Hyperspectral Imaging honey dataset

Tessa Phillips, Ary Noviyanto, Waleed Abdulla

April 22, 2020

1 Dataset Imagery Capturing

The images of honey have been captured with a hyperspectral imager SOC-710 from Surface Optic. The hypercubes captured are between 400 - 1000nm in wavelength with a 5nm increment and have a 520 x 696 spatial resolution.

The system used to capture these images has been developed using rigorous testing and evaluation to create a standard for the hyperspectral imagery capture of honey [1][2].

The dataset comprises the features (128 spectral bands) of the hyperspectral images after a well-designed procedure for calibration, pre-processing and segmentation. This process produces a consistent dataset through all trials, and so it is comparable to the initial work on feature reduction and classification. Figure 1 shows the procedure of preparing the samples. There are six acquisitions of each honey sample; these are then captured with the hyperspectral imager and calibrated, pre-processed and segmented to get the final spectral information.

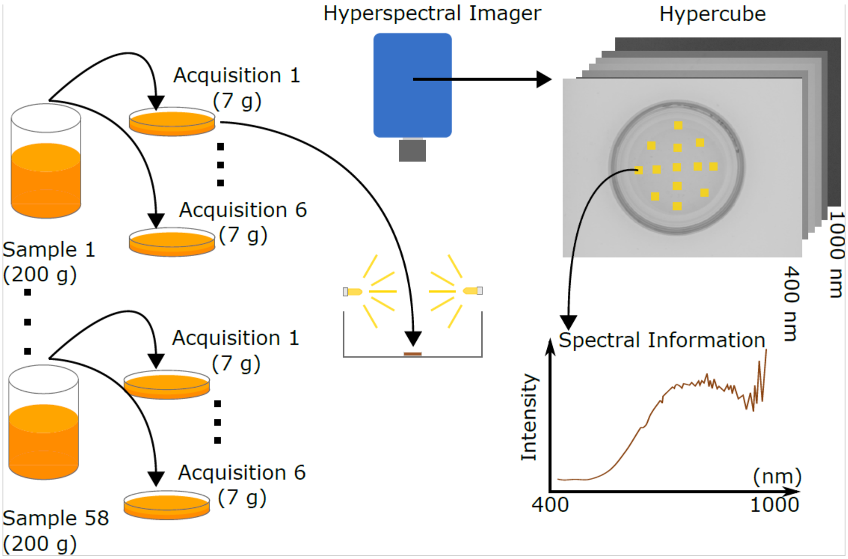

Figure 1: The steps for preparing the hyperspectral imagery samples and extracting spectral information.

2 Imagery Calibration

White balancing is a standard calibration procedure for hyperspectral images, as it ensures all the images are based on a consistent amount of light [1][3]. Equation 1 describes the standard white balancing process for the position (x, y), and wavelength λ. Where $\hat{I}(x, y, \lambda)$ is the calibrated hypercube, $I(x, y, \lambda)$ is the original hypercube, and $D(x, y, \lambda)$ and $W(x, y, \lambda)$ are the dark and white references, the dark reference is found by capturing with the camera cap on, and the white reference is found by capturing the calibration material with our standard lighting system.

$$\hat{I}(x, y, \lambda) = \frac{I(x, y, \lambda) - D(x, y, \lambda)}{W(x, y, \lambda) - D(x, y, \lambda)} \tag{1}$$

The dynamic white balancing method extends from the standard white balancing approach and uses a reference value from each row of the hyperspectral image instead of a separate reference for each pixel in the image. Equation 2 shows the dynamic white balancing process.

$$\hat{I}(x, y, \lambda) = \frac{I(x, y, \lambda) - D(x, y, \lambda)}{W(y, \lambda) - D(x, y, \lambda)} \tag{2}$$

This is important in this work because we are using a push-broom hyperspectral imager so the lighting conditions could change between the rows of the images. The dynamic white balancing approach is what has been used for this dataset [3].

Although in capturing the images the lighting has been kept as consistent as possible, there are some slight inconsistencies due to external lights from the surrounding environment. This needs to be compensated by the proposed white balancing approach.

3 Imagery Segmentation

The segmentation is performed by first cropping the image to remove the background, selecting only the honey sample in the centre of the image as a region of interest. The region of interest is split into 25 segments using a five by five grid [1], as shown in figure 2.

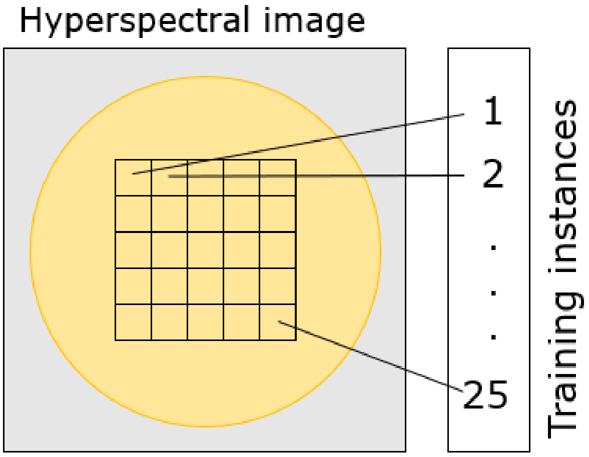

Figure 2: Segmentation of honey imagery through extracting 25 instances from each hyperspectral image of honey. The honey sample yellow circle depicts the honey sample.

The segmentation procedure provides many training examples from each acquisition [3]. The spatial information from the hyperspectral images does not vary significantly as the honey samples are prepared to be homogenous. Yet there are slight perturbations in the data within the grids. In this case, we obtain a lot of training and testing examples sufficient for preparing and testing the predictive models. The training and testing sets do not split up these segments; rather, each image is either part of the training or testing set. The same approach is followed in the cross-validation.

4 Pre-processing

The pre-processing step follows the segmentation and uses a simple normalisation approach, where the mean of the spectra is forced to be at 0, and the standard deviation to be 1. Other normalisation approaches were also considered; however, this was experimentally found to be the best approach for this dataset [1].

This pre-processing step improves the training convergence, and performance of many machine learning algorithms, such as support vector machines (SVM) and neural networks.

5 Dataset Overview

This dataset consists of segmented and pre-processed hyperspectral images of honey samples. The full dataset contains 21 different classes of honey, where the class label represents the botanical origin. The honey samples came from a range of different brands, and the brand name was also included in the data. The full dataset of honey samples contains 8700 total instances. Each instance is the result of spatial segmentation of the hyperspectral imagery. It contains 128 features - representing the spectral wavelengths of the hyperspectral camera. The classes of honey included in this dataset are:

- Blue Borage (BB)
- BorageField
- Clover
- Field+Tawari
- Honeydew
- Kamahi
- Manuka (This means Manuka honey that is not UMF rated so might not be pure)
- ManukaBlend (This means it is a blend of Manuka honey and other honey types)
- ManukaUMF10
- ManukaUMF12
- ManukaUMF13
- ManukaUMF15
- ManukaUMF18
- ManukaUMF20
- ManukaUMF22
- ManukaUMF5
- Multifloral (This is a mix of several different honey types)
- Pohutakawa (Pohu)
- Rata
- Rewarewa
- Tawari

5.1 Manuka Honey Subset

For the Manuka honey subset, we only include Manuka honey labelled with the unique Manuka factor (UMF) system. The Manuka and Manuka blend classes are therefore not included in this subset. The Manuka subset includes:

- ManukaUMF10
- ManukaUMF12
- ManukaUMF13
- ManukaUMF15
- ManukaUMF18
- ManukaUMF20
- ManukaUMF22
- ManukaUMF5

5.2 Attribute Description

The features from the hyperspectral images are the working wavelengths of the hyperspectral camera. Two additional attributes have been considered and can be useful for splitting the dataset into training and testing sets, as well as testing the generalisation ability of the algorithms.

The 'Brand' of honey represents the manufacturer that has supplied the honey. This attribute is included because it can be useful to test if a system can classify all honey types within a brand against each other. It can also be useful when developing general systems to check if we can exclude a brand from the training set and still have a good performance with the testing set. The brands have been anonymised for confidentiality reasons, the brand labels have been renamed as $C1, C2, ..., C11$.

The 'Acquisition' attribute represents the different sampling of images for the same type and brand of honey. As portrayed in figure 1, for each unique jar of honey, there have been six samples taken and captured by the hyperspectral imaging system. Each image captured is numbered with an acquisition number between one and six. This attribute allows us to split the training and testing sets such that we obtain a balanced distribution of all the honey types. This also ensures that we do not have an instance in the testing set that comes from a segment included in the training set. For testing, we use acquisition number six, and for training, we use acquisitions one to five.

Finally, the class attribute indicates the class of honey, which is the botanical origin, and the UMF value if it is UMF rated Manuka honey. Botanical origins have a huge impact on the value of honey, where some types are precious such as pure Manuka honey, and others are much more common and not considered as valuable, such as multi-floral honey.

6 Existing work on the dataset

There has been work done on this database with the classical machine learning and statistical algorithms [1]-[5].

A comprehensive, comparable work is found in [5], where a testing accuracy of 90.52%, and cross-validation accuracy of 90.65% were achieved using a Class Embodiment Autoencoder and the KNN classifier for the full dataset.

The Manuka honey subset has not had any work that can be directly compared to, as the work on Manuka honey reported in [4] has focussed on splitting the dataset based on the brand of honey.

References

[1] A. Noviyanto, "Honey botanical origin classification using hyperspectral imaging and machine learning," Ph.D. dissertation, The University of Auckland, 2018.

[2] A. Noviyanto and W. H. Abdulla, "Honey dataset standard using hyperspectral imaging for machine learning problems," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 473–477.

[3] ——, "Segmentation and calibration of hyperspectral imaging for honey analysis," *Computers and Electronics in Agriculture*, vol. 159, pp. 129–139, 2019.

[4] ——, "Honey botanical origin classification using hyperspectral imaging and machine learning," *Journal of Food Engineering*, vol. 265, p. 109684, 2020.

[5] T. Phillips and W. Abdulla, "Class embodiment autoencoder (ceae) for classifying the botanical origins of honey," in *Image and Vision Computing New Zealand (IVCNZ)*, 2019.