame rate almost equal to that of a video. This technology offers significant advantages over other imaging sonars, thanks to the use of advanced transducer arrays that allow multiple acoustic signals to be captured simultaneously and displayed in a 2D image. In addition, no constant speed or time is required to generate an image from individual pings, as is the case with SSS, for example. Of course, there are some challenges that can complicate the direct use of conventional optical algorithms for acoustic imaging, namely the weak texture of features, the uneven intensity of pixels, acoustic reverberation leading to multiple ghost objects, and acoustic shadows caused by other objects blocking the passage of waves. However, with the advent of much higher resolution sonars, it is now possible to introduce and apply new techniques from the field of optical image processing, and so they have been introduced into optical sensing for underwater vision-based applications. On the other hand, these sensors have good tracking performance at close range.

As water turbidity impairs the ability to extract features from optical images, these acoustic image sensors have been used as a promising alternative in such situations. Several commercially available FLS solutions were analysed to assess the feasibility of these sonars in difficult visibility conditions and to select the most suitable one for the intended application. Teledyne, Tritech and Imagenex were initially identified as the three leading manufacturers of acoustic sensor.

In short, Tritech offers the smallest imaging multibeam echosounders in the world. The 720im and 720ik Micron Gemini models are suitable for the intended application as they can also be used at shorter distances. The main differences concern the field of view and the resolution of the returned data, which is higher for the 720ik model. Teledyne has two different segments of acoustic sensors that can be used as forward-looking sonar devices and are adapted to the specific context: Blue View P900-130 and M-Series Mk2. The first solution provides high-quality sonar images in real time, but offers a low resolution. For the M-Series MK2 model, a compact sonar device with higher resolution and lower power consumption, there are two possible models: 900Mk2 and 900-2250Mk2. The main difference between the two models is the frequency, as the second model offers two possible frequencies that allow it to be used at close range, as it works up to 10 metres away from structures. Imagenex offers two different models to meet the requirements of any application, offering high speed and resolution. As the main difference between the two models is the operating frequency and field of view, the 965A model is considered superior due to its higher maximum range (100 metres instead of 30 metres).

Table 7 shows an overview of the individual solutions.

Table 7: Commercial Forward-Looking Sonar solutions.

	720im	***720ik***	***M900 Mk2***	***M900-2250 Mk2***	***965A***
Group	Tritech		Teledyne		Imagenex
Frequency (KHz)	720		900	900/2250	675
Rate (Hz)	20	97	25		40
FOV (º)	90	120	130		120
Range (m)	50	120	100	100/10	100
Resolution (cm)	0,8	0,4	1,3	0,6	20
Depth (m)	300/750	350	1000		1000
Power (V)	12-48	19-74	12-48		22-32
Current (mA)	360		410	540	780
Weight (g)	244	430	1200	1000	450
Size (cm)	11.9x4.0	12.5x15.0	19.2x10.1	20.6x12.7	26.4x9.2

With a small FOV it is more difficult to find a sufficient match of features between successive images. On the other hand, a higher FOV provides faster images of the environment. The sensors must have a high resolution in order to utilise target identification techniques such as optical imaging. The resolution of a sonar image defines the size of the smallest object that can be detected. Minimising this length is important because the information that can be extracted from a survey depends on the size of the smallest details that can be observed: If the resolution is too coarse, small but potentially important details will not be recognisable or may even be missed altogether. In other words, range resolution is the ability to distinguish between targets based on distance from the sonar device. The lower this value is, the higher the resolution of the image. At shorter distances, this is a particularly important factor as it significantly affects the resolution of the images you see on the screen.

The choice therefore fell on Gemini 720ik (Tritech)[3]. It is a real-time system with a data refresh rate of 97 Hz that delivers a high-quality, video-like imaging. In addition, it has been specially developed for use in shallow waters and has an integrated Velocity of Sound (VoS) sensor that ensures extremely accurate imaging and measurement. This model produces images of outstanding clarity and benefits from 512 beams, resulting in an angular and range resolution of 0.25^{o} and 4 mm respectively. Its small size and light weight, as well as its proven, reliable performance, make it the ideal sonar for use on underwater vehicles for navigation, obstacle avoidance and target detection, as well as for diving operations.

Therefore, this solution for the perceptual setup comprises:

- 1 x Osela green 3-line laser (520nm), up to about 3m of range operation;
- 1 x Mako G-125 with a 6mm lens at 30 fps and a resolution of 1292× 964;
- 1 x Gemini 720ik FLS, 97 Hz, 4mm of range resolution, measuring distance up to 120 m and fied of view of 120^{o} horizontal and 20^{o} vertical.

A schematic representation of the system for capturing perception data can be found in Figure 30. In short, the "Acquisition System" consists of an Odroid to which the camera and the Forward Looking Sonar sensors are connected. The acquisition starts with a command received from an external PC via an Ethernet connection. The data is then sent to the external PC for storage.

3 FLS: https://www.tritech.co.uk/files/Datasheets/Gemini-720ik-2023-datasheet-rev9.pdf

A command is sent to the Odroid when data acquisition is to be stopped. The "Laser" module contains a laser that is connected to an Arduino. From the external PC, the line projector can be switched on and off via a network connection if it is to be detected by the visual sensor.

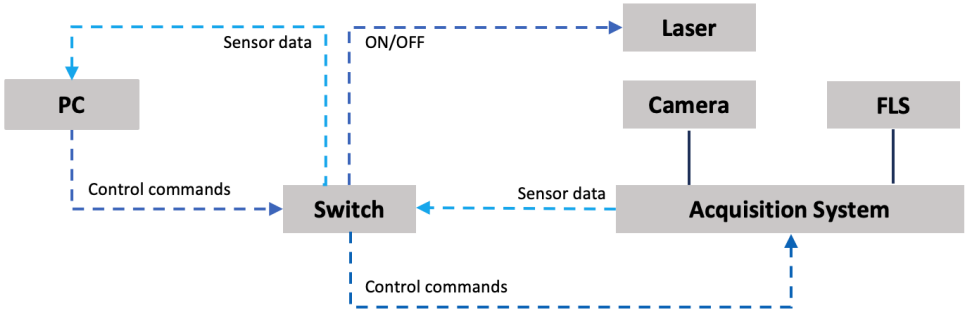

Figure 30: Schematic representation of the perception acquisition system.

The underwater GPS-G2 system[4] is used to obtain the real-time positioning of the mission during data acquisition. It is a SBL system that provides reliable, accurate and robust acoustic localisation of a single tracking device. Particularly noteworthy is the reliable positioning in the most difficult environments, i.e. in shallow water, in tanks/pools, near installations, etc. The G2 Topside is connected to the G2 antenna, which must be placed in the water with a direct line of sight to the locator. A Wi-Fi signal emitted from the G2 Topside's internal web server allows the user to track the locator via the Water Linked web-based GUI. An integrated GPS and Inertial Measurement Unit (IMU) provide an absolute GPS position of the locator, so no additional third-party sensors are required for accurate navigation. The G2 system is individually configured at the time of order to meet the specific requirements of the user and the target to be tracked, with a range of 300 metres required. The G2 system is supplied as standard with a battery powered U1 locator so that the G2 system can be used to track any target (e.g. ROV or diver). The tracking device does not need to be integrated, only physically attached to the target.

Later, it is then possible to establish a ground truth for the loop closure situations explained in section 5.2.2 and thus to evaluate well the performance of the proposed methods, either in terms of localisation over the studied structure or in terms of identification of similarities between the same locations in the different timestamps.

Figure 31 shows the sensor models used for the proposed perceptual data acquisition system.

4 GPS:https://waterlinked.com/web/content/7539?unique=0e0957583b670554b27abb667a70927bcf35c6a7

In order to collect the data according to the intended context, this proposed acquisition system is coupled with the laboratory's ROV, the so-called Customisable Remotely-operated Aquatic Bot, see Figure 32.

Figure 31: Sensor models of the perception acquisition system.

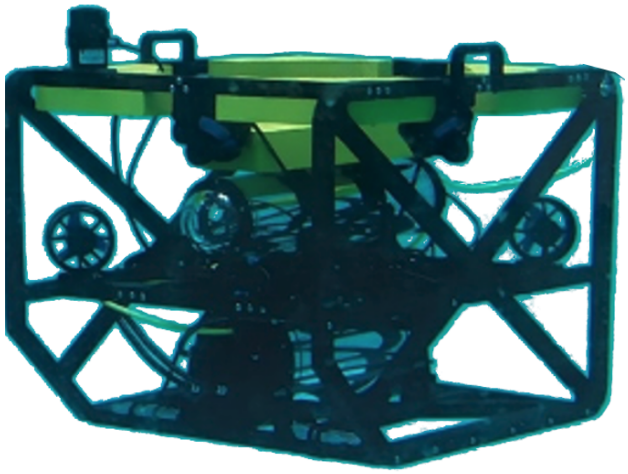

Figure 32: Remote-controlled underwater vehicle: CRAB.

As the name suggests, this is a remote-controlled vehicle with a 150 metre cable. It consists of 8 brushless motors that allow 6 degrees of freedom and has a 470 Wh battery that allows an operating time of more than 3 hours. It is capable of reaching a maximum depth of 60 metres. The vehicle has an IMU, a pressure sensor and an altimeter as standard. The operator interacts with the CRAB via the control console. This software makes it possible to send commands to the robot and receive measurements of variables related to the robot, such as telemetry, and to display the visual or acoustic data in real time to control the vehicle according to the planned mission.

Figure 33 shows the sensor system installed in CRAB and highlights the individual sensors.

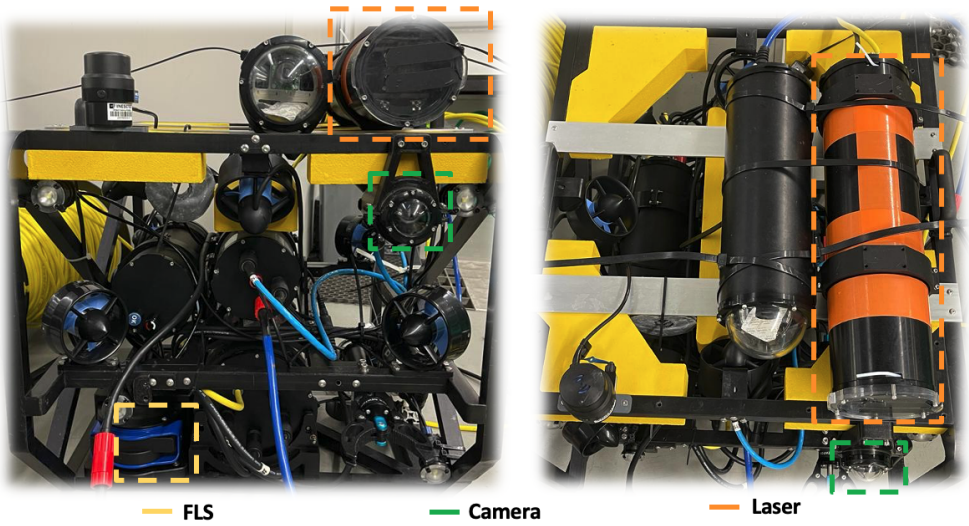

Figure 33: Perception acquisition system mounted on CRAB.

5.3.2 Datasets

In order to evaluate the different perception algorithms in a real environment, the proposed setup is used to create multiple underwater datasets that consider different scenarios/paths.

A special focus is placed on an optical data campaign recorded in the port of Ponta Delgada, Azores, during the Advanced Vehicles for Exploration of the Seas (WAVES)[5] from 5 to 7 July 2023, which is described below:

- PortAzoresCAM: CRAB performs an L-shaped trajectory with a return to the starting position near harbour structures such as piles and 2 floating piers at a fixed height, z.

  It is therefore a closed loop in which the robot changes the angle of view. The camera was pointed at the structures with a view angle of 60^{o} and each image consists of 640×480 pixels. The camera recorded data at 30 fps and the positioning sensor ensures a rate of at least 30Hz. The trajectory comprises about 4000 images and was recorded under difficult visibility conditions, i.e. in turbid water, as working in a factory affects the water quality, and under the influence of sunlight, which should be clear compared to the water in the harbour of Leixões in Porto.

5 Workshop organised by CRAS-INESC TEC, CIBIO-Azores and IS2E in Ponta Delgada, Azores, to promote advances in marine and underwater robotics for various institutions/companies and to explore problems, challenges and solutions for use in the Azores Sea.

Figure 34 shows examples of images on missions in the Port of Ponta Delgada, Azores.

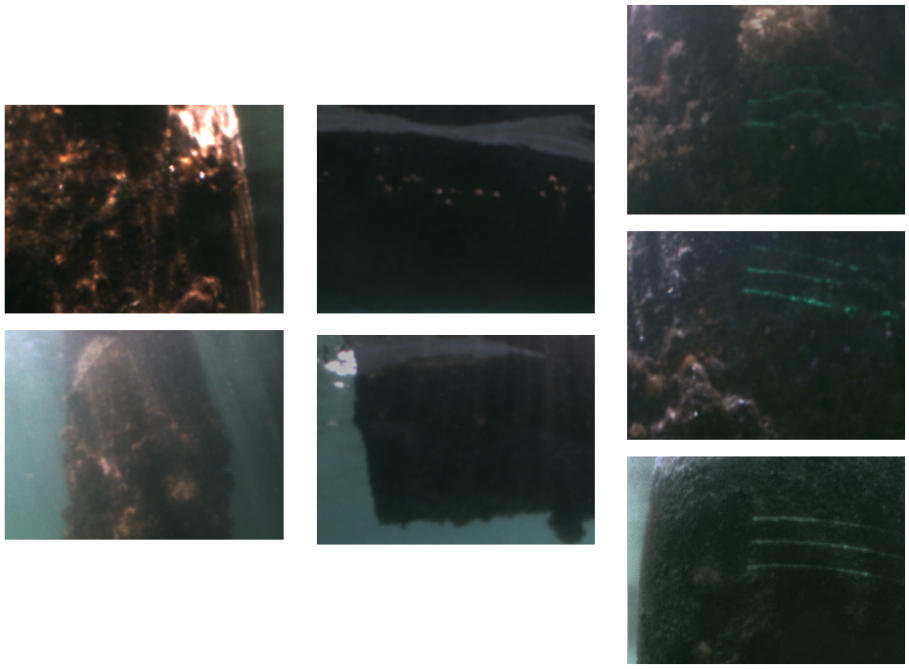

Figure 34: Illustrative example of optical images taken by CRAB during an operation in the Port of Ponta Delgada, Azores.

6 VISUAL PLACE RECOGNITION

Harbour facilities have a number of structures that need to be inspected regularly. To successfully navigate and perform these tasks, robotic vehicles must be able t recognise previously seen places and deal with the inherent perceptual problems of these underwater environments. Therefore, this chapter presents a behavioural evaluation of a method for visual place recognition for harbour scenarios under typically poor visibility conditions.

The chapter is organised as follows: Section 6.1 gives an overview of the problems of visual place recognition in underwater environments. Section 6.2 presents the architecture of the visual place recognition approach and describes its behaviour based on underwater input images under different visibility conditions and under variation of the different configuration parameters. Then, in section 6.3, an enhancement technique is proposed as a pre-processing step to improve the performance of the place recognition approach and to understand to what extent the cameras can overcome perceptual problems. Finally, the main conclusions of this chapter are discussed in section 6.4.

6.1 INTRODUCTION

For the regular inspection of port facilities, underwater vehicles must be able to navigate in a practicable manner and thus recognise a previously visited place, as this is an essential aspect to compensate for the accumulated pose deviations. This mechanism of place recognition requires effective identification of landmarks between images taken at different timestamp, i.e. the selection of very robust features is crucial for similarity detection. Traditional machine learning techniques focus on identifying previous observations, which makes the navigation principles very stationary and dependent on the representativeness of the collected observations. Although it is challenging to make this decision independently of the environment, i.e. through unsupervised learning, it is important that the vehicle continuously learns the environment during the inspection mission to adapt its behaviour accordingly, especially in the underwater scenario where the appearance may change over time. Navigation near structures is still a challenge due to perceptual limitations. Machine vision systems are an attractive solution for environmental detection in rugged close-range operations

as they operate at distances of less than 3m, provide comprehensive information and are easy to use.

Given the advantages of using binary features and the strengths of BoW techniques described in the literature, a behavioural evaluation of an unsupervised learning method based on a binary BoW was performed on data containing scene changes, perceptual aliasing conditions or dynamic elements, as described in chapter 4. Full-indexing vocabularies were used for continuous modelling of the scene, so that the recognition of revisited areas is analysed as an incremental learning approach. It shows a good balance in handling such difficult conditions at low computational cost. To take full advantage of these sensors for short-range navigation, this chapter presents an evaluation of this technique in an underwater environment where localisation and mapping is more difficult because this environment is naturally still unknown and uncontrolled and the appearance of a place can change over time.

6.2 IMPACT ASSESSMENT OF ENVIRONMENT CONDITIONS

The nature of the underwater environment presents a number of challenges for the accurate positioning of vehicles: It is dynamic and lacks structural features. In addition, harbour scenarios are strongly influenced by light incidence (shallow waters) and turbidity, which often complicates the behaviour of navigation and mapping tasks by cameras and makes the perception of the environment a challenge, as the perception range of optical devices is severely limited in very poor visibility. Such conditions lead to a lack of image detail in such images, which affects the selection of very robust features and makes it difficult to detect loop closures, and the vehicle may fail to recognise some loops correctly or detect some erroneous loops so that the trajectory is not adjusted or adjusted incorrectly.

So, "*are the visual approaches able to deal with the inherent challenge of the underwater environment of harbour scenarios?*" This section therefore evaluates the behaviour of the binary BoW technique in harbour scenarios to analyse the feasibility of a purely visual solution for similarity detection under strict visibility conditions.

6.2.1 Methodology

Based on experiments in challenging scenarios where the appearance changes similarly to underwater and a comparative study of different place detection methods [64], a tree-based place detection based on

the DBoW2 and DLoopDetector libraries is also proposed here [65]. Following the experiments in challenging outdoor scenarios presented in section 4.3 and based on a state-of-the-art evaluation of different feature detectors/extractors for underwater [37], ORB features are also considered for the experiments in underwater environments. Figure 35 shows the process overview for image similarity detection.

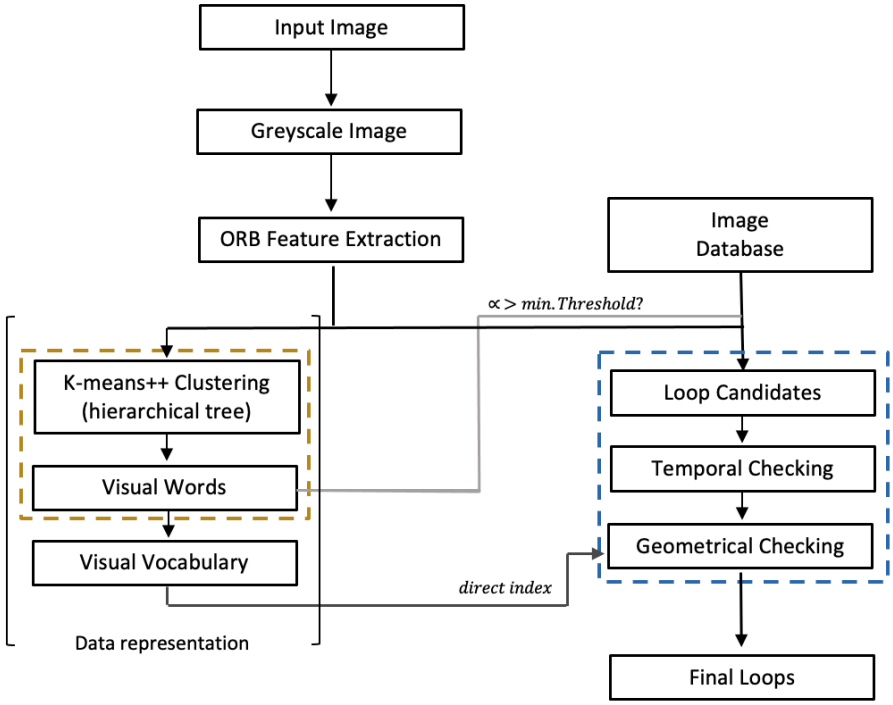

Figure 35: Block diagram of the proposed VPR algorithm based on optical images.

All input images are converted to greyscale colour space and then for each image, ORB features and descriptors are extracted for each image. Based on the features, the vocabulary tree is then built based on the agglomerative hierarchical K-Means++, where the depth factor L is set to 6 and the maximum number of nodes or groups for each level, K, is set to 10, based on the features of the underwater images as explained previously. So, based on the Manhattan distance, the clustering steps for each level are performed based on the clusters obtained in the previous level until L is reached. You will then get a tree with W leaves - words of the vocabulary. Each word is assigned a weight based on its relevance in the training corpus. An inverted and a direct file are also kept along the bag-of-words: For each word, the weight and a list of images in which it occurs (inverted index), and for each image, the nodes at the l level that are ancestors of the words occurring in the image, as well as the list of local features associated with each node (direct index) are stored. The visual vocabulary is then stored, as indicated by the yellow block in Figure 35.

Next, as indicated in the blue block (Figure 35), a greyscale colour space conversion is applied to each current image to extract ORB features and then recognise similar places: Based on the Hamming distance, the features are converted into a BoW vector v_t and the database is searched for v_t. The result is a list of matching candidates based on the weight of the words and their normalised scores (L1 score, i.e. Manhattan distance), represented by the light grey connection. Only matches whose score exceeds a minimum threshold α are considered. To prevent images that are close in time from competing with each other when querying the database, they are grouped into islands and treated as a single match, i.e. only the island with the highest score is selected as a match group. If an island passes the time constraint - more than $k = 3$ consistent islands - only the match (database image) with the higher normalised score is considered as a candidate for closing the loop. To obtain more consistent results, this candidate must then be accepted as a loop by the geometric check by finding a fundamental matrix between the matching images with RANSAC supported by at least 12 correspondences, *minFPoints*. In order to calculate these matches, a direct index is used that analyses the vocabulary, with the comparison being made only between the features associated with the same nodes at a level $l = 2$ in the vocabulary tree represented by the dark grey connection.

6.2.2 Experimental Results

In this way, the behaviour of the proposed BoW scheme based on ORB features is evaluated for image similarity detection in underwater scenarios under challenging appearance conditions, including some common severe visibility conditions caused by external and intrinsic environmental factors. For each scenario, full-indexing vocabularies previously created with images of the trajectory performed are created to model the operating environment and evaluate the performance in the recognition of seen places similar to online approaches. All experiments are based on the extraction of up to 1500 features per image and were performed on a workstation with an AMD Ryzen 7 2700X @ 3.7Ghz with 16GB RAM and NVIDIA RTX 3060.

Datasets

To evaluate the performance of the BoW technique in underwater environments, the PortSeafloorA and PortSeafloorB datasets described in section 5.2.3 are used. They were recorded with the Stonefish simulator from a camera installed on the AUV and pointed at a harbour floor. As a reminder, Figure 36 shows the complete trajectories and highlights the respective loops.

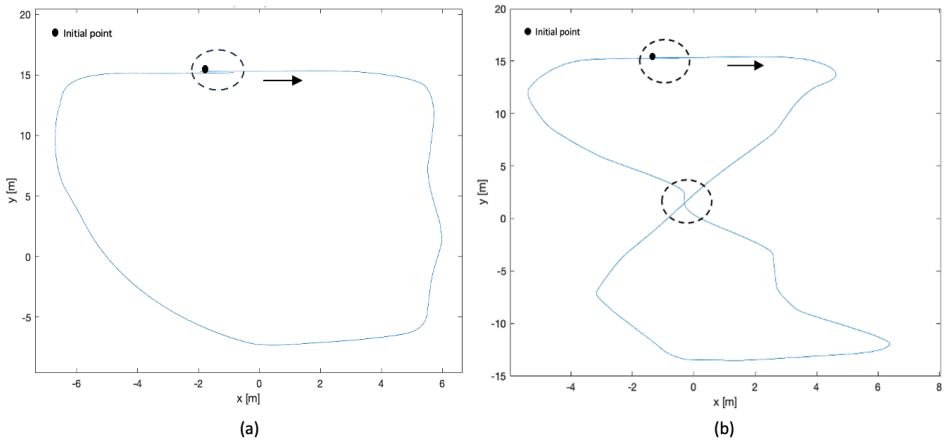

Figure 36: Predefined trajectories followed by the AUV for PortSeafloorA (a) and PortSeafloorB (b) datasets.

Different parameter configurations were then considered for each trajectory to assess the impact of changes in scene appearance on visual place recognition. As previously mentioned, some water types (wt) were tested using three Jerlov measurements to simulate clear water (wt = 0.1), very low (wt = 0.25) and low turbidity (wt = 0.28), as suggested in section 5.2.1. To simulate the effects of light incidence, three sun angles (also referred to as sun position or sun orientation) were tested to create dark (sun position = -3.08°), light (sun position = 50.77°) and bright scenes (sun position = 70°). Suspended particles were also included in the scenario to simulate a certain dinamism. Figure 37 shows an illustration of the effects of these environmental parameters on the captured images.

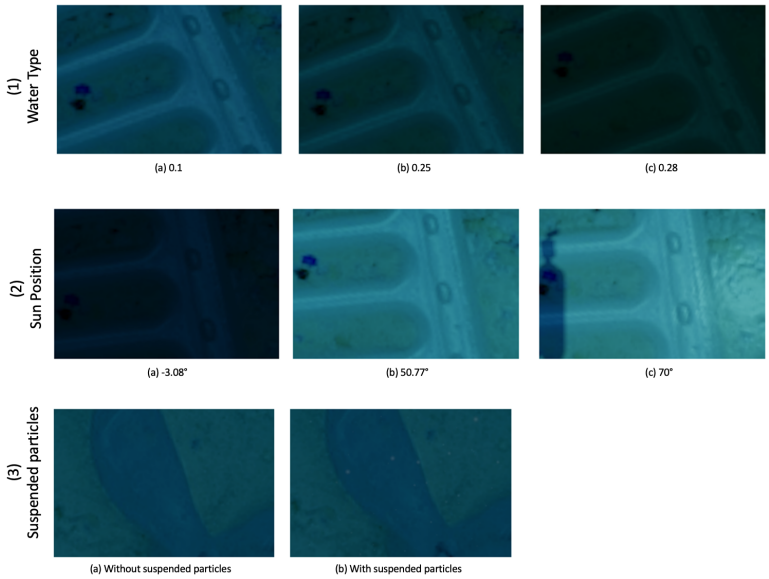

Figure 37: Illustrative example of the effects of turbidity (1), the position of the sun (2) and the presence of suspended particles (3) on the appearance of the images.

Loop Closure Detection

To evaluate the behaviour of the algorithm, the performance metrics Precision, Recall and F1-score are calculated to measure its strength and robustness in detecting previously seen places, as described in section 4.3. For this purpose, a threshold of approximately 1m was used for placing two images at the same location when creating the ground truth for loop closure, based on the footprint of the camera. TP, FP and FN are then also considered for the evaluation against the ground truth. Loops detected by the algorithm whose distance is less than 0.3m are regarded as TP; loops whose distance is more than 0.3m are regarded as FP. If there are loops in the ground truth that do not match the loops recognised by the algorithm, these are considered FN.

The precision-recall curves for all scenarios in the PortSeafloorB data that vary the similarity threshold (α) between 0.2 and 0.35 are shown in Figure 38.

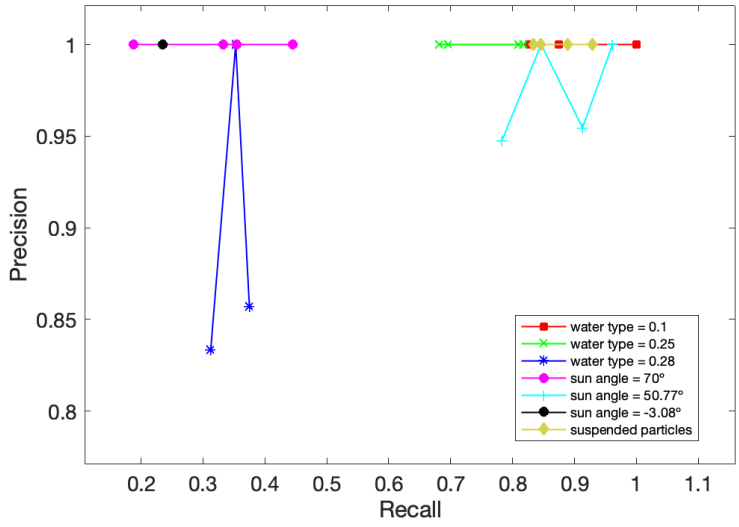

Figure 38: Precision-recall curves in all visibility scenarios for the PortSeafloorB dataset when varying the similarity threshold α.

For low turbidity and light scenes, the behaviour of the algorithm is not consistent: for some similarity values, a precision of 100% is not achieved (False Positive (FP) are detected), which is not desirable in a navigation context, as this can lead to inappropriate trajectory adjustments and thus to incorrect pose estimation. In general, a similarity threshold of 0.2 or 0.3 achieves better results in retrieval accuracy. Nevertheless, the similarity requirements are low at α=0.2, as distant features are considered as the same points and an overfitting result can be obtained in some scenarios. Therefore, α=0.3 proved to be the most suitable and was therefore used for all subsequent experiments.

Thus, the three water types (wt) mentioned, i.e. clear water, very low and low turbidity, were tested for both trajectories. As the image gets darker, the image contrast decreases and the different elements in the image are no longer easily recognisable, as you can see below in Figure 37 (1).

For PortSeafloorA in clear water, the algorithm only fails to recognise a loop closure (FN), as can be seen in Figure 39. An accuracy of 100% and a recall of 85.71% were achieved, which means that all loops were recognised well. Together with the incorrectly unrecognised situation, this also resulted in an F1-score of 92.31%. In this scenario, the performance of the algorithm only decreases by 8% with increasing water type, which was not expected given the difficult visibility conditions visible to the naked eye. This can perhaps be explained by the fact that this is a simple trajectory with only one long-term loop recorded with the same viewpoint at different times. However, looking at the number of extracted features and the final matches between the images, a decrease in this factor can be observed with increasing turbidity. It is therefore more difficult to recognise robust features that could hinder adequate navigation under these difficult conditions.

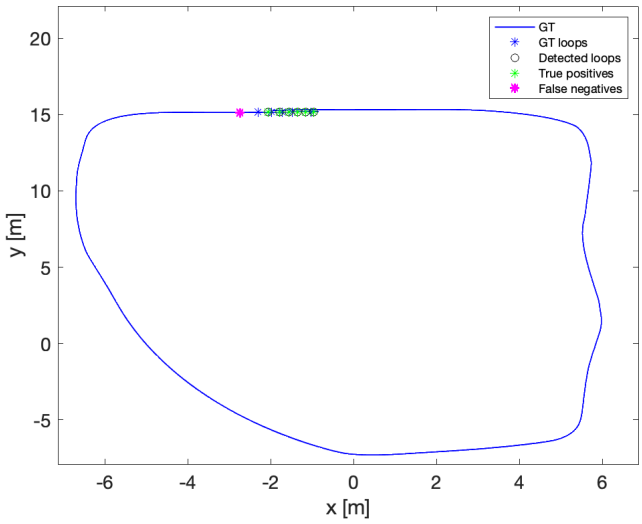

Figure 39: Appearance-based loop closure for the PortSeafloorA dataset in the clear water scenario.

For the PortSeafloorB data, the quality of the algorithm's performance deteriorated as expected with the increase in the water type parameter. Figure 40 shows the results obtained for each simulated water type. In this trajectory, there are two situations where the loop closes, making it more difficult to achieve generally good performance. As can be seen in this case, the decrease in image quality led to a delay in loop detection, and in the worst case the algorithm is not able to recognise the first loop situation, see Figure 40 (c).

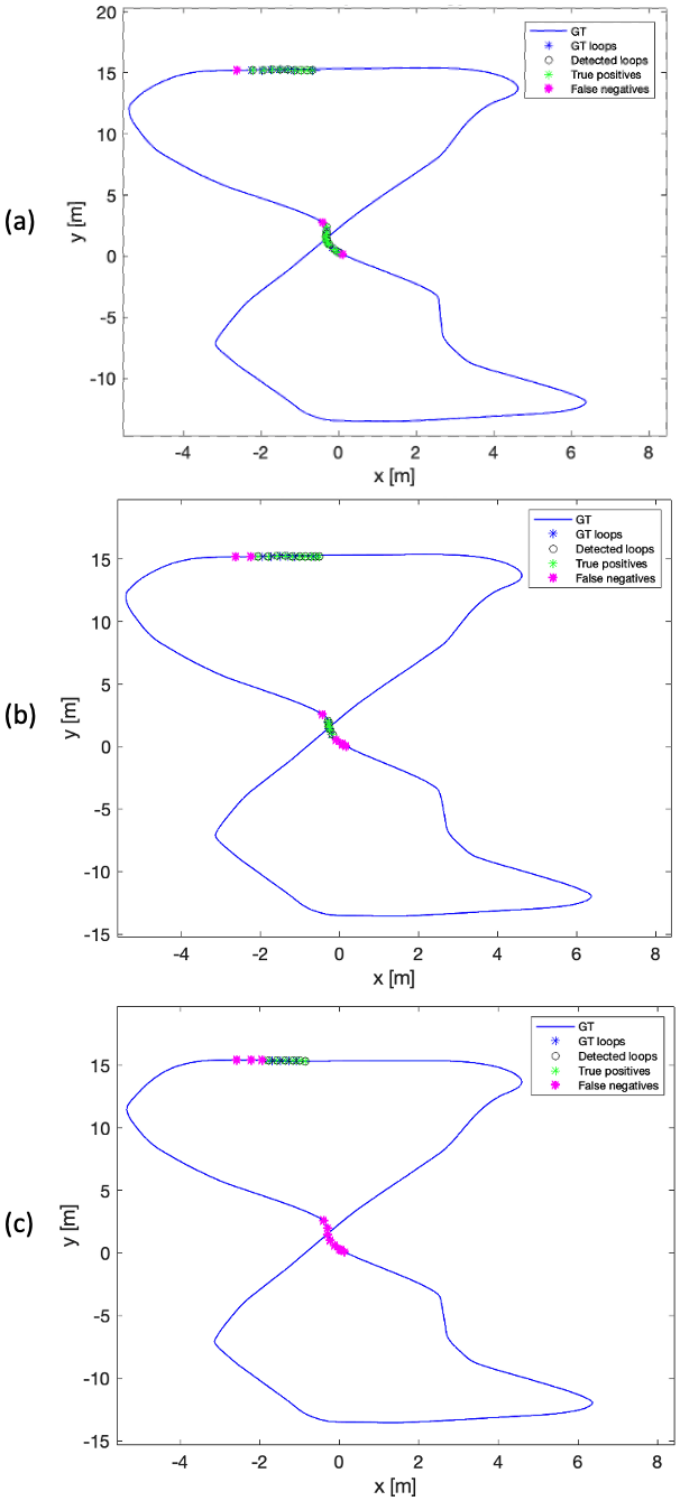

Figure 40: Appearance-based loop closure for the PortSeafloorB dataset for clear (a), very low turbid (b) and low turbid (c) waters.

For a quantitative view, Table 8 shows the performance of the algorithm for recognising similar places. It achieves an accuracy of 100% in all cases, which means that no loops were falsely recognised, i.e. there are unique features.

Table 8: Performance of the VPR algorithm for the PortSeafloorB when varying the water type (wt), i.e. the turbidity level.

	wt = 0.1	**wt = 0.25**	**wt = 0.28**
Precision (%)		100	
Recall (%)	87.50	69.57	35.29
F1-score (%)	93.33	82.05	52.17

In contrast, recall decreased during the experiments when the environmental conditions changed, making it more difficult to extract the matching features. For example, with very low turbidity, the algorithm does not recognise 7 situations in which a loop closes, which means a decrease in recall by 20%. Nevertheless, it is able to deal with different viewpoints in both cases. In the last scenario, a greater decrease in recognisability (recall) can be observed, as no loop was recognised in the intersection area. The algorithm is therefore not able to cope with several challenges at the same time: dark environment, short loops (small overlap between images and small similarity area) and change of viewpoint. This behaviour is not reliable in a navigation context, as a false negative result can lead to an incorrect adjustment of the trajectory and thus to an unreliable pose estimation.

Next, the sun angle was changed between the three mentioned tilt angles, i.e. dark, light and bright scenes, to evaluate the effects of illumination variation on similarity detection between the images. This test makes it possible to simulate the various common scenarios in underwater operations, namely shallow waters and even deep areas. Figure 37 (2) shows the effects of the individual sun angle inclinations on the captured images. Table 9 shows the quantitative results for the trajectories performed to acquire the PortSeafloorA data.

Table 9: Performance of the VPR algorithm for the PortSeafloorA when varying the position of the sun, i.e. the angle of incidence.

	angle = -3.08°	**angle = 50.77°**	**angle = 70°**
Precision (%)		100	
Recall (%)	50.00	62.50	62.50
F1-score (%)	66.67	76.92	76.92

For dark scenes, the algorithm is similar to that for low turbidity, as visibility is also very poor. However, there are more than two false non-detections (FN) in this sun experiment, which corresponds to a 27% reduction in recall. In light scenes, the algorithm achieves a higher performance with a 12.5% increase in recall. In the last scenario, the calculated recall is the same compared to the previous result (light scene), but the number of extracted features and identifiable matches is

lower. Although the image quality is higher for humans, the algorithm is more sensitive to this last configuration - bright scene - than to the worst visibility condition in terms of water type, i.e. the low turbidity case, and shows lower recall - 77% versus 62.5%. Bright scenes indeed have a lower image contrast, which makes feature extraction and thus similarity detection more difficult.

Figure 41 shows the results of the same experiment, but now for PortSeafoorB data. It can be seen that the algorithm leads to better detection of loops as the illumination level increases, but when the sun tilt is high, the performance decreases and some loops are not detected. This can also be explained by the fact that a higher sun angle leads to more unstructured images, which makes feature extraction more difficult. Another problem caused by a higher sun position is the changes in appearance caused by the shadow of the vehicle. These changes are dynamic and can therefore affect the similarity detection between images taken at different times. Table 10 therefore shows the significant drop in performance in place recognition at extreme sun inclinations. In dark scenes, the algorithm starts with a delay in similarity detection - one more false negative. Nevertheless, neither case is suitable for real-world use, as there are many unrecognised loop closures. In short, in the best cases, i.e. a light scene and clear water, the algorithm behaves similarly and fails to recognise 3 loop closures.

Table 10: Performance of the VPR algorithm for the PortSeafloorB when varying the position of the sun, i.e. the angle of incidence.

	angle = -3.08°	**angle = 50.77°**	**angle = 70°**
Precision (%)		100	
Recall (%)	23.53	84.62	33.33
F1-score (%)	38.10	91.67	50.00

The effects of the suspended particles are therefore shown in Figure 42. It was to be expected that the algorithm would incorrectly recognise situations in which loops close (precision $< 100\%$) and would not recognise others (FN), as the dynamics change the appearance of the scene. But only the last situation occurs, perhaps because the set of particles to be simulated is small and sparse. Compared to the experiment without particles (clear water), recall only decreases by 3%. In addition, the number of final matches is also lower, indicating that the particles did not add any value to the feature extraction - no matching features.

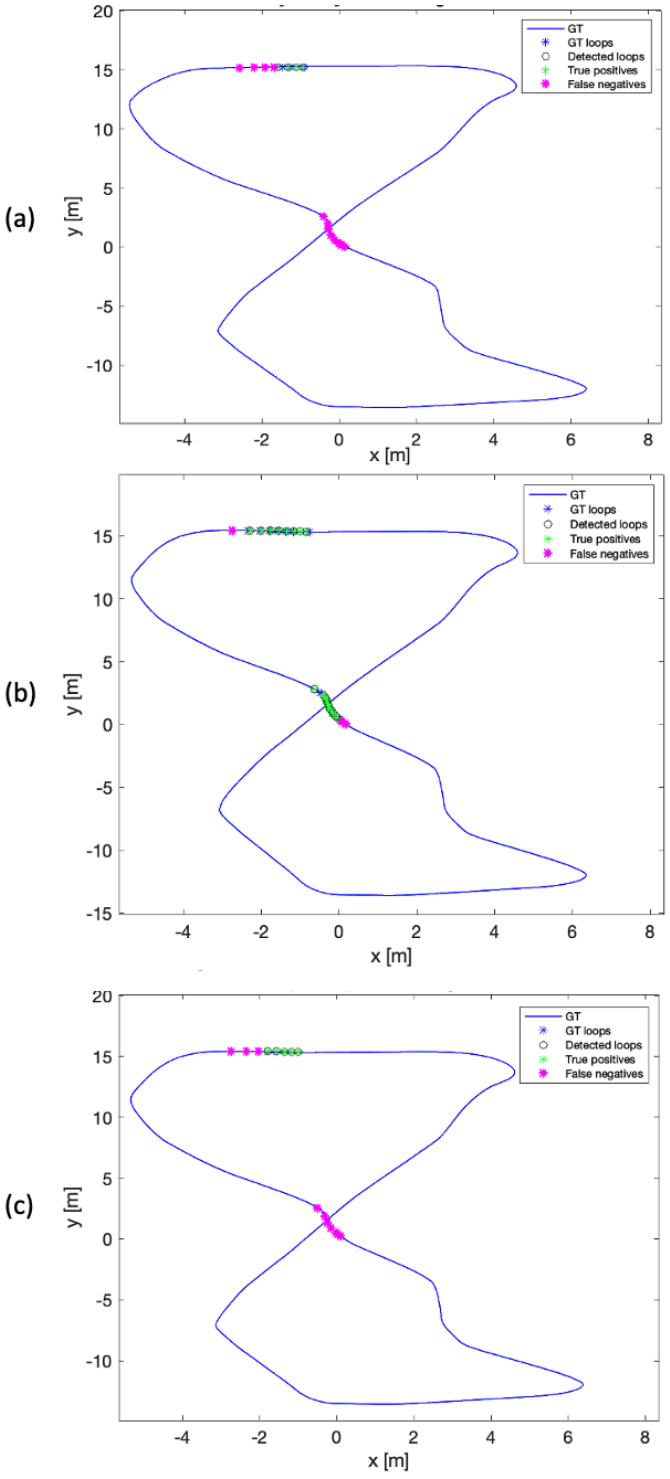

Figure 41: Appearance-based loop closure for the PortSeafloorB dataset for dark (a), light and bright (c) scenes.

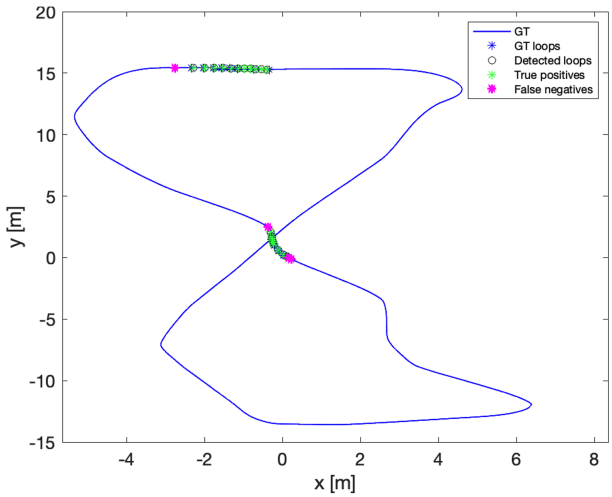

Figure 42: Appearance-based loop closure for the PortSeafloorB for water with suspended particles.

6.2.3 Discussion

In this section, therefore, a preliminary behavioural evaluation of a BoW technique based on ORB features was performed for two port scenarios. For this purpose, datasets from the Stonefish simulator were used, which contain some typical problems of these environments that affect the quality of the captured images. For the PortSeafloorA data, where the AUV performed a simple trajectory with a one-time and long-term loop closure situation, all loops recognised by the algorithm were correctly identified. In this data, the changes in the scene did not affect the overall performance too much, as the loop closure situation was always recognised, albeit with a delay of one detection in the worst case of water type and sun position, achieving 77.78% and 50% of recall respectively. For PortSeafloorB, the performance of the algorithm decreases with increasing water type. In very low turbidity, the recall rate drops by 20% as the algorithm is unable to perform 7 detections. In poorer visibility, i.e. with low turbidity, the algorithm does not recognise the situation of the first loop closure, which can be dangerous in a navigation context as there are no trajectory adjustments and consequently an incorrect pose estimate is obtained. Experiments with sun tilt showed poor performance in place recognition for extreme values, i.e. for dark and bright scenes. The first case can illustrate a depth operation where a darker environment produces scenes without intensity differences, i.e. recognisable features. In contrast, shallow waters, i.e. bright scenes, can produce unstructured images and shadows, which also complicates feature extraction. It has thus been shown that visual similarity detection is nevertheless sensitive to some visual conditions that are common in uncontrolled and difficult scenarios.

6.3 IMPACT ASSESSMENT OF CLAHE ON ENVIRONMENT CONDITIONS

Based on the feasibility analysis of a purely visual solution for underwater similarity detection presented in the previous section 6.2, visual sensors are, as expected, sensitive to some challenging conditions: high brightness and turbidity. They result in missing image details in such images, image features are more difficult to extract and even if extracted, the features are not well recoverable. In such situations, the vehicle can either recognise incorrect loop closures or even fail to recognise some situations correctly, which is dangerous in a navigation context. Image enhancement algorithms optimised for specific feature measurement methods are a promising approach that is often used as a pre-processing step for image modification to overcome the lack of matching features [66]. Commonly used methods include Histogram Equalization (HE), the wavelet transform and the Retinex algorithm [67]. However, although these algorithms can process underwater images, they have some inherent shortcomings: with HE, some details of the processed image may disappear and the peaks of the histogram may be excessively enhanced; the wavelet transform fails in deep waters; and the Retinex algorithm provides edge enhancement and colour constancy but does not cope well with large differences in brightness [68]. Later, two advanced techniques were introduced in addition to the usual HE: Adaptive Histogram Equalization (AHE) and Contrast Limited Adaptive Histogram Equalization (CLAHE). Both calculate several histograms, each corresponding to a specific image section, to increase the local contrast and emphasise the edges in certain regions. CLAHE avoids excessive contrast enhancement. HE is generally regarded as a valuable image pre-processing technique that can be used to extract additional data from low-contrast images. In this context, some conventional enhancement methods have being compared with respect to the three challenging aspects/tasks in underwater image enhancement: Visibility degradation, colour cast and higher level detection/classification. CLAHE proved to be the better choice and the most competitive method for poor visibility, i.e. degraded images, and for colour correction [69] [70].

In order to fully utilise the advantages of these sensors for close-range, an evaluation of CLAHE is presented in this section. The aim is to find out whether CLAHE helps the algorithm to overcome the mentioned inherent perception problems in harbour scenarios by addressing the detection of loop closures.

6.3.1 Methodology

To decide whether an image of a place is a known or new location, CLAHE is applied as pre-processing to improve the contrast.

Figure 43 shows an overview of the method for recognising image similarities.

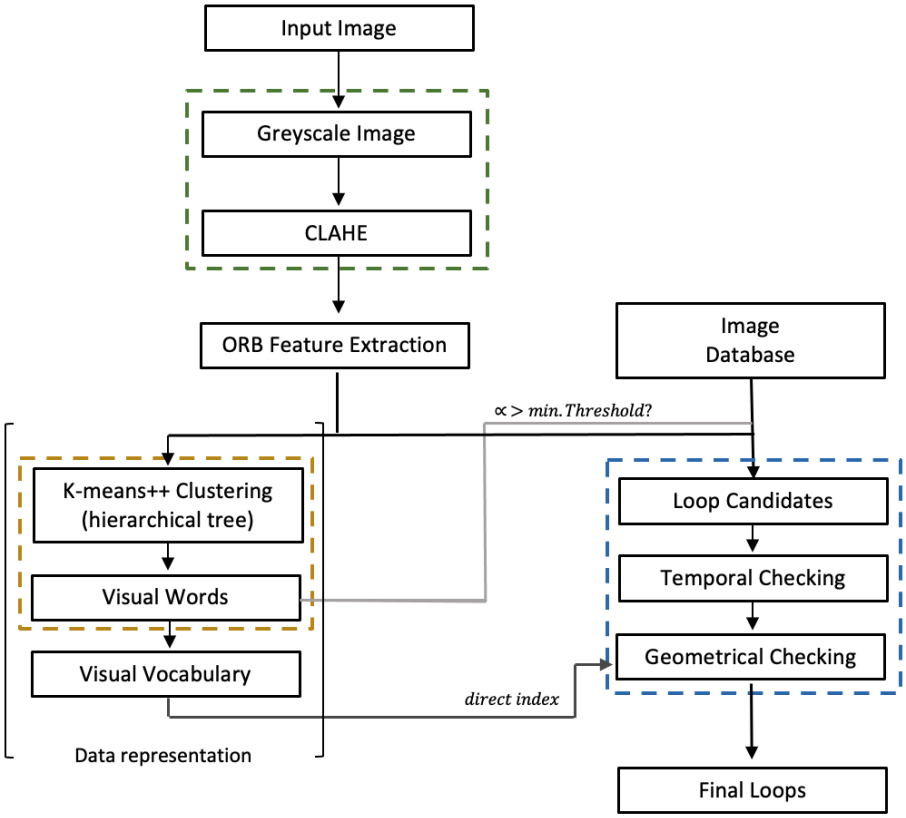

Figure 43: Block diagram of the proposed VPR algorithm based on optical images with CLAHE enhancement technique.

CLAHE is then first applied to images [71], based on the OpenCV 3.4 (C++) library. An input image is thus divided into several non-overlapping regions, which are referred to as sub-images, blocks or tiles. In each tile, the histogram is cropped and redistributed to avoid overemphasis. In addition, a bilinear interpolation is performed between the mapping functions of the neighbouring tiles to avoid the occurrence of block artefacts, as shown in Figure 44.

Thus, there are two key parameters: the Number of Tiles (NT) and the Clip Limit (CL), which mainly control the quality of the image to be enhanced. NT is the integer value that controls the number of non-overlapping areas, i.e. the image is divided into equally sized areas (usually square areas) based on this value. CL is the number that controls the noise gain.

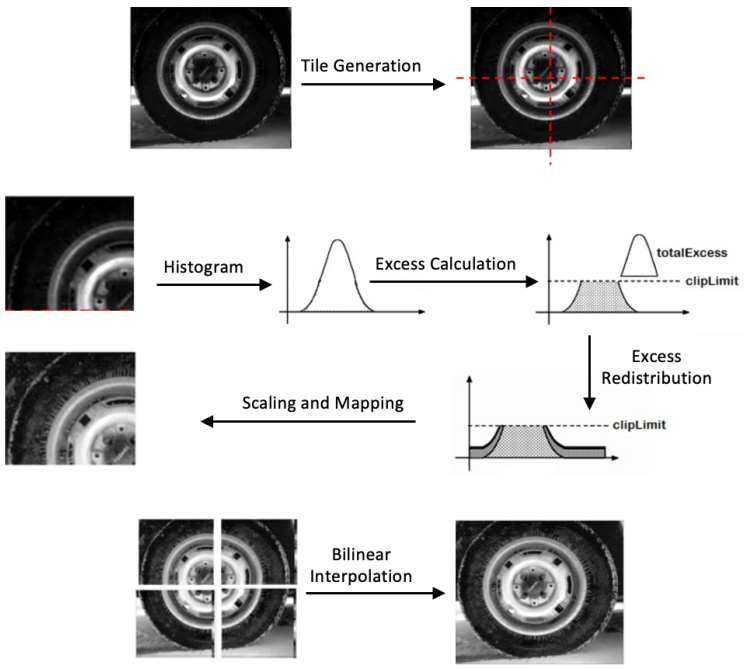

Figure 44: Overview of the CLAHE-based enhancement process.

If CL is increased, the image becomes brighter because the input image has a very low intensity and a higher CL makes its histogram fllater. If NT is increased, the dynamic range increases and the image contrast increases. Nevertheless, the choice of suitable parameters is a difficult task because NT depends on the image resolution on the one hand and the clip limit (or contrast factor) is defined as a multiple of the average histogram content to limit the peak of the histogram, depends on a certain "strength" and can be determined by:

$$\text{ClipLimit (CL)} = fClipLimit \times \frac{XSize \times YSize}{NBins}, \tag{9}$$

where *fClipLimit* defines the clip factor in the range [0,1], *XSize* and *YSize* are the number of pixels in each sub-block and de *NBins* are the number of grey/colour levels.

This step is indicated by the green block in Figure 43. From this data, i.e. all enhanced images, the features and descriptors for each image are first extracted and then the process of creating the vocabulary tree and recognising similar images begins, as previously described in section 6.2.1.

6.3.2 Experimental Results

Therefore, the performance of the proposed purely visual place recognition algorithm for image similarity detection in underwater harbour scenarios is described. Experiments were conducted on data contain-

ing some common poor visibility conditions. First, an analysis of the use of CLAHE enhancer on the images is performed. Then, the behaviour of the binary VPR in detecting previously seen places under different operating conditions is investigated. Full-indexing vocabularies are also used for each scenario. As mentioned before, this way the operating environment is modelled and the performance in detecting previously seen places is evaluated similarly to online approaches. All experiments are based on the extraction of up to 1500 features per image and were performed on a workstation with an AMD Ryzen 7 2700X @ 3.7Ghz with 16GB RAM and NVIDIA RTX 3060.

Datasets

The evaluation of the behaviour of the BoW technique in difficult underwater environments was performed with some datasets created with the Stonefish simulator described in section 5.2.3.

To evaluate the effectiveness of CLAHE, the PortSeafloorB dataset is therefore used. In this case, the camera was pointed downwards to the seafloor and performed a trajectory in the shape of an "8" as shown in Figure 45 (a). The scenarios with very low turbidity and light scenes were considered and the trajectories include 2527 and 2475 images, respectively. The HarbourArea dataset, which simulates an inspection mission in a harbour infrastructure, was used to evaluate the place detection algorithm. To understand how similarity detection behaves in such scenarios when only visual data is used for navigation, strict visibility constraints are considered. In such a case, the vehicle drives at a distance of about 2-3 metres from the structures and the camera was pointed at the structures. Figure 45 (b) shows the trajectory with a loop closure in the final. Very low turbidity and light scenes were also considered, containing 1440 and 1457 images respectively.

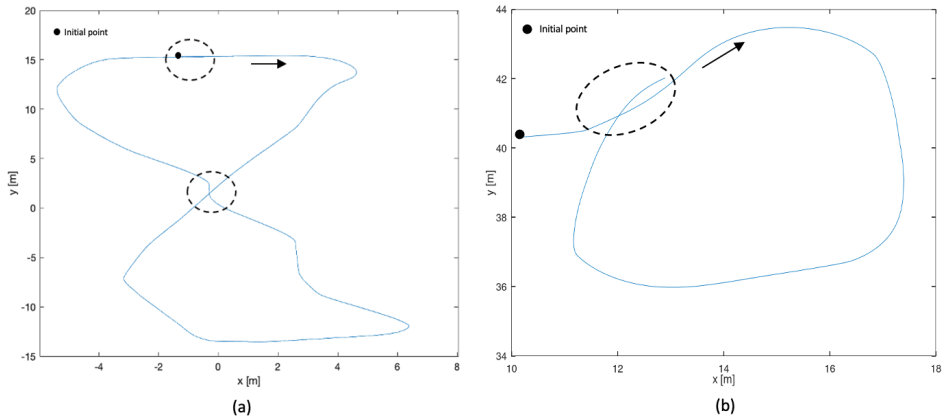

Figure 45: Predefined trajectories of the AUV for PortSeafloorB (a) and HarbourArea (b) datasets.

CLAHE Effectiveness

Visual data is often used in underwater environments to support close-range operations. However, image acquisition in poor environments can lead to various problems, including low contrast, poor lighting and noise. The poor image quality can affect subsequent image processes, e.g. the extraction of features that differ in certain properties such as brightness, colour or texture, i.e. depend on intensity differences. Contrast enhancement is achieved by changing the pixel intensity and increasing the differences between objects and their background in the image. Enhancing image contrast can therefore improve image quality, including edges, details, brightness and boundaries.

Perfomance Metrics

Although the performance evaluation of an image enhancer is still a difficult problem, there are some performance metrics that consider both perceptual and functional aspects [72] [73]. The perceptual evaluation is based on a human perspective and takes into account features related to the human perception of image content and details. The functional evaluation is instead done from a machine perspective, i.e. by evaluating the enhancer effects with a specific machine application. Both metrics are used to quantitatively measure the image features, usually in comparison to an ideal image, the so-called reference image. Therefore, some of these metrics are applied to the intended context and then used to evaluate the effectiveness of CLAHE in relation to the input image.

The **perceptual evaluation** was performed by comparing the following metrics, which were initially calculated based on the original image and the enhanced version of the image:

- Mean Squared Error (MSE), which describes the mean square error between the filtered and the original image, i.e. it quantifies the differences between the values - intensities - implied by an estimate and the true quality that is certified. If two images are identical, the MSE is zero;
- Root Mean Squared Error (RMSE), which measures the amount of change per pixel due to processing. It is given by the square root of MSE;
- Peak Sginal to Noise Ratio (PSNR), which calculates the ratio between the maximum possible power of a signal and the power of the interfering noise that affects the accuracy of its representation. It is used as a measure of the quality of the reconstruction of an image. It is usually expressed on a decibel scale, with a high value indicating high image quality;

- Structure Similarity Index Method (SSIM), which provides a normalised value for the structural similarity between the two images. Image degradation is considered as a change in the perception of structural information: strongly interdependent pixels or spatially closed pixels. It is based on the "perception" of luminance (visible edges) and contrast (texture). Its value ranges from -1 to 1, where 1 stands for perfect similarity, 0 for no similarity and -1 for perfect anti-correlation.

Full reference image quality assessment (FR-IQA) or even Reduced reference image quality assessment (RR-IQA) refers, as the name suggests, to the evaluation of the quality of a distorted image by comparing it with the original, supposedly undistorted version of the same image. The degree of distortion is calculated by measuring the deviation of the distorted image from the reference image. In this case, however, there is no reference image that could limit the quality assessment. Therefore, quality assessment without reference, i.e. No reference image quality assessment (NR-IQA), is becoming increasingly important to monitor image quality in quality control systems and is therefore also considered [74] [75]. These algorithms provide the quality of an image without the need for a reference image or its features. This problem is of course much more difficult than the two problems mentioned above: Since there is no reference image, one has to model the statistics of the reference image, the nature of the human visual system, and the effects of distortions on the image statistics in an unsupervised way. Furthermore, it is very difficult to evaluate the effectiveness of a quality measure with a particular distorted image if there is no reference image available that should correspond well to the human perception of image quality. Therefore, **perceptual evaluation** is then also based on:

- Natural Image Quality Evaluator (NIQE), which calculates the score for a natural image for images that may be distorted or of low perceptual quality, i.e. it calculates the naturalness of the images based on a "quality-aware" collection of statistical features based on a simple and successful statistical model for natural scenes in space [76]. The value ranges from 0 to 10 and a lower value means better perceptual quality;

- Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE), which uses scene statistics with locally normalised luminance coefficients to quantify possible losses of "naturalness" in the image due to distortion, resulting in a holistic measure of quality. However, no distortion-specific features such as ringing, blurring or blocking are calculated [77]. The score is in the range [0, 100], with lower scores reflecting better perceptual quality;

- Perception-based Image QUality Evaluator (PIQUE), which quantify distortion by extracting local features to predict quality. To mimic human behaviour, the quality is estimated only for the perceptually relevant spatial regions [78]. The value ranges from 0 to 100, with a lower value indicating better perception.

- Image Super Resolution Evaluator, which is based on three types of low-level statistical features in both spatial and frequency domains to quantify super-resolution artefacts and learn a two-stage regression model to predict the quality scores of super-resolution images without reference to reference images, as proposed by [79]. Most super-resolution images are evaluated with a full reference metric, but the effectiveness is not clear and the required reference images are not always available in practise. Here, this metric is called *ISRE*. Its score is in the range [0, 10], with higher scores reflecting better perception;

The **functional evaluation** was carried out by comparing the description and matching performance of the ORB algorithm with and without the use of an enhancer as a pre-processing step (original image). Here too, the ORB algorithm is quantified based on the original image and the enhanced version using the following measures:

- Number of ORB keypoints (Kp) recognised on each image, which depends on the image content. In general, an enhancer is expected to increase Kp. However, if Kp_enhanceimage (Kp_EnhImg) >>Kp_originalimage (Kp_OrigImg) means that the enhancer may generate noisy pixels that are falsely recognised as keypoints. If Kp_EnhImg << Kp_OrigImg, then the enhancer describes the image poorly;

- Number of good matches (gm), i.e. the Kp of a query (enhancement) image that matches the corresponding Kp of the original image. In this case (ORB keypoints) the measurement is based on the Hamming distance. The matcher only returns the pairs (i,j) where the j-th descriptor in the matcher's collection is closest to the i-th query descriptor and vice versa, i.e. it only returns consistent pairs;

- Inliers, i.e. the effective matches between images. For this purpose, RANSAC is used to find a number of inliers from the correspondences. The higher the value, the more similar the images are.

Behavior analysis

As described in the previous section 6.3.1, CLAHE is based on manual parameters: NT and CL. For the PortSeafloorB dataset, each image consists of 300x200 pixels and was therefore divided into equal squares of 2500 pixels each, i.e. NT = [6,4]. NT = [3,2] could also be used as equal square areas, i.e. 100x100 pixels, are generated. However, experiments have been carried out dividing the images into more blocks to analyse areas with fewer pixels, as this is the configuration closest to the standard of the OpenCV technique ([8,8]). On the other hand, a higher CL increases the image contrast. An enhancer is naturally expected to increase the brightness of the image and the visibility of details, smooth the colour distribution or remove unevenness and at the same time maintain or increase the naturalness of the image. However, the exact extent of these "properties" and their changes after enhancement depends on the image content. Therefore, for each scenario - haze and illumination scenarios - different CL were tested, depending on the strength of the clip factor (CL = 1, 2, 3 and 4). All experiments are based on the extraction of up to 1500 features per image. Figure 46 illustrates the changes in appearance caused by each CL in images with very low turbidity and light scenes for the two loop closure situations shown in this data: Crossing (LC-1) and Final Straight (LC-2). As you can see, the image contrast increases with CL. This improvement process requires a calculation time of less than 0.7 ms per image, which does not affect real-time operation.

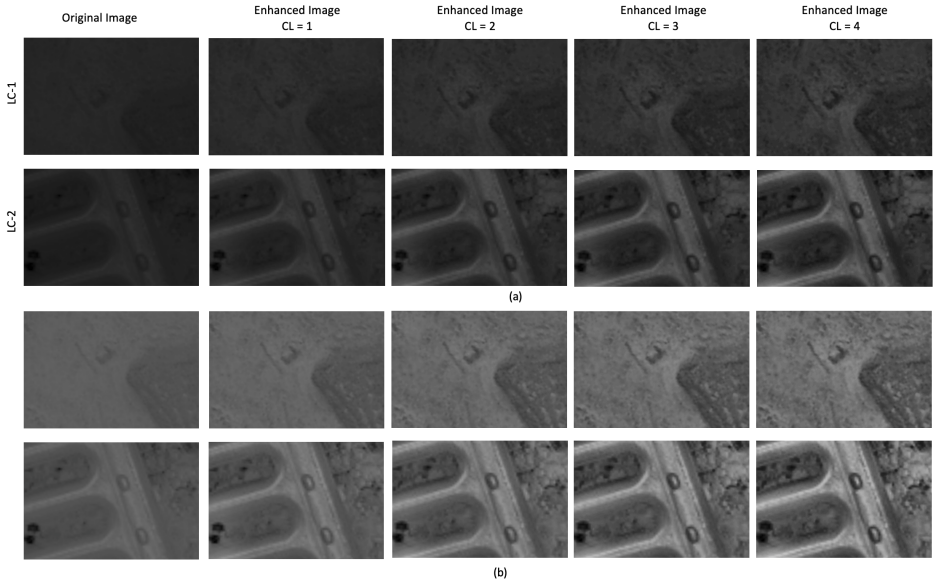

Figure 46: Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied at very low turbidity (a) and light scenes (b) in the PortSeafloorB dataset.

To understand these changes based on the input image, i.e. the FR-IQA for each clip boundary in both visibility conditions is shown in

Table 11 and Table 12. For each scenario, a set of images from LC-1 and LC-2 is considered so that each score is determined by an arithmetic mean of a sequence of 6 images.

MSE and RMSE refer to pixel changes, i.e. differences between intensities due to the processing of the input images. Since we need to investigate the enhancement technique to improve the image quality, these errors should be high, which means that the image differs in appearance from the input image. But *"how much must the images differ?"* There must be an "upper limit", but there is no reference image... In contrast, PSNR and SSIM, which both refer to the fidelity/quality of the image reconstruction, usually need to be high to be close to the reference image (input image). In this context, however, the goal is only to obtain an image with improved quality - an increase in image contrast - and therefore these values should be lower. But again, *"how much?"* Inappropriate values can lead to poor looking images with exaggerated contrast or a "stretching" of the colour channels. With this in mind, an average of the values obtained for each CL was calculated, with CL = 2 standing out for both visibility scenarios. CL = 3 is close to the values obtained for some metrics, in particular MSE and SSIM.

Table 11: FR-IQA and functional evaluation for very low turbidity scenes using different CL.

		LC-1			
		CL = 1	**CL = 2**	**CL = 3**	**CL = 4**
Perceptual Evaluation	**MSE**	35.87	75.31	153.33	265.65
	RMSE	4.00	8.68	12.38	16.30
	PSNR	32.59 dB	29.36 dB	26.28 dB	23.89 dB
	SSIM	0.99	0.97	0.95	0.92
Functional Evaluation	**Kp_OrigImg**	199			
	Kp_EnhImg	799	1140	1184	1204
		LC-2			
		CL = 1	**CL = 2**	**CL = 3**	**CL = 4**
Perceptual Evaluation	**MSE**	212.50	621.34	1081.80	1564.97
	RMSE	14.23	24.90	32.87	39.55
	PSNR	25.09 dB	20.22 dB	17.80 dB	16.19 dB
	SSIM	0.96	0.90	0.84	0.79
Functional Evaluation	**Kp_OrigImg**	230			
	Kp_EnhImg	854	1055	1130	1048

Table 12 shows that MSE/RMSE is smaller and PSNR/SSIM is larger, for example for the lighting scenario. This can be explained by the fact that the input image is clearer and therefore there are not as many discrepancies in terms of pixel/image representation, as can also be seen in Figure 46 (b). Nevertheless, the quality of the input images is improved, as can be seen from the increase in the number

of recognised keypoints after the CLAHE improvement (Kp_EnhImg). This is also true for both scenarios when comparing the crossover and final straight loop closure situations, as the first loop region does not show as many intensity differences - the images are more uniform in colour. Consequently, the effect of image enhancement is not as noticeable, i.e. the input images and the enhanced images are more similar: MSE/RMSE smaller and PSNR/SSIM larger. As expected, the keypoints recognised in the original images (poor visibility) are smaller in both scenarios than in the enhanced images. It is intended that an enhancer increases this number, but if it is much higher, many preserved keypoints could also be noise.

Table 12: FR-IQA and functional evaluation for light scenes using different CL.

		LC-1			
		CL = 1	**CL = 2**	**CL = 3**	**CL = 4**
Perceptual Evaluation	**MSE**	23.07	72.66	130.10	184.35
	RMSE	4.80	8.52	11.40	13.57
	PSNR	34.51 dB	29.53 dB	27 dB	25.48 dB
	SSIM	0,99	0.98	0.96	0.93
Functional Evaluation	**Kp_OrigImg**	503			
	Kp_EnhImg	1083	1169	1195	1211
		LC-2			
		CL = 1	**CL = 2**	**CL = 3**	**CL = 4**
Perceptual Evaluation	**MSE**	186.95	410.83	713.64	1069.66
	RMSE	13.67	20.27	26.70	32.69
	PSNR	25.42 dB	22 dB	19.60 dB	17.85 dB
	SSIM	0.99	0.96	0.92	0.88
Functional Evaluation	**Kp_OrigImg**	632			
	Kp_EnhImg	1084	1166	1184	1202

So the question arises: *"which is the right enhancer?"* Therefore, an average of the obtained ORB features is calculated for each CL, where CL = 2 is emphasised for all cases. The goal of image enhancement in this work is to increase the number of correct feature matches in poorly lit scenes.

To understand whether the enhancement provides a better description of an image, the number of matches and inliers was calculated. Based on the sequence LC-2, images taken at different times are compared considering the original (OI) and the enhanced scenario (CL). Table 13 shows the results for turbidity (wt) and illumination (sang) conditions. As you can see, the enhancer increases the number of consistent pairs that become more or less stable above a certain clip threshold, i.e. the number is in the same order of magnitude. For these functional metrics, the improvements are also more significant in

the cloudy scenario (referred to as the very low turbidity scenario), as the image content in the original image is practically unrecognisable compared to the bright images - poor visibility. In this context, we are looking for a high number of inliers, which indicates that the images are more similar. If the overall mean values are then measured, CL = 2 generally shows good behaviour. Figure 47 illustrates this behaviour for the turbidity scenario.

Table 13: Functional evaluation between images taken at different times with very low turbidity and light scenes.

			LC-2				
			OI	**CL = 1**	**CL = 2**	**CL = 3**	**CL = 4**
wt = 0.25	**Functional Evaluation**	**gm**	109	351	431	464	494
		Inliers	93	269	328	358	374
sang = 50.77°		**gm**	197	360	387	386	401
		Inliers	107	174	189	190	205

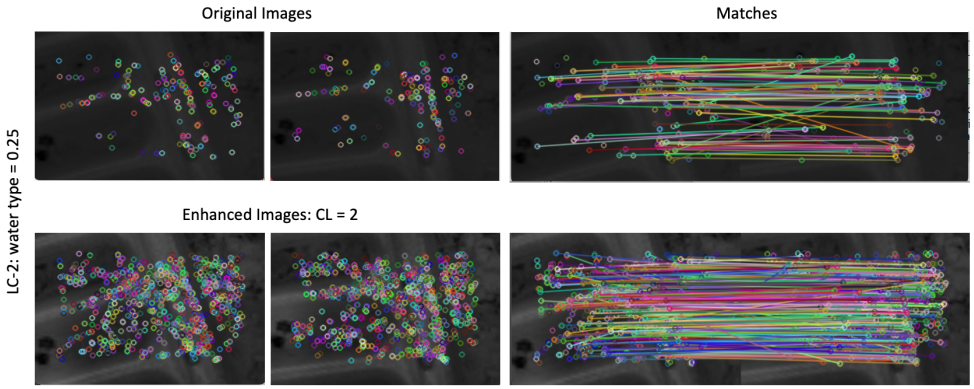

Figure 47: Comparison of keypoint matches with and without CLAHE-based image enhancement for a scene with very low turbidity. First row: original images; second row: CLAHE-enhanced images.

It shows the detected features and matches (represented as dots or lines) in the original and CLAHE-enhanced (CL = 2) images taken at different times in one place. As you can see, the enhanced image pair has a larger number of features and correct matches (represented by linear lines, i.e. shorter distances between image features). These results show the effect of image enhancement to improve image quality and consequently feature extraction and matching between images.

Table 14 and Table 15 show the NR-IQA algorithms for each clip boundary in both visibility conditions. For each scenario, a set of images from LC-1 and LC-2 is also considered, so that each score is determined by an arithmetic mean of a sequence of 6 images.

Table 14: NR-IQA for very low turbidity scenes using different CL.

		LC-1			
		CL = 1	CL = 2	CL = 3	CL = 4
Perceptual Evaluation	NIQE	5.39	5.27	5.58	6.02
	BRISQUE	45.71	43.05	42.63	43.64
	PIQUE	61.91	56.61	52.63	48.67
	ISRE	2.77	2.99	3.03	3.12
		LC-2			
		CL = 1	CL = 2	CL = 3	CL = 4
Perceptual Evaluation	NIQE	5.87	5.52	5.51	5.50
	BRISQUE	49.97	44.74	41.99	40.21
	PIQUE	54.13	47.54	45.97	46.97
	ISRE	2.13	2.26	2.40	2.68

Table 15: NR-IQA for light scenes using different CL.

		LC-1			
		CL = 1	CL = 2	CL = 3	CL = 4
Perceptual Evaluation	NIQE	5.29	5.54	6.04	6.61
	BRISQUE	50.24	49.31	48.46	47.59
	PIQUE	42.39	32.99	30.45	28.00
	ISRE	2.94	2.94	3.14	3.36
		LC-2			
		CL = 1	CL = 2	CL = 3	CL = 4
Perceptual Evaluation	NIQE	6.82	6.43	6.45	6.41
	BRISQUE	55.34	52.32	51.19	50.59
	PIQUE	62.15	54.35	53.50	51.56
	ISRE	2.52	3.04	3.40	4.03

The original images actually have a poor appearance and the four clip boundaries improve the image quality, i.e. the evaluation metrics used are improved. Based on an average of the values of each CL, CL = 2 stands out for the turbidity scenario. For the lighting scenario, CL = 2 and CL = 3 stand out. But in these cases, there is an indicator that shows the quality of the image. Therefore, in general, BRISQUE, PIQUE and ISRE with a higher CL are better. However, PIQUE increases again in a turbidity scenario with CL = 4. NIQE, on the other hand, hardly changes or increases with CL, i.e. the quality of the images is poor. In any case, the images lose their naturalness from CL = 2, images lose their naturalness on the basis of this algorithm. For the remaining metrics, CL = 4 highlights all cases except the LC-1 situation in the turbidity scenario, where BRISQUE in CL = 3 is better.

Loop Closure Detection

To measure the performance of the proposed VPR method, a ground truth for loop closure was then created for both visibility scenarios in the HarbourArea dataset. Based on each pose file, a threshold of approximately 0.85m is used for placing two images at the same location based on camera footprint. As described earlier, place detection performance is evaluated using precision and recall metrics. Therefore, TP, FP and FN are considered for evaluation against ground truth. As mentioned above, the loops detected by the algorithm are considered as TP or FP based on a threshold of 0.3m. If there are loops in the ground truth that do not match the loops recognised by the algorithm, these are considered FN. To find an ideal combination of precision and recall, the F1-score - the harmonic mean of both metrics - is also calculated. To consider images as loop candidates, a similarity threshold (α) of 0.3 was used. This value proved to be the most suitable for the similarity requirements to avoid distant features being considered as the same points.

Taking into account the effect of CLAHE described above in section 6.3.2, CL=2 and CL=3 were used to evaluate the performance of visual place recognition in the HarbourArea dataset. Again, the contrast of the structures increases with CL in the scenarios turbidity (very low turbidity) and sunlight (bright scene), as shown in Figure 48.

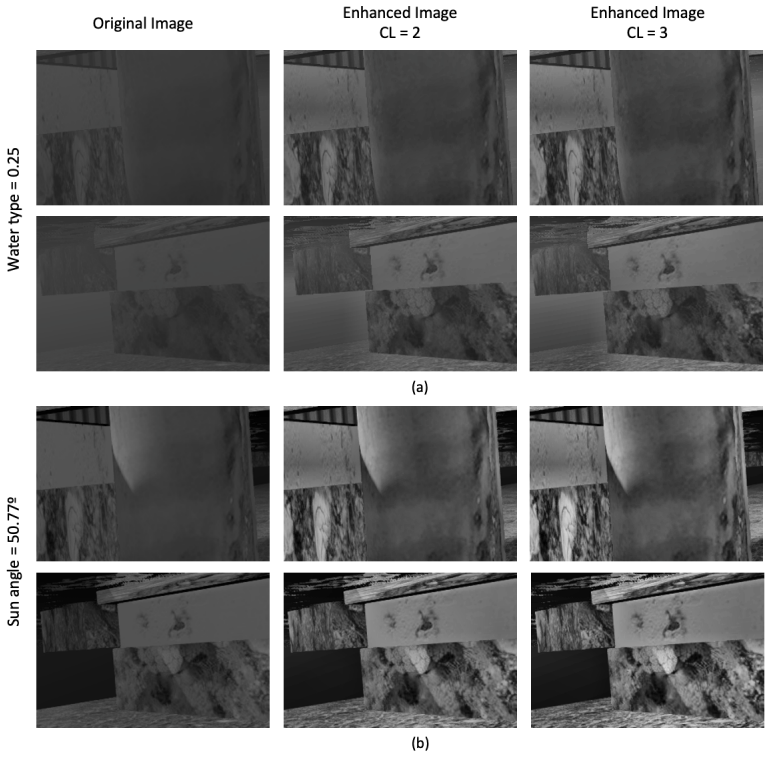

Figure 48: Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied for very low turbidity (a) and light (b) scenes in the HarbourArea dataset.

The behaviour of the place recognition method in the clouding scenario was tested with original images and two different CLAHE configurations of the clipping limit. As can be seen in Figure 49, there are improvements with increasing CL.

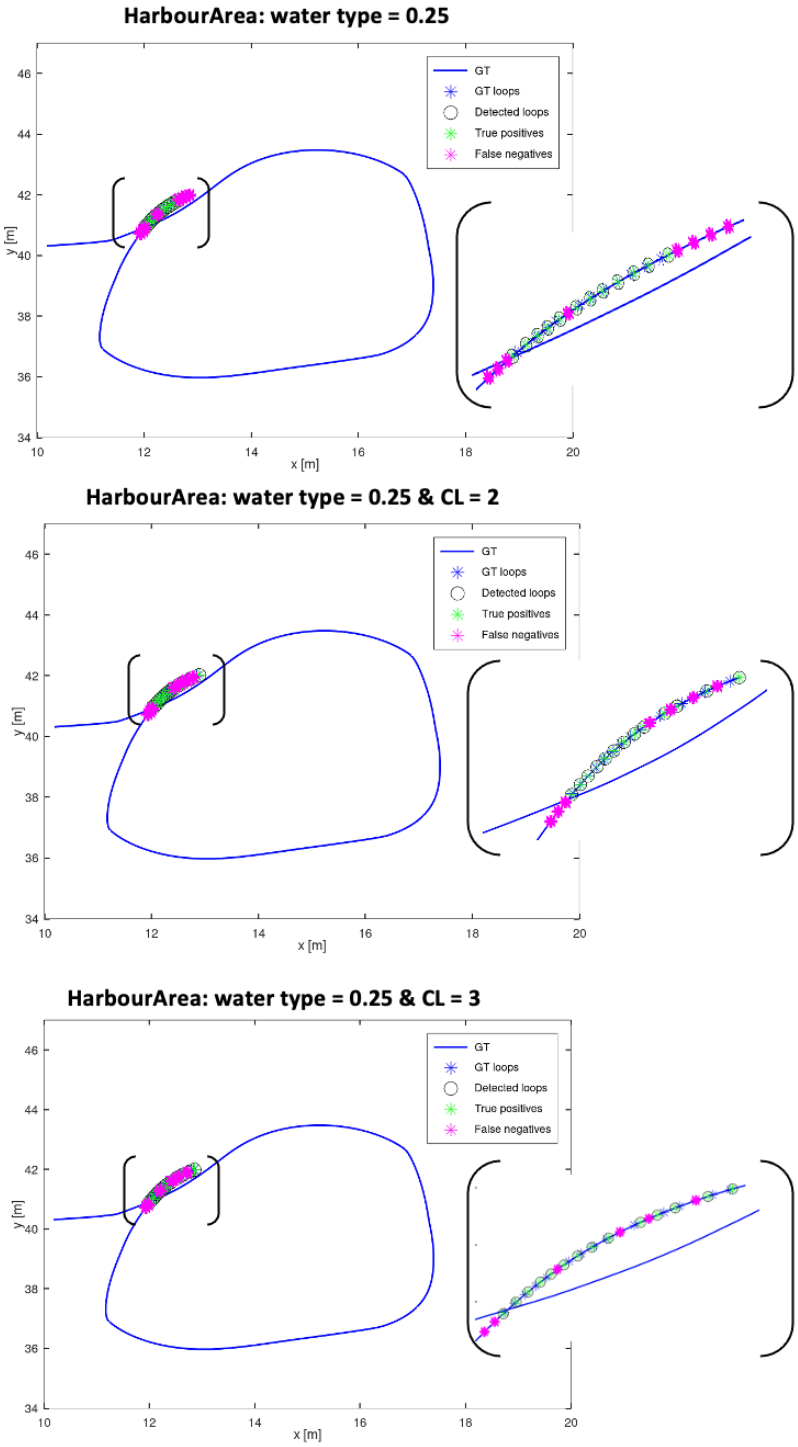

Figure 49: Appearance-based loop closure for the HarbourArea dataset in a scenario with very low turbidity when varying CL.

In general, it is difficult to see these improvements, but looking at the region of loop closures (enlarged), it is noticeable that the enrichment

technique provides more consistent keypoints, as the algorithm is less prone to lose loops from the moment it starts to detect them, although the number of corrected loops (TP) detected is not very high. In addition, unlike the originally captured images, similarities along the trajectory can be recognised, especially in the last part. With CL = 3, loops can be recognised if the algorithm starts earlier and approaches the ground truth. In this case, 14 loops are recognised correctly (TP) and 6 loops are not recognised (FN). Table 16 shows the performance of the algorithm when recognising similar places. It achieves an accuracy of 100% in all cases, which means that no loops were falsely recognised FP, i.e. there are distinguishing features. The recall increases during the experiments when the environmental conditions change (higher image contrast).

Table 16: Performance of the VPR algorithm for the HarbourArea dataset in the very low turbidity scenario with varying CL.

			CL = 2	CL = 3
Water Type = 0.25	100	**Precision (%)**	100	
	60	**Recall (%)**	65	70
	75	**F1-Score (%)**	78.79	82.35

In addition, it is known that there is an increase of up to 40 images recognised as a previously visited place, with an average of about 40 inliers, proving the consistency that the enrichment level provides in this type of challenging environments. In general, a performance improvement of almost 10% is achieved. In a navigation context with an accuracy of 100%, a higher recall is required as an unrecognised loop closure situation does not adjust the trajectory and affects the ability to obtain an accurate position estimate.

Next, the scenario was tested with the position of the sun to evaluate the effects of lighting on recognising similarities between images and to understand how two different clipping boundaries help. This test is crucial as it replicates a real harbour scenario - shallow water and therefore heavily influenced by sunlight. As can be seen in Figure 50, there are also improvements with increasing CL, although in this case they are not as significant, i.e. there is an increase in performance of around 4% as can be seen in Table 17. This fact can be explained by the fact that the original image has a higher visibility from the beginning than the scenario with turbidity. Incidentally, the faster detection allows us to recognise that the robot is entering an area that has already been visited, which was only the case with a higher image contrast in the previous case. Nevertheless, the incidence of light can smooth the images - an unstructured image - and the enhancement technique achieves more consistent similarity detection, correctly recognising 14 loops.

With CL = 3, better performance is achieved along the trajectory, with 7 loops not being recognised, but at the same time some noise may occur - erroneous keypoints - which may cause a delay in detecting the loop closure. In all cases, a precision of 100% is achieved.

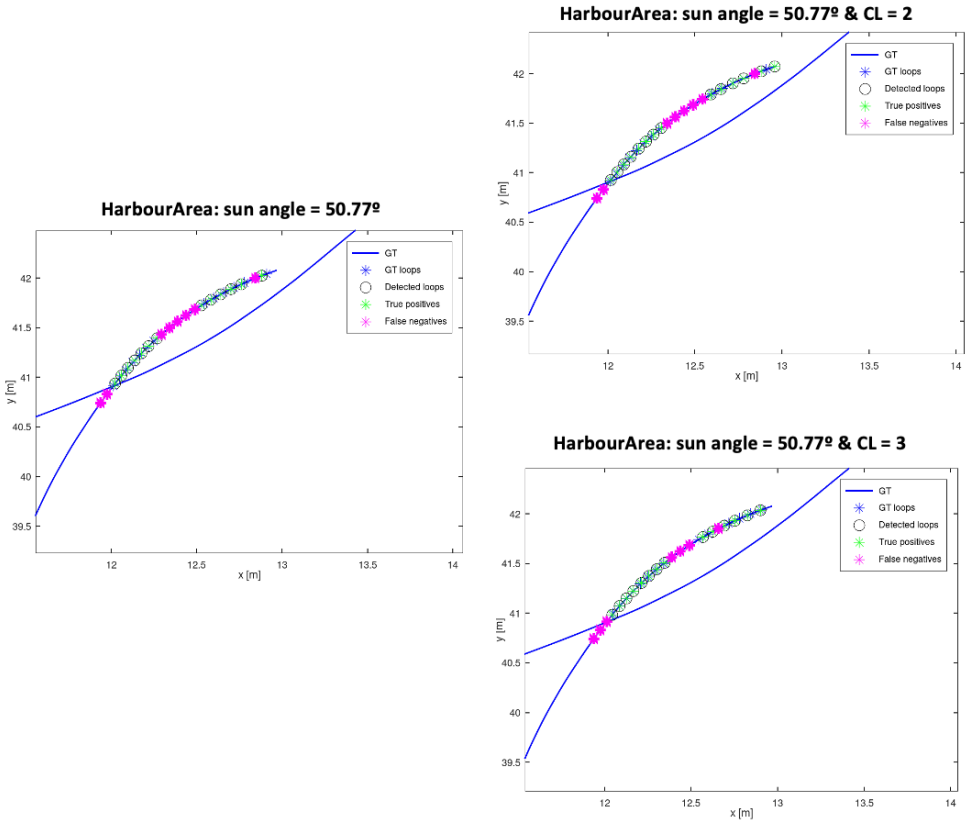

Figure 50: Appearance-based loop closure for the HarbourArea dataset in the light scenes when varying CL.

Table 17: Performance of the VPR algorithm for the HarbourArea dataset in the light scenario with varying CL.

			CL = 2	CL = 3
	100	**Precision (%)**	100	
Sun angle = 50.77º	61.90	**Recall (%)**	63.64	66.67
	76.47	**F1-Score (%)**	77.78	80

Given this behaviour with CL = 3 in the turbidity and brightness scenarios and considering NR-IQA above in 6.3.2, the effect of applying CLAHE with CL = 4, i.e. twice the value of the average of all cases, is therefore also considered to understand what the increase of about 15 inliers in both scenarios means. In this way, it is possible to determine whether these new features are effective and therefore improve the perception of the scene or whether, on the contrary, they represent noise and can therefore impair place recognition.

Table 18 shows the results achieved for both visibility scenarios. As you can see, for the very low turbidity scenario, this clip limit increases the performance of the algorithm (an increase of 5%), which means that CL = 4 better describes the scene. This increase is due to the detection of another loop closure, which indicates that a higher clip limit offers further advantages under these conditions. For the brightness scenario, CL = 4 offers no advantage, as the algorithm behaves as if it were using CL = 3 for images. This shows that the largest number of features is not sufficient to improve the texture of images. However, it also shows that this number does not represent noise and does not jeopardise the robustness of the algorithm (TP = 0).

Table 18: Performance of the VPR algorithm for the HarbourArea dataset in the waters with very low turbidity and light scenes at CL = 4.

			CL = 4
Water type = 0.25	100	**Precision (%)**	100
	60	**Recall (%)**	75
	75	**F1 score (%)**	85.71
Sun angle = 50.77°	100	**Precision (%)**	100
	61.90	**Recall (%)**	66.67
	76.47	**F1 score (%)**	80

6.3.3 Discussion

In this section, therefore, a purely visual approach to place recognition was presented that deals with perception problems in a harbour scenario. Since the poor image quality may affect subsequent image processes such as feature extraction and consequently the matching of images with each other, the proposed method includes a pre-processing step to increase the contrast of the captured images in order to fully exploit the advantages of these sensors operating effectively at close range. The CLAHE technique was applied to the captured images with different CL. It has been shown to be effective for an adequate perception of the environment by increasing the inliers between images of the same place taken at different times to a minimum of about 60%. In terms of similarity detection performance, all loop closure situations recognised by the algorithm under both visibility conditions were correctly identified. It is noticeable that the pre-processing step provides more consistent keypoints that allow detection along the trajectory, especially in the last part. In the illumination scenario, the improvements are not as significant as the original image has a higher visibility and therefore recognises faster that the robot is entering an already visited area. As the image is brighter, the increase in CL

can create some noise points. On the other hand, this can also lead to unstructured images, so the enhancement method provides better results. However, it was found that the number of extracted features is much lower for walls that have only some corrosion. A performance of 85.71% and 80% was obtained for turbidity and illumination scenarios, respectively. The proposed VPR algorithm thus shows good performance at low computational costs. Nevertheless, it is important to know the limitations of the environmental constraints of the cameras and even consider the possibility of relying on other sensors under such conditions.

6.4 SUMMARY

In this chapter, the behaviour of a BoW technique based on ORB features was evaluated in underwater scenes of harbours with different environmental conditions. It was found that the detection of loop closures based on the raw input image data is impaired.

The performance of the algorithm decreases with increasing water type as it is not able to recognise the first situation of a loop closure at lower turbidity, which is dangerous in a navigation context as there are no trajectory adjustments and consequently an incorrect pose estimation is obtained. Quantitatively, the algorithm achieves a recall of 69.57% at very low turbidity (a decrease of about 20% compared to suitable visibility conditions), with 7 loop closures not being recognised. With increasing turbidity, i.e. at low turbidity, the recognisability drops significantly to 35.29%. Experiments with solar tilt also showed poor performance in place recognition for extreme values, with a maximum recall of 62.50% being achieved. Visual similarity detection is thus sensitive to some visual conditions that are common in uncontrolled and difficult scenarios, as the poor image quality affects feature extraction and thus the matching of images to each other.

Therefore, a pre-processing step is proposed to increase the contrast of the captured images in order to take full advantage of these sensors working effectively at close range. The CLAHE technique was applied to the captured images (raw data). It has been shown to be effective for an adequate perception of the environment by increasing the inliers between images of the same place taken at different times to a minimum of about 60%. In addition, all loop closure situations recognised by the algorithm under both visibility conditions were correctly identified: The pre-processing step provides more consistent keypoints that allow detection along the trajectory, especially in the last part. In the best case, the algorithm achieves a recall of 75% and 66.77% in very low turbidity and bright scenes with a precision of 100%. The results show that a higher threshold (CL) does not bring any further advantages.

However, it has been shown that visual similarity detection is still sensitive to some changes in scene appearance that are common in uncontrolled and challenging scenarios. Therefore, it is important to improve the performance of place recognition to ensure successful autonomous navigation even when features are missing or the camera is already starting to fail to detect recognisable features. In this way, it could be interesting to perform an evaluation to verify if the sonar images are able to overcome the above-mentioned underwater challenges to understand when this imaging sonar can be helpful in an operational mission, e.g. for structures with missing features or

to assist in cases where the camera is already starting to fail to recognise distinguishable features. Therefore, chapter 7 proposes a hybrid place detection system that can adapt to the typical changes in the underwater environment.

7 HYBRID PLACE RECOGNITION

Visual sensors work effectively at close range. However, in harbours, visibility can be poor, making place recognition difficult as the visual appearance of a local feature may be impaired. Under these operating conditions, imaging sonars can be a promising solution as these sensors do not suffer from haze, which is an advantage. Therefore, an acoustic approach for unsupervised recognition of similar images based on FLS data is proposed in this chapter to solve the perception problems in harbour facilities.

The chapter is organised as follows: Section 7.1 gives an overview of the visual similarity detection problem in underwater environments and highlights the help of sonars in the proposed deployment scenarios. Section 7.2 presents a boundary characterisation for visual place recognition from optical images under perceptual challenges. Section 7.3 describes the operating principles of FLS devices and analyses the behaviour of acoustic image-based similarity detection to understand when this imaging sonar can be used as an aid in an operational mission. This evaluates the feasibility of a hybrid place detection system that can adapt to the typical changes of the underwater environment. Finally, section 7.4 describes the main conclusions of this chapter.

7.1 INTRODUCTION

To perform an inspection mission, underwater vehicles must accurately navigate and recognise revisited places, i.e. loop closure detection to compensate for cumulative pose deviations [80]. In an image-only retrieval model, the decision whether the sensor has reached a revisited scene during its movement is based on similarity measures between images stored on maps. However, due to perceptual limitations, navigation in the vicinity of structures is still a challenge. Image processing systems are an attractive solution for close-range environment recognition as they work at distances of less than 3m, provide rich information and are easy to use. However, the underwater environment is dynamic and lacks structural features. In addition, harbour scenarios are often affected by turbidity or illumination (shallow waters), which often complicates the behaviour of navigation and mapping tasks performed by cameras, as the perception range of optical devices is severely limited in very poor visibility. Such conditions make the detection of

loop closures difficult and the vehicle may fail to detect some loops correctly or detect some erroneous loops so that the trajectory is not adjusted or is adjusted incorrectly. In a previous study, the efficiency of a purely visual system for recognising similar images was analysed. The enhancement method improves image quality and provides more consistent keypoints. Nevertheless, the analysis presented in chapter 6 showed that the cameras are still susceptible to strong haze and brightness conditions and at best achieve a recognition rate of 75%. So the question is, *"what are the limits of environmental conditions that cameras can overcome to recognise similarities between images?"* On the other hand, active sonars today emit an acoustic wave and receive the backscatter, so they provide acoustic images with which they can perceive the environment - imaging sonars. Although these sensors suffer from distortion and masking effects due to their physical characteristics, they do not suffer from clouding effects as the images are based on the sound emitted and returned. Therefore, this category of sonars, namely FLS, is seen as a promising solution for these challenging environmental conditions in real-world applications. The performance of these sonars has greatly improved and the resolution of their images continues to increase, allowing FLS to provide comprehensive underwater acoustic images.

Therefore, developing efficient approaches to extract visual data from sonar images and understanding their performance remains critical to ensure that FLS data is suitable for vehicle perception of the environment. It is also important to understand whether this data can later be matched with the camera data when environmental conditions change. The aim of this chapter is therefore to understand the extent to which the camera can overcome the inherent perceptual difficulties of harbour scenarios and to evaluate the behaviour of FLS data under such conditions in order to understand how robots can be successfully assisted in continuous navigation.

7.2 LIMIT CHARACTERISATION OF VISUAL IMAGES ON ENVIRONMENT CONDITIONS

Based on section 6.3, in which the efficiency of a purely visual system for recognising similar images was analysed, it became clear that the underwater environment provides images without structural features, even when CLAHE is used: This environment is often affected by turbidity or illumination, which often complicates the behaviour of navigation and mapping tasks performed by cameras, as the perception range of optical devices is severely limited in very poor visibility. Such conditions make the detection of loop closures difficult and the

vehicle may not recognise some loops correctly or incorrectly, resulting in the trajectory not being adjusted or being adjusted incorrectly. Nevertheless, cameras are considered to be a more attractive solution for close range. In this case, an investigation of the camera's operational limitations in terms of visibility is conducted to understand to what extent the proposed visual approach to recognising similar places can overcome the inherent perceptual problems in harbour scenarios. As shown in the previous chapter 6, cameras can be sensitive to either heavy haze or bright scenes, producing images with no appearance or texture respectively. However, the effects are the same: there are no robust features that allow for stable and efficient matching. A boundary characterisation of turbidity for the use of cameras in port facilities is evaluated here.

7.2.1 Methodology

To decide whether an image is a known or new place, the method proposed in section 6.3 is applied, following the overview shown in Figure 43. First, the CLAHE technique is applied to all greyscale images and the vocabulary tree (K = 10 and L = 6) is generated based on all ORB features of the enhanced images using K-Medians clustering with k-means++ seeding. An inverted and a direct index are created along the vocabulary tree: Thus, for each word, the weight and a list of images in which it occurs are stored (inverted file). In addition, for each image, the nodes at the l level that are ancestors of the words occurring in the image and the list of local features associated with each node are stored (direct file). To detect similar images, each current image is converted into a bag-of-words vector after conversion to greyscale and the application of CLAHE and searched in the database. You will receive a list of eligible matching images calculated by a normalised L1 score between the BoW vectors of each current image and the database image based on the weighting of the words. Only matches whose score exceeds a minimum similarity threshold α are considered. Each candidate must undergo a temporal and geometric check to obtain more consistent results. In the geometric check, the comparison is only performed between the features that are connected to the same nodes at the level $l = 2$ in the vocabulary tree using the vocabulary.

7.2.2 Experimental Results

Therefore, the performance of the proposed purely visual place recognition system for recognising image similarities in underwater harbour scenarios is described. The experiments were conducted with data under different visibility conditions.

For each scenario, full-indexing vocabularies previously created with images of the performed trajectory are also created here to model the operating environment and evaluate the performance in recognising seen places similar to online approaches. All experiments are based on the extraction of up to 1500 features per image and were performed on a workstation with an Intel i7-1270P @ 2.2Ghz with 16GB RAM.

Datasets

To characterise the limits of visibility for operation with cameras for similarity detection between images under challenging appearance conditions, the HarbourMissionCAM dataset captured by the Stonefish simulator is used. This data was then acquired with a camera configured to capture images of the bottom near the structures of interest in an appropriate downward direction, as FLS.

Figure 51 illustrates the AUV mission and highlights the area where the loop around the concrete wall is closed.

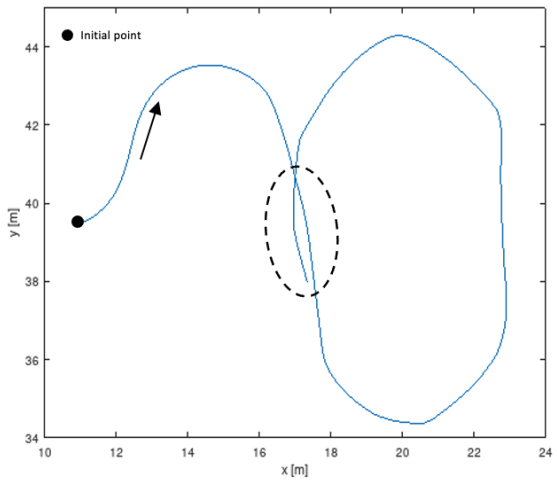

Figure 51: Predefined trajectory of the AUV for the HarbourMissionCAM dataset.

Here, scenarios with considerable (wt = 0.35), high (wt = 0.4) and very high turbidity (wt = 0.45) were considered, as labelled in 5.2.1. The trajectories comprise 857, 840 and 805 images, respectively. The dataset is explained in more detail above in section 5.2.3.

Figure 52 shows an illustration of the effects of turbidity on the captured images.

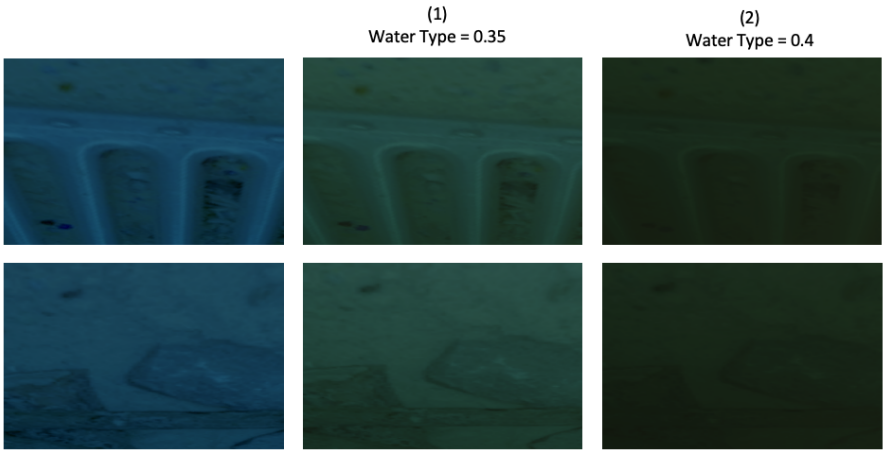

Figure 52: Illustrative example of the effects of considerable (1) and high (2) turbidity on the appearance of images.

Loop Closure Detection

To measure the performance of the proposed similarity detection method, a ground truth for loop closure was created for each visibility scenario of the HarbourMissionCAM dataset. Each pose file is used to create the ground truth, using a threshold of approximately 0.5m to consider two images equal based on the camera footprint. As described above, loop detection performance is evaluated using precision and recall metrics. Thus, TP, FP and FN are used to evaluate the performance against the ground truth. The F1-score is also calculated to find an ideal combination of precision and recall. Loops detected by the algorithm whose distance is less than 0.3m are considered TP; loops whose distance is more than 0.3m are considered FP. If there are loops in the ground truth that do not match the loops recognised by the algorithm, these are assumed FN.

To avoid considering distant features as equal points, a similarity threshold α of 0.3 was used as a default value to consider images as loop candidates. CLAHE is based on the manual configuration of the NT and CL parameters to increase the brightness of the image and the visibility of details, smooth the colour distribution or remove imperfections while preserving or increasing the naturalness of the image. However, the exact extent of these "properties" and their changes after enhancement depends on the image content, as described in section 6.3.1. Each image consists of 300x200 pixels and therefore NT = [3,2] and NT = [6,4] are considered here (instead of only NT = [6,4] as in the previous case) to understand the effects of creating histograms from smaller or larger blocks in terms of pixels and even to establish a possible correlation with the nature of the appearance of underwater scenes. This dataset contains images of the seabed, which means that the features are a priori not as similar as in the previous case

(HarbourArea dataset), as the images show structures with less texture or with a texture that can be easily confused (e.g. plants) - perceptual aliasing. Therefore, CL = 1 and CL = 2 were tested for each of the three visibility scenarios. These CL values are therefore considered for each NT. In some cases, CL = 3 was also considered.

Figure 53 illustrates the effects of applying the individual CL values in the individual water types (wt). Here, too, the image sharpness increases with the CL value, although, as you can see, it becomes difficult to improve the appearance of the image at very high turbidity.

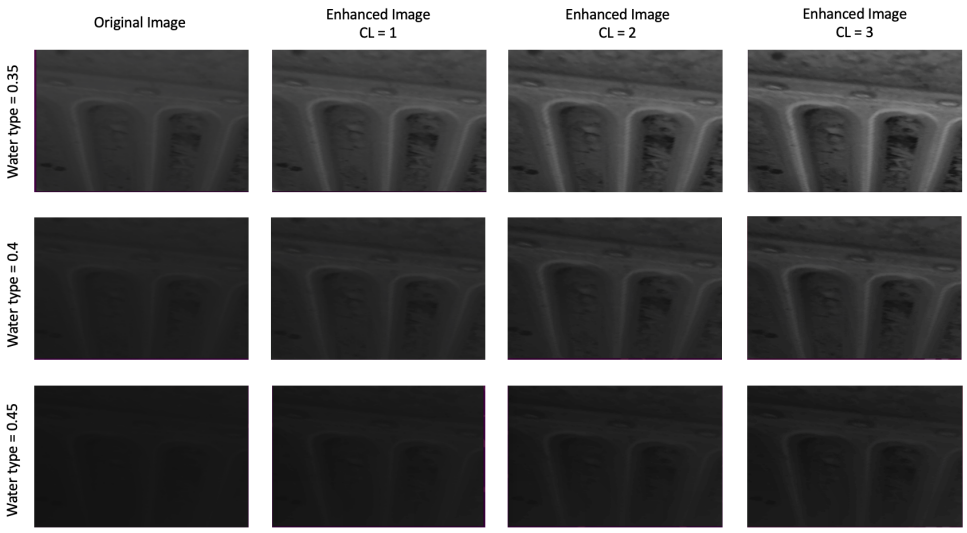

Figure 53: Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied in scenes with considerable (top), high (middle) and very high (bottom) turbidity in the HarbourMissionCAM dataset.

In order to understand to what extent the cameras can handle inherent underwater scenes so that the robot can perform inspection missions, i.e. navigate accurately in the vicinity of structures, the behaviour of the place recognition method is analysed for different turbidity conditions. First, a scenario with considerable turbidity was analysed, as the scene still appears perceptible to the "naked eye".

Table 19 shows the performance of the similarity detection approach considering different CL for each NT parameterisation, where CL = 1 and CL = 2 are considered. As expected, the algorithm is able to recognise all loop situations with a precision of 100% under these environmental conditions, i.e. without FP. Nevertheless, the algorithm incorrectly recognises a loop closure when it divides images into 100×100 pixel blocks (NT = [3,2]) when it applies either CL = 1 or CL = 2 to the input images. With CL = 2, the loop closure is recognised early, which may mean that a higher CL value describes the worst image.

Table 19: Performance of the VPR algorithm for the HarbourMissionCAM dataset at different NT and CL in the considerable turbidity waters.

	NT = [3,2]		NT = [6,4]	
	CL = 1	**CL = 2**	**CL = 1**	**CL = 2**
Precision (%)	93.33		100	
Recall (%)	100		100	
F1 score (%)	96.55		100	

Even if you increase the minimum number of matches according to RANSAC to 13, the algorithm recognises a faulty loop, but later than with CL = 2 and 12 matches according to RANSAC (defined default value), i.e. you get the same result as with CL = 1. This shows that a higher value produces more confusing features. This can be explained by the fact that the input image is not very degraded and therefore we generate noise features when using a higher CL value. At NT = [6,4], the algorithm achieves the best performance for both CL by recognising all loop closure situations without false detections, as shown in Figure 54.

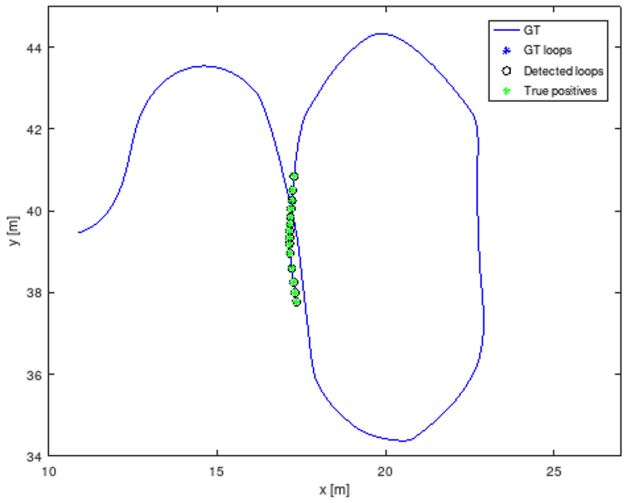

Figure 54: Appearance-based loop closure for the HarbourMissionCAM dataset in the considerable turbidity at NT = [6,4].

CL = 3 was also considered, but 3 FP were detected, which means that this value is too high given the input image and leads to confusing features, i.e. noise features. In this case, the images are divided into several blocks and more detailed/local histograms are created to better describe the scenes.

The turbidity is then increased and Table 20 shows the performance of the algorithm in recognising similar places at high turbidity.

The loss of visibility is noticeable because at CL = 1, the algorithm is no longer able to recognise some loop situations FN = 4.

Table 20: Performance of the VPR algorithm for the HarbourMissionCAM dataset at different NT and CL in the high turbidity waters.

	NT = [3,2]		NT = [6,4]	
	CL = 1	**CL = 2**	**CL = 1**	**CL = 2**
Precision (%)	100	90.91	100	75
Recall (%)	71.43	71.43	71.43	64.29
F1 score (%)	83.33	80	83.33	69.23

This means that this clip limit is not sufficient to improve the appearance of the scene. If you increase CL, the algorithm retains the same errors in recognising similar situations and false loop closures are detected, i.e. the strength of the algorithm is maintained and it loses robustness in recognising known places without errors (precision $< 100\%$). In this way, the results for both NT are balanced, but for CL = 2, FP = 1 and FP = 3, using NT = [3,2] or NT = [6,4] respectively. In this case, it is therefore not advantageous to create more histograms. Even if the minimum matches allowed by RANSAC are taken into account, the algorithm achieves a precision of 100%, but is not able to obtain correctly recognised loops (it fails on 5 instead of 4 loops). The best behaviour is achieved with CL = 1 when the images are divided into blocks of either 100×100 or 50×50 pixels, as shown in Figure 55. As you can see, the algorithm is more unstable when NT = [6,4]. Thus, given the input data, more detailed histograms describe the worst images: We make local enhancements when we have visually degraded images, i.e. we highlight low quality features a priori.

The performance of the algorithm in recognising similar places at very high turbidity is shown in Table 21. Given the previous experiments, CL = 3 is also considered for this case.

Table 21: Performance of the VPR algorithm for the HarbourMissionCAM dataset at different NT and CL in the very high turbidity waters.

	NT = [3,2]		NT = [6,4]	
	CL = 2	**CL = 3**	**CL = 2**	**CL = 3**
Precision (%)		100		66.67
Recall (%)	15.39	53.85	7.69	15.39
F1 score (%)	26.67	70	14.29	25

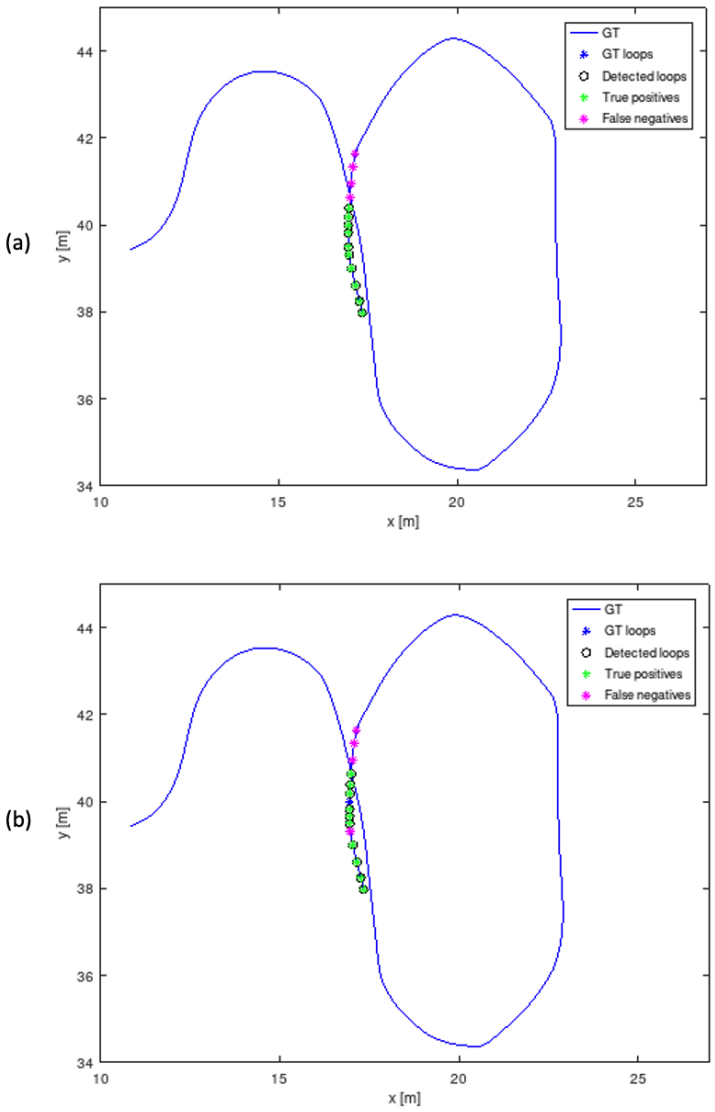

Figure 55: Appearance-based loop closure for the HarbourMissionCAM dataset in the high turbidity waters at NT = [3,2] (a) or NT = [6,4] (b) and CL = 1.

The application of CL = 1 and CL = 2 to these scenes is too low a value: with CL = 1 no loop closures are recognised, regardless of the NT. With CL = 2, 1 and 2 loops are recognised if NT = [6,4] or NT = [3,2] is used. If you then increase the clip limit for CL = 3, the algorithm can recognise 7 loops in the best case if you use NT = [3,2].

In contrast, splitting the images into 50x50 pixel blocks means that, given the inferior quality of the images, it is advantageous not to perform detailed histograms. This leads to indistinguishable and confusing features, so the algorithm loses strength and robustness. The best performance under this severe visibility conditions is then achieved with NT = [3,2] and CL = 3: TP = 7, FP = 0 and FN = 6.

Figure 56 illustrates this behaviour.

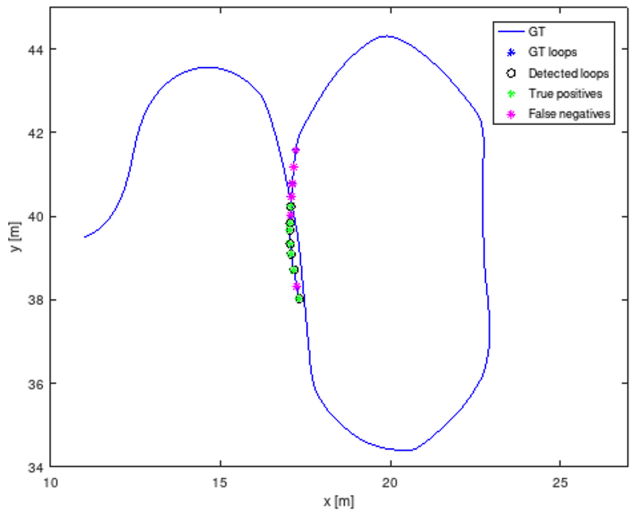

Figure 56: Appearance-based loop closure for the HarbourMissionCAM dataset in waters with very high turbidity at NT = [3,2] and CL = 3.

7.2.3 Discussion

This section has therefore presented a purely visual approach to place recognition that addresses perception problems in harbour scenarios to understand how far cameras can go due to their effective operation at close range. Therefore, a boundary characterisation of water types where cameras can be used to accurately detect loop closures was presented. Since the poor image quality may affect subsequent image processes such as feature extraction and consequently the matching of images with each other, the proposed method includes a preprocessing step to increase the contrast of the captured images to take full advantage of these sensors. CLAHE was applied to the captured images with different CL. Here, two different NT were used to understand the effect of creating fewer or more blocks, i.e. when more local enhancements are applied or not. This enhanced technique has been shown to be effective for an adequate perception of the environment. With considerable turbidity, the algorithm performs well and is able to recognise all loop closure situations without failures and without false detection of loops: 100% recall is achieved with 100% precision: TP = 14, FP = 0 and FN = 0. With a reasonable perception of the scene, a higher CL is not required and splitting the images into more blocks, i.e. NT = [6,4], is advantageous. By creating more detailed, i.e. local histograms (smaller areas of the image), we describe the image better, i.e. we emphasise robust features. With reduced image quality, i.e. scenes with high turbidity, the loss of visibility is noticeable: In this case, CL = 1 is not able to recognise all situations of loop closure, as in the previous visibility. If you increase CL, the FN are maintained and there are also FP in any NT configuration. Nevertheless, the behaviour

of the algorithm at NT = [3,2] is more stable in the sense that once you are able to detect that you are in the same place, you will not miss any loop situation along the straight line, which means that the generation of the histogram will generally produce slightly more robust features given the degraded input images. The algorithm can then achieve a recall of 71.43% with 100% of precision under these conditions: TP =10, FP = 0 and FN = 4. The case with very high turbidity is also tested. In this situation it is necessary to use CL = 3 to recognise more than 2 loops. Here it becomes even clearer that splitting the image into more blocks is not at all beneficial given the quality of the input image (visually degraded). The best performance under this stringent condition is then obtained based on CL = 3 and NT = [3,2], where the algorithm achieves a recall of 53.85% at 100% of precision: TP = 7, FP = 0 and FN = 6. The decrease in the performance of the algorithm is obvious with the degradation of the scene: greater turbidity. At high turbidity (water type = 0.4), the algorithm is already sensitive and has some difficulty in recognising robust and distinguishable features. At very high turbidity (water type = 0.45), it is impossible to recognise loops correctly and stably, even with larger image enhancements (CL = 3). Therefore, it is crucial to evaluate other sensor solution for an adaptive place detection system, for helping in cases where the camera is already starting to fail to recognise distinguishable features so that the underwater vehicles can perform accurate inspection missions.

7.3 BEHAVIOUR ASSESSMENT OF ACOUSTIC IMAGES

Harbour environments are often affected by turbidity or illumination, which often complicates the behaviour of navigation and mapping tasks performed by cameras, as the perception range of optical devices is severely limited in very poor visibility. Based on the previous section 7.2, in which the efficiency of a purely visual system to detect similar images with increasing turbidity was analysed, it became clear that the underwater environment can provide images without structural features under such conditions, even when CLAHE is used: Loop closures may be detected that are not suspected, and the vehicle may fail to recognise some loops correctly or incorrectly, resulting in no or incorrect trajectory adjustment. Today, imaging sonars emit acoustic waves and receive the backscatter, creating acoustic images that can be used to perceive the environment. Although these sensors suffer from distortion and occlusion effects due to their physical properties, they do not suffer from clouding effects because the images are based on the emitted and returned sound. Therefore, FLS can be a promising

solution for robust loop closure detection. FLS is characterised by the fact that it provides a representation of the environment in front of the robot and allows overlapping images during motion. Due to the characteristics of FLS data, namely low signal-to-noise ratio, low/inhomogeneous resolution and weak feature textures, which affects the visual appearance of images with weak feature textures, conventional feature-based registration methods are not yet used for acoustic imaging. However, the performance of these sonars has greatly improved and the resolution of their images continues to increase, so that FLS can provide comprehensive underwater acoustic images. Therefore, developing efficient methods for extracting visual data from sonar images and understanding their performance remains crucial to ensure that the FLS data is suitable for vehicle perception of the environment. In addition, this data can later be compared with the camera data if required as environmental conditions change. In this section, a method for acoustic place recognition is proposed. The idea is to apply the knowledge of visual to the acoustic data to evaluate whether it is effective in detecting loops at close range, and then exploit its potential in conditions where cameras can no longer provide such distinguishable information.

The most important points of this study are therefore:

- the evaluation of the effectiveness of extracting visual data, i.e. features from sonar images;
- understanding the needs and behaviour of sonar devices deployed near structures;
- the evaluation of the proposed visual approach to place recognition from sonar data.

FLS basics

Two-dimensional FLS is a new category of sonar devices that can provide acoustic images at a high frame rate and is therefore also referred to as an acoustic camera. The various operating specifications, such as acoustic beamwidth, operating frequency, acquisition rate and beamforming, are always associated with the sensor models. However, the operation of sonars remains the same: the sonar emits acoustic waves that cover its FOV in azimuth (θ) and elevation (ϕ), and the intensity of the returning beam is then determined based on a range (r) and bearing (θ), as shown in Figure 57. FLS imaging projects a 3D scene into a 2D image, just like an optical camera. The depth of objects is not lost, but in a sonar image it is not possible to uniquely determine the elevation angle at a given r and θ, i.e. the reflected echo can originate from any point in the reference elevation arc. The images are arranged and mapped in polar coordinates.

In this way, the measurements of a raw image correspond to the beams in the angular direction and the range samples in the distance axis. For easier interpretation, the resulting image is then mapped to 2D Cartesian coordinates, resulting in images with uneven resolution.

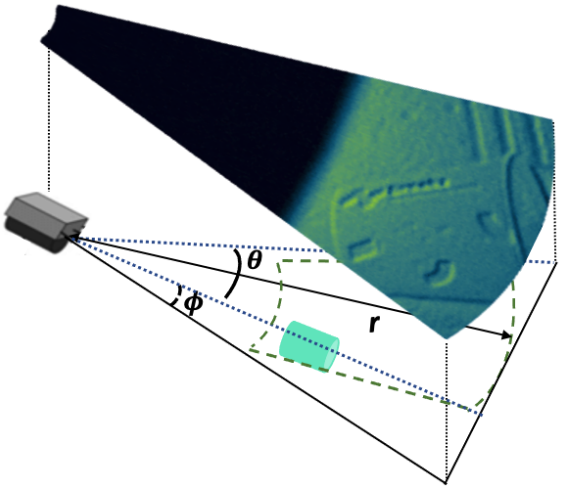

Figure 57: Illustration of FLS operation: The sound energy is returned based on r and θ and viewed as a map of 3D points on the null plane.

Acoustic images are able to see through cloudy environments, but at the cost of a much more difficult kind of data. Therefore, there are some issues associated with this type of sonar imagery that can be challenging for inspection tasks, such as:

- Low resolution: FLS are often categorised as high-resolution sonars, but their image resolution falls far short of modern cameras, which typically have millions of pixels. Of course, cross and down resolution is critical for image quality and distinguishing between closely spaced objects. However, the sparsity of measurements increases with distance when displayed in Cartesian space. This leads to uneven resolution, which affects the visual appearance of images with weak feature textures;

- Low signal-to-noise ratio: Even with a large FOV, sonar images have a high noise level caused by the mutual interference of sampled acoustic echoes, underwater motors near the surface or other acoustic sensors;

- Inhomogeneous insonification: FLS normally have a Time Varying Gain (TVG) mechanism with the aim of equalising transmission losses so that similar targets located at different distances can be perceived with similar intensity. However, changing the angle of incidence or tilt can lead to variations in image illumination and other effects that depend on the varying sensitivity of the transducers or lens, which in turn depends on their position in the sonar's FOV.

These inhomogeneous intensities can influence the image matching step and, of course, the pose estimation and loop closure phases. It is recommended to configure the forward-looking sonar so that there is a small angle between the imaged plane and the bore line (grazing angle), as this allows for the largest possible volume [81]. Of course, a small angle results in a larger area without reflected echoes in the image (black area), which reduces the effective imaging area, but this configuration allows vehicles to perform inspection tasks for structures that require close-range navigation. This is because it also avoids shadows in the images caused by occlusions and significant changes in visual appearance;

- Other artefacts: Interfering content can occur in the sonar images, which can lead to ambiguities when compared: acoustic reflections from the surface, artefacts due to reverberation or ghost artefacts. However, these interferences can usually be reduced by a suitable configuration and image composition.

7.3.1 Methodology

In order to decide whether an image is a new or not (known) place, a comparison is made between the queried images and the database images using a similarity measure. The extraction of features is therefore the first step in CBIR in order to obtain a numerical description. These features must then be aggregated and stored in a data structure in an abstract and compressed form to facilitate the similarity search. These measures indicate which places (image contents) are most similar to the current place. This is an important step that affects CBIR performance because an inappropriate measure will recognise fewer similar images and reduce the accuracy of the CBIR system. The proposed tree-based approach is therefore used for similarity detection for the intended context. This method was introduced for underwater optical images as previously presented in chapter 6 and is now adapted for acoustic images.

Figure 58 shows the process of place recognition based on FLS images. Based on this data, a Region of Interest (ROI) is used for each current image to discard the area without reflected echoes. Therefore, each input image is first "cut" into a rectangular region to account for the effective visual information captured by FLS. Based on the geometry model schematically shown in Figure 59, each FLS image - 514×750 pixels - covers a width of about 2m (A) and a height of 3m (B), i.e. it represents an area of about $6m^2$, as mentioned earlier. However, there is an area without reflected echoes, i.e. a black area on the images.

Therefore, based on the OpenCV (C++) library, a ROI is selected for each FLS image, which represents the effectively imaged area of the FLS output. Therefore, each image is first resized to 500×750 and a bounding box of 250×350 pixels is used, which means that each FLS image maps an area of about 1.5m^2, as shown in Figure 60. Based on all ROI images, ORB features and descriptors are extracted to then create the visual vocabulary and recognise similar places as previously described in section 6.2.1.

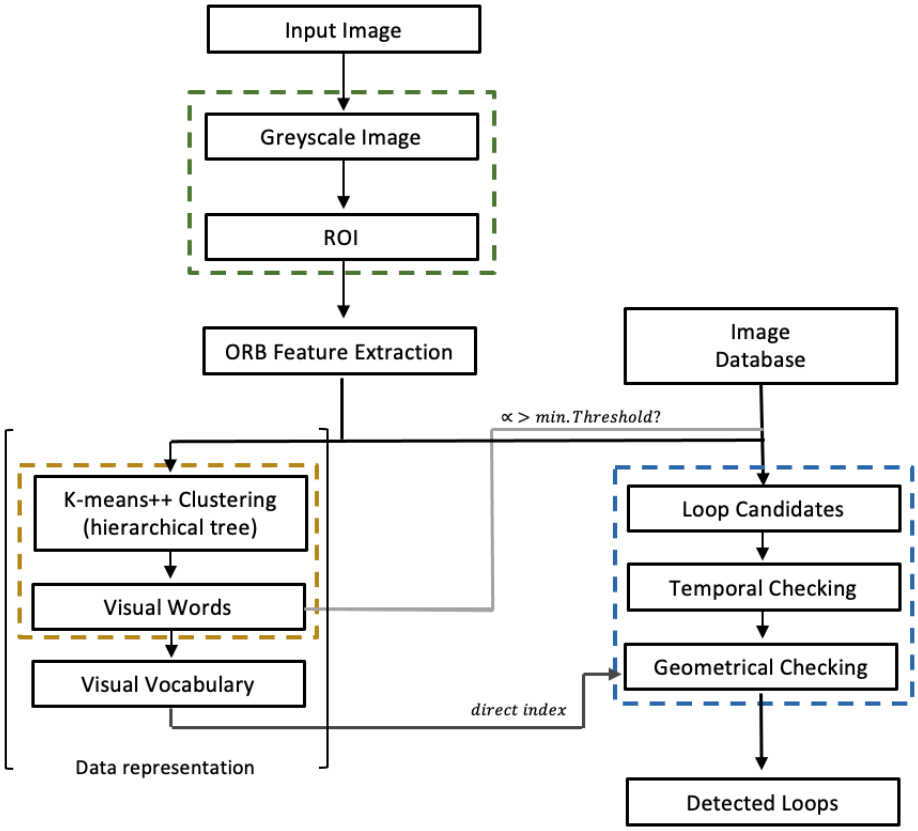

Figure 58: Block diagram of the proposed VPR algorithm based on acoustic (FLS) images.

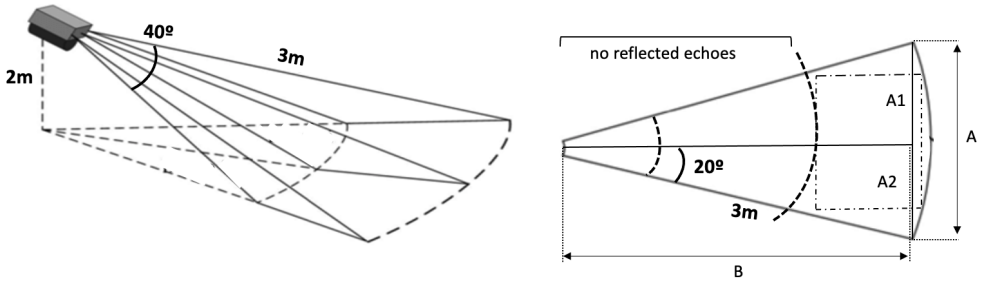

Figure 59: Illustration of the geometry model used for FLS operation.

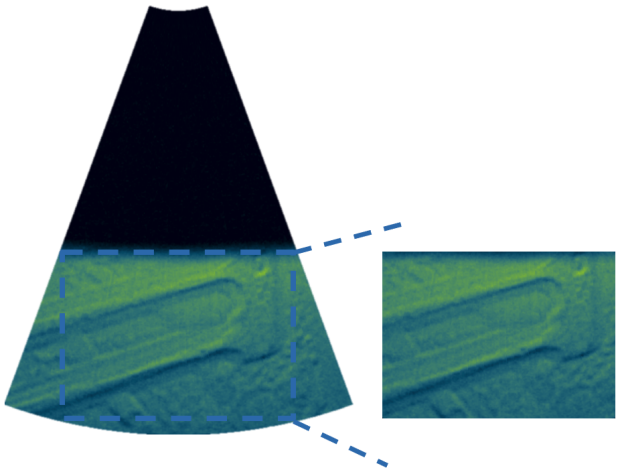

Figure 60: Illustrative example of the effectively mapped area of each FLS image.

7.3.2 Experimental Results

The operation of the algorithm for acoustic place recognition in an underwater harbour environment is described here. First, the performance measures used to evaluate the ORB on FLS images and their effectiveness are described and illustrated. The performance measures used to evaluate the detection of binary loop closures are then described. Their behaviour in detecting revisited places with different configurations is also shown. As with the optical data, a fully indexed vocabulary is created. Thus, the scenes are continuously modelled and then the recognition behaviour of the revisited places is analysed as an incremental learning approach. All experiments are based on up to 1500 features per image for the extraction step and were performed on a workstation with an Intel i7 7700K @ 4.5 GHz with 16 GB RAM and an NVIDIA GTX 1080.

Datasets

For this preliminary assessment of using FLS data to overcome camera issues in underwater environments, the HarbourMissionFLS dataset collected with the Stonefish simulator is used. The FLS is a top-down sensor. As it is intended to provide data on the ground near the structures of interest, its design and configuration were adapted accordingly. For this proposal, this sensor was then configured as a camera in the HarbourMissionCAM dataset in order to obtain a footprint as similar as possible and to analyse the behaviour of both sensors under the same conditions. Therefore, the AUV follows a trajectory close to the structures and closes a loop around the concrete wall with a total of 1067 FLS images. As a reminder, Figure 61 illustrates the AUV mission and highlights the area of the loop closure. The data is explained in more detail above in section 5.2.3.

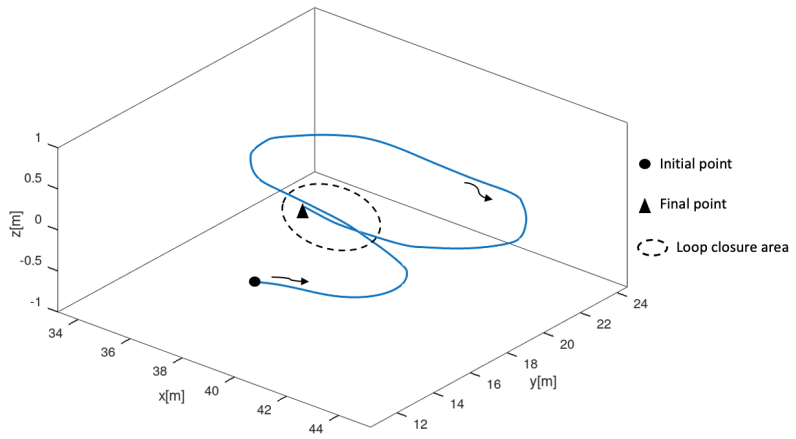

Figure 61: Predefined inspection route by the AUV for HarbourMissionFLS dataset.

Acoustic Features Effectiveness

Imaging sonars provide increasingly rich acoustic images that allow us to perceive the environment. However, inherent features such as weak textures, interfering content, and low signal-to-noise ratio can affect subsequent imaging phases, especially feature extraction and matching. Therefore, a functional analysis is considered to analyse the behaviour of the ORB on acoustic images based on the number of ORB keypoints (Kp), the number of matches (Nm) and inliers. As described in section 6.3.2, the recognised Kp on the input images are linked to the visual content, as a higher value usually indicates a better description of the scene. Nevertheless, in some situations there may be noisy pixels that lead to incorrectly recognised keypoints. Nm stands for the number of Kp of a query image that correspond to those in the current image, based on the Hamming distance. Finally, the inliers are the correct correspondences between two images. For this, RANSAC is used from the previously obtained matches: the higher the value, the more similar the compared images.

Different textures were tested under standard environmental conditions to analyse the behaviour of feature extraction and matching algorithms on FLS images. Table 22 shows the performance evaluation for each scenario. A set of 30 consecutive images is used, where each value is determined by a simple arithmetic mean.

Table 22: Functional evaluation of grid and floor regions for clear waters based on FLS images.

		Grid Area	Bottom Area
Functional Evaluation	**Kp**	326	
	Nm	326	
	Inliers	90	41

As you can see, the descriptor finds the same number of keypoints for grid and bottom areas. However, the effective matches are higher in the grid region because the bottom region contains only a few stones, an area with little texture. An example of this behaviour for both scenarios can be found in Figure 62, which is illustrated by a high number of linear lines (correct matches). Table 23 shows the obtained functional score considering optical images for both scenarios, based on the same sequence of 30 images. It is evident that optical images have more keypoints and correct matches between consecutive images when visibility is appropriate, as these sensors provide images with more details and have a stronger texture perception, since FLS images are based on emitted and returned sounds.

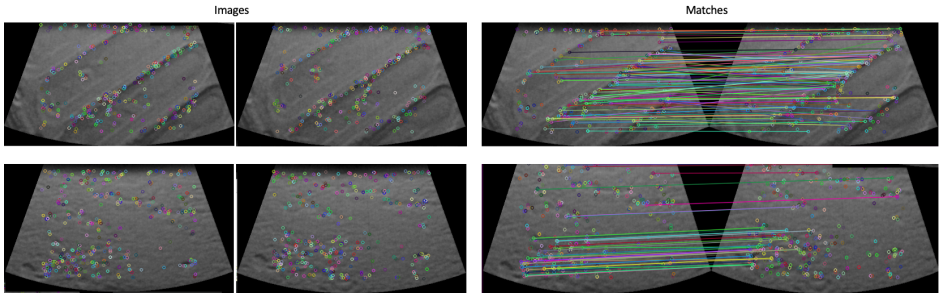

Figure 62: Comparison of detected keypoints (shown as dots) and matches (shown as straight lines) in FLS images on scenes with (top) and without texture (bottom).

Table 23: Functional evaluation of grid and floor regions for clear waters based on optical images.

		Grid Area	Bottom Area
Functional Evaluation	**Kp**	935	332
	Nm	914	313
	Inliers	643	151

However, the cameras react sensitively to poor texture scenarios, as shown by the decline in all functional measurements (bottom area). In fact, the performance in such areas was not that far from that obtained with the acoustic images: only three times more instead of seven. Moreover, the FLS do not suffer from visibility problems, while the cameras lose performance in such scenarios, as can be seen in Table 24, where the camera detects about 80 fewer inliers under turbid conditions. It was expected that the performance of the optical images would be worse than that of the acoustic images, but the turbidity set in the simulator was not severe. Our aim was to understand the impact of this effect on the performance of the optical image-based methods. Nevertheless, this suggests that this behaviour is not sufficient to

obtain inliers between images acquired at different times, i.e. with different appearances, and thus to detect loops under these conditions.

Table 24: Functional evaluation based on inliers for floor areas in turbid waters based on acoustic and optical images.

		Bottom Area
Inliers	**FLS**	41
	Camera	70

On the other hand, Table 25 takes into account poor lighting conditions, and as can be seen, the cameras are also not able to provide many distinguishable features and thus find many matches between consecutive images. As expected, the FLS can fulfil these conditions without any loss of performance. This behaviour is evident when looking at consecutive images, which show that the cameras do not consider loop closures in such environment scenarios, i.e. in low light conditions; the scene becomes darker or brighter, resulting in unstructured images.

Table 25: Functional evaluation based on inliers for floor areas in light scenes based on acoustic and optical images.

		Bottom Area
Inliers	**FLS**	41
	Camera	18

Loop Closure Detection

Given this drop in performance of visual data compared to FLS images in poor visibility conditions based on consecutive images, it is being investigated whether this acoustic sensor can help in such situations to find similar images to recognise that the robot has already visited a certain place and thus provides a correct estimate of the vehicle's position.

In order to evaluate the behaviour of the proposed acoustic place detection algorithm, a ground truth of loop closures were created for the HarbourMissionFLS data. For this purpose, a boundary of about 1m was drawn to consider two images as the same place, based on the effectively mapped FLS area. In this way, 16 real situations of loop closure were considered. The precision and recall metrics are used to assess the performance of location detection. Thus, TP, FP and FN cases are considered for the evaluation against the ground truth. For this purpose, each loop determined by the place recognition algorithm is compared with the defined ground truth based on a maximum

allowable distance. If the detected end loop has a distance of less than 0.3m, it is classified as TP. Otherwise, the loop is categorised as FP. The remaining loops that are considered true in the ground truth are then defined as FN, since the algorithm could not identify them.

For visual images, a similarity threshold (α) of 0.3 was used to assign images as loop candidates, as this value was best suited to avoid distant features being considered the same. Thus, if we also consider this similarity threshold when searching for loop candidates in FLS images, we can see the behaviour of the place detection method in Figure 63. As you can see, the algorithm does not recognise false loop closures, i.e. there is no FP. This means that the recognised features are distinguishable, taking into account the similarity threshold used. However, the algorithm tends to lose loops along the trajectory as only 6 loop closures are correctly recognised. This fact may mean that the similarity threshold for the appearance provided by the acoustic images can be very high.

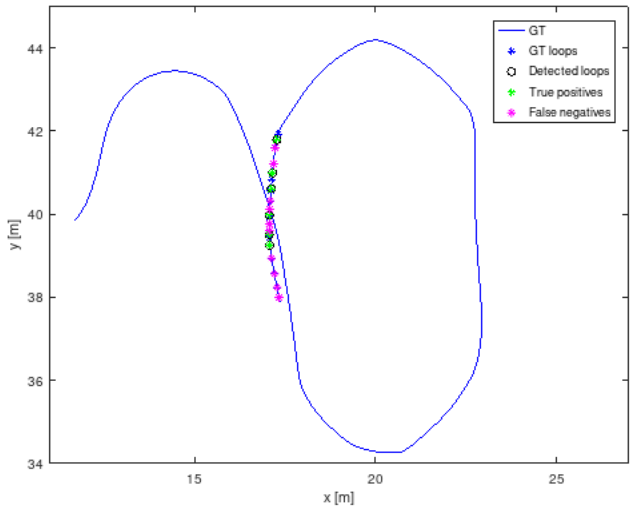

Figure 63: Appearance-based loop closure for the HarbourMissionFLS dataset at α = 0.3 with minFPnts = 12.

Table 26 shows the performance values achieved by the algorithm in place recognition considering different similarity thresholds. It achieves a precision of 100% in all cases, i.e. there is no FP, which is crucial in the navigation context. If you lower the similarity requirements, more situations where loops close are recognised. Nevertheless, the performance improvements between iterations are not very significant. Moreover, the best results are obtained with α = 0.15, which seems to be too low a threshold and leads to each image being wrongly considered as a loop candidate, since points far away from each other are treated as the same features. Nevertheless, the algorithm fails to close 5 loops, i.e. it presents five FN and achieves a maximum recall of 68.75%. This behaviour probably indicates that the features are not robust enough to benefit from the benevolence of the requirements.

Table 26: Performance of the VPR algorithm for the HarborMissionFLS dataset when varying the similarity threshold α, with minFPnts = 12.

	α = 0.3	α = 0.25	α = 0.2	α = 0.15
Precision (%)		100		
Recall (%)	37.50	43.75	50.00	68.75
F1-Score (%)	54.55	60.87	66.67	81.48

To decide whether an image (loop candidate) actually represents the same place of a database image, a geometric check is performed between image pairs, as described in section 6.2.1. For optical images, this check must be supported by at least 12 correspondences. But, the inherent properties of FLS sonar can affect the visual appearance of the images (weak textures). Therefore, Table 27 shows the effects of reducing the minimum number of correspondences, minFPnts, to consider two images as the same location. The method allows a minimum of 5 correspondences. A similarity threshold α of 0.3 is considered for all cases.

Table 27: Performance of the VPR algorithm for the HarbourMissionFLS dataset when varying the threshold for minimum correspondences minFPnts, at α = 0.3.

	minFPnts = 10	minFPnts = 8	minFPnts = 6
Precision (%)		100	
Recall (%)	62.50	81.25	87.50
F1-Score (%)	76.92	89.66	93.33

By reducing the number of inliers between the images of the query (loop candidate) and the database, the algorithm recognises more situations in which a loop closes. The best results are obtained when only 6 correspondences between images are considered. This is an inappropriate value for comparing images taken with different timestamps and considering that the method allows at least five correspondences as a lower limit.

Figure 64 illustrates the behaviour of the algorithm for two initial cases (considering 10 and 8 correspondences). Above a certain value, however, the increase is no longer significant: 10, 13 or 14 loop closures are correctly recognised and there is still not much difference in the variation of this parameter. There are therefore no more matches that fulfil the requirements of RANSAC.

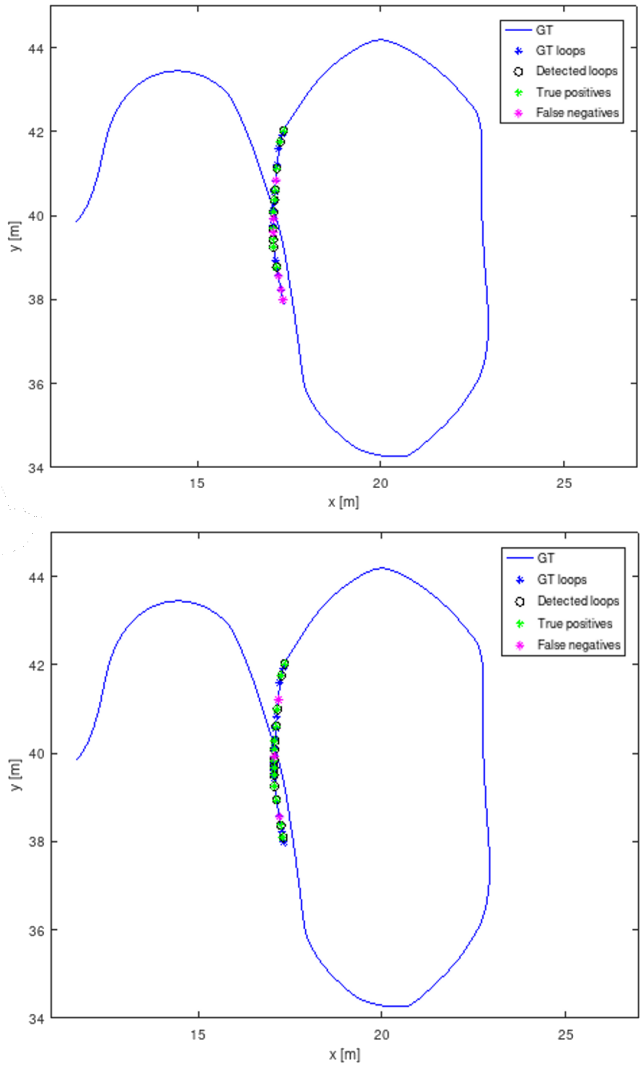

Figure 64: Appearance-based loop closure for the HarbourMissionFLS dataset at $\alpha = 0.3$ and minFPnts = 10 (top) or minFPnts = 8 (bottom).

Comparing the results for $\alpha = 0.3$ and using 12 or 10 as the minimum number of matches to consider the images as equally local, the performance of the algorithm increases; the number of unrecognised loop closures, i.e. 10 FN, becomes the number of correctly recognised loops, i.e. 10 TP. Therefore, it is important that the similarity threshold requirements are higher for images from different locations that are not recognised as similar. Geometric verification is a final check (like filtering) and therefore it is better to be a bit more benevolent at this stage. Nevertheless, it is obvious that the features and matches are indistinguishable in order to benefit from the benevolence of the requirements. The simplification of the two conditions, i.e., similarity and matches, is impractical because distant features are considered as the same point and many images are mistakenly considered as candidates for a loop. This leads to incorrect detection of a loop closure, which is

dangerous for the navigation context as the trajectory is adjusted in an inappropriate way, resulting in incorrect pose estimation. Considering the experiments and some assumptions, the behaviour based on this parameterisation, i.e. based on $\alpha = 0.3$ and minFPnts = 10, seems to be the most suitable for FLS images, achieving a precision of 100% and a recall of 62.50%.

Image Enhancement Procedure

To counteract image degradation, i.e. the poor visibility that leads to missing image details on optical images, enhancement techniques are used to correct the lack of matching features. There are some common methods, but for underwater scenes, CLAHE has proven to be the better choice to extract additional information from low-contrast images, as presented in section 6.3. Figure 65 shows an overview of the process.

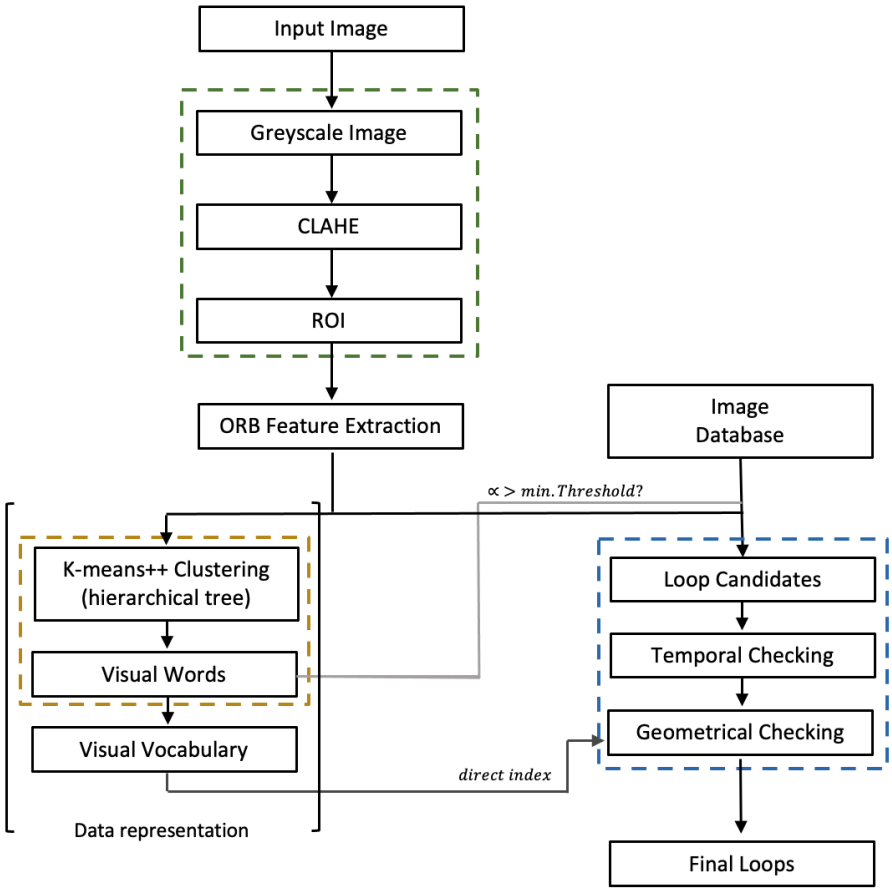

Figure 65: Block diagram of the proposed VPR algorithm based on acoustic (FLS) images with CLAHE enhancement technique.

In short, this technique emphasises edges in specific regions and enhances local contrast by dividing each image into distinct, non-overlapping sub-images or tiles. For each block, the histogram is cropped and remapped to avoid excessive emphasis. To avoid block

artefacts, a bilinear interpolation is then performed between neighbouring tiles. There are therefore two key parameters that mainly determine the quality of the image enhancement: the NT, which determines the number of blocks that divide the images into squares of equal size, and the CL, which controls the noise gain and is also known as the contrast factor. If you increase the CL value, the resulting image will generally be brighter, as a higher CL value flattens the image histogram. A higher NT value increases the dynamic range and contrast of the image. However, selecting a suitable configuration of these parameters is a challenge as they are related to the resolution of the image and the content of the histogram. The described experiments show that FLS has inherent properties that can degrade data and provide images with weak texture, i.e. features that are not very robust. Therefore, based on optical images and using the OpenCV 3.4 (C++) library, CLAHE is applied to FLS images to understand whether these changes can improve image quality and consequently feature extraction and matching. For this purpose, each image is divided into blocks of 62,500 pixels, i.e. NT = [2,3]. After applying an enhancer, the brightness and details of the images are expected to increase while maintaining the naturalness of the image, but this depends on the visual content. Therefore, CL = 1 and CL = 2 were tested. Figure 66 illustrates the changes in the appearance of the images by each CL, without and with texture, and shows that the CL increases the image contrast and the edges and details are improved. Consequently, the keypoints detected on the enhanced images are higher than on the original images in both cases.

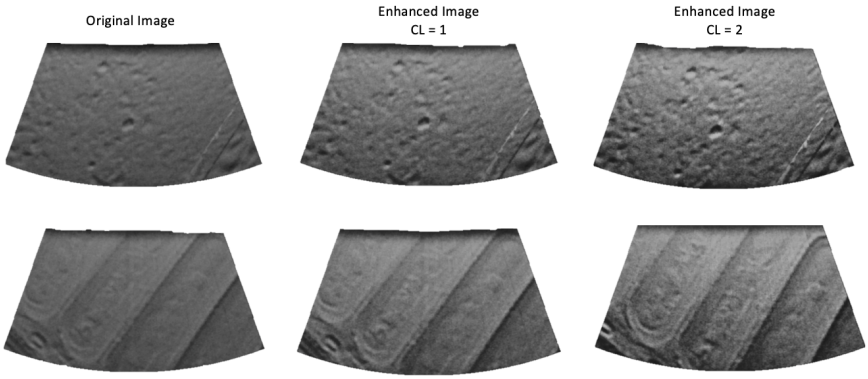

Figure 66: Illustrative example of the effects of the CLAHE-based enhancement method, where CL is varied in scenes with (bottom) and without texture (top) in the HarbourMissionFLS dataset.

But "*are these features strong enough to improve the accuracy of the matching phase?*" Therefore, considering the effect of CLAHE, the behaviour of the place detection method was tested for both CL to understand whether the enhancement provides a better description of the images. Table 28 shows the achieved performance in recognising similar places,

based on the same test conditions and assumptions from the above experiments. In this context, a similarity threshold α of 0.3 and only 12 or 10 matches (minFPnts) are considered the minimum value to consider two images as similar. As you can see, the algorithm achieves a precision of 100% in every case, i.e. there are no misidentified loops. This means that CLAHE increases the image detail and maintains the distinctness of the features for the similarity threshold used. One would expect recognition to increase with contrast, i.e. as CL increases. However, this is only at the expense of reducing the geometric test requirements, i.e. the minimum number of inliers to consider a candidate as an effective loop.

Table 28: Performance of the VPR algorithm for the HarbourMissionFLS dataset with varying CL and threshold for minimum correspondences minFPnts, at NT = [2,3].

	CL = 1		**CL = 2**	
	minFPnts = 12	**minFPnts = 10**	**minFPnts = 12**	**minFPnts = 10**
Precision (%)	100			
Recall (%)	87.50	81.25	75.00	87.50
F1-Score (%)	93.33	89.66	85.71	93.33

Moreover, the algorithm achieves the same recall in two cases, but again with a lower value of inliers at CL = 2. This shows that more keypoints do not necessarily mean distinguishable points; the enhancer may generate points that are falsely recognised as keypoints because they may be noise, for example. So it seems that the algorithm shows a more stable behaviour when CL = 1 and minFPnts = 12 is used. It achieves a recall of 87.5%, recognises 14 loops and only drops two loops, as can be seen in Figure 67.

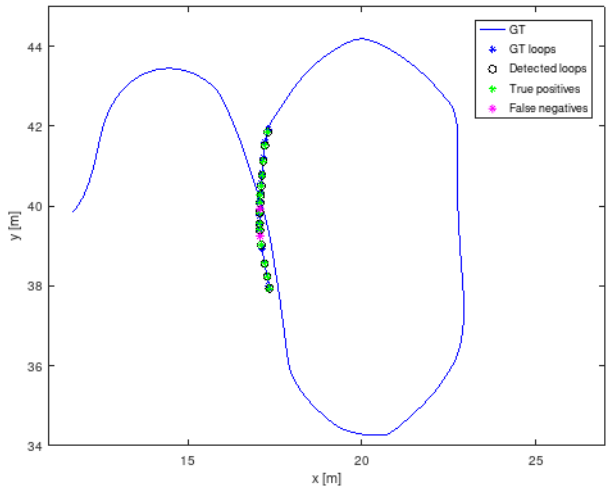

Figure 67: Appearance-based loop closure for the HarbourMissionFLS dataset based on α = 0.3 and minFPnts = 12 at CL = 1 and NT = [2,3].

Next, NT = [4,6] is used to find out whether more subdivision of the images, i.e. more detailed histograms, has an influence on the image description and improves the similarity detection behaviour of the algorithm. Table 29 shows the behaviour of the algorithm in detecting previously visited places, also based on a similarity threshold α of 0.3 and only 12 or 10 matches as minFPnts as the minimum value to consider two images similar.

Table 29: Performance of the VPR algorithm for the HarbourMissionFLS dataset with varying CL and threshold for minimum correspondencesminFPnts, at NT = [4,6].

	CL = 1		CL = 2	
	minFPnts = 12	**minFPnts = 10**	**minFPnts = 12**	**minFPnts = 10**
Precision (%)	100			
Recall (%)	68.75	87.50	62.50	93.75
F1-Score (%)	81.48	93.33	76.92	96.75

In all cases, a precision of 100% is achieved, which means that splitting the images maintains the distinctness of the features. On the other hand, increasing CL only increases recall at the expense of decreasing the minimum matches for two images considered similar in content, as in the previous case. Looking at CL, the behaviour with CL = 2 is better at minFPnts = 10, where another loop closure situation is detected. In this case, CL = 1 achieves the same performance as in the previous case, i.e. a recall of 87.50%, but allowing for 10 inliers for two images representing the same location, which leads us to assume that in this case increasing the number of subdivisions does not describe the image better. With CL = 2 and 10 inliers, the algorithm achieves a recall of 93.75%, with only one loop closure detection failing. The generated image with this value of the contrast threshold therefore does not provide any robust keypoints.

For both experiments, i.e. the tables 28 and 29, and considering CL = 1 and CL = 2, it is obvious that a better description of the image is obtained with NT = [2,3] considering at least 12 inliers. On the other hand, the performance of the algorithm is higher when NT = [4,6] and 10 inliers are taken into account to consider images as similar, and in both cases the difference is due to the detection of another loop closure.

7.3.3 Discussion

This section therefore presents a purely acoustic method of place recognition that can overcome the inherent limitations of perception in harbour scenarios. Poor visibility can complicate the behaviour of navigation and mapping tasks performed by cameras. Apply what

is known about visual data to acoustic data to assess whether it is effective in detecting loops at close range, and then utilise its potential in conditions where cameras can no longer provide such distinguishable information. In general, sonar data provides fewer features than camera images under normal visibility conditions. However, as the visual images become more demanding, the performance of the FLS remains the same, while that of the optical sensor decreases, reaching 18 points, of which the FLS can recognise 41 in low light conditions. The turbidity has increased, even at not too strong values, and the matching performance between successive images decreases sharply. The algorithm detects about 80 fewer inliers compared to normal visibility conditions, suggesting that optical-based methods may not respond effectively to such severe conditions and fail to detect loop closures. Considering this degradation of visual data performance in low visibility conditions, it was investigated whether FLS can help to find similar images in such conditions to recognise that the vehicle is already in a certain area and thus enable a correct estimation of the vehicle position. Based on the usual thresholds for optical images, i.e. considering $\alpha = 0.3$ and minFPnts = 12, the algorithm does not falsely recognise situations where a loop closes, i.e. FP = 0 achieves 100% of precision. However, the strength of the algorithm is low as it tends to lose loops along the trajectory. It only recognises 6 situations in which loops close and thus achieves a recall of 37.5%. This performance is unsuitable for navigation purposes and the requirements for the similarity threshold and the minimum number of inliers were therefore analysed. The lower the similarity threshold, the more situations in which the loop closes are recognised by the method. Nevertheless, the performance improvements are not very significant. If the minimum value of matches between the images is lowered so that they are considered to be the same location, the algorithm can obviously recognise more loop closures. However, a balance must be found to achieve a quality trade-off. With 10 inliers as minimum threshold and $\alpha = 0.3$, the algorithm recognises 10 loops and achieves 62.5% recall and 100% precision. However, the experiments show that the features are not robust and therefore the matches cannot be distinguished to benefit from the benevolence of the requirements. Therefore, the effect of an enhancement technique was evaluated to understand whether it provides a better description of the images. The CLAHE method was used as it was found to be the better choice for extracting additional information from low contrast images. To improve contrast and highlight image edges, it depends on the parameterisation of two parameters: NT, which divides the images into equal square regions, and CL, which controls noise amplification. If we look at the resulting images, we can see that the contrast has been effectively increased and the details improved, so that more keypoints can be recognised com-

pared to the original images. To understand whether there are noisy pixels and whether these features are robust enough to increase the number of correct matches, we tested the behaviour of place detection for CL = 1 and CL = 2 as well as NT = [2,3] and NT = [4,6]. Initially, one would expect the recall to increase with contrast, but this only happens at the expense of reducing the geometric test requirements, i.e. by resorting to 10 inliers. Furthermore, the algorithm achieves the same maximum recall in two cases, but again with a lower value of inliers at CL = 2. Thus, it can be seen that more keypoints do not necessarily mean distinguishable points. Next, the number of image tiles was also increased, i.e. for NT = [4,6], to find out whether creating more detailed histograms has an impact on the image description and consequently improves the similarity detection behaviour of the algorithm. In this case, increasing CL only increases recall at the expense of reducing the minimum matches for two images to be similar in content. Moreover, in this case the algorithm achieves the same performance as in the previous case, but taking into account 10 inliers. It can therefore be assumed that an increase in the number of subdivisions is not accompanied by a better description of the images. In all cases, a precision of 100% is achieved, which means that there are no misidentified loops, i.e. CLAHE increases the image details and maintains the discriminative power of the features for the similarity threshold used. The experiments thus show that increasing both parameters does not lead to a better description of the scene. Under the same assumptions, the most balanced result was obtained with CL = 1 and NT = [2,3]. Under these conditions, the proposed FLS-based approach fails only for two loop closures and achieves a recall of 87.5%, which corresponds to an increase of 50% compared to the original result. Typically weak texture of FLS images shows to contribute to the algorithm achieving a precision of 100% in all cases. Even if you lower the similarity threshold, this behaviour is not affected. This performance is crucial in a navigation context where the goal is to maximise recall at 100% of precision, as this way no incorrect trajectory adjustments are made that can lead to incorrect pose estimation of the vehicle. This fact also complicates the strength of the algorithm not to miss any loops during the trajectory, as a minimum of consistent matches between the images is required. Therefore, CLAHE improves the behaviour of the algorithm, i.e. it increases the image contrast and details (without noisy pixels) and more effective features are detected. For this reason, there are no features that could confuse the algorithm and thus affect its robustness in correctly recognising a place. Since the images of the FLS are based on emitted and returned sounds, the behaviour of the FLS is the same regardless of the environmental conditions. This shows that the FLS supports loop detection in poor visibility when the camera no longer provides detailed information.

7.4 SUMMARY

The place recognition mechanism is therefore essential to avoid cumulative positioning errors. Indeed, visual loop detection is widely used as cameras provide a large amount of information and are very robust at short distances. However, these sensors are very sensitive to visibility conditions, i.e. changes in appearance. Therefore, a feasibility analysis for adaptive place recognition in harbour scenarios was carried out in this chapter. For this purpose, the behaviour of the binary loop closure was first evaluated using optical data under different visibility conditions and different water types and then using acoustic data.

To understand the extent to which cameras can solve perception problems in harbour scenarios, we first presented a boundary characterisation of water types where cameras can be used to accurately detect loop closures. With considerable turbidity, the algorithm performs well and is able to recognise all loop closure situations without failures and without false detection of loops: 100% recall is achieved with 100% of precision. With a reasonable perception of the scene, a higher CL is not required and splitting the images into more blocks, i.e. NT = [6,4], is advantageous. With increasing water turbidity, i.e. a deterioration in image quality, the loss of visibility was actually noticeable at high turbidity: In this case, CL = 1 is not able to detect all loop closure situations as in the previous visibility. On the other hand, the FN are preserved even if CL is increased, and there is also FP in each NT configuration. Nevertheless, the behaviour of the algorithm is more stable when NT = [3,2]. Under these conditions, the algorithm can then achieve a recall of 71.43% with a precision of 100%. With very high turbidity, it is necessary to use CL = 3 in order to recognise more than 2 loops. Here it becomes even clearer that splitting the image into several blocks is not at all favourable. The best performance under this strict condition is then achieved based on CL = 3 and NT = [3,2], where the algorithm achieves a recall of 53.85% at 100% of precision. The performance degradation of the algorithm is obvious with the degradation of the scene: greater opacity, which is dangerous in a navigation context.

Therefore, a purely acoustic method was proposed to understand whether the robots can fulfil their inspection mission regardless of the challenges of the environment. To this end, the knowledge of the visual data in the acoustic data was used to assess whether it is effective in detecting loops at close range, and then to exploit its potential in conditions where the cameras can no longer provide such distinguishable information, thus enabling a correct estimation of the vehicle position. Based on the same thresholds for optical images, i.e. considering α = 0.3 and minFPnts = 12, the algorithm falsely fails

to recognise situations where a loop closes, i.e. FP = 0 achieves a precision of 100%. However, the strength of the algorithm is low, as it tends to lose loops along the path. It recognises 6 situations in which loops close and thus achieves a recall of 37.5%. This performance is unsuitable for navigation purposes as no adjustments are made to the trajectory. With 10 inliers as a minimum threshold, the algorithm recognises 10 loops and achieves 62.5% recall with 100% precision. However, the experiments show that the features are not robust and therefore the matches cannot be distinguished to benefit from the benevolence of the requirements. Therefore, the effect of an enrichment technique was also evaluated here to understand whether enrichment provides a better description of the images for acoustic image-based addressee recognition. The CLAHE method was used as it was found to be the better choice for extracting additional information from low contrast images. The contrast was effectively increased and the details improved so that more keypoints can be recognised compared to the original images. Increasing the two configuration parameters, i.e. CL and NT, does not lead to a better description of the scene. Under the same assumptions, the most balanced result was obtained with CL = 1 and NT = [2,3], where the proposed FLS-based approach only fails at 2 loop closures and achieves a recall of 87.5%, an increase of 50% over the original result.

Based on experiments, the similarity detection method therefore performs worst from a considerable turbidity (water type = 0.4) for visual images compared to acoustic images in any CLAHE configuration, which means that we no longer take advantage of the texture of visual images: In the best case, the algorithm fails at 4 loop closures, while it only fails to recognise 2 loops for FLS images. This indicates that the FLS supports loop detection in poor visibility when the camera no longer provides detailed information. This chapter has opened new horizons and shown that they can be of great help for underwater vehicles to perceive the environment in low visibility conditions and thus not to jeopardise the task of navigation in the vicinity of structures. Therefore, an adaptive navigation method improves the ability of vehicles to navigate in real underwater scenarios to perform accurate inspection missions.

8 HYBRID POSE ESTIMATION

The precise localisation of a vehicle is an essential prerequisite for inspection work: For autonomous navigation to be successful, the robot must know its position over time. Given the effectiveness of cameras operating at close range, visual odometry is a robust technique used for this purpose. However, the perception problems typical of harbour environments can make it difficult to determine an accurate vehicle position, as is the case with place recognition. In such situations, the FLS data has proven to be a promising solution to help underwater vehicles perceive their surroundings in poor visibility conditions. Therefore, a Visual Odometry (VO) approach for harbour scenarios based on optical and acoustic data is proposed in this chapter to solve the perception problems in harbour facilities and prevent the vehicle from losing its position during the inspection mission.

The chapter is organised as follows: Section 8.1 provides an overview of VO, highlighting the typical problems encountered in underwater environments and the help that imaging sonars can provide in understanding the scenes under certain conditions. Section 8.2 describes the proposed method of visual odometry and illustrates its effectiveness based on optical images. A boundary characterisation of the visual conditions is presented to use this sensor for accurate vehicle positioning. Section 8.3 evaluates the behaviour of odometry based on acoustic images to understand if this data can be used as a tool for an inspection mission. Finally, section 8.4 describes the main conclusions of this chapter.

8.1 INTRODUCTION

Localisation is the most important task for autonomous vehicles so that they can follow their path and correctly identify and avoid obstacles. Therefore, to perform inspection tasks in semi-structured environments, such as harbour facilities, underwater vehicles must be able to navigate in a feasible and stable way near and/or around the structure to be inspected. Among the various techniques for positioning mobile robots, VO stands out as a robust technique for vehicle localisation, namely for close-range operations in underwater scenarios, as the images also contain a sufficient amount of meaningful information such as colour, texture, shape, etc. Based on this technique, a robot

can localise itself using only a stream of images captured by one or more cameras: a process of estimating the ego-motion, i.e. the translation and orientation of the robotic vehicle. However, the performance of VO is directly related to the imaging conditions, which can interfere with the pixel displacement calculation and lead to errors in the displacement estimation. Considering that port scenarios are affected by turbidity or illuminations that can produce images with missing structural features, it is clear that the inherent perception problems of these environments can easily directly affect the behaviour of the localisation task and consequently compromise the inspection mission. This sensitivity of optical sensors to strict visibility conditions was demonstrated in the previous chapter 7, where cameras are unable to provide detailed information even when CLAHE is applied and the robustness to recognise similar loops was lost. Although imaging sonars provide images with weak texture, FLS data proved to be a good solution to support vehicle perception in poor visibility conditions.

Therefore, the development of efficient scene perception approaches is suitable for vehicle localisation. Thus, the aim is to understand to what extent the camera can cope with difficult conditions and to evaluate the behaviour of FLS data under such conditions. Therefore, under the same assumption for both visual data, a hybrid odometry approach is proposed in this chapter to provide consistent and stable localisation for successful continuous navigation of underwater vehicles depending on the environmental conditions.

8.2 VISUAL POSE ESTIMATION

As already mentioned, it became clear that the cameras lose performance in poor visibility conditions, it is no longer possible to utilise the detailed information they provide, even if CLAHE is applied. Nonetheless, these solutions are effective for close-range operations and so the aim here is to understand where the limits are for utilising visual data, i.e. that the vehicle is able to detect robust features in order to use this information to accurately position the vehicle without compromising the performance of AUV inspection missions.

The method for estimating the position and motion of robots is based on analysing consecutive visual data, i.e. each frame is considered as a step measurement: a movement to the right or left, up or down indicates a change in the scene from frame to frame, which can be challenging in underwater environments, especially in inspection missions where the robot has to make short movements along a structure.

Using computer vision algorithms, these changes can be interpreted as pose changes. Specifically, for each pair of images, we need to find the rotation matrix R and the translation vector t that describe the movement of the vehicle between the two images based on feature extraction and tracking. VO can be broadly categorised into a monocular or a stereo system depending on whether data from one or multiple cameras is used. A monocular system suffers from scale uncertainty and the vector t can only be computed up to a scale factor, but it mitigates the effects of calibration errors in motion estimation, is essentially suitable for small robots, and is effective when the distance to the scene is much larger than the baseline (stereo). Therefore, monocular feature-based odometry is proposed to analyse the behaviour of pose estimation under different visibility conditions.

8.2.1 Methodology

In underwater environments, VO is a major challenge but plays an important role as ordinary sensors are usually impractical for vehicle positioning in these hostile environments. To understand the behaviour of this technique in challenging underwater scenes, a monocular VO approach is proposed based on a modified version of mono-vo[1], an OpenCV 3 (C++)-based implementation of an VO algorithm. Feature tracking is used to extract features in the source images and then try to retrieve the same features in the same view instead of extracting them independently and then matching them by comparing descriptors, i.e. feature matching, to obtain high temporal resolution. It is based on FAST features, but as mentioned before, when using image processing for navigation tasks, it is crucial to extract reliable image features.

Given the advantages of using binary features and the proven efficiency of ORB features in underwater environments, these features are also considered for vehicle localisation. On the other hand, the experiments in Chapter 6 have shown that the challenging conditions of these environments can lead to missing image details, making robust feature extraction and matching difficult. It has been shown that the enhancement technique (CLAHE) enables an appropriate perception of the scenes by increasing the inliers between the images. This preprocessing step is also considered here for all greyscale images before feature extraction.

The approach is therefore modified accordingly: The process overview for the proposed visual odometry is shown in Figure 68.

1 https://github.com/avisingh599/mono-vo

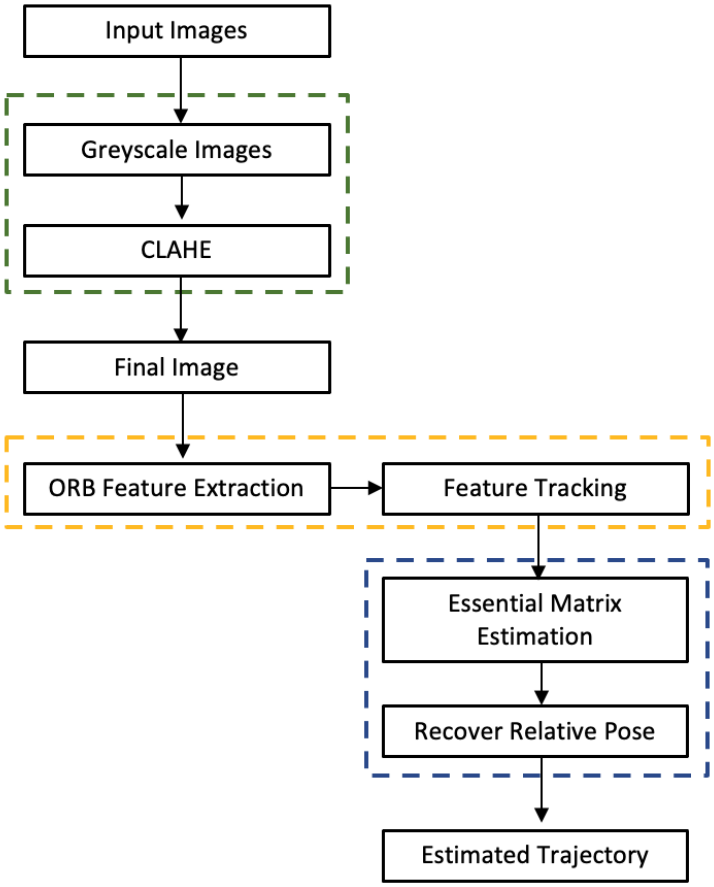

Figure 68: Block diagram of the proposed VO algorithm based on optical images with CLAHE enhancement tecnhique.

Thus, starting from the captured images I^t and I^{t+1}, features are extracted in I^t and these features are then tracked in I^{t+1}, after CLAHE (highlighted by the green block in Figure 68) has been applied: ORB features are used and the corners detected in the previous step are fed into the next step, which uses the KLT tracker, indicated by the yellow block in Figure 68. It looks for a sparse pixel-wise correspondence that basically circles each corner to be tracked, and uses this local information to find the corner in the next image: It assumes a point is nearby and uses image gradients to find the best possible feature point movement. In KLT tracking, some points may be lost (as they move out of the field of view) and re-detection is triggered when the total number of features falls below a certain threshold. As the blue block (in Figure 68) shows, once we have point correspondences, the essential matrix is estimated to obtain the relationship, i.e. to compute the relative camera positions between two consecutive images based on the camera focal length and the principal point obtained by the calibration process. Nister's 5-point algorithm is used for this purpose. Since feature tracking algorithms are not perfect and therefore we have several erroneous correspondences, RANSAC is used to handle outliers: At each iteration, five points are randomly selected

from the set of correspondences, the essential matrix is estimated, and then this matrix is used to check whether the other points are inliers. The algorithm ends after the fixed number of iterations and the essential matrix with which the maximum number of points matches is used. The relative pose is then restored, i.e. R and t are estimated with Singular Value Decomposition (SVD) of the essential matrix. Finally, we can follow the trajectory by associating the translation vectors and the rotation matrices with the following equations:

$$R_{pos} = RR_{pos} \tag{10}$$

$$t_{pos} = t_{pos} + tR_{pos}\,, \tag{11}$$

where R_{pos}, t_{pos} indicates the pose of the camera.

8.2.2 Experimental Results

Therefore, the performance of the proposed visual odometry system for vehicle localisation in underwater harbour scenarios is described. Experiments were conducted with data under different visibility conditions. First, an analysis of the effects of applying CLAHE to the images on the ego-motion is performed. Each image/movement means a change in the scene, so the estimation of the scaling factor may become erroneous if the change is large, leading to an incorrect estimation of the resulting trajectory. Since vehicle control in underwater is unstable, an analysis of the camera frame rate is also considered to minimise the influence of unintended vehicle movements on the pose estimation. Subsequently, the behaviour of VO in scenes under different operating conditions is investigated based on these assumptions. All experiments are based on the extraction of up to 2000 features per image and were performed with an AMD Ryzen 7 2700X @ 3.7Ghz with 16GB RAM computer.

Datasets

The evaluation of the behaviour of ORB-based odometry in underwater scenarios was performed with some datasets created with the Stonefish simulator (for more details see section 5.2.3): some trajectory parts of the datasets PortSeafloorA and HarbourMissionCAM were considered. In this context, both datasets are captured here at a rate of 30 Hz.

Figure 69 shows each trajectory as a reminder and highlights the samples for each dataset used.

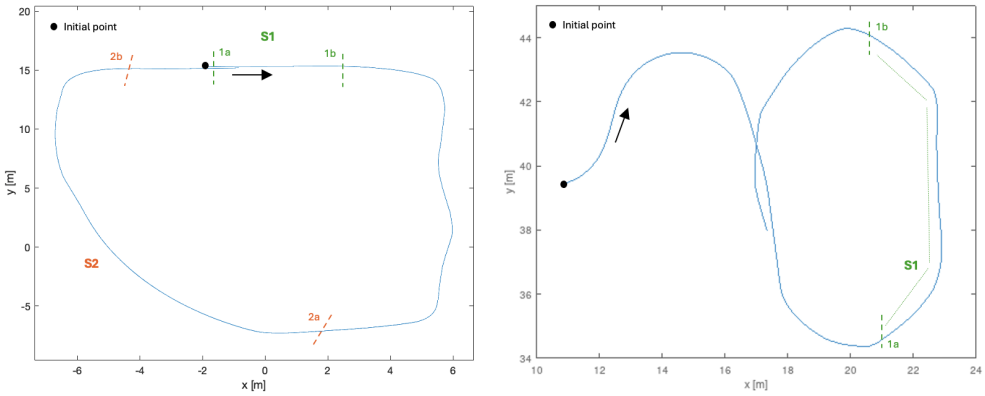

Figure 69: Samples of trajectories of the AUV for PortSeafloorA (a) and HarbourMissionCAM (b) datasets, used for optical odometry.

To understand the effectiveness of CLAHE and the impact of frame rate on vehicle positioning accuracy as a preliminary assessment, the PortSeafloorA dataset is used. In this case, the camera is pointed at the seabed and the vehicle follows a simple trajectory with a loop closure and few direction changes, as shown in Figure 69 (a). The scenario with clear water, whose trajectory comprises 1789 images, is then considered for these first experiments. The very low turbidity scenario, comprising 1782 images, is also considered to understand the impact of deteriorating quality on pose estimation. The first example, labelled S1, comprises 120 images in a straight motion. Track S2 represents a curve with two directional changes of direction and comprises 770 images. To evaluate the visual odometry algorithm under strict visibility conditions and to understand where the limits of visibility for accurate navigation with cameras lie, the HarbourMissionCAM dataset is used. In this case, the AUV simulates an inspection mission around a concrete wall, where the camera captures images of the ground near the structures, see Figure 69 (b). Taking into account the results of the experiments in section 7.2, scenarios with considerable and high turbidity were considered, which here comprise 3910 and 3938 images respectively. The selected sample (S1) contains directional changes and comprises about 1500 images.

Odometry

Firstly, the impact/susceptibility of using consecutive images when displaying possible erroneous changes in the scene was analysed on the basis of different frame rates. For this purpose, using the example of the beginning of the trajectory (S1) of PortSeafloorA in clear water spanning over 100 frames, the behaviour of the VO system in estimating this trace based on all captured frames, $\frac{1}{3}$ and $\frac{1}{6}$ of the captured frame rate was investigated, see Figure 70.

As you can see, only with $\frac{1}{3}$ of frames (Figure 70 (b)) it is possible to obtain a better perception of the linear motion and avoid that small deviations of the vehicle motion influence the ego motion.

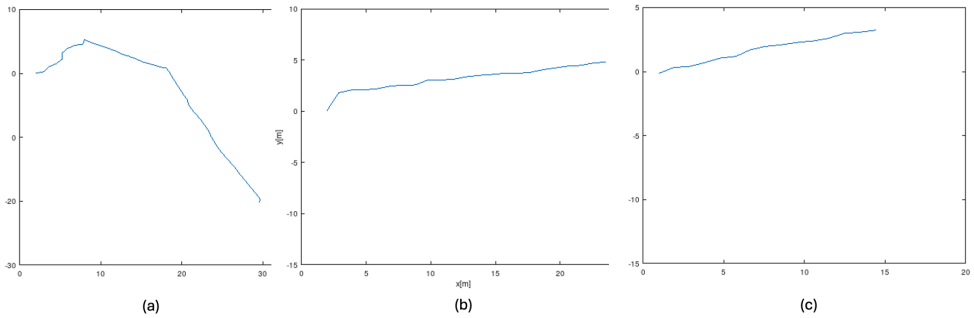

Figure 70: Appearance-based odometry of a straight line for the PortSeafloorA dataset in clear water at CL = 1 and NT = [3,2] at acquired frame rate (a) and at $\frac{1}{3}$ (b) or $\frac{1}{6}$ (c) of frame rate.

Nevertheless, the effect of CLAHE is evaluated in the pose estimation for this track where the robots move in a straight direction. CLAHE is effective to perceive the environment. However, it is based on the manual configuration of the CL and NT parameters to increase the naturalness of the images. As each image consists of 300×200 pixels, NT = [3,2] and NT = [6,4] are taken into account. CL = 1 and CL = 2 were tested for each NT, as the images do not have poor visibility.

Figure 71 illustrates the more appropriate behaviour achieved with CL = 1 and NT = [3,2].

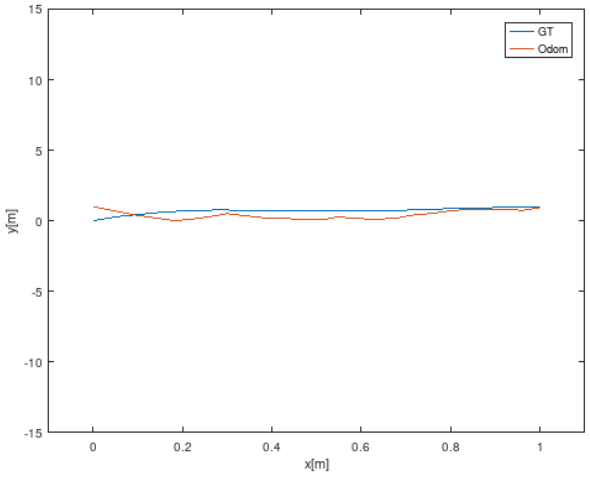

Figure 71: Appearance-based odometry of a straight sample (S1) for the PortSeafloorA dataset in clear water at CL = 1 and NT = [3,2].

Although there is no scale information in this case, the visual odometry is able to estimate a linear motion despite some deviations. With CL = 2, the algorithm is not able to estimate the executed movement of the

vehicle. This may be due to the fact that the images have sufficient visibility and therefore the enhancer is higher, resulting in poor looking images with excessive contrast or "stretching" of the colour channels. Performing a more detailed histogram, i.e. using NT = [6,4], is not beneficial as the algorithm is not able to determine the correct motion whether CL = 1 or CL = 2 is used: The estimated motion contains some deviations or emphasised slopes, which means that the enhancement method produces many noisy keypoints.

The behaviour of the algorithm when changing direction was then analysed using the S2 path of the PortSeafloorA dataset, see Figure 72 (a), initially with clear water. Figure 72 (b) shows the estimate obtained with the VO algorithm, where the camara follows the ground truth (GT), although not with such a circular motion. This behaviour is based on CL = 1 and NT = [3,2], as expected from the above analysis of the effectiveness of CLAHE under these water visibility conditions. In this case, however, the images had to be resized to 480×320 pixels to achieve this motion perception. This can possibly be explained by the fact that the trajectory area is less structured: the robots move in the sand area for most of the track and encounter two less structured objects, namely an anchor and a plastic cup (in contrast to track S1, where the AUV sees a grid, stones and tree trunks). In this case, only a qualitative analysis is possible, because it is not possible to quantify the movement (there is no scaling factor). Moreover, there is no certainty about the estimated movement, i.e. the shape of the movement could be an "illusion", since the last pose is maintained if there is no estimate.

In this way, the behaviour of the algorithm is described in terms of the deviation of the orientation (from the tangent to the trajectory) between successive positions to allow a more concrete analysis of the proposed odometry algorithm. This measure is insensitive to scaling and is calculated as follows:

$$\theta\ (^{o}) \ = \ atan2(\Delta y, \Delta x) \times \frac{180}{\pi}, \tag{12}$$

where θ is the angle measure in degrees in the range [-180°, 180°] between the changes of the poses (Δy and Δx). If there are no estimates, the previous pose is retained.

Therefore, Figure 72 (c) shows a comparison between the estimated and the actual orientation of the robot during the trajectory. As you can see, the algorithm can start as intended being that 60% of the samples approach the ground truth. The behaviour worsens when the robot starts to shift its motion forward and then orientate to the right (opposite to the initial position).

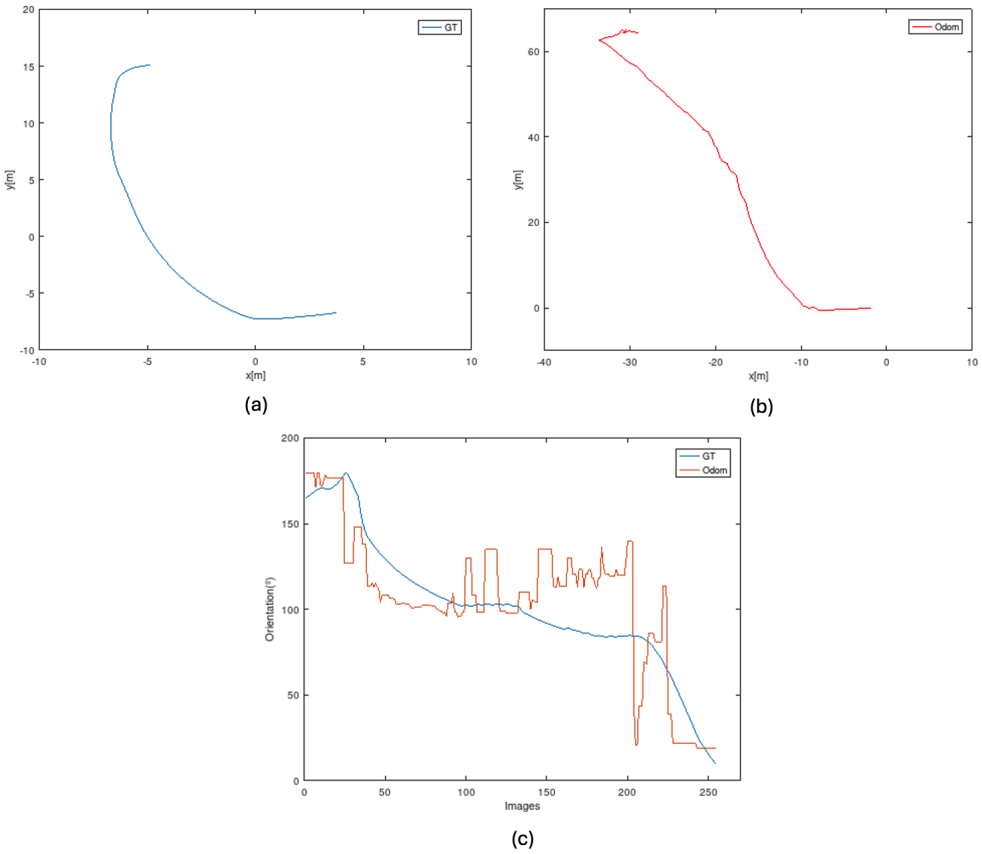

Figure 72: Appearance-based odometry of a curve sample (S2) for the PortSeafloorA dataset in clear water at CL = 1 and NT = [3,2].

For a quantitative analysis, the integral of absolute error (IAE) between ground truth and visual odometry, which in words represents the area of the two curves, is calculated on the basis of their orientation deviations, denoted by θgt and θodom:

$$\theta e\ (^{o}) \ = \ \int_{1}^{samples} \ | \ (\theta odom - \theta gt) \ | \ dx \tag{13}$$

In this case, the average total error is 20.2°. If we consider up to 150 samples, the average error decreases by 5°, which refers to the 60% of the trajectory. The maximum peak value (error) is more or less 60°, which is therefore reached at the last change of direction to reach 0°.

Figure 73 shows the estimate obtained for the same track, i.e. S2 of the PortSeafloorA data, to preliminarily understand the behaviour of the cameras when estimating the pose in poor visibility. Although it is already a challenging track as the robot navigates in regions with little texture, VO is able to detect the directional changes even in this case, but it is obvious that it was more difficult to perceive the curvilinear motion of the robot, see Figure 73 (a).

This behaviour is due to image enhancement: CL = 2 and NT = [6,4]. Nevertheless, the algorithm is more sensitive compared to the previous case, i.e. there are some fluctuations as the features are not as robust, making the transformation between the images not as clear and reducing the performance of the estimated trajectory. This can be seen in Figure 73 (b), where the visual algorithm is able to correctly estimate the initial orientation, but soon loses the notion of orientation because it thinks that it should practically maintain the direction (rectilinear movement), while in reality it should orient itself in the opposite direction, i.e. decrease the orientation. However, with the second change of direction, it can better understand the true motion. Based on (13), the average total error is 32.5°. If we consider up to 150 samples, the average error is, as expected, twice as high as in clear water (33° instead of 15°). From sample 100 onwards, the average error decreases and is 27.5°, as this is the moment of the second direction when the robot starts to shift its motion forwards and then orientates itself to the right, towards the end of the trajectory.

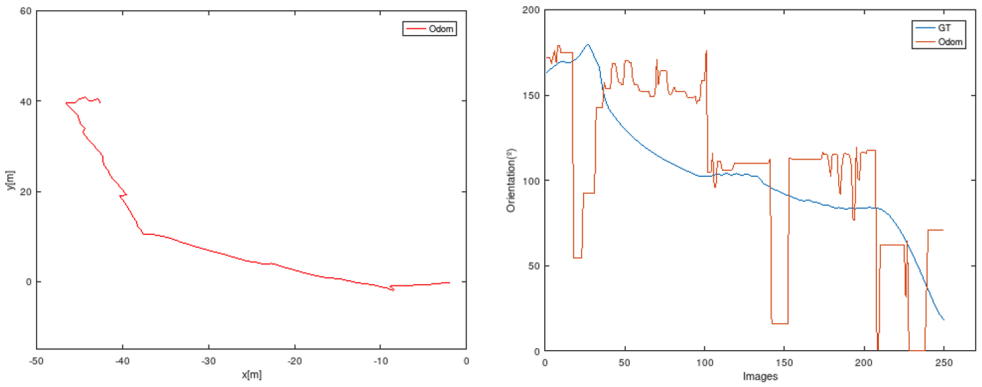

Figure 73: Appearance-based odometry of a curve sample (S2) for the Port-SeafloorA dataset at very low turbidity at CL = 2 and NT = [6,4].

Next, to validate the limits of visibility for operation with cameras for pose estimation between consecutive images under challenging appearance conditions, the stretch S1 of the HarbourMissionCAM dataset captured by the Stonefish simulator is used, as shown in Figure 69 (b). This data was captured using a camera configured to capture images of the ground near the structures of interest in an appropriate downward direction. In addition, the camera is tilted 35°, which allows for a wider area and preemptive navigation when the robot has a medium/long range perspective, such as at FLS. However, this configuration means that the images are not planar to the movement of the vehicle, which means that you have a false perception of movement (tilted). In this respect, some areas are "seen" by the robot as very close (zoom) and others as very far away, i.e. the appearance of the same object looks different and the extracted features are very different and

therefore not matchable in the feature tracking step. Different parameterisations were considered, but for the reasons mentioned it was not possible to reach a configuration where the algorithm could follow the robot's movement. More specifically, there were many estimation errors and those that could be predicted were mostly wrong, leading to inappropriate changes in the estimate. This was to be expected considering how difficult the tracking process is, which makes a correct transformation between consecutive images impossible. For example, the algorithm loses 455 features in a sequence of 15 images in straight forward motion, which means that the appearance of the objects is not considered similar.

In order to understand the behaviour of visual odometry under different visibility conditions, this dataset was therefore simulated with a camera pointing at the ground (planar), placed far enough in front of the robot to simulate the most preemptive navigation possible. In this case, the algorithm loses 25 features in straight forward motion in the same sequence of 15 frames mentioned above as an example of tilted data acquisition. Figure 74 shows scenes recorded with the camera tilted (a) and planar to the motion (b).

Figure 74: Illustration of the scene of the HarbourMissionCAM dataset, recorded with a titled (a) and a planar (b) camera.

The visual positioning algorithm in this S1 track of the inspection dataset, see Figure 75 (a), was then first evaluated in water type between clear and with very low turbidity. Figure 75 (b) therefore shows the behaviour of the algorithm with respect to the orientation deviations according to (12). This behaviour was determined on the basis of CL = 2 and NT = [3,2]: The visibility conditions are not as strict as in the previous case and the track generally has more texture as the robot navigates near a wall that encloses a paper cup, the grid and some tree trunks. The position algorithm can only estimate the movement from frame 60 onwards, which is perhaps justified by the fact that the scene is textureless: sand/sand with the side of a tyre. Nevertheless, from this point onwards, visual odometry is able to perceive the movement of the robot, even if it is not so smooth, i.e. with some orientation fluctuations. However, there are two phases in which the algorithm recognises the correct orientation of the movement but is subsequently unable to estimate the position and thus maintain the previous pose, as you can see from the straight lines between images 150-300 and 360-430, which correspond to the textureless region of the

track. More specifically, in the first case, the image consists of sand with the side of a tyre, which can make feature tracking difficult due to the constant appearance of the object. In the second case, although it is a sand region, it corresponds to the change of the scene, i.e., it shows the end of the tyre and only sand, which may cause the algorithm to temporarily fail to track the features. Based on (13), the overall average error is 28.7°. From frame 60, the average error decreases to 27.1°. The visual positioning method performs better from frame 480 to the end corresponding to the last direction change, as it can correctly estimate the orientation of the executed robot movement and achieves an average error of 14.3°.

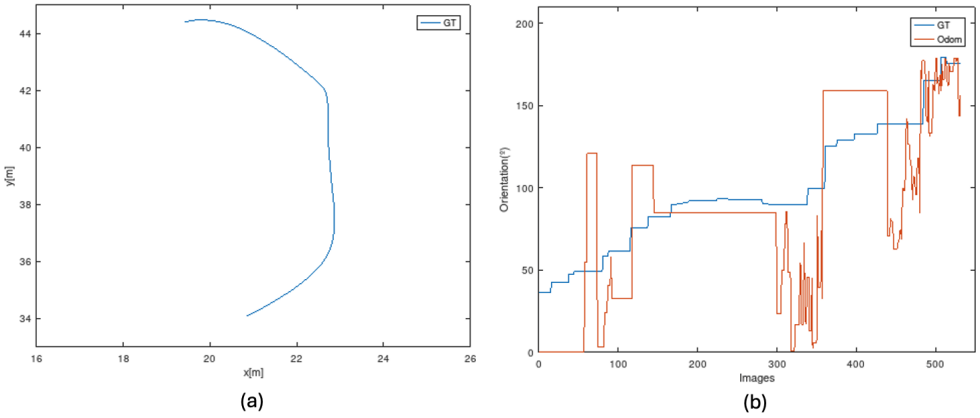

Figure 75: Appearance-based odometry of an inspection sample (S1) for the HarbourMissionCAM dataset in clear water at CL = 2 and NT = [3,2].

In order to understand to what extent the cameras can handle inherent underwater scenes so that the robot can perform inspection missions, the behaviour of the visual positioning method is analysed for different visibility conditions, namely for waters with considerable and high turbidity. First, the behaviour of the algorithm was investigated for considerable turbidity based on CL = 2 and NT = [3,2], as in the previous case. One would expect that the image must be enhanced compared to the previous visibility, but these effects resulted in the motion not being fully perceived, probably because the greenish hue does not hinder the full perception of the scene with the naked eye, which means that when using a higher enhancer, the images were therefore unnecessarily emphasised, resulting in non-robust features. On the other hand, in this case, fewer features need to be activated for the re-tracking mechanism, i.e. the matches between the features decrease more. This means that the features are less robust, as expected. This prevents the algorithm from doing too many re-tracking steps, which makes the estimate less accurate. Therefore, the visual algorithm can follow the robot trajectory under these conditions, but with even more orientation fluctuations as the features are not as ro-

bust and the perspective changes are less noticeable, see Figure 76 (a). As in the previous case, there are the same two phases in which the algorithm cannot estimate the pose, but at no time can it estimate the turn performed by the robot with this visibility. The overall average error is 49.1°.

With high turbidity, CL = 6 and NT = [3,2] are required. In this case, the visibility is so strong that the images need to be enhanced to extract the features. Nevertheless, the features are not robust, i.e. they lead to indistinguishable and confusable features, which makes the visual algorithm very sensitive, i.e. with many oscillations given the actual motion of the robot (motion relatively linear), as can be seen in Figure 76 (b).

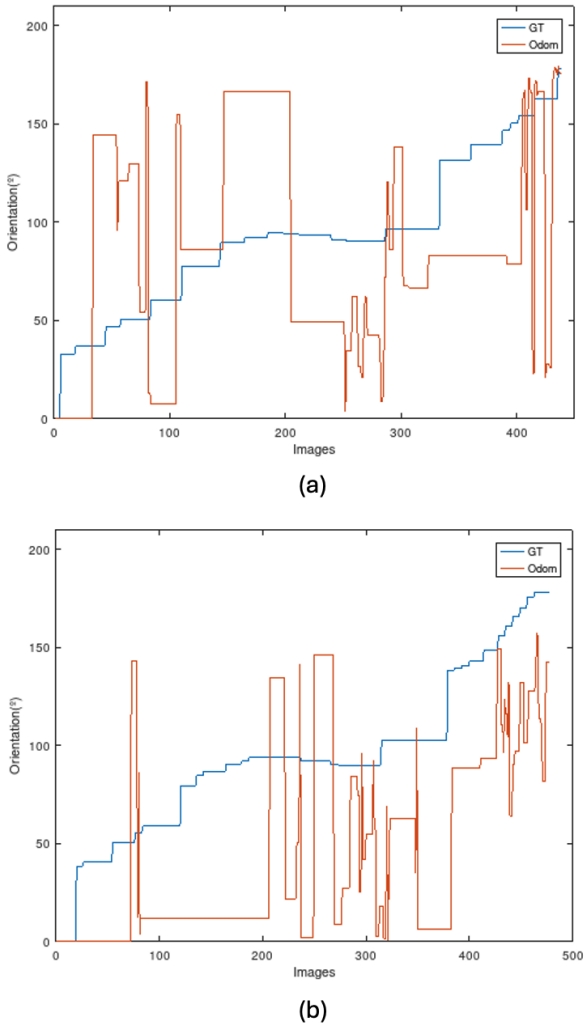

Figure 76: Appearance-based odometry of an inspection sample (S1) for the HarbourMissionCAM dataset at considerable (a) and high (b) turbidity at CL = 2 or CL = 6 and NT = [3,2].

Moreover, under these conditions, the position algorithm can only estimate the motion from frame 210, which corresponds to the area where the texture is preserved: The robot hits a paper cup, a grid and

tree trunks on the ground. Therefore, the average total error for this degree of opacity is 57.7°.

Table 30 shows an overview of the overall average error based on (13), denoted here as AIAE, of the visual positioning algorithm as a function of visibility conditions for the HarbourMissionCAM dataset. As expected, the performance of the method decreases as the turbidity of the water increases, making the identification and tracking of robust and distinguishable features more difficult.

Table 30: Performance of the VO algorithm for the HarbourMissionCAM dataset under different water visibility conditions.

	clear water	**considerable turbidity**	**high turbidity**
AIAE (°)	27.13	49.05	57.73

8.2.3 Discussion

Cameras are effective when it comes to operate closeness, but there are some environmental conditions that can produce images with poor texture, which can affect the detection of robust features and consequently the performance of pose estimation. In this section, monocular feature-based odometry is introduced to analyse the behaviour of pose estimation under different visibility conditions and to understand where the limitations lie for accurate use of this data. Firstly, the effect of using all frames according to the capture rate is considered. Therefore, this preliminary assessment is based on clear water and on the first frames of the PortSeafloorA dataset, where the robots move in a straight direction. Only with $\frac{1}{3}$ of frames it is possible to get a better perception of the movement (linear). Nevertheless, the effect of CLAHE is evaluated and it is proved that this enhancement technique is effective to perceive the environment. NT = [3,2] and NT = [6,4] are considered for CL = 1 and CL = 2. Performing a more detailed histogram, i.e. using NT = [6,4], is not beneficial: The estimated motion contains some deviations or accentuates slopes, which means that the improvement technique produces many noisy keypoints. With CL = 2, the algorithm is not able to estimate the executed movement of the vehicle. This means that the gain is too high with sufficient visibility for the perception of the scene, resulting in poor looking images with excessive contrast or "stretching" of the colour channels. The behaviour of the algorithm when changing direction was analysed using a path, also from the PortSeafloorA dataset, as a preliminary assessment. Initially, in clear water, the best performance was obtained based on CL = 1 and NT = [3,2], as expected from previous analyses under these conditions. However, in this case, the images had to be upscaled to 480×320 pixels for the algorithm to perform all directional

changes, perhaps due to the trajectory region being less structured. Using the scale-independent measure, i.e. the deviations of the orientations, it can be seen that the algorithm approaches ground truth in 60% of the samples and achieves an overall average error of about 20°. Next, this track was performed in poor visibility to gain an initial understanding of the behaviour of the cameras and enable accurate pose estimation in more challenging conditions. The algorithm is able to recognise changes in directional, but it was obviously more difficult to recognise the curvilinear movements of the robot. As expected, under these visibility conditions, it was necessary to further improve the images, i.e. to resort to CL = 2 and NT = [6,4]. Nevertheless, the features are not as robust as the algorithm is more sensitive to losing the orientation movement shortly after the estimates begin: It thinks that it should keep the direction in principle, while in reality it orients in the opposite direction. Under these conditions, the method reaches an average total error of about 32°. Therefore, the limits of visibility for operation with cameras to estimate vehicle position between successive images under difficult appearance conditions were validated using a trace of the HarbourMissionCAM dataset. Although a titled camera was essential for preemptive navigation, the same object looks different in this configuration, so there are no matching features in the tracking phase. Consequently, the positioning algorithm is unable to estimate the movement of the robot. This data was then simulated using a camera pointing at the ground (planar), placed far enough in front of the robot to enable the most preemptive navigation possible. Based on images taken in a water type between clear and with very low turbidity, the best performance was determined based on CL = 2 and NT = [3,2], as there were no severe visibility conditions and the tracks are generally structured when the robot navigates close to a wall. Despite some orientation deviations, the algorithm is able to estimate the intended movement. There are two phases where the algorithm is not able to estimate the position and maintain the previous pose, perhaps because they are textureless regions and especially because the change in the scene causes the features to be temporarily untrackable. The overall average error is around 29°. The visual positioning method performs better from the last direction change onwards, achieving an average error of 14°. To understand to what extent the cameras can process inherent underwater scenes in such a way that the robots can carry out inspection missions, scenes with water with considerable and high turbidity are then considered. For considerable turbidity, the algorithm performs better when it is based on CL = 2 and NT = [3,2] as in the previous case. The method can follow the intended trajectory, but with more orientation fluctuations, which means that the features are not robust and therefore the perspective changes are perceived less, as expected. Therefore, in scenes with considerable

opacity, visual odometry reaches an average total error of 49°. At high opacity, CL = 6 and NT = [3,2] are required, which was to be expected given the strict visibility of the input images. However, this extension also leads to confusing features that make the algorithm very sensitive when it comes to correctly estimating the perspective changes. For example, the positioning method can only estimate the last 40% of the trajectory, perhaps because this corresponds to the area where the texture is preserved. At this level of opacity, optical-based odometry therefore achieves an average total error of around 58°. The visual odometry algorithm is reacts sensitively to changes in appearance that are typical of the underwater environment: Pose estimation performance decreases as water turbidity increases, making it more difficult to detect and track robust and distinguishable features. It is therefore important to understand whether acoustic data can help with positioning.

8.3 ACOUSTIC POSE ESTIMATION

Harbour environments are usually affected by turbidity or light incidence, which often complicates the behaviour of navigation and mapping tasks performed by cameras, as the perception range of image processing-based devices is severely limited in very poor visibility or bright scenes, as these effects provide images without structural features, even when CLAHE is used, as shown above. Nevertheless, it became clear in section 7.3 that FLS can be a promising solution for robust feature extraction in such poor visibility conditions. Therefore, a feature-based odometry based on acoustic images is proposed to analyse the behaviour of pose estimation.

8.3.1 Methodology

As mentioned before, to estimate the movement of the vehicle in such a context, the VO method is emphasised, where a comparison between the two consecutive images is performed using feature tracking. The proposed method has already been introduced for underwater optical images (see section 8.2) and is now adapted for acoustic images. Therefore, based on this data, a ROI is used for each current image to discard the area without reflected echoes to account for the effective visual information captured by FLS. Based on the geometry model, a bounding box of 250×350 pixels is then used, which means that each FLS image maps an area of about 1.5 m^2, see Figure 59. Based on all ROI images, ORB features are extracted and tracked to then estimate the position of the vehicle as previously described in section 8.2.1.

Figure 77 shows the process of ego-motion based on FLS images.

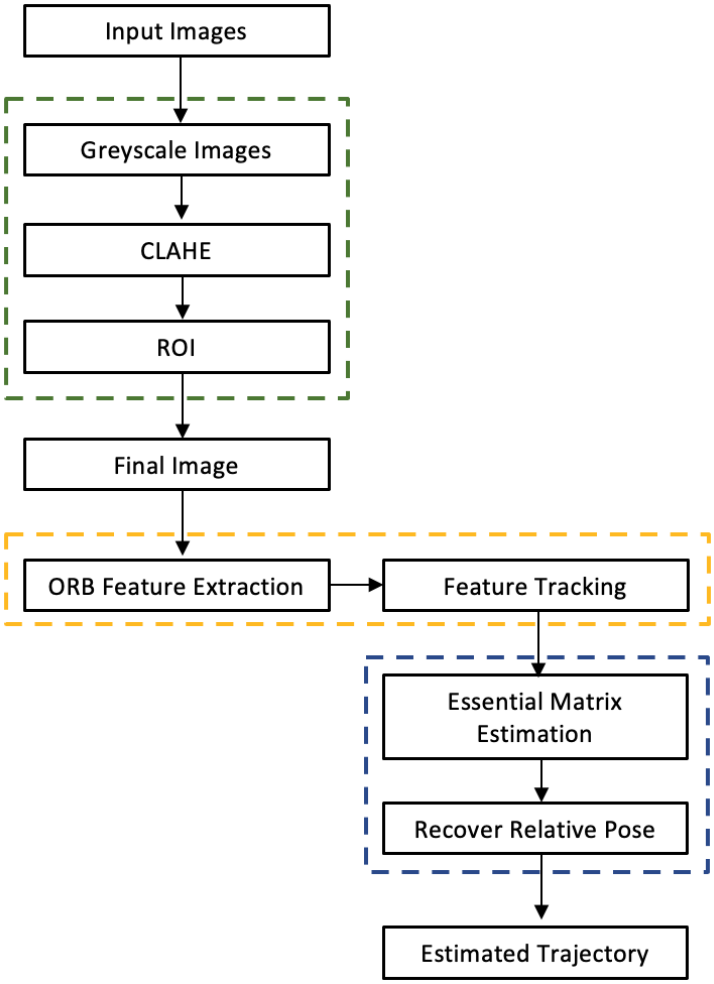

Figure 77: Block diagram of the proposed VO algorithm based on acoustic (FLS) images with CLAHE enhancement technique.

8.3.2 Experimental Results

In the following, the functionality of the proposed acoustic-based odometry system in underwater harbour scenarios is therefore described. First, an analysis of the behaviour of ego-motion using FLS images is performed. Then, the performance of the system is compared with optical data under different visibility conditions. All experiments are based on the extraction of up to 2000 features per image and were performed using an AMD Ryzen 7 2700X @ 3.7Ghz with 16GB RAM computer.

Datasets

For this preliminary assessment of the use of FLS data for AUV localisation to overcome camera problems in underwater environments, parts of the trajectory performed by HarbourMissionFLS dataset are therefore used to evaluate the behaviour of ORB-based odometry on acoustic images, here at a rate of 30 Hz. The AUV follows a trajectory

near the structures and around the concrete wall, i.e. a loop closure comprising 4578 images (for more details see section 5.2.3), as shown in Figure 78. For comparison, the same HarbourMissionCAM sample was used, which is highlighted (S1) and comprises approximately 1500 images.

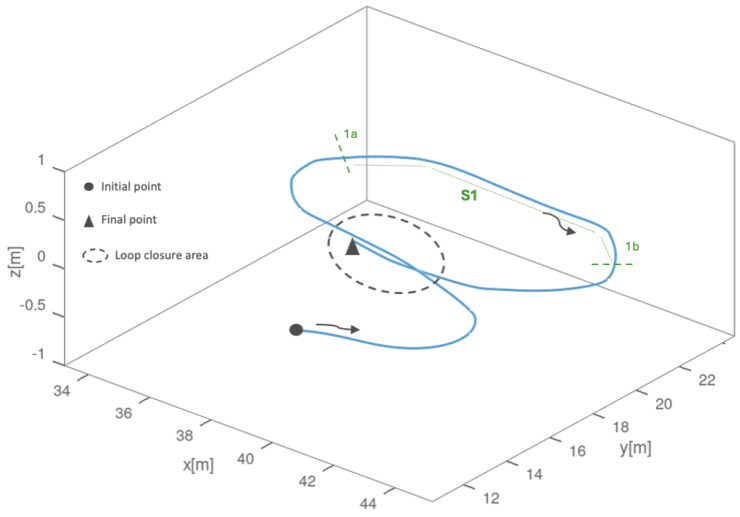

Figure 78: Sample of the trajectory performed by AUV for HarbourMission-FLS dataset, used for acoustic odometry.

Odometry

To assess whether the odometry approach is able to estimate motion from acoustic images under the same conditions as the optical data, an image section (S1) representing an approximately rectilinear motion was considered, see Figure 75 (a). Therefore, the performance of the algorithm is analysed based on the deviation of orientation (from the tangent to the trajectory) between successive positions, as described in equation (12). This allows a quantitative analysis based on the integral of the absolute error between the two curves, according to (13).

Figure 79 thus shows the behaviour of the algorithm. This behaviour was determined on the basis of CL = 5 and NT = [2,3]: Although the track generally has a texture as the robot navigates near a wall, i.e. a paper cup, a grid and some tree trunks, the input images are less representative due to the inherent properties of FLS. This is probably also the reason why the RANSAC requirements for the tracking phase had to be reduced compared to the optical-based approach. As you can see, the positioning algorithm is able to start the motion estimation, although not as smooth as intended due to an area with little texture caused by the presence of sand on the side of a tyre. Nonetheless, the acoustic odometry from frame 30 onwards is able to detect the robot's movement despite some orientation deviations, albeit more sensitive than the behaviour of the camera data in clear water. Perhaps this is

due to the weak texture of the acoustic data, which can provide fewer matching features.

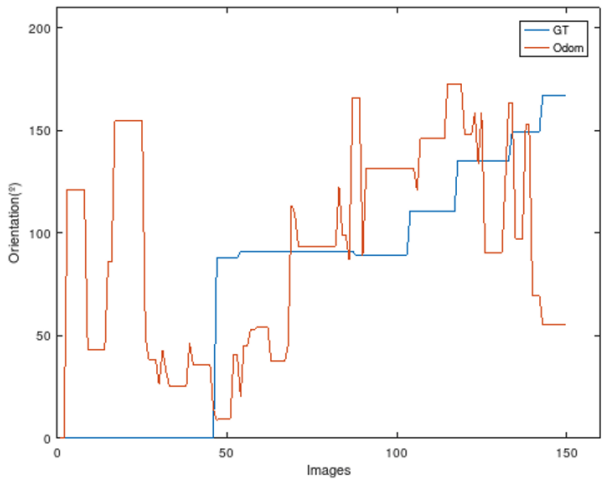

Figure 79: Appearance-based odometry of an inspection sample (S1) for the HarbourMissionFLS dataset at CL = 5 and NT = [2,3].

Nonetheless, the acoustic estimation is more consistent than in clear water because there is only a short period where the algorithm is not able to estimate the position and maintain the previous position, as you can see from the straight lines between frames 90-100, corresponding to the transition from an area with the tyre sidewall to pure sand, which can lead to a temporary inability to track the features. Acoustic odometry performs similarly well compared to the visual algorithm in scenarios with considerable turbidity, although the algorithm with acoustic images can perceive the intended movement earlier. Nevertheless, it is more susceptible in the last part of the trajectory, i.e. it has more difficulties in correctly estimating the changes in perspective. This approach is of great importance compared to visual performance in scenes with high turbidity: FLS images have a weak texture, which may lead to a smaller number of features for the representation of the scene. On the other hand, only the robust features are used effectively, so the algorithm is less prone to perspective changes, even if these are caused by the natural vibrations of the underwater vehicle, and the ego-motion estimation therefore does not assume as many false drifts. The average total error is therefore 50.8°. From frame 30 onwards, the average error drops to 41.1°. As you can see from the qualitative analysis, the optical-based approach starts to become sensitive at considerable turbidity: The acoustic method approaches the optical method under these conditions and is characterised by an error decrease of about 7° at high turbidity.

8.3.3 Discussion

Harbour environments are often affected by turbidity or illuminations that provide images without structural features, which often complicates the behaviour of navigation tasks performed by cameras, even when CLAHE is used. Therefore, feature-based odometry based on acoustic images is proposed in this section to analyse the behaviour of pose estimation. It was then investigated whether these data, even if weakly structured compared to optical images, can help the robot in localisation, especially under strict visibility conditions, as they are not affected by these phenomena. To evaluate whether the odometry approach is able to estimate motion from acoustic images under the same conditions as the optical data, the same image section (S1) was used for the HarbourMissionFLS dataset. The analysis is also based on the deviation of the orientation (from the tangent to the trajectory) between successive positions, which also allows a quantitative assessment based on the integral of the absolute error between the two curves. The behaviour was determined on the basis of CL = 5 and NT = [2,3], as the input images are less representative due to the inherent properties of FLS. For this reason, it was probably also necessary to reduce the RANSAC requirements for the tracking phase based on this data. The initial motion estimation is not as smooth as intended due to an area of low texture caused by the presence of sand on the side of a tyre. Nevertheless, from frame 30 onwards, the robot is able to recognise the movement of the robot despite some orientation deviations. Perhaps this is due to the weak texture of the acoustic data, which can provide less suitable features. Nevertheless, the acoustic estimation is more consistent than that in clear water: there is only a short period of time in which the algorithm cannot estimate the position and maintain the previous position. Compared to the visual algorithm, this approach performs similarly well in the presence of considerable turbidity, although it can recognise the intended movement earlier based on these acoustic images. On the other hand, it has more difficulties in correctly estimating perspective changes. This approach proved to be very important compared to the visual performance obtained in scenes with high turbidity: The weak texture of the acoustic images ensures that only the robust features are used effectively, so the algorithm is probably less susceptible to perspective changes and therefore the ego-motion does not assume as many false deviations. Therefore, the acoustic odometry has achieved an overall average error of about 51^{o}. Based on these preliminary results of the qualitative analysis, the acoustic method shows that it is able to provide positional information based on appearances close to that of the optical method under considerable turbidity conditions (about 1^{o} larger error) and that it has an error reduction of about 7^{o} at high turbidity.

8.4 SUMMARY

The precise localisation of a vehicle is an essential prerequisite for inspection work: For autonomous navigation to be successful, the robot must know its position over time. Given the effectiveness of cameras operating at close range, visual odometry is a robust technique used for this purpose. However, the perception problems typical of harbour environments can make it difficult to determine an accurate vehicle position. The FLS data is a promising solution to help underwater vehicles perceive their surroundings under poor visibility conditions. Therefore, a Visual Odometry (VO) approach for harbour scenarios based on optical and acoustic data is proposed in this chapter.

To gain an initial understanding of the behaviour and limitations of cameras for accurate positioning under different visibility conditions, a monocular feature-based odometry based on optical images was presented. A preliminary evaluation is therefore based on clear water in which the robot moves in a straight direction to analyse the effect of CLAHE. This enhancement method proved to be effective in the perception of the environment. In addition, it was found that with sufficient visibility, using CL = 2 and creating a more detailed histogram, i.e., using NT = [6,4], is not beneficial, so the motion estimates contain some deviations or emphasise slopes, which means that the enhancement technique produces many noisy keypoints. This means that the gain is too high, resulting in poor looking images with excessive contrast or "stretching" of the colour channels. The behaviour of the algorithm when changing direction was then analysed using a path in clear water. The best performance was obtained based on CL = 1 and NT = [3,2], as expected from previous analyses under the same conditions. The images here had to be upscaled to allow the algorithm to recognise all directional changes, perhaps due to the trajectory region being less structured. When searching for orientation deviations, the algorithm approaches ground truth in 60% of the samples and achieves an average total error of about 20^{o}. In poor visibility, the algorithm is able to recognise directional changes, but it is more difficult to recognise the curvilinear movements of the robot. Under these visibility conditions, it was necessary to further improve the images and therefore resort to CL = 2 and NT = [6,4]. Nevertheless, the features are not as robust as the algorithm is more sensitive to the loss of orientation movements shortly after the start of the estimates. Under these conditions, the method increases the overall average error by about 12^{o}.

Next, therefore, the limits of visibility were validated for operation with cameras to estimate vehicle position between successive images under challenging appearance conditions. In a water type

between clear and with very low turbidity, the best performance was determined based on CL = 2 and NT = [3,2], as there were no severe visibility conditions and the tracks are generally structured. Despite some orientation deviations, the algorithm is able to estimate the intended movement. The overall average error is about 29°. The visual positioning method performs better from the last change of direction onwards, achieving an average error of 14° in this case. In the case of considerable turbidity, the algorithm also performs better when based on CL = 2 and NT = [3,2]. The method can follow the intended trajectory, but with more orientation fluctuations, which means that the features are not robust and therefore the perspective changes are less perceived. Therefore, in these scenes with considerable opacity, the visual odometry reaches an average total error of 49°, i.e. an increase of 20°. At high turbidity, CL = 6 and NT = [3,2] are required. However, this extension leads to confusing features that make the algorithm very sensitive when it comes to correctly estimating the perspective changes: the positioning method can only estimate the last 40% of the trajectory. At this level of turbidity, the optical-based odometry therefore reaches an average total error of about 58°, i.e. an increase of about 30° compared to the performance in clear water. The performance of pose estimation therefore decreases with increasing water turbidity, which makes the detection and tracking of robust and distinguishable features more difficult. In particular, the optical estimation is sensitive to the appearance of scenes with considerable turbidity, which in the real context corresponds to water type II in the Jerlov measurements, i.e. the penultimate type of open ocean.

Therefore, feature-based odometry based on acoustic images has also been proposed to analyse the behaviour of pose estimation and to understand whether the robot can navigate with this data, even if it is weakly structured, regardless of the conditions encountered. To test whether the odometry approach is able to estimate motion from acoustic images under the same conditions as the optical data, the same image slice (S1) was used for the HarbourMissionFLS dataset. The behaviour was determined based on CL = 5 and NT = [2,3], as the input images are less representative due to the inherent properties of FLS. In addition, the requirements for the tracking phase had to be reduced based on this data. The initial motion is not estimated to be as smooth as intended, perhaps due to the small texture area. Nevertheless, from frame 30 onwards, the robot is able to recognise the movement of the robot despite some orientation deviations. Compared to the visual algorithm, this approach performs similarly well in the presence of considerable opacity, although it is able to recognise the intended motion earlier based on the acoustic images. On the other hand, it has more difficulties in correctly estimating perspective changes. This approach proved to be very important compared to the visual

performance obtained in scenes with high opacity: The algorithm is probably less susceptible to perspective changes and therefore the ego-motion does not assume as many false deviations. Therefore, the acoustic odometry has achieved an average total error of about 51°. The acoustic method has thus shown that it is able to provide pose information close to that of the optical method under considerable turbidity conditions and is characterised by an error reduction of about 7° compared to the results at high turbidity.

Part III

FINAL CONSIDERATIONS

9 CONCLUSION AND FUTURE WORK

This thesis focuses on the development of a hybrid solution for close-range localisation and loop closure in port scenarios.

The main conclusions are therefore described below according to each topic:

- **Appearance-based PR**

  To validate the effectiveness of BoW algorithms for fast and robust place recognition, a behavioural evaluation of these traditional learning techniques based on binary features was performed for uncontrolled environments with challenging scenes: Appearance changes, perceptual aliasing conditions or dynamic elements. It was found that a more representative visual vocabulary contributes to the robustness of the loop closure algorithm the more challenging an environment is. The BoW techniques show that they are able to handle challenging conditions, albeit at the expense of the number of extracted features.

  Nevertheless, the results have shown that the required computational effort is lower and that this approach can be used in a real application. As the outdoor environment is also unknown and can be as dynamic as underwater, it is assumed that this technique can also cope with the challenges of the underwater environment and was therefore applied accordingly to this data.

- **Underwater Experimental Setup**

  Inspection tasks require close-range navigation, which reduces the number of sensors that can be used due to limited perception. Despite technological developments that reduce costs, the collection of data on the marine environment is still a very costly, time-consuming and technically demanding activity. Therefore, there is a gap in available data for the intended context. Nevertheless, it is important to validate the algorithms before applying them in a real context. In this respect, environmental simulations are a powerful tool for initial validation and pre-testing. They allow the variation of environmental or vehicle parameters, sensor configurations, different operating scenarios, etc.

  Therefore, the Stonefish simulator was used to develop some datasets for the preliminary evaluation of the planned proposal. As real data is crucial to validate the efficiency, perception solutions were selected depending on the context, namely a camera,

a laser, a FLS and a positioning system as ground truth. The acquisition system is mounted on a ROV to collect data in real scenarios in the harbour.

- **Visual PR**

  A preliminary assessment of the behaviour of a BoW technique based on ORB features was performed for two port scenarios. For PortSeafloorA data, the changes in the scene did not affect the overall performance too much as the loop closure situation was always recognised and 77.78% and 50% of recall was achieved. For PortSeafloorB, the performance of the algorithm decreases with increasing turbidity. At very low turbidity, the recall rate drops by 20% as the algorithm is unable to perform 7 detections. At low turbidity, the algorithm does not recognise the first loop closure situation, which can be dangerous in a navigation context. Experiments with solar tilt showed poor performance in place recognition for extreme values. The first case can illustrate a depth operation where a darker environment produces scenes without intensity differences, i.e. recognisable features. In contrast, shallow waters, i.e. bright scenes, can produce unstructured images and shadows, which also complicates feature extraction.

  The system has thus been shown to be sensitive to some visual conditions that are common in uncontrolled and difficult scenarios.

- **CLAHE for Visual PR**

  A purely visual approach to place recognition has been proposed, which includes a pre-processing step to fully exploit the advantages of these sensors, which operate effectively at close range. Thus, the CLAHE technique was chosen, which proved to be effective by increasing the inliers between images of the same place to a minimum of about 60%. In terms of similarity detection performance, all loop closures recognised by the algorithm under both visibility conditions were correctly identified.

  A performance of 85.71% and 80% was achieved in scenarios with very low turbidity and light scenes, respectively. The proposed VPR algorithm thus shows good performance at low computational cost.

- **Limit characterisation for Visual PR**

  A boundary characterisation of water types where cameras can be used to accurately detect loop closures was presented to understand how far cameras can go due to their effective operation at close range. At considerable turbidity, the algorithm performs

well and is able to recognise all situations of loop closure without failures and without false detection of loops: 100% recall is achieved at 100% precision: TP = 14, FP = 0 and FN = 0. At high turbidity, the loss of visibility is noticeable: the algorithm can achieve a recall of 71.43% at 100% of precision under these conditions: TP =10, FP = 0 and FN = 4. The very high turbidity scenario is also tested and under this stringent condition the algorithm achieves a recall of 53.85% at 100% of precision: TP = 7, FP = 0 and FN = 6.

Therefore, it is crucial to evaluate other place detection solutions that help in cases where the camera is already starting to fail to recognise distinguishable features so that the underwater vehicles can perform accurate inspection missions.

- **Acoustic PR**

  A purely acoustic method of place recognition was presented that can overcome the inherent limitations of perception in harbour scenarios. In general, sonar data provides fewer features than camera images under normal visibility conditions. However, as the visual images become more challenging, the performance of the FLS remains the same, while that of the optical sensor decreases, reaching 18 points, of which the FLS can recognise 41 in low light conditions. The proposed algorithm does not falsely recognise situations in which a loop closes, i.e. it achieves a precision of 100%. However, the strength of the algorithm is low as it tends to lose loops along the trajectory and achieving a recall of 37.5%. Therefore, the effect of CLAHE was evaluated to understand whether the enhancement provides a better description of the images for acoustic image-based place recognition.

  The proposed approach based on FLS can achieve a recall of 87.5%, which is an increase of 50% over the original result. The enhancement technique thus improves the behaviour of the algorithm, i.e. it increases the contrast and details of the image (without noisy pixels), and more effective features are detected. This behaviour indicates that the FLS supports loop detection in poor visibility when the camera no longer provides detailed information.

- **Visual Odometry**

  A visual odometry method based on optical and acoustic images is proposed to enable vehicle navigation independent of environmental conditions. Therefore, the limits of visibility were first validated for operation with cameras to estimate the vehicle position between successive images under challenging visibility conditions. In a water type between clear and with very low

turbidity, the algorithm has proven to estimate the intended movement despite some orientation deviations. The average total error is 28.7°. It performs better from the last change of direction, achieving an average error of 14.3°. In scenes with considerable turbidity, the method can follow the intended trajectory, but with more orientation deviations, which means that the features are not robust and therefore the perspective changes are perceived less. In these scenes, the visual odometry reaches an average total error of 49°, i.e. an increase of 20° compared to the previous case. A higher CLAHE was required for scenes with high turbidity, but this increase also leads to confusing features that make the algorithm very sensitive: The positioning method can only estimate the last 40% of the trajectory. Therefore, visual odometry reaches an average total error of about 57.8°, i.e. an increase of about 30° compared to the performance in clear water.

The visual odometry algorithm is sensitive to changes in appearance, making it difficult to detect and track robust and distinguishable features.

Therefore, a feature-based odometry based on acoustic images was also proposed, based on the same assumptions as the visual data. These data required a higher CLAHE value compared to the visual data, as expected given the less representative images due to the inherent characteristics of FLS. The requirements for the tracking phase therefore had to be reduced. The initial motion is not estimated as smoothly as intended, but after this start (about 6% of the trajectory) the robot is able to recognise the robot's motion despite some orientation deviations. This approach performs similarly well in the presence of considerable turbidity, although it is able to recognise the intended motion earlier from the acoustic images. However, it has more difficulties in correctly estimating changes in perspective. This approach proved to be very important compared to visual performance in scenes with high turbidity. Thus, this approach achieved an average total error of 50.8°.

Thus, the acoustic method has shown that it is able to provide position information based on appearance information. Furthermore, it approximates the optical method under considerable turbidity conditions and is characterised by an error reduction of 7° compared to the results at high turbidity (57.8° to 50.8°).

Future Work

Despite the relevance of the work done in this thesis and the results obtained, there are still some important aspects to contribute for a relevant scientific work and which form the core objectives we are pursuing in our ongoing and future work.

The relevant aspects are listed below:

- **New data campaigns.** It is intended to conduct new real data campaigns that include different types of trajectories to inspect structures in order to increase the robustness of the developed methods.

- **Improvement of the odometry algorithm for preemptive localisation.** The idea is to incorporate a pre-processing step in optical images so that the necessary configuration of the camera, i.e. with a tilt angle that allows the robot to navigate preemptively is not a problem for feature-based motion estimation. In this sense, we intend to evaluate the feasibility of transformation between different image planes.

- **Improving the odometry algorithm for accurate localisation.** Given the large influence of loop closure detection in correcting the accumulated drifts on the position of the vehicle, it is considered to include this mechanism in the odometry approach to enable accurate relative short-range localisation. In this way, the vehicle will also know when the inspection of a particular structure has been completed.

- **Evaluating the impact of appearance changes for accurate navigation.** It is important to understand whether the change in appearance of a place during navigation due to various causes, such as changes in illumination, weather or turbidity, will result in mismatched features and thus affect pose estimation and loop detection. Should this be the case, a suitable solution to improve the methods for pose estimation or place recognition will be presented and evaluated to enable reliable navigation for autonomous vehicles.

- **Adaptive perception navigation solution.** Considering the effectiveness of the cameras at close range and the fact that the acoustic sensors do not suffer from these clouding effects, it will be interesting that the autonomous robotic vehicles can activate the most appropriate sensor during an inspection mission, depending on the conditions encountered, i.e. visibility or texture. Therefore, some variable thresholds should be defined to indicate to the vehicle the appropriate switching of sensors.

- **Evaluating the feasibility of a laser-based structured light system.** Given the possibilities offered by these systems in 3D reconstruction, the intention is to transfer these solutions to the navigation context to understand whether they can help robots to perceive the environment in scenarios with less texture or repetitive patterns. It will also be investigated to what extent we can utilise this additional information in the image. *"Does it fit within the operating limits of the camera?" "Or is this sensor even less tolerable?"*

Part IV

APPENDIX

A PLACE RECOGNITION IN REAL ENVIRONMENTS

Some experiments in underwater scenes were considered to analyse the robustness of the similarity detection algorithm under difficult conditions. Based on the optical data described in section 5.3.2 - PortAzoresCAM dataset - parallel trajectories were performed over two floating piers labelled A and B, with the return to the starting point, comprising 623 and 212 images, respectively.

Figure 80 shows the inspection structures.

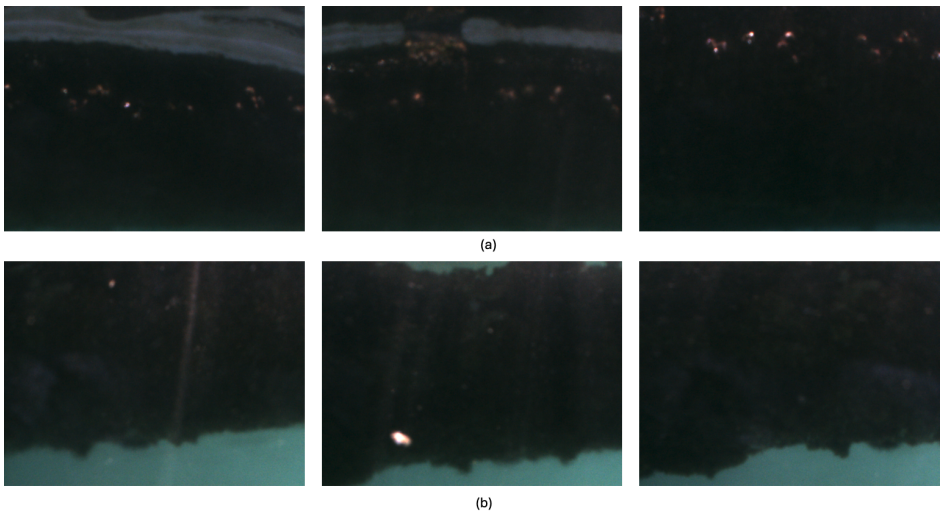

Figure 80: Illustrative example of recorded scenes of the floating piers A (a) and B (b) in the harbour of Ponta Delgada, Azores.

Since there is no position information, a novel metric for real loop closure decision was considered to evaluate the performance of the algorithm under real conditions in the same way. Therefore, there is a high proportion of loop closures corresponding to traversing a larger number of frames, which implies a slower movement of the robot. For this purpose, a failure rate was determined and 457 and 111 real loops were obtained for the trajectory in the floating piers A and B, respectively.

Table 31 and Table 32 show the behaviour of the algorithm in detecting previously visited places for each trajectory about floating piers. Based on raw data, there was a high failure rate (case A) or no detected similarity (case B). Using CLAHE based on CL = 2, chosen by a comparision between BRISQUE, NIQE and PIQUE in a sequence

of 20 images, in the first case some confusable features are created: FP increases and so, the precision decreased 5%. In the second case, although there is a little increase in performance, it is not possible to operate with feasibility (TP = 4 and FP = 14). Given the ambiguity of the scenes (perceptual aliasing) and the severe visibility conditions, such as turbidity, suspended particles and the effects of the sunlight, this behaviour was expected.

Table 31: Performance of the VPR algorithm for the PortAzoresCAM dataset - floating pier A - with CLAHE at NT = [4,3].

		CL = 2
Failure rate (%)	32	37
Precision (%)	68	63
Recall (%)	16	38

Table 32: Performance of the VPR algorithm for the PortAzoresCAM dataset - floating pier B - with CLAHE at NT = [4,3].

		CL = 2
Failure rate (%)	-	78
Precision (%)	-	22
Recall (%)	-	4

In addition, a trajectory was performed near a pile but with a different viewpoint and distance to the structure in the second pass. The decision metric for loop closure was retained so that 71 real loop situations had to be obtained. However, as expected, lower performance was achieved, even when using the enhancement technique, see Table 33, as the structure's appearance is very different in the two-time acquisitions, as visible in Figure 81.

This example is intended to show the importance of autonomous navigation to ensure the consistency of the inspection missions.

Table 33: Performance of the VPR algorithm for the PortAzoresCAM dataset - pile - with CLAHE at NT = [4,3].

		CL = 2
Failure rate (%)	71	83
Precision (%)	29	17
Recall (%)	3	4

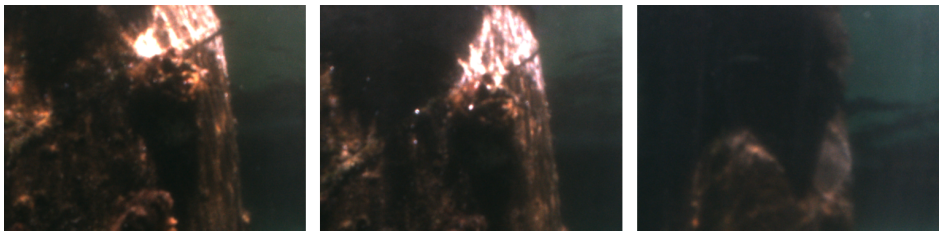

Figure 81: Illustrative example of recorded scenes of a pile in the harbour of Ponta Delgada, Azores.

Based on the acoustic data provided by SoundMetrics, as described in section 3.3, a sequence of 186 FLS images of concrete blocks with the return to the initial point was selected, as visible in Figure 82. An ROI with 250x650 pixels was considered to extract features in the effective area mapped by the sonar. The metric for true loop closure decision was retained, and there are 94 similarity situations.

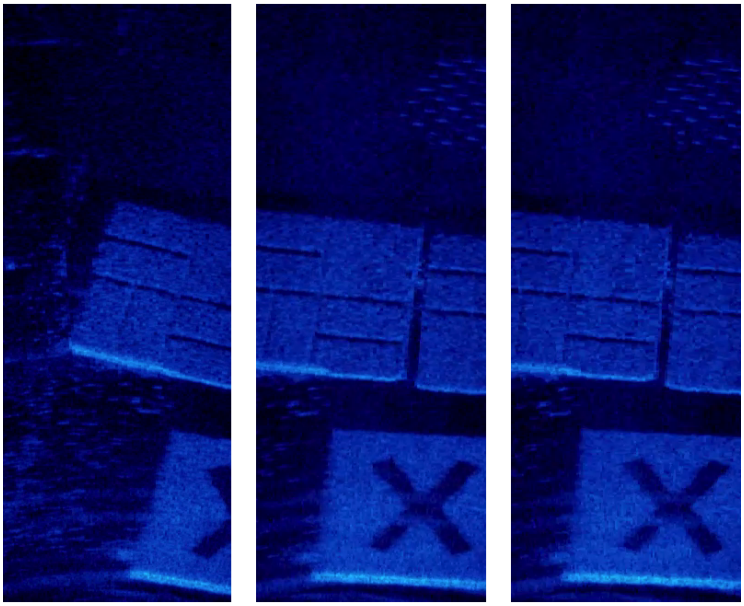

Figure 82: Illustrative example of recorded acoustic images of concrete blocks, recorded by SoundMetrics.

As visible in Table 34, the results are generally good by the obtained failure rate measurements.

Table 34: Performance of the VPR algorithm for the SoundMetrics dataset - concrete blocks - with CLAHE at NT = [5,13].

		CL = 1	CL = 2
Failure rate (%)	17	12	16
Precision (%)	83	88	84
Recall (%)	16	16	17

Highlights the performance obtained resorting to CLAHE based on CL = 1 (FP = 2). The increase of the image contrast (CL = 2) degradates the image, so more features don't mean better quality: noisy pixels were generated, affecting the algorithm's robustness to recognise a place without error (the precision decreases 4%).

B | POSE ESTIMATION IN REAL ENVIRONMENTS

Some experiments in underwater scenes were also considered to analyse the robustness of the relative orientation algorithm under difficult conditions. Based on the optical data described in section 5.3.2 - PortAzoresCAM dataset - the data of the floating pier A were considered, but here only the first movement from left to right was performed, consisting of 106 optical frames. So the robot performs a linear movement parallel to the structure, but of course there are some aspects that can influence this movement, such as currents or the use of the human-controlled cable.

The Figure 84 shows the behaviour of the algorithm with respect to the variation of the orientations for raw (a) and CLAHE (b) data.

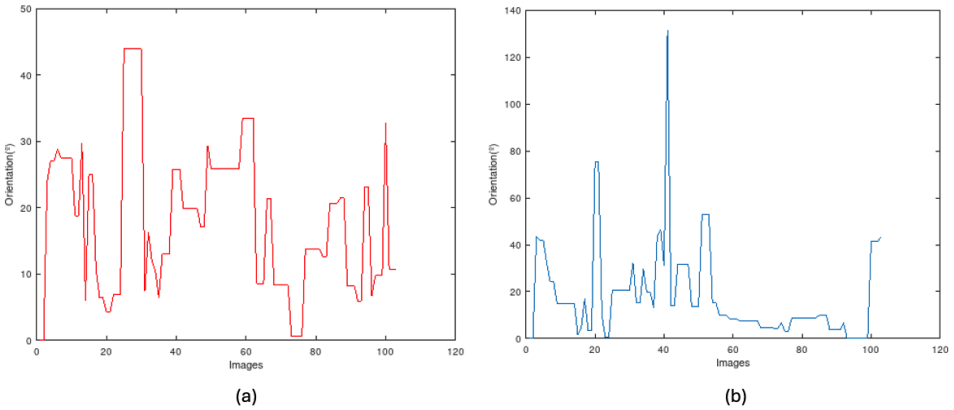

Figure 83: Appearance-based odometry of an inspection sample (Foating pier A) for the PortAzoresCAM dataset at CL = 1 and NT = [4,3].

The arithmetic means of the orientation variation is 17.9° and 16.7° for each case, respectively. Thus, it was evident that the features were not very robust as they required new recurrent extractions. The enhancement technique improves the consistency of the estimation. Nevertheless, an increase of about 1° was achieved, perhaps because we are already facing unfavourable conditions and therefore highlighting some already "noisy" pixels. In this case, the approach requires 244.2 ms for each image.

Based on the acoustic data provided by SoundMetrics, as described in section 3.3, a sequence of 92 FLS images of concrete blocks moving from left to right is considered here. As can be seen in Figure (a), the

presence of stable features has been demonstrated and a practically rectilinear motion with an arithmetic mean of orientation variations of 10^{o} has been obtained as intended. The approach requires 63.6 ms for each image. However, in this case, the application of the enhancement technique adds noisy pixels (see Figure (b)), extracting even more features - bottom graph - and decreasing the performance by up to 20^{o}.

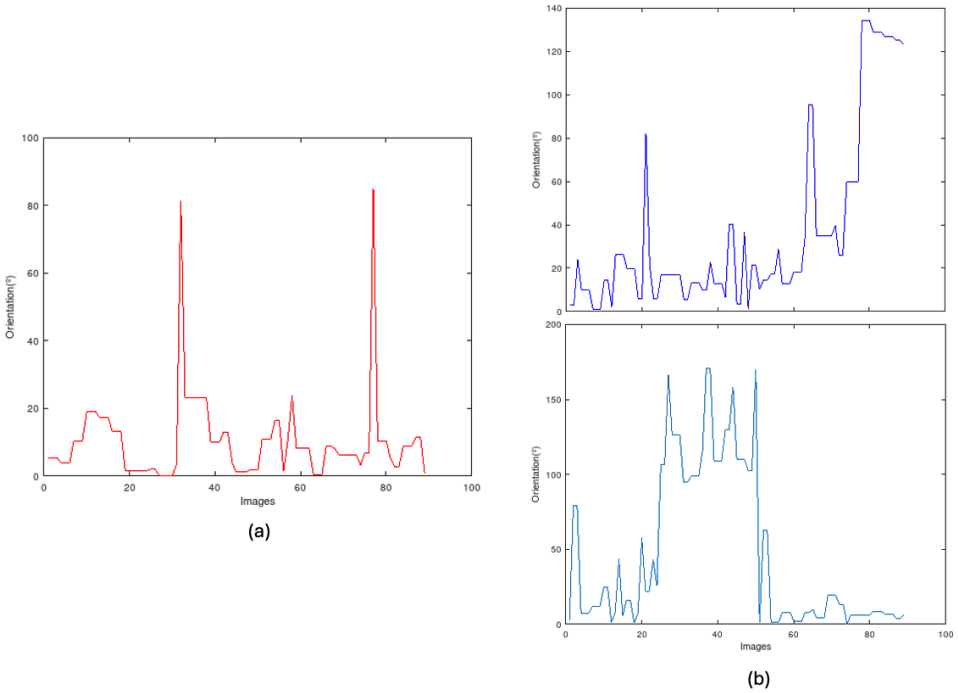

Figure 84: Appearance-based odometry of an inspection sample (concrete blocks) for the SoundMetrics data for raw (a) and using CLAHE (b) data at CL = 1 and NT = [5,13].

BIBLIOGRAPHY

[1] J. Melo and A. Matos. «Survey on advances on terrain based navigation for autonomous underwater vehicles». In: *Ocean Engineering* 139 (2017), pp. 250–264. ISSN: 00298018. DOI: 10.1016/j.oceaneng.2017.04.047 (cit. on p. 12).

[2] X. Wang, X. Fan, P. Shi, J. Ni, and Z. Zhou. «An Overview of Key SLAM Technologies for Underwater Scenes». In: *Remote Sensing* 15.10 (2023). ISSN: 2072-4292. DOI: 10.3390/rs15102496 (cit. on pp. 12, 17, 42).

[3] H. Lu, Y. Li, Y. Zhang, M. Chen, S. Serikawa, and H. Kim. «Underwater Optical Image Processing: a Comprehensive Review». In: *Mobile Networks and Applications* 22 (2017), pp. 1204–1211. ISSN: 15728153. DOI: 10.1007/s11036-017-0863-4. arXiv: 1702.03600 (cit. on p. 13).

[4] J. Aulinas, Y. Petillot, J. Salvi, and X. Lladó. «The SLAM problem: A survey». In: *Frontiers in Artificial Intelligence and Applications*. Vol. 184. 1. 2008, pp. 363–371. ISBN: 9781586039257. DOI: 10.3233/978-1-58603-925-7-363 (cit. on p. 13).

[5] M. Massot-Campos and G. Oliver-Codina. «Optical Sensors and Methods for Underwater 3D Reconstruction». In: *Sensors* 15.12 (2015), pp. 31525–31557. ISSN: 1424-8220. DOI: 10.3390/s151229864 (cit. on pp. 14, 36).

[6] F. Bruno, G. Bianco, M. Muzzupappa, S. Barone, and A. V. Razionale. «Experimentation of structured light and stereo vision for underwater 3D reconstruction». In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.4 (2011), pp. 508–518. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2011.02.009 (cit. on pp. 15, 36).

[7] M. Massot-Campos and G. Oliver-Codina. «Underwater Laser-based Structured Light System for one-shot 3D reconstruction». In: *SENSORS, 2014 IEEE*. 2014, pp. 1138–1141. DOI: 10.1109/ICSENS.2014.6985208 (cit. on pp. 18, 19, 21, 36).

[8] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha. «Visual simultaneous localization and mapping: a survey». In: *Artificial Intelligence Review* 43 (2015), pp. 55–81. DOI: 10.1007/s10462-012-9365-8 (cit. on pp. 20, 26, 27).

[9] F. Hidalgo and T. Braunl. «Review of Underwater SLAM techniques». In: *Conference on Automation, Robotics and Applications (ICARA)*. 2015, pp. 306–311. ISBN: 9781479964666. DOI: 10.1109/ICARA.2015.7081165 (cit. on p. 20).

[10] A. Palomer. «3D Underwater SLAM Using Sonar and Laser Sensors». PhD thesis. 2018. URL: https://www.tdx.cat/handle/10803/665536 (cit. on p. 20).

[11] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. «Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age». In: *IEEE Transactions on Robotics* 32.6 (2016), pp. 1309–1332. ISSN: 15523098. DOI: 10.1109/TRO.2016.2624754. arXiv: 1606.05830 (cit. on pp. 21, 27, 31).

[12] D. Scaramuzza and F. Fraundorfer. «Visual Odometry Part I: The First 30 Years and Fundamentals». In: (cit. on p. 21).

[13] S. Zhang, S. Zhao, D. An, J. Liu, H. Wang, Y. Feng, D. Li, and R. Zhao. «Visual SLAM for underwater vehicles: A survey». In: *Computer Science Review* 46 (2022), p. 100510. ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2022.100510 (cit. on p. 21).

[14] M. O. A. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail. «Review of visual odometry: types, approaches, challenges, and applications». In: *SpringerPlus* 5 (2016) (cit. on p. 22).

[15] M. Ferrera, J. Moras, P. Trouvé-Peloux, and V. Creuze. «Real-time monocular visual odometry for turbid and dynamic underwater environments». In: *Sensors* 19.3 (2019), pp. 1–20. ISSN: 14248220. DOI: 10.3390/s19030687. arXiv: 1806.05842 (cit. on pp. 23, 35).

[16] E. Pepperell, P. Corke, and M. Milford. «Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints». In: *International Journal of Robotics Research* 35.9 (2016), pp. 1057–1179. ISSN: 17413176. DOI: 10.1177/0278364915618766 (cit. on p. 24).

[17] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. «Visual Place Recognition: A Survey». In: *IEEE Transactions on Robotics* 32.1 (2016), pp. 1–19. ISSN: 15523098. DOI: 10.1109/TRO.2015.2496823. arXiv: 1612.00593 (cit. on pp. 25, 27).

[18] M. T. Law, N. Thome, and M. Cord. «Bag-of-Words Image Representation: Key Ideas and Further Insight». In: *Fusion in Computer Vision - Advances in Computer Vision and Pattern Recognition*. Springer International, 2014. Chap. 2, pp. 29–52. DOI: 10.1007/978-3-319-05696-8_2 (cit. on pp. 25, 27, 30, 41).

[19] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós. «An image-to-map loop closing method for monocular SLAM». In: *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*. 2008, pp. 2053–2059. ISBN: 9781424420582. DOI: 10.1109/IROS.2008.4650996 (cit. on p. 26).

[20] N. Kejriwal, S. Kumar, and T. Shibata. «High performance loop closure detection using bag of word pairs». In: *Robotics and Autonomous Systems* 77 (2016), pp. 55–65. ISSN: 09218890. DOI: 10.1016/j.robot.2015.12.003 (cit. on pp. 27, 30).

[21] S. Sharma, V. Gupta, and M. Juneja. «A Survey of Image Data Indexing Techniques». In: *Artif. Intell. Rev.* 52.2 (2019), 1189–1266. ISSN: 0269-2821. DOI: 10.1007/s10462-018-9673-8 (cit. on p. 28).

[22] S. Kulkarni, S. Kulkarni, D. Bormane, and S. Nalbalwar. «RANSAC Algorithm for Matching Inlier Correspondences in Video Stabilization». In: *European Journal of Applied Sciences* 5.1 (2017), pp. 1–30. DOI: 10.14738/aivp.51.2692 (cit. on p. 30).

[23] N. R. Gracias, S. Van Der Zwaan, A. Bernardino, and J. Santos-Victor. «Mosaic-based Navigation for Autonomous Underwater Vehicles». In: *IEEE Journal of Oceanic Engineering* 28.4 (2003), pp. 609–624. ISSN: 03649059. DOI: 10.1109/JOE.2003.819156 (cit. on p. 34).

[24] J. Aulinas, M. Carreras, X. Llado, J. Salvi, R. Garcia, R. Prados, and Y. R. Petillot. «Feature extraction for underwater visual SLAM». In: *OCEANS 2011 IEEE - Spain*. 2011, pp. 1–7. DOI: 10.1109/Oceans-Spain.2011.6003474 (cit. on p. 34).

[25] R. Pérez-Alcocer, L. A. Torres-Méndez, E. Olguín-Díaz, and A. A. Maldonado-Ramírez. «Vision-Based Autonomous Underwater Vehicle Navigation in Poor Visibility Conditions Using a Model-Free Robust Control». In: *Journal of Sensors* 2016 (2016), pp. 1–16. DOI: 10.1155/2016/8594096 (cit. on p. 34).

[26] M. M. Nawaf, D. Merad, J.-P. Royer, J.-M. Boï, M. Saccone, M. Ben Ellefi, and P. Drap. «Fast Visual Odometry for a Low-Cost Underwater Embedded Stereo System †». In: *Sensors* 18.7 (2018). ISSN: 1424-8220. DOI: 10.3390/s18072313 (cit. on p. 35).

[27] F. Bo Miquel Nordfeldt, F. Bonin-Font, C. Gabriel Oliver, and C. Yolanda Gonzalez. «Evolving Visual Odometry for Autonomous Underwater Vehicles». In: *IFAC-PapersOnLine* 55.31 (2022). 14th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2022, pp. 381–386. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2022.10.458 (cit. on p. 35).

[28] M. Massot-Campos, G. Oliver-Codina, H. Kemal, Y. Petillot, and F. Bonin-Font. «Structured light and stereo vision for underwater 3D reconstruction». In: *OCEANS 2015 - Genova*. 2015, pp. 1–6. DOI: 10.1109/OCEANS-Genova.2015.7271433 (cit. on p. 36).

[29] M. Massot-Campos, G. Oliver, A. Bodenmann, and B. Thornton. «Submap bathymetric SLAM using structured light in underwater environments». In: *2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*. 2016, pp. 181–188. DOI: 10.1109/AUV.2016.7778669 (cit. on p. 37).

[30] A. Palomer, P. Ridao, and D. Romagós. «Inspection of an underwater structure using point-cloud SLAM with an AUV and a laser scanner». In: *Journal of Field Robotics* 36 (Sept. 2019). DOI: 10.1002/rob.21907 (cit. on p. 37).

[31] J. Li. «Place Recognition and Localization for Multi-Modal Underwater Navigation with Vision and Acoustic Sensors». PhD thesis. 2017, pp. 1–122. ISBN: 9780355656541. URL: https://search.proquest.com/docview/2019923453?accountid=49007{\%}0Ahttp://www.yidu.edu.cn/educhina/educhina.do?artifact={\&}svalue=Place+Recognition+and+Localization+for+Multi-Modal+Underwater+Navigation+with+Vision+and+Acoustic+Sensors{\&}stype=2{\&}s=on{\%}0Ahttp://s (cit. on p. 37).

[32] F. Mandic, I. Rendulic, N. Miković, and D. Nad. «Underwater Object Tracking Using Sonar and USBL Measurements». In: *J. Sensors* 2016 (2016), 8070286:1–8070286:10 (cit. on p. 38).

[33] K. Siantidis. «Side scan sonar based onboard SLAM system for autonomous underwater vehicles». In: *IEEE/OES Autonomous Underwater Vehicles (AUV)*. 2016, pp. 195–200. ISBN: 9781509024421. DOI: 10.1109/AUV.2016.7778671 (cit. on p. 38).

[34] M. Franchi, A. Ridolfi, and L. Zacchini. «2D Forward Looking SONAR in Underwater Navigation Aiding: an AUKF-based strategy for AUVs». In: *IFAC-PapersOnLine* 53.2 (2020). 21st IFAC World Congress, pp. 14570–14575. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2020.12.1463 (cit. on p. 38).

[35] S. Rahman, A. Q. Li, and I. Rekleitis. «Sonar Visual Inertial SLAM of Underwater Structures». In: *IEEE International Conference on Robotics and Automation*. 2018, pp. 5190–5196. ISBN: 9781538630815. DOI: 10.1109/ICRA.2018.8460545 (cit. on p. 39).

[36] S. A. K. Tareen and Z. Saleem. «A Comparative Analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK». In: *International Conference on Computing, Mathematics and Engineering Technologies*. 2018. DOI: 10.1109/ICOMET.2018.8346440 (cit. on pp. 40, 51, 57).

[37] F. Hidalgo and T. Bräunl. «Evaluation of Several Feature Detectors/Extractors on Underwater Images towards vSLAM». In: *Sensors* 20.15 (2020). ISSN: 1424-8220. DOI: 10.3390/s20154343 (cit. on pp. 40, 43, 89).

[38] D. Gálvez-López and J. D. Tardós. «Real-Time Loop Detection with Bags of Binary Words». In: *International Conference on Intelligent Robots and Systems - San Francisco*. 2011, pp. 51–58. ISBN: 9781612844541. DOI: 10.1109/IROS.2011.6048525 (cit. on p. 40).

[39] R. Mur-Artal and J. D. Tardós. «Fast relocalisation and loop closing in keyframe-based SLAM». In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 846–853. DOI: 10.1109/ICRA.2014.6906953 (cit. on p. 40).

[40] T. Nicosevici and R. Garcia. «Automatic Visual Bag-of-Words for Online Robot Navigation and Mapping». In: *IEEE Transactions on Robotics* 28.4 (2012), pp. 886–898. ISSN: 15523098. DOI: 10.1109/TRO.2012.2192013 (cit. on p. 40).

[41] N. Kejriwal, S. Kumar, and T. Shibata. «High performance loop closure detection using bag of word pairs». In: *Robotics and Autonomous Systems* 77 (2016), pp. 55–65. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2015.12.003 (cit. on p. 41).

[42] P. L. Negre, F. Bonin-Font, and G. Oliver. «Cluster-Based Loop Closing Detection for Underwater Slam in Feature-Poor Regions». In: *IEEE International Conference on Robotics and Automation*. IEEE, 2016, pp. 2589–2595. ISBN: 9781467380263. DOI: 10.1109/ICRA.2016.7487416 (cit. on p. 41).

[43] E. Garcia-Fidalgo and A. Ortiz. «Hierarchical Place Recognition for Topological Mapping». In: *IEEE Transactions on Robotics* 33.5 (2017), pp. 1061–1074. DOI: 10.1109/TRO.2017.2704598 (cit. on p. 41).

[44] E. Garcia-Fidalgo and A. Ortiz. «iBoW-LCD: An Appearance-Based Loop-Closure Detection Approach Using Incremental Bags of Binary Words». In: *IEEE Robotics and Automation Letters* 3 (2018), pp. 3051–3057 (cit. on p. 41).

[45] A. Teran Espinoza. «Acoustic-Inertial Forward-Scan Sonar Simultaneous Localization and Mapping». MA thesis. KTH, School of Electrical Engineering and Computer Science (EECS), 2020, p. 58 (cit. on p. 42).

[46] N. H. Vilarnau. «Forward-looking sonar mosaicing for underwater environments». In: 2014 (cit. on p. 42).

[47] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson. «Pose-Graph SLAM Using Forward-Looking Sonar». In: *IEEE Robotics and Automation Letters* 3.3 (2018), pp. 2330–2337. DOI: 10.1109/LRA.2018.2809510 (cit. on p. 42).

[48] M. M. Santos, G. B. Zaffari, P. O. Ribeiro, P. L. Drews-Jr, and S. S. Botelho. «Underwater place recognition using forward-looking sonar images: A topological approach». In: *Journal of Field Robotics* 36.2 (2019), pp. 355–369 (cit. on p. 42).

[49] D. Neupane and J. Seok. «A Review on Deep Learning-Based Approaches for Automatic Sonar Target Recognition». In: *Electronics* 9.11 (2020). ISSN: 2079-9292. DOI: 10.3390/electronics9111972 (cit. on p. 43).

[50] S. Lee, I. Seo, J. Seok, Y. Kim, and D. S. Han. «Active Sonar Target Classification with Power-Normalized Cepstral Coefficients and Convolutional Neural Network». In: *Applied Sciences* 10.23 (2020). ISSN: 2076-3417. DOI: 10.3390/app10238450 (cit. on p. 43).

[51] X. Zhou, S. Yuan, C. Yu, H. Li, and X. Yuan. «Performance Comparison of Feature Detectors on Various Layers of Underwater Acoustic Imagery». In: *Journal of Marine Science and Engineering* 10.11 (2022). ISSN: 2077-1312. DOI: 10.3390/jmse10111601 (cit. on p. 43).

[52] A. C. Duarte, G. B. Zaffari, R. T. S. da Rosa, L. M. Longaray, P. Drews, and S. S. C. Botelho. «Towards comparison of underwater SLAM methods: An open dataset collection». In: *OCEANS 2016 MTS/IEEE Monterey*. 2016, pp. 1–5. DOI: 10.1109/OCEANS.2016.7761315 (cit. on p. 44).

[53] A. Mallios, E. Vidal Garcia, R. Campos, and M. Carreras. «Underwater caves sonar data set». In: *The International Journal of Robotics Research* 36 (Oct. 2017), p. 027836491773283. DOI: 10.1177/0278364917732838 (cit. on p. 45).

[54] M. Ferrera, J. Moras, P. Trouvé-Peloux, V. Creuze, and D. Dégez. *The Aqualoc Dataset: Towards Real-Time Underwater Localization from a Visual-Inertial-Pressure Acquisition System*. 2018. arXiv: 1809.07076 [cs.RO] (cit. on p. 46).

[55] B. Joshi, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, A. Q. Li, N. Vitzilaios, and I. Rekleitis. «Experimental Comparison of Open Source Visual-Inertial-Based State Estimation Algorithms in the Underwater Domain». In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019, pp. 7227–7233. DOI: 10.1109/IROS40897.2019.8968049 (cit. on p. 46).

[56] D. Arthur and S. Vassilvitskii. «k-means++: The Advantages of Careful Seeding». In: *SODA '07: Proceedings of the eighteenth annual ACMSIAM symposium on Discrete algorithms*. 2007, pp. 1027–1035 (cit. on p. 53).

[57] H. Chatoux, F. Lecellier, C. Fernandez-maloigne, H. Chatoux, F. Lecellier, C. F.-m. Comparative, and H. Chatoux. «Comparative Study of Descriptors with Dense Key points». In: *23rd International Conference on Pattern Recognition*. 2016 (cit. on p. 57).

[58] A. Geiger, P. Lenz, and R. Urtasun. «Are we ready for autonomous driving? The KITTI Vision Benchmark Suite». In: *Conference on Computer Vision and Pattern Recognition*. 2012. DOI: 10.1109/CVPR.2012.6248074 (cit. on p. 57).

[59] M. Warren, D. McKinnon, H. He, and B. Upcroft. «Unaided Stereo Vision Based Pose Estimation». In: *Australasian Conference on Robotics and Automation*. 2010. ISBN: 9780980740417 (cit. on p. 57).

[60] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. «The New College Vision and Laser Data Set». In: *International Journal of Robotics Research* 28.5 (2009), pp. 595–599. ISSN: 02783649. DOI: 10.1177/0278364909103911 (cit. on p. 57).

[61] P. Ciéslak. «Stonefish: An Advanced Open-Source Simulation Tool Designed for Marine Robotics, With a ROS Interface». In: *OCEANS 2019 - Marseille*. 2019, pp. 1–6. DOI: 10.1109/OCEANSE.2019.8867434 (cit. on p. 66).

[62] C. A. Williamson and R. C. Hollins. «Measured IOPs of Jerlov water types». In: *Appl. Opt.* 61.33 (2022), pp. 9951–9961. DOI: 10.1364/AO.470464 (cit. on p. 68).

[63] S. Schubert and P. Neubert. *What makes visual place recognition easy or hard?* 2021. arXiv: 2106.12671 [cs.CV] (cit. on p. 69).

[64] H. Wang, C. Wang, and L. Xie. «Online Visual Place Recognition via Saliency Re-identification». In: *IEEE RSJ International Conference on Intelligent Robots and Systems*. 2020, pp. 5030–5036. arXiv: 2007.14549 (cit. on p. 88).

[65] D. Gálvez-López and J. D. Tardós. «Real-time loop detection with bags of binary words». In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2011, pp. 51–58. DOI: 10.1109/IROS.2011.6094885 (cit. on p. 89).

[66] J. Verne. «Image Pre-Processing». In: *Computer Vision Metrics: Survey, Taxonomy, and Analysis*. Ed. by OREILLY. Apress, 2014, pp. 39–83. ISBN: 9781430259299 (cit. on p. 99).

[67] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo. «Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light». In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.12 (2020), pp. 4861–4875. DOI: 10.1109/TCSVT.2019.2963772 (cit. on p. 99).

[68] W. Luo, S. Duan, and J. Zheng. «Underwater Image Restoration and Enhancement Based on a Fusion Algorithm With Color Balance, Contrast Optimization, and Histogram Stretching». In: *IEEE Access* 9 (2021), pp. 31792–31804. DOI: 10.1109/ACCESS.2021.3060947 (cit. on p. 99).

[69] Y. Xu, H. Wang, G. D. Cooper, S. Rong, W. Sun, and H. Reza Karimi. «Learning-Based Dark and Blurred Underwater Image Restoration». In: *Complex.* 2020 (2020). ISSN: 1076-2787. DOI: 10.1155/2020/6549410 (cit. on p. 99).

[70] W. Luo, S. Duan, and J. Zheng. «Underwater Image Restoration and Enhancement Based on a Fusion Algorithm With Color Balance, Contrast Optimization, and Histogram Stretching». In: *IEEE Access* 9 (2021), pp. 31792–31804. DOI: 10.1109/ACCESS.2021.3060947 (cit. on p. 99).

[71] K. Zuiderveld. «VIII.5. - Contrast Limited Adaptive Histogram Equalization». In: *Graphics Gems*. Ed. by P. S. Heckbert. Academic Press, 1994, pp. 474–485. ISBN: 978-0-12-336156-1. DOI: https://doi.org/10.1016/B978-0-12-336156-1.50061-6 (cit. on p. 100).

[72] D. Asamoah, E. Oppong, S. Oppong, and J. Danso. «Measuring the Performance of Image Contrast Enhancement Technique». In: *International Journal of Computer Applications* 181 (Oct. 2018), pp. 6–13. DOI: 10.5120/ijca2018917899 (cit. on p. 103).

[73] M. Lecca, A. Torresani, and F. Remondino. «Comprehensive evaluation of image enhancement for unsupervised image description and matching». In: *IET Image Processing* 14.16 (2020), pp. 4329–4339. DOI: https://doi.org/10.1049/iet-ipr.2020.1129. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-ipr.2020.1129 (cit. on p. 103).

[74] V. Kamble and K. Bhurchandi. «No-reference image quality assessment algorithms: A survey». In: *Optik* 126.11 (2015), pp. 1090–1097. ISSN: 0030-4026. DOI: https://doi.org/10.1016/j.ijleo.2015.02.093 (cit. on p. 104).

[75] Y.-H. Liu, K.-F. Yang, and H.-M. Yan. «No-Reference Image Quality Assessment Method Based on Visual Parameters». In: *Journal of Electronic Science and Technology* 17.2 (2019), pp. 171–184. ISSN: 1674-862X. DOI: https://doi.org/10.11989/JEST.1674-862X.70927091 (cit. on p. 104).

[76] A. Mittal, R. Soundararajan, and A. C. Bovik. «Making a "Completely Blind" Image Quality Analyzer». In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212. DOI: 10.1109/LSP.2012.2227726 (cit. on p. 104).

[77] A. Mittal, A. K. Moorthy, and A. C. Bovik. «No-Reference Image Quality Assessment in the Spatial Domain». In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708. DOI: 10.1109/TIP.2012.2214050 (cit. on p. 104).

[78] V. N, P. D, M. C. Bh, S. S. Channappayya, and S. S. Medasani. «Blind image quality evaluation using perception based features». In: *2015 Twenty First National Conference on Communications (NCC)*. 2015, pp. 1–6. DOI: 10.1109/NCC.2015.7084843 (cit. on p. 105).

[79] «Learning a no-reference quality metric for single-image super-resolution». In: *Comput. Vis. Image Underst.* 158.C (2017), 1–16. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2016.12.009 (cit. on p. 105).

[80] S. Garg, T. Fischer, and M. Milford. «Where Is Your Place, Visual Place Recognition?» In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2021. DOI: 10.24963/ijcai.2021/603 (cit. on p. 119).

[81] J. Su, X. Tu, F. Qu, and Y. Wei. «Information-Preserved Blending Method for Forward-Looking Sonar Mosaicing in Non-Ideal System Configuration». In: *2023 IEEE Underwater Technology (UT)* (2022), pp. 1–5 (cit. on p. 132).