Multi-source Data Collection for State-of-the-art Data Analysis from Ground-proximate Images in Sea Ice Classification

Lucas Woltmann[1], Rune Dalmo[2], and Raymond Kristiansen[3]

[1] Faculty of Computer Science, Technische Universität Dresden, Dresden, Germany
[2] Faculty of Engineering Science and Technology, UiT The Arctic University of Norway, Narvik, Norway
[3] Northern Research Institute (Norut) Narvik, Narvik, Norway

Abstract. In modern data analysis it is imperative to use well maintained data sources with curated content. This publication gives an approach for research areas, where there is no such central facility. The specific area used here is sea ice classification from images. The publication is split into two parts. The first part describes the integration of discontinuous sources for different aspects of data needed. The second part describes a simple approach for getting the sea ice concentration from images taken by an ice breaker. The classification is based on the above mentioned multi-source data. We can illustrate that it is possible to combine data from various sources to a central repository and use this repository to obtain the sea ice concentration from images without using any other inputs.

1 Introduction

When working with data in classification or regression one will usually find well-formed sources. These sources contain of a feature space describing every data point and a label giving the true value for the classification or regression. The truth value is called the Gold Standard G.

With the recent expansion of classical data analysis into other research areas there is a lack of central repositories with a focus towards machine-readable information. Many sources, especially in science, give either human-readable data or information for a partial aspect only. This makes the data acquisition an essential part of data analysis. Often, it is referenced as the Extraction-Transformation-Load (ETL) process.

We found that most data in Polar research does not contain a Gold Standard. Getting experts for labelling the data can be expensive and not feasible for large collections of data. Another solution is to combine multiple sources. But combining these sources needs a common point of reference, like a common position or a common timestamp.

To prove that it is already possible to perform state-of-the-art data analysis, we took the following steps.

1. Collected data from public online data repositories,
2. Integrated the collected data into a central repository,
3. Used the labelled data from the central repository to read the sea ice concentration from images.

Arctic research has a high demand for data analysis. In this particular instance we are talking about sea ice classification. This includes sea ice extent, sea ice concentration, sea ice roughness and sea ice geometry. Whereas traditional methods include satellite images for analysis, this publication relies on images taken on board an ice breaker. The pictures are used for obtaining the sea ice concentration. This could be helpful for the navigation of ships in the Arctic sea, where one could utilize the automated information extraction from the images for direct control input. Compared to the satellite images, the images from the ship can give a more detailed look onto the ice and have a higher resolution. Where with resolution, we mean there are more images per surface area unit.

2 Related Work

Data integration and collection is done in a vast variety of areas. The ETL process is used commonly for combining heterogeneous sources. Vassiliadis, Simitsis et al.[18] give a common approach for the ETL process.

Most of the data analysis for sea ice is done via satellite images [19][15][17]. There is some research on images taken in low orbit, like from unmanned aerial vehicles (UAVs) in low-level flights [11]. This research concentrates on sea ice operations [4] and marine navigation [2][8]. Using ground-proximate images is an essential part for Arctic research. This can be an advantage when analysing geometry or roughness of the sea ice. A major problem is the lack of centralised data sources for such images. Whereas satellite images can be obtained in public repositories [3] [5], there are no such well-maintained collections for other kinds of image data. Like most data collected on Arctic research cruises, the image data and meta data is highly sparse. Not only are there values missing, but they also only give a small frame either in global coverage or diversity of measurements. Sparse data always needs high effort in data integration. Most of the time, research data must be obtained by detailed combination of various sources. It is not trivial to combine the inhomogeneous information suppliers. With this publication, we want to show the potential of an integrated data repository for new data sources, emerging from the open data movement.

The before mentioned publications use a wide spectrum of computer vision algorithms for image analysis. These include Neural networks, Bayesian classifier and others. For the prove of concept, we opted for a simpler approach for image segmentation. All the other algorithms are much more complex and can be described and customized for this problem in a different publication each.

3 Work

This chapter will give an overview over the methods for collecting machine-readable base data to be used by the the ice concentration estimator. The connection made between the inhomogeneous sources are shown. Additionally a method for estimating sea ice concentration from image data is used to verify the benefit of the collected data.

3.1 Data Description

The data collection uses several input sources from different providers. Given their individual purpose, they differ immensely in type and format. In general we used three repositories:

- Rolling deck to repository (R2R) of the US Navy for the USCGC Healy [7],
- Ice extent data from NASA DAAD [13],
- Image repository of the USCGC Healy [1].

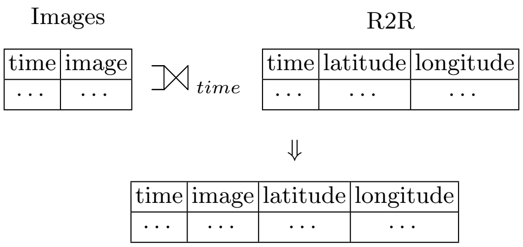
The R2R is a central repository for collecting and distributing underway data of research vessels. Maintainer and owner is the US Navy. It is one of the central points of reference for scientific data exchange in oceanic research. For the purpose of this paper, the R2R can deliver R2NAV data tables [6]. These contain navigation data, including longitude, latitude, heading and speed of a ship for a given timestamp (UTC). The measures are taken every minute of a cruise.

The ice extent data is stored in a 720x720 EASE2-grid [10]. This results in an edge length of 25km for each grid square. Each cell contains an integer label for the sea ice concentration. The ice extent is measured daily.

The image repository contains images taken on deck of the USCGC Healy, facing frontwards. Each image is only labelled with a timestamp. Images are taken hourly.

3.2 Data Collection

Given the images taken on board the ship, there are key elements missing for a successful classification. The ice extent measured by the satellite contains a grid giving the extent at a position for a given timestamp. For using this information with the images, we need to add a position and a timestamp to each image. The positions are available through the R2R. The R2R has a tabular structure with an entry for each position of the ship for every minute of the cruise. By downloading information from the catalogue via the R2R's API and putting together all cruises from 2010 to 2016, we obtained a complete table with the time and the positions of the ship over the last six years. This table can now be joined with the image data. The image data already uses timestamps for the photographs, which we can use as the join predicate. Figure 1 details the left outer join and the resulting table. We now have access to the positions of the images.



Images

time	image
...	...

$\bowtie_{time}$

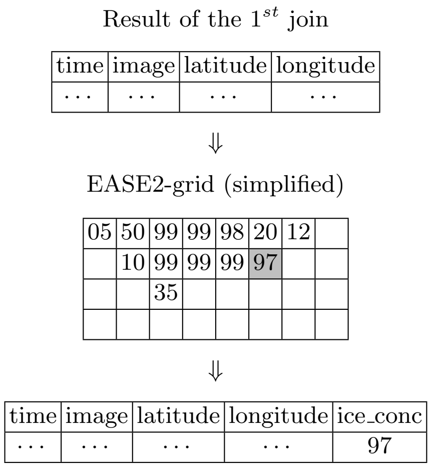
R2R

time	latitude	longitude
...	...	...

⇓

time	image	latitude	longitude
...	...	...	...

Fig. 1. Join of the images with the R2R data



Result of the 1^{st} join

time	image	latitude	longitude
...	...	...	...

⇓

EASE2-grid (simplified)

05	50	99	99	98	20	12	
	10	99	99	99	97		
		35					

⇓

time	image	latitude	longitude	ice_conc
...	...	...	...	97

Fig. 2. Selection of the closest grid cell

With this table we can make use of the satellite measurements. The percentage of the ice is given in a 720x720 grid for the Northern hemisphere. Because the image positions are much finer than the grid for the ice extent, there is no one-to-one connection between position of the image and a grid cell. We need to estimate the image's position onto the grid. Additionally, the time of the image must be correlated to the day of the ice extent measurement. So each image uses its position and timestamp to find the nearest center of a grid cell. To find the closest cell on the curved surface of the Earth, we used the Haversine distance [12] between the position of the image and the grid cell center. The image gets labelled with the ice extent value of that grid cell at the day the image was taken. Figure 2 gives an overview of the process.

After removing images with dead links or other erroneous properties, we obtain a collection of approximately 15000 labelled images. This collection is called *central repository*.

3.3 Sea Ice Concentration Classification

To access the quality and advantages of the central repository, the image data is used for a task, which would not be possible without the integration step. We used a image segmentation algorithm to obtain the sea ice concentration and compare it to the labels. For a successful classification we have chosen 3000 images with no occlusions and with decent brightness. The images are converted to a gray scale range. This returns the image I as an matrix with the dimensions of the original image. The interval for each matrix cell is $[0, 255]$. The first step cuts off the parts of the images above the horizon line and below the tip of the bow of the ship. This leaves only the sea and the sea ice in the image. The second step applies the Sobel operator [16] to receive an elevation map of the image. The elevation map details the changes of intensity in the image. The Sobel operator S on the image I is described as

$$S = \sqrt{(H * I)^2 + (V * I)^2} \tag{1}$$

The matrices H and V are given statically by directional derivative gradient estimation [16].

$$H = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{2}$$

$$V = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{3}$$

With the elevation map and some specific gray values as markers, we are able to use the Watershed transformation [9]. The result is an image with two partitions. Black areas indicate water and bright areas indicate sea ice. The percentage of sea ice concentration p_{ice} on an image with width w, height h and number of bright pixels c_{bright} is given by

$$p_{ice} = \frac{c_{bright}}{w \cdot h} \tag{4}$$

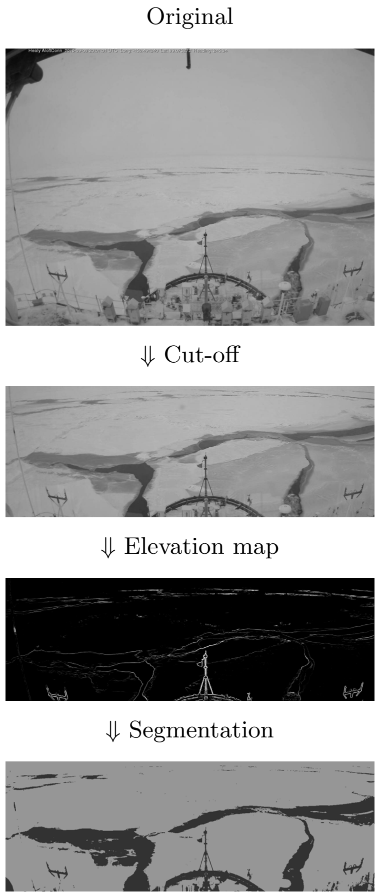
$$p_{ice} \in [0, 1];\ p_{ice} \in \mathbb{R} \tag{5}$$

Figure 3 gives the simplified structure of the algorithm.

The chapter has shown the divers approach from collecting from multiple sources to describing the combination of the Sobel operator and the Watershed transformation for image segmentation. With the integrated data set, the algorithm is able to label a given image with a percentage of sea ice.

4 Evaluation

For evaluating the collected data, we used the image segmentation. If the segmentation algorithm gives results with a low error rate, it can be assumed that



Original

⇓ Cut-off

⇓ Elevation map

⇓ Segmentation

Fig. 3. Segmentation of an image

the data is well-formed. Using up to 3000 test images, the algorithm obtained the sea ice concentration. The classification got compared to the labels from the satellite data. We used the mean square error (MSE) between the classifications and the labels for all images. The MSE for classifications p and labels l for a number of images N is

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - \frac{l_i}{100}\right)^2 \tag{6}$$

For comparison we also used the mean absolute error (MAE). The MAE for classifications p and labels l for a number of images N is

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|p_i - \frac{l_i}{100}\right| \tag{7}$$

N	MSE	MAE
500	0.0150	0.107
600	0.0149	0.103
1000	0.0171	0.112
2000	0.0175	0.112
3000	0.0167	0.108

Table 1. MSE and MAE for different numbers of images

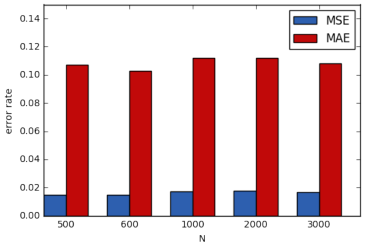


Fig. 4. MSE and MAE in comparison

With

$$\forall p, l : p \in [0,1];\ p \in \mathbb{R};\ l \in [0,100];\ l \in \mathbb{N} \tag{8}$$

for both error rates. Table 4 and Figure 4 detail different MSE and MAE for various N. The images were chosen consecutively.

The taken measurements have reasonably low error rates. Both error types show that the average error on the 0%-100% sea ice concentration scale is about 10%. The low error rates both in MSE and MAE show a high reliability for the algorithm. This indirectly shows the good quality, and in a certain way the usefulness, of the integrated data.

It should be noted that the algorithm highly relies on the quality of the on-board and satellite data. Higher quality images are going to reduce the error rate even more. Less noise in all images should give a better base line for the image segmentation through equalizing different lighting for the algorithm. Currently, noise reduction is one of the major challenges for the algorithm. If we use a higher resolution[4] for the satellite data, the Gold Standard will also be more reliable. It is conceivable to use different data sources and compare the different error rates.

5 Conclusion

We have shown that modern data analysis in Arctic research is already possible with today's available data. The right methods and tools make it possible to

[4] again: a finer grid, not the image resolution

integrate inhomogeneous sources and use them for new advances. The chosen example algorithm is capable of classifying the sea ice concentration from an image. For this it only uses the pixel information and the integrated data labels from a third party source. The algorithm can also be applied to images from low-level flight drones, putting the research into a more general perspective. The results of this research could also be an aid for the navigation of ships in ice-covered waters. By automatically extracting sea ice information from a bow-mounted camera, it can supply vital information for operations in the Arctic.

There are still remaining problems to tackle. One major problem is the polar day and night cycle. In the polar night, the camera can not take any pictures, which are not completely black. Even the artificial lighting on the ship is not enough to raise the quality of the images. In the polar day, one gets a few images every day which face the sun directly. This results in overexposed images. Bright reflections on the water can be confused with ice. A camera capturing more than the visible light spectrum (full spectrum camera) could be useful in both instances. Additional, several spectra from different cameras can be combined.

In general, there are some images, having occlusions or faults in them. The occlusions or faults include cracks and drawings on the visor, antennas blocking the view or the camera fallen off its mounting. This renders most of these erroneous images unusable.

The last difficulty are the images themselves. There is still a part of the bow visible after cropping, biasing the percentage to be insignificantly lower than in reality. Another problem is the fish eye distortion of every image, curving the horizon line. The algorithm currently does only a straight cut on the highest point of the distorted line, leaving an insignificantly part of the sky in the image.

All the before mentioned problems could be eradicated in further research. It would also be desirable to get more image sources, be that from other ships or UAVs (drones). For a common repository, it will be mandatory to integrate other forms of data. Physical measurements from ships, like thrust or the counterforce on the hull, would be one possibility.

At last, the research can be broaden to cover other types of sea ice properties, like roughness and geometry. There are already some endeavours getting additional properties from satellite images [14]. This could be transferred into this research as well.

References

1. Deck Images USCGC Healy. http://icefloe.net/Aloftcon_Photos/. Accessed: 2017-01-03.
2. MarineUAS – Autonomous Unmanned Aerial Systems for Marine and Coastal Monitoring. http://www.itk.ntnu.no/itn/. Accessed: 2017-01-03.
3. National Snow & Ice Data Center. http://nsidc.org/data/resources/collections. Accessed: 2017-01-03.
4. OpSIce – Operations in Sea Ice-Covered Waters. http://opsice.com/. Accessed: 2017-01-03.
5. PolarView. http://www.polarview.aq/arctic. Accessed: 2017-01-03.

6. R2NAV data description. http://get.rvdata.us/format/100002/format-r2rnav.txt. Accessed: 2017-01-03.
7. Rolling Deck to Repository (R2R). http://www.rvdata.us/catalog/Healy. Accessed: 2017-01-03.
8. F. Balampanis, I. Maza, and A. Ollero. Area decomposition, partition and coverage with multiple remotely piloted aircraft systems operating in coastal regions. In *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 275–283, June 2016.
9. S. Beucher. The Watershed Transformation Applied To Image Segmentation. In *Scanning Microscopy International*, pages 299–314, 1991.
10. M. J. Brodzik and K. W. Knowles. EASE-Grid: A Versatile Set of Equal-Area Projections and Grids. In *M. Goodchild (Ed.) Discrete Global Grids*. National Center for Geographic Information & Analysis, 2002.
11. M. Eckerstorfer, E. Malnes, H. Vickers, S. A. Solbø, and A. Tøllefsen. Avalanche Debris Detection Using Satellite- and Drone Based Radar and Optical Remote Sensing. *AGU Fall Meeting Abstracts*, December 2014.
12. Don Josef de Mendoza y Rios. F.R.S. Recherches sur les principaux Problemes de l'Astronomie Nautique. In *Proceedings of the Royal Society*, 1796.
13. A. Nolin, R. L. Armstrong, and J. Maslanik. Near-Real-Time SSM/I-SSMIS EASE-Grid Daily Global Ice Concentration and Snow Extent. Version 4. Ice Extent. NASA DAAC at the National Snow and Ice Data Center, 1998.
14. Anne W Nolin, Florence M Fetterer, and Theodore A Scambos. Surface roughness characterizations of sea ice and ice sheets: Case studies with misr data. *IEEE Transactions on Geoscience and Remote Sensing*, 40(7):1605–1615, 2002.
15. B. Scheuchl, R. Caves, I. Cumming, and G. Staples. Automated sea ice classification using spaceborne polarimetric SAR data. In *IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217)*, volume 7, pages 3117–3119 vol.7, 2001.
16. I. Sobel and G. Feldman. A 3x3 Isotropic Gradient Operator for Image Processing. Never published but presented at a talk at the Stanford Artificial Project, 1968.
17. L. K. Soh and C. Tsatsoulis. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):780–795, Mar 1999.
18. Panos Vassiliadis, Alkis Simitsis, and Spiros Skiadopoulos. Conceptual Modeling for ETL Processes. In *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP*, DOLAP '02, pages 14–21, New York, NY, USA, 2002. ACM.
19. N. Y. Zakhvatkina, V. Y. Alexandrov, O. M. Johannessen, S. Sandven, and I. Y. Frolov. Classification of Sea Ice Types in ENVISAT Synthetic Aperture Radar Images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2587–2600, May 2013.