ScanBathy: A new solution to digitize depth data from historic survey sheets

Ronan Créach, Sébastien Boschi, Laurent Boutry, Jonathan Genevier, Paul Claverie and Alexandre Badez

SHOM gathered over the past centuries more than 75,000 analogue field sheets, some of the oldest dating from the early 1800s. Before the digital era, soundings were recorded as hand-plotted numerals on bathymetric sheets that were used to produce nautical charts. Data recovery and rescue are crucial in most cases, for example in areas that were not mapped with modern technologies or to improve understanding of seabed changes in nearshore areas.

Context
SHOM gathered over the past centuries more than 75,000 analogue field sheets, some of the oldest dating from the early 1800s. Before the digital era, soundings were recorded as hand-plotted numerals on bathymetric sheets that were used to produce nautical charts. Data recovery and rescue are crucial in most cases, for example in areas that were not mapped with modern technologies or to improve understanding of seabed changes in nearshore areas.

Technical environment
ScanBathy was developed by the French company Magellium, tailor-made for optimizing SHOM digitization of historical survey sheets and replacing RETINE software (Louvart, 1996).

The software is mainly built on GDAL, OpenCV, Deeplearning4J, Qt and Eclipse RCP libraries.

1- Project Initialization
The main interface is used for:
- project management (create, modify, archive);
- project workflow (import, georeferencing, digitize and control);
- data export.

The field sheet is loaded from a file or a URI in TIF, JPEG2000 or PDF format at 300dpi or higher.

Metadata associated with the digitization project are created through this interface.

Print and export functions are used to generate the digitizing outputs that will populate SHOM bathymetric database.

2- Georeferencing
ScanAdjust module is an efficient tool dedicated to georeferencing tasks. It is segmented so that only the necessary functions are displayed at each step of the process.

It uses cartographic annotations, like crosses or grids, to compute geometric models. The user either manually picks up control points on these marks or uses the automatic acrosses detection module. Moreover, special control points are used to compute local deformation models, which greatly correct local deformations induced by the foldings of the sheets.

As the geometric deformations are viewed on the fly on screen, the validation is done with great ease and accuracy.

3- Soundings Digitizing
If there is not enough samples to train the ConvNet classifier, the software allows to create a learning dataset from the soundings sheets. Once the learning set is created, a SVM classifier may be trained (Arandjelovic and Zisserman, 2012).

The learning is done from the projection of the pixel values in the Histogram of Oriented Gradient space.

4- Optical character recognition (OCR) and machine learning
Specialized OCR techniques have been developed in order to cope with handwritten soundings detection. Some challenging digit recognition issues have been tackled such as automatic orientation detection and digits clustering in dense soundings areas.

Depending on the maps and training datasets, ScanBathy is able to switch between a convolutional neural network classification approach and a Support Vector Machine (SVM) one. Both methods offer significant results and improve substantially the soundings detection rate.

Thresholding and preprocessing are of great importance to feed the classifier with clear and normalized inputs. A bilateral filter and morphological operators are combined while keeping straightforward settings for users. Digits preprocessing normalizes their dimensions, radiometry and orientation.

The software is provided with ConvNet classifier trained on the NIST special database 19 (S019) (Cireşan et al., 2012) containing 430,000 handwritten digits. Despite its amount of samples, S019 is not fully representative of the studied data.

Trainings with S019 reach 99.4% of accuracy but only 80% over bathymetric, fieldsheets data. Further optimization of the Convolutional Neural Networks implementation generated better detection score on the fieldsheets reaching 90% of accuracy.

5- Production control
A dedicated selection mode allows to quickly check and potentially edit the values of the soundings.

A validation grid allows to bound the inspected areas. A grid cell can be identified as checked only once the user has controlled every sounding inside.

It facilitates an extensive validation without omission.

A layer map with a hypsometric lookup table (LUT) helps to visually identify erroneous values.

Conclusion
ScanBathy is a new quality assessed and time-efficient solution to tackle the massive job of digitizing soundings from historic survey sheets. It has been tailor-made for this specific task and benefits of years of user experience from the SHOM operators.

The advanced OCR and machine learning techniques implemented greatly speed up the production with a gain up to 5x. The exhaustive settings will allow to tune at no cost the digitizing interface to deal with every kind of survey sheet.

References