

# From measurement to material – Preparing hyperspectral signatures for classification

Jennifer Walocha<sup>1</sup> and Matthias Richter<sup>1,2</sup>

<sup>1</sup> Karlsruhe Institute of Technology, Institute for Anthropomatics and Robotics  
Adenauerring 4, D-76131 Karlsruhe, Germany

<sup>2</sup> Fraunhofer Institute of Optronics, System Technologies and Image  
Exploitation (IOSB), Fraunhoferstr. 1, D-76131 Karlsruhe

**Abstract** Due to the possibility of classifying unknown materials fast and accurately the industries interest in spectroscopy is growing. However, reliable classification is a matter of suitable preprocessing. Existing solutions found in the literature are often very specific a particular combination of materials. In this paper we present a method to preprocesses hyperspectral data in order to enables general classification of many materials. The system is divided into five modules: selection, transformation, reduction, decorrelation and classification. We demonstrate our method in a demonstrator system that is available as both web- and standalone application.

## 1 Introduction

Classification of materials is prevalent in industry, especially in the field of quality assurance, but also finds application in other areas, e.g., in mining and food safety. Here the difficulty is to classify fast and accurately on the basis of the provided measurement.

Due to the continuing progress in sensor technology, companies are now able to afford improved optical sensors which generate images with increasing quality. Previously high-priced hyperspectral sensors, primarily used in remote sensing and meteorology, gradually replace conventional RGB- and multispectral cameras. The data from these systems is used to obtain highly detailed measurements.

However, raw or incorrectly processed measurements may prevent meaningful analysis of the data. Furthermore, statistical analysis and

machine learning techniques may become slow when the data is not properly processed. Normalization of the measurements is also needed to ensure that data from different sources (e.g. different camera sensors) is comparable.

In this paper, we propose a procedure to preprocess hyperspectral data in order to achieve comparability and to improve speed and accuracy of the subsequent classification stage. We measure great importance towards the possibility of classifying numerous spectral signatures from different materials corresponding to a very large number of classes. We demonstrate our methods with our software *QueryMe*.

In the following, we give a short overview of common procedures for classifying hyperspectral images and compare them to our method. In Section 2, we describe our method and the different steps taken in order to process the measurements. Acquisition and storage of hyperspectral signatures and our software *QueryMe* are briefly outlined in Section 3. Finally, we summarize our results and give an outlook towards further research in Section 4.

## 1.1 Related work

Several procedures for classifying hyperspectral images can be found in the scientific literature. However, most of these methods are only concerned with a very specific application scenario. For instance, Serranti et al. developed a hyperspectral imaging system utilizing the near infrared spectrum to detect polyolefin, polyethylene and polypropylene in building and construction waste [1]. In addition to the classification of a particular substance, objects are often compared with respect to specific quality criteria, for example, to measure the ripeness or health of fruits and vegetables, like tomatoes [2] and mushrooms [3]. In many applications only a small subset of problem-specific wavelengths is selected. For instance, Elmasry et al. employed partial least squares regression and stepwise discrimination analysis to select three discriminative wavelengths with the goal to detect developing bruises on McIntosh apples [4]. Liu et al. processed hyperspectral images with a Gabor filter bank to assess the quality of pork meat [5]. Using this technique, they achieved perfect classification in their experiments.

All these methods are engineered to very specific needs. However, in this paper we consider a much larger problem domain: classification

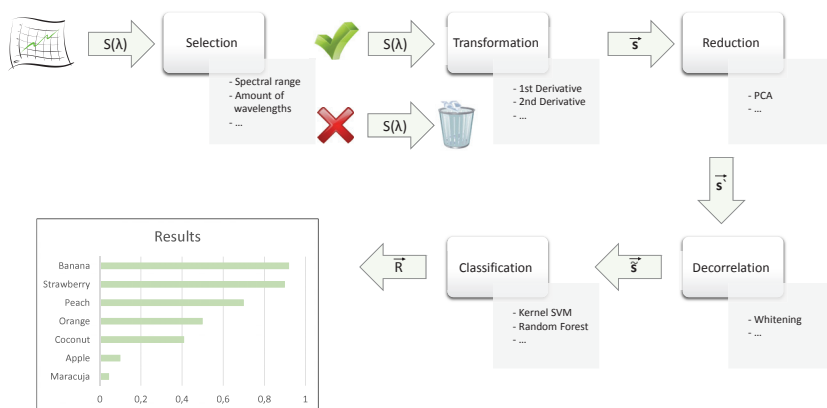
of many (order of hundreds) materials. Therefore, such methods will likely not provide the expected classification performance.

## 1.2 Contributions

In this paper, we present a system to enable general classification of hyperspectral data without regard to a specific problem domain. In our analysis, we found it beneficial to divide the process into separate functional blocks, each with a dedicated task: selection, transformation, reduction, decorrelation and classification. We analyzed these components and give recommendations for concrete methods that we find most useful. Moreover, with *QueryMe* we provide a prototypical implementation of our methods. The software is implemented as a library; we provide access to the functionality in form of both a web- and a stand-alone application.

## 2 Methods

The methods proposed in this paper are applicable to different types of measurements. The only requirement is that all measurements represent the same type of information, e.g. all measurements represent



**Figure 11.1:** Overview of the preprocessing and classification steps.

absorbance-spectra or all measurements represent reflection spectra. Furthermore, we assume that the data is normalized to a common baseline. That is the influence of the sensor, lighting, etc. are removed from the measurements (see also Sec. 3).

We define a spectrum to be a mapping

$$\mathcal{S} : \mathbb{R}_+ \rightarrow [0, 1] \lambda \mapsto \mathcal{S}(\lambda) = I, \quad (11.1)$$

where  $\lambda$  denotes a wavelength and  $\mathcal{S}(\lambda) = I$  denotes the measured normalized intensity at that wavelength. In practice, a measurement is defined only on certain sampling points  $\lambda_i$ ,  $i = 1, \dots, N$  within an interval (or support)  $\mathcal{D} = [\lambda_{\text{low}}, \lambda_{\text{high}}]$ . We simply perform linear interpolation between the sampling points  $\lambda_i$ . Sometimes the sensor may contain dead or defective pixels that do not provide usable measurements. To handle these cases augment the definition of  $\mathcal{S}$  and assign a special value  $\mathcal{S}(\lambda) = \perp$ .

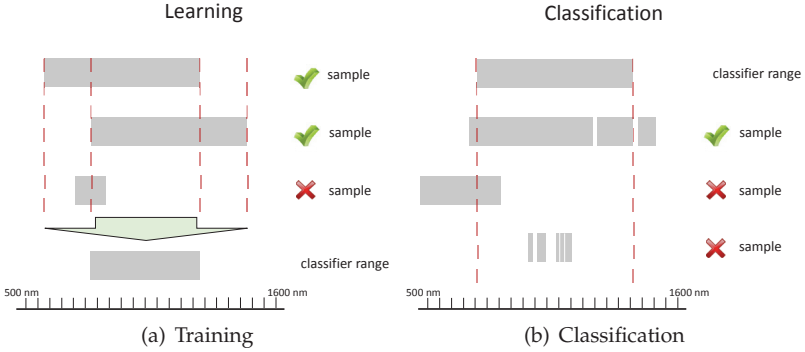
As seen in Figure 11.1, we identified four major preprocessing steps a measurement has to pass before entering the classification (or learning) stage. The first step, selection, either accepts a measurement to further processing or drops it if it is not suitable for classification. If accepted, a measurement undergoes a transformation stage where discriminative regions in the spectrum are accentuated and a feature vector is derived. This is followed by a feature reduction, where the dimensionality of the feature vector is reduced and redundant information is removed. Finally, the feature vectors are decorrelated to reduce runtime and increase classification performance of certain machine learning algorithms. Each step is explained in more detail in the following.

## 2.1 Selection

The selection stage determines whether to accept or deny an unknown spectrum. In the learning phase, this is to ensure that all training data is valid. Here a spectrum is only accepted if the spectrum's support matches or exceeds a user-definable range that the classifier should be trained on. Spectra that are defined over a larger support are truncated to match the desired range (see Fig. 11.2(a)).

In classification, it ensures that the classifier is able to correctly classify the measurement. Here it is verified that the classifier was learned

using spectra that have the same or smaller support than  $\mathcal{S}(\lambda)$ . If the support is smaller, the spectrum is truncated to the range requested by the classifier (see Fig. 11.2(b)).



**Figure 11.2:** Selection of spectra based on the support.

In both cases, only spectra that contain no invalid values, i.e. only spectra with  $\mathcal{S}(\lambda) \neq \perp$  for all  $\lambda \in \mathcal{D}$ , are allowed. Other “sanity checks” are also possible, e.g., whether the measurement contains a minimum number of sample points or if the maximum intensity of the spectrum exceeds a threshold.

## 2.2 Transformation

To accentuate the discriminative parts in a spectrum, it is processed by function  $\mathcal{T}(\mathcal{S}(\lambda)) = \mathcal{S}'(\lambda)$ , where  $\mathcal{T}$  is an arbitrary transformation.

In our case, we chose the first derivative

$$\mathcal{T}(\mathcal{S}(\lambda)) = \frac{d\mathcal{S}}{d\lambda}(\lambda) \quad (11.2)$$

to emphasize sudden changes in the spectral signature. The drawback is that this also increases the influence of additive noise and therefore decreases the signal to noise ratio. Other possible transformations include power-normalization  $\mathcal{T}(I) = \text{sgn}I|I|^\alpha$  or normalization by mean and standard deviation over the intensities,  $\mathcal{T}(I) = (I - \bar{I})/s_I$ . After normalization, the spectrum is sampled into a feature vector

$\mathbf{s} = (s_1, \dots, s_K)^\top$  of size  $K$ . Each entry  $s_k = \mathcal{S}'(\lambda_k)$  corresponds to a value of the transformed spectrum, where the sample points  $\lambda_k$  are evenly spaced over the support.

## 2.3 Reduction

In the reduction step, the dimensionality of the feature vector is reduced and redundant information is eliminated resulting in a lower-dimensional feature vector  $\mathcal{R}(\mathbf{s}) = \mathbf{s}' \in \mathbb{R}^D$ , with  $D < K$ . Common choices for  $\mathcal{R}$  include principal component analysis, partial least squares regression or feature selection methods.

## 2.4 Decorrelation

Finally, the feature vector  $\mathbf{s}'$  is decorrelated through whitening,  $\tilde{\mathbf{s}} = W\mathbf{s}'$ , where  $W$  is the whitening matrix. This is done in order to speed up the subsequent machine learning algorithms and increase the performance of certain classification methods. If the machine learning algorithm is not sensitive to correlated features, this step may be skipped.

## 2.5 Classification

Using the processed features, we train one kernel SVM (RBF-kernel) for each material in a one-vs-all scheme. In classification, this allows to reduce the number of evaluated classifiers by exploiting hierarchical structures in the database (see [6]): The user may select certain groups of materials that the unknown material likely belongs to, thereby reducing the possibilities for misclassification.

# 3 Implementation

In this section, we detail the implementation of our methods explained in Section 2. Prior to this, we summarize our procedure of acquisition and normalization as well as the storage of hyperspectral images.

### 3.1 Acquisition

A general solution for the classification problem bears the difficulty not only to classify different kinds of materials but also to classify measurements acquired by various sensors in many places. For this reason measurements are normalized to a common reference frame and stored in our hierarchical database of hyperspectral material signatures in order to be available at all times. As a first step, the measurements need to be normalized to a common reference “white” to remove the influence of lighting and varying dynamic range of the different sensors. Moreover, dark current in the sensors manifests as noise, which also has to be removed from the measurements. To achieve both, we follow the approach by Irgenfried and Negara [6], which we briefly outline in the following.

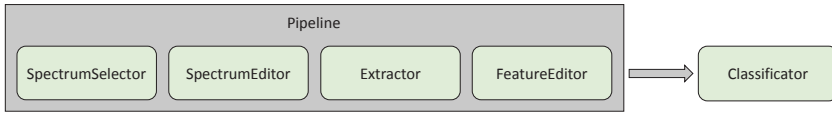
Each time before a measurement is acquired, a mean dark current spectrum  $D(\lambda)$  is determined by taking a measurement with a closed shutter. Then, the spectrum of a white thin teflon tile is measured to obtain a reference white spectrum  $W(\lambda)$ . Finally, the raw measurement  $I_{\text{raw}}(\lambda)$  is recorded and the final, normalized measurement is computed according to

$$I(\lambda) = \frac{I_{\text{raw}}(\lambda) - D(\lambda)}{W(\lambda) - D(\lambda)}. \quad (11.3)$$

Afterwards, the foreground of the image is separated from the background in a semi-automatic segmentation step. Finally, the resulting meta-data is imported into the relational database while the measurements, segmentation and registration files are kept in compact binary data files and are accessed through a common data access layer. This composition allows high performance with a huge amount of data stored. Moreover, as mentioned in Section 2, the measurements are organized into material groups which is used to restrict the number of classifiers to evaluate in the classification stage.

### 3.2 QueryMe

QueryMe implements our methods mentioned in Section 2 in a pipeline model (see Fig. 11.3). A separate pipeline is created for the training



**Figure 11.3:** Overview of the processing pipeline used for both learning and classification.

and classifying phases. In both pipelines each processing step is implemented in a modular fashion and can be exchanged or even entirely omitted. The modules are named *SpectrumSelector*, *SpectrumEditor*, *Extractor*, *FeatureEditor* and *Classifier* and correspond to the processing steps detailed in Section 2. The system allows to introduce further modules at will, e.g. to normalize measurements from different sources. Moreover, the algorithms of those modules are exchangeable so that the whole system is not rigid and different kinds of mathematical methods may be applied. This is most useful in the final module, *Classifier*, as the optimal solution for this module is not known at the present. The entire pipeline can be serialized and saved in a single so that a user may swap several, specialized pipelines depending on the tasks.

## 4 Conclusion

In this paper, we presented a procedure to preprocess hyperspectral measurements in order to facilitate classification of different materials. Moreover, we implemented our method as a demonstration system that is available as web- or standalone application.

However, there is still room for improvement. Once learned, a classifier can not be updated and must therefore be trained from scratch when the database is updated. Recent developments in online-learning methods offer solutions to that aspect.

We would also like to investigate which preprocessing steps are most suited in the context of classification with hyperspectral datasets, especially with regard to the reduction step.

Finally, it would be interesting to adapt the existing method to support not only classification of materials, but also chemometric regression techniques, to e.g. grade ripeness of fruits, and spectral unmixing, to discover the composition of unknown objects.



## References

1. S. Serranti, A. Gargiulo, and G. Bonifazi, "Classification of polyolefins from building and construction waste using nir hyperspectral imaging system," *Resources, Conservation and Recycling*, vol. 61, pp. 52–58, 2012.
2. G. Polder, G. W. A. M. van der Heijden, I. T. Young, "Hyperspectral image analysis for measuring ripeness of tomatoes." in *2000 ASAE Annual International Meeting*, ASAE, Ed., vol. 003089, 2000.
3. M. Taghizadeh, A. A. Gowen, and C. P. O'Donnell, "The potential of visible-near infrared hyperspectral imaging to discriminate between casing soil, enzymatic browning and undamaged tissue on mushroom (*agaricus bisporus*) surfaces," *Computers and Electronics in Agriculture*, vol. 77, no. 1, pp. 74–80, 2011.
4. G. Elmasry, N. Wang, C. Vigneault, J. Qiao, and A. ElSayed, "Early detection of apple bruises on different background colors using hyperspectral imaging," *LWT - Food Science and Technology*, vol. 41, no. 2, pp. 337–345, 2008.
5. L. Liu, M. O. Ngadi, S. O. Prasher, and C. Gariépy, "Categorization of pork quality using gabor filter-based hyperspectral imaging technology," *Journal of Food Engineering*, vol. 99, no. 3, pp. 284–293, 2010.
6. S. Irgenfried and C. Negara, "A framework for storage, visualization and analysis of multispectral data," in *OCM 2013 - Optical Characterization of Materials*, 2013, pp. 203–214.