Ocean Sci. Discuss.,
https://doi.org/10.5194/os-2020-62-RC2, 2020

Ocean Science
Discussions

Interactive comment on "Towards operational phytoplankton recognition with automated high-throughput imaging and compact convolutional neural networks" *by* Tuomas Eerola et al.

Anonymous Referee #2

Received and published: 14 August 2020

The paper "Towards operational phytoplankton recognition with automated high-throughput imaging and compact convolutional neural networks2"presents how a conventional neural network method can be used to automatically recognise phytoplankton classes in real time (or near real time) in the Baltic Sea. This approach uses images collected by an in situ optical sensor. The study is in the scope of the special issue of JERICO-RI where the biological component has been investigated using different techniques such as the image recognition using different bio-optical sensors. It is a very clear, and easy to understand even without deep background in mathematical

processes and operations. One of the main limitations to use the Deep learning approach is the need of a large number of images and equal quantity per group to build the training sets. Therefore, the authors explore different techniques to increase the performance of the approach. I recommend the publication. However, the article will gain in mentioning also: 1) the current efforts of imaging data management with the establishment of current international biodiversity data standards, such as Darwin Core (DwC), used OBIS (EUROBIS) and GBIF. 2) what are the benefits to use the approach in the article versus the established tools such as ECOTAXA based on Deep Machine Learning also.

Comments:

Line 62: "FlowCytobot is among the most frequently used imaging flow cytometer". Until now, there was only one or two groups of American scientists using the FlowCytobot. I do not think that we can say most frequently used in this case.

Line 80: what are the practical implications to aquatic research which are mentioned? This needs clarification.

Line 100: it will be worthwhile here to mention the principal of a FAIR data: findability, accessibility, interoperability, and reusability.

Line 109: "FerryBox"

Line126: is it testing set or training set is equal to 25% Needs clarification (see Table2)? Why 25% has been chosen

Line 211: "CNN performs significantly better than the Random Forest implementation". It should be mentioned that that the two methods used different attributes with a higher number used for CNN which explains a better performance for CNN.

Line 226: in what identifying the planktonic species is important for the Baltic Sea ecosystem? I understood that the authors want to relate their mathematical approach to an ecological interest but it will be relevant to have some information about why monitoring the species is important, particularly for those who do not know the Baltic Sea.

Line 253: “ecological relevance” should be better to mention human health concern?

Line 294: “there exists”, replace by there is.

Line 304: “ It is impossible to create classes for all images. . ..” This sentence underlines the lack of information concerning the percentage of phytoplankton recognised compared to the on those which are not recognised and potentially included in “ small roundish or elongated objects”.

Interactive comment on Ocean Sci. Discuss., https://doi.org/10.5194/os-2020-62, 2020.