Self-supervised multimodal representations for marine robotics

Azamat Kaibaldiyev [1]

Supervisors: Oscar Pizarro, Ricard Marxer

[1]Norwegian University of Science and Technology [2]University of Toulon [3]Laboratoire d'Informatique et Systèmes

Co-funded by the Erasmus+ Programme of the European Union

MIR MARINE & MARITIME INTELLIGENT ROBOTICS

Abstract

This thesis work explores multimodal learning techniques for habitat classification using remotely sensed and visual data. Autonomous Underwater Vehicles (AUVs) play a vital role in marine scientific surveys, providing efficient data collection and observations over marine ecosystems. Benthic habitat mapping, which involves classifying seabed sites into different habitat categories, is a key objective in marine ecology. AUVs capture visual imagery of the seabed, while multibeam sonars collect bathymetry data. By correlating visual imagery with features from the bathymetry data, reliable habitat classification models can be developed. This study investigates self-supervised learning approaches, particularly contrastive learning, to enable robust classification and image-content prediction. Results show that contrastive learning on bathymetry data achieves test accuracy rates of approximately 59% and 63% for patch sizes 16x16 and 32x32, respectively. In contrast, visual imagery achieves over 86% accuracy. Multimodal learning, combining visual images with bathymetry patches, yields accuracies of about 71% and 72% for different patch sizes. Separate networks with shared loss achieve accuracies of over 71%. This work demonstrates the feasibility and effectiveness of multimodal learning techniques in habitat classification, leveraging the strengths of both visual and bathymetry data.

Motivation

- Nowadays, Autonomous Underwater Vehicles (AUVs) are integrated into marine scientific survey tasks for efficient data collection in inaccessible areas, offering extended operation time and excellent data extraction capacity.
- Benthic habitat mapping is a key objective in marine ecology, involving the classification of sea-bed sites into different habitat categories (e.g., coral reefs, sand, seagrasses) [2].
- AUVs perform benthic imaging, extracting high-resolution visual imagery with geo-references along planned paths. Visual imagery from AUVs provides interpretable data used for habitat classification and benthic research.
- Deploying AUVs to obtain complete sea-bed imagery is impractical due to time and cost constraints. The marine seafloors of interest can cover thousands of square kilometers, requiring substantial human and machine resources. AUV image sensors typically cover a limited area (1-10 square meters) within a few meters range.
- Multi-beam sonars efficiently obtain large-scale seabed bathymetry data, although at lower resolution, and are less affected by water turbidity.
- Underwater robots often carry multiple sensing modalities (e.g., sonar and visual camera) to complement strengths and overcome limitations.
- Benthic habitats correlate with underlying bathymetry, enabling the use of machine learning predictive models for habitat identification. Correlation involves combining ground-truth imagery data with features extracted from seabed bathymetry data.
- Reliable and robust models are needed to learn correlations between scarce visual imagery and features from acoustic data.
- This thesis work investigates self-supervised learning approaches with multiple modalities for robust classification and image-content prediction.

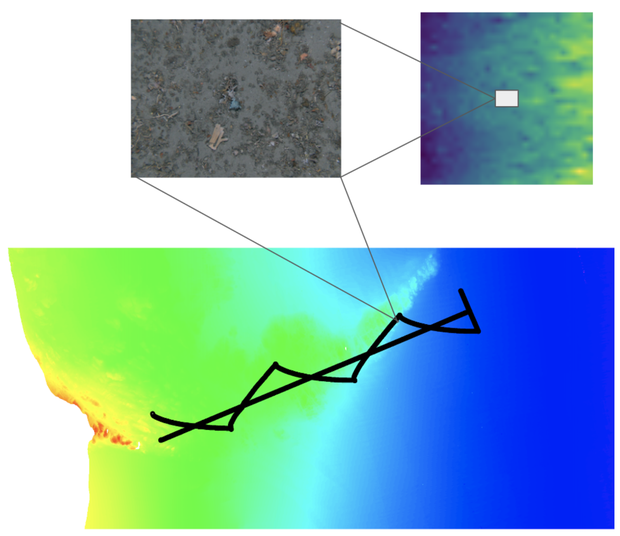

Figure 1. An example demonstrating the matching of photos from an AUV survey to the appropriate bathymetry. Pairs of 16 by 16 and 32 by 32 patches of gridded bathymetry are derived for each picture point along the AUV route which is depicted in black. The patch extends much beyond the size of the image.

SimCLR - A Simple Framework for Contrastive Learning of Visual Representations.

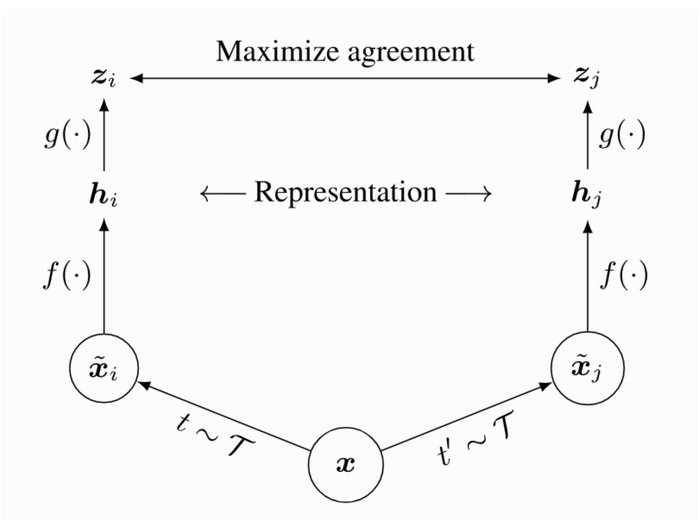

Figure 2. SimCLR network setup.

[1]

Approach

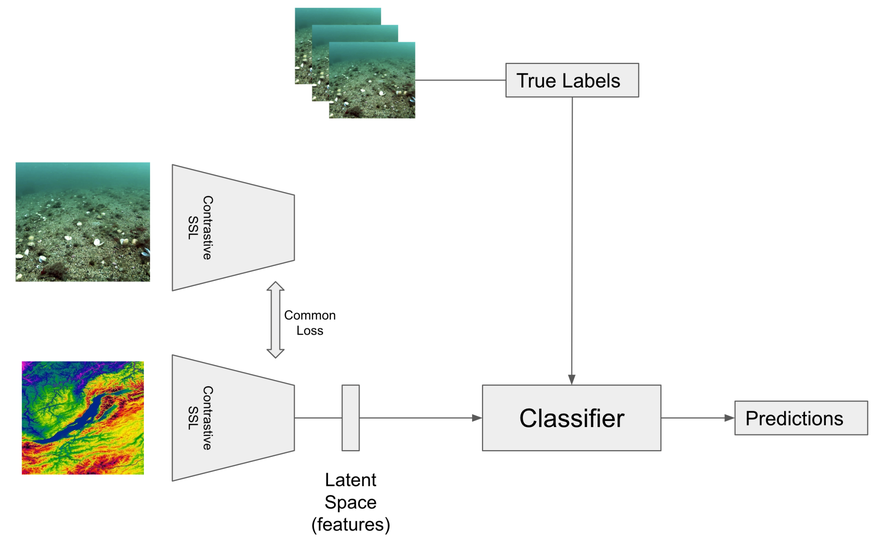

Figure 3. Overall model for multimodal training of single bathymetry plus bathymetry and optical image pairs

Results

SimCLR model	Split	Train accuracy (%)	Test accuracy (%)
Bathymetry 16	Line	77,21	63,97
Bathymetry 32	Line	65,22	59,79
Visual Images	Line	91,03	86,10
Single network bathymetry 16 and image pairs	Line	76,83	71,12
Single network bathymetry 32 and image pairs	Line	77,84	71,89
Double network bathymetry 16 and image pairs on bathymetry	Line	76,36	71,59
Double network bathymetry 32 and image pairs on bathymetry	Line	76,25	71,23
Double network bathymetry 16 and image pairs on images	Line	76,34	74,08

Figure 4. Table of performance results of training classifier on feauture representations from contrastive learning

Discussion

Conducting contrastive learning on a large amount of bathymetry data and subsequently performing classification using the extracted features yielded test accuracy rates of approximately 59% and 63% for patch sizes 16x16 and 32x32, respectively. This lower performance can be attributed to the lower resolution of the bathymetry data. In contrast, performing the same contrastive learning on visual images resulted in a test accuracy performance of over 86%. This signifies the superior performance of visual images, which possess high angular resolution and offer detailed spatial information about the observed scene. Regarding the multimodal learning approach, where visual images were contrasted with bathymetry patches within a single network, it yielded accuracies of about 71% and 72% for patch sizes 16x16 and 32x32, respectively. Furthermore, employing separate networks for visual and bathymetry data, with a shared loss function to preserve their representations, achieved accuracies of approximately over 71%. These results highlight the benefits of multimodal learning, as it enables one modality with extensive coverage capabilities to learn rich features from the other modality, which may have a narrower scope. In this context, synchronous learning of the feature space of visual images, with their high angular resolution and detailed spatial information, and bathymetry patches, providing valuable depth information about the underwater topography and seafloor characteristics despite their lower resolution, proved to be advantageous. Thus, the utilization of multimodal learning with the contrastive learning technique demonstrated its feasibility and effectiveness in this study.

Conclusion

In conclusion, this thesis aimed to explore multimodal learning techniques involving remotely sensed and visual data in the context of habitat classification. Gathering visual imagery of large underwater areas is a resource-intensive task, while bathymetry data obtained from multibeam sonar devices provides a more easily accessible representation of the underwater terrain. Combining these modalities allows for efficient and reliable habitat classification. The limited availability of underwater visual imagery can serve as ground truth, while the widespread availability of bathymetry data can be used to estimate and predict habitat classes based on the imagery data. By leveraging both modalities, they can complement each other and contribute to a common task and goal.

In this master's thesis work, self-supervised multimodal learning techniques, specifically contrastive learning, were investigated, implemented, and tested on terrestrial and underwater datasets. The results demonstrated the mutual benefit of each modality. Multimodal learning enhanced the performance of models in predicting bathymetry data by leveraging the contrastive learning of features from abundant but low-quality bathymetry data and scarce but rich visual data.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton.
A simple framework for contrastive learning of visual representations.
In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[2] Stefan B Williams, Oscar Pizarro, Michael Jakuba, and Neville Barrett.
Auv benthic habitat mapping in south eastern tasmania.
In *Field and Service Robotics: Results of the 7th International Conference*, pages 275–284. Springer, 2010.

azamat.kaibaldiyev@alumni.nu.edu.kz