Classification of Zooplankton Images Using Descriptors

Popular image databases like Ecotaxa provide computer-aided annotation of images:

- the system suggests a classification of an image by means of a machine-learning method,
- the scientist validates / corrects the automatically generated classification in order to produce the final classification.

For the automatic classification, Ecotaxa uses descriptors / features extracted from the images, e.g. shape area, convexity, etc.

This project has two objectives:

1. To explore machine learning methods based on image descriptors and compare them with methods based on deep learning,
2. To understand how these methods can scale when applied to very large data sets.

For training our models we have used the same dataset used with the Deep Learning methods. The input data is stored in comma-separated files exported from Ecotaxa.

Training the model

You can check out the model from gitlab with

```
git clone git@gitlab.awi.de:gbusatto/de.awi.analytics.lokides.git
```

In order to train the scikit-learn model, change to the project directory and run

```
python src/starred-randomforest-classifier.py
```

Preliminary results of image classification with descriptors and random forests show an accuracy comparable to that obtained by CNNs working directly on the image data. The training time for this model were in the order of magnitude of a few minutes. These result must be further evaluated and validated on larger datasets.

In order to train the pyspark model, you have to install spark.

Comparison of different machine learning frameworks: pyspark, Scikit-Learn, Tensorflow. In particular, evaluation of Spark for ML:

1. Is it possible to use the same ML models with Scikit-learn and Spark?
2. Do these model scale with large data sets?

Original data set: about 120000 samples, 25 classes.

Reduced data set (with the purpose of reducing bias): 7000 samples, 7 classes (1000 samples per class)