HLB DETECTION USING HYPERSPECTRAL RADIOMETRY

Jose Gonzalez-Mora, Carlos Vallespi, Cristian S. Dima

National Robotics Engineering Center
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Reza Ehsani

Citrus Research and Education Center
University of Florida
Lake Alfred, Florida

ABSTRACT

The need for sustainable agriculture requires the adoption of low input, long-term and cost-effective strategies to overcome the adverse impact of disease and nutritional deficiencies on citrus groves. In this context, early detection of diseased trees has become an important topic in the citrus industry. Multiple factors make field assessment of disease conditions a challenging task: the non-specific nature of many symptoms, the possibility of having localized affections in only certain areas of the tree or the correlation with other factors such as tree age. In this paper we investigate hyperspectral sensing as an effective approach to detect the Huanglongbing disease in citrus trees. We analyze the visible and near infrared spectral responses from the leaves to discriminate infected trees from healthy samples. The accuracy of the diagnosis is improved by means of feature selection techniques that prevent overfitting problems due to the high dimensionality and collinearities in the data. We provide experimental results illustrating the performance of the proposed techniques using data collected in the field.

Keywords: Disease detection, Huanglongbing, citrus greening, hyperspectral sensing, feature selection

INTRODUCTION

Detecting plant health conditions plays a key role in crop protection and farm pest management. Large amounts of resources are used every year to control various diseases common to citrus crops. This involves the extensive use of fungicides and pesticides, which evokes serious concerns over deteriorating groundwater quality and over the high costs involved and the consequent profit loss. Furthermore, for some diseases the early detection of the disease is even more important to avoid the extension of the pathogen.

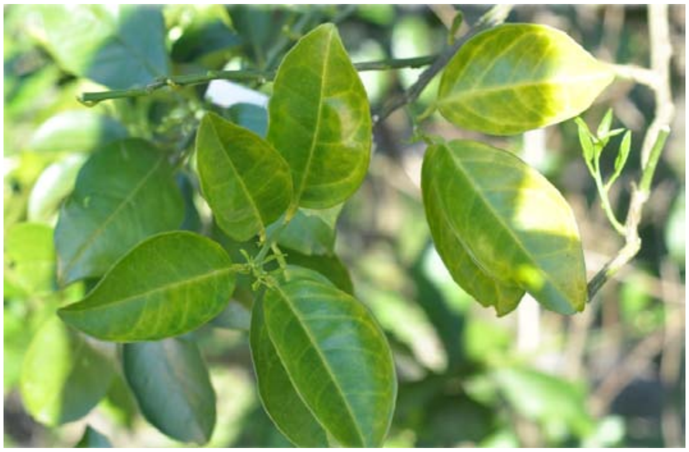


Fig. 1. Orange tree branch containing visible symptoms of Huanglongbing

Although there are several disorders that affect citrus crops, Huanglongbing (HLB), also known as citrus greening, is considered one of the most devastating diseases (Floyd and Krass, 2006). It is a severe disease that threatens the citrus industry due to the non-specific nature of its symptoms and the fact that there is no treatment or prevention at this time that can completely eradicate the disease in the infected areas. The disease is caused by several species of the genus Candidatus Liberibacter [Zhao, 1981; Da Graca and Korsten, 2004]. The bacteria have not been cultured in the laboratory and do not survive outside the host cells, making them difficult to study.

Visually, the most characteristic symptom in leaves is vein yellowing or a blotchy mottling of all or part of the leaf, resulting in an overall yellow appearance (see Fig. 1). However, field assessment based on these characteristics is difficult since they resemble other diseases (such as stubborn disease and tristeza) and nutritional deficiencies (i.e. zinc-like deficiency). Early symptoms of HLB include a yellowing of only one limb or sector of the tree canopy. The disease is more difficult to detect in older trees that lack vigor or are under stress due to other problems. Chronically infected trees display extensive limb dieback, tend to drop fruit prematurely and are sparsely foliated with small leaves that point upward. HLB-infected fruit are frequently small, underdeveloped, and misshapen, with curved columella and aborted seeds. They tend to remain green at least in part, and, unlike healthy fruit that color up from the stylar end, coloring starts at the stem (peduncle) end. The juice is high in acid, and abnormally bitter, rendering the fruit inedible.

Diseased trees can be identified as suspect in the field by their foliar and fruit symptoms but a Polymer Chain Reaction (PCR) test is required to verify HLB in the laboratory (Floyd and Krass, 2006). Since its first introduction in 1983 by the two Nobel prize winners M. Smith and K.B. Bullis, PCR has become a powerful technique for detection and identification of plant pathogens. Although PCR methods are sensitive and specific, consistent detection of HLB pathogens in infected plants is generally thought to be problematic, presumably because of the low concentration and the uneven distribution of the pathogens in host plants

(McClean 1970). Consequently, molecular detection protocols have generally been limited to the confirmation of visible infections.

The current emphasis is on avoiding further infections, which makes detection of the disease in early stages critical. Technologies assisting in early HLB identification could have a significant positive impact in managing the disease. Optical sensing technologies (e.g. machine vision and spectroscopy) have shown great potential for food safety, quality evaluation and disease detection in crops. One of the core technologies that have proven to be effective is hyperspectral radiometry.

Spectral reflectance properties of leaves in the visible and near-infrared (VNIR) bands have been shown to be highly correlated with their chemical composition. Researchers have reported connections between spectral signatures and chlorophyll concentration, plant stress or crops diseases (Lu and Chen, 1998; Carter and Knapp, 2001; Keulemans et al., 2007; Liu et al., 2007; Mishra et al., 2007).

Methods for processing and analyzing hyperspectral reflectance spectrometry have to accommodate data that are high-dimensional, with a number of wavelengths that typically exceeds by far the number of available samples, and exhibit a high degree of interband correlation leading to excessive data redundancy and poor generalization of the results. One way to mitigate this problem is to use nonparametric methods such as neural networks, because they are able to learn complex decision boundaries that are difficult to capture with parametric methods in the classification of high-dimensional data. An alternative is to seek lower dimensional representations of the data by selecting only those spectral bands useful for the specific problem being considered. Preprocessing the spectral data so the number of bands is less than the number of samples is a commonly adopted approach for many methods (Jain and Zongker, 1997; Doak, 1992; Schmidth et al., 2004). To this end, we analyze different feature selection techniques to handle the ill-posed nature of the reflectance spectrometry.

In this paper we describe a procedure for collecting hyperspectral data in field conditions using a VNIR radiometer. Several problems which occurred during the data capture are discussed, along with the methods used to tackle them. We estimate the performance of a logistic regression classifier for HLB classification at the leaf level and finally, we evaluate different feature selection algorithms to detect the most discriminative wavelengths for disease detection.

The outline of the paper is as follows. First, we describe previous work in the areas of hyperspectral radiometry and disease detection. Second, we present the methodology used for data collection and the dataset analyzed in our experiments. We then introduce our analysis on VNIR spectra classification and feature selection for HLB detection. Finally, we compile some conclusions and future work.

PREVIOUS WORK

The structure and physiological status of a plant is expressed in its reflectance pattern. Incident light is partly reflected by the plant and the amount of reflected light depends on different factors, such as pigment concentration and internal organization of biochemical elements, the external leaf morphology or its internal structure. The reflectance spectra of leaves and fruits undergo remarkable changes under deficiency of mineral nutrition, different stress conditions or pollution, during adaptation to variable solar irradiation, and in the course of senescence.

Traditional techniques for plant analysis in physiological and biochemical studies usually involve wet chemical methods requiring the destruction of the tissue. They are time consuming and can be affected by different artifacts due to impurities in the tissue extracts, incomplete pigment extraction or instability of the components. The application of nondestructive optical methods is becoming a popular alternative since they allow rapid measurements on a large number of samples, which thereafter remain intact and could be used for further analysis. Recently, commercially available reflectometers suitable for field measurements from plants are designed, providing reliable spectral data.

A plant leaf represents a complex optical system. It consists of several structures with different refraction indices that contain high amounts of pigments. The development of nondestructive techniques for plant analysis requires the understanding of their in vivo spectroscopy, localization of pigments in leaves and the structure and patterns of their changes during physiological processes in plants. Although detailed investigations of leaf optical properties have appeared in the literature (Vogelmann, 1993), many extended approaches for quantitative pigment analysis in situ consider the leaf as a “black box”. These techniques conduct a supervised analysis correlating sensed properties of the leaf (i.e. spectra) with some categorization of its properties (i.e. pigment level or health status).

The reflectance spectrum of a leaf is determined to a great extend by absorptions due to water and pigments. Different authors have shown how visible and near infrared (VNIR) spectrum of a leaf contains information on leaf moisture content, plant pigment concentration and leaf cellular structure.

Carter, 1991 analyzed different effects of leaf *water content* on reflectance. The primary effect reported is a decrease of the reflectance from approximately 1300 to 2500 nm due to the strong water absorption at these wavelengths. Between approximately 700 and 1300 nm, absorption by water is relatively weak and, in general, leaves do not contain other substances with elevated absorption in this range producing a higher reflectance in this area of the spectrum. Throughout the visible spectrum, the water absorption is much weaker than in the infrared, but chlorophyll and accessory pigments absorb strongly between 400 and 700 nm resulting in a typical low diffuse reflectance in this range. When water is lost from a leaf, absorption decreases and consequently reflectance tends to increase in the 1300-2500 nm range. However, Carter observed that reflectance also increases in the 400-1300 nm range. This secondary effect was explained as an influence of water content on absorption by other substances in the leaf, such as pigments and on wavelength-independent processes, particularly multiple reflections inside the leaf.

The absolute concentrations of pigments as well as their ratios are also important properties of the leaf, whole plants and plant communities. There are changes in the pigment content in the course of plant growth, during adaptation to unfavorable environmental conditions and under various stress conditions, damages and diseases. Both qualitative and quantitative changes in pigment content of plants are reflected in the tissue optical properties. Merzlyak et al., 2003 presented different algorithms for pigment analysis using visible and near infrared remote sensing. The content of chlorophylls, the dominant pigment of green leaves, determines to a great extent the amount of Photosynthetically Active Radiation (PAR) absorbed by the leaf, the photosynthetic rate and plant productivity. Carotenoids are involved in light harvesting and other physiologically important functions, preventing, via several mechanisms, the damages to plants caused by excessive fluxes of visible radiation. Experimental results show that the 550 and 700 nm wavelengths are highly sensitive to Chlorophyll content. Carter and Knapp, 2001 correlated physiological stress conditions with the optical response in plants finding consistent changes in the reflectance patterns in the green-yellow spectrum and near the 700 nm. This common optical response was connected with the reduction of the chlorophyll concentration in leaves.

The shift of the red edge in the reflection spectra of vegetation (reflectance between 680-760 nm) is a known phenomenon indicating changes in the biological status of plants. Boochs et al., 1990 analyzed the variability of the reflectance in this area of the spectrum concluding that the red edge is not fully described by the shift of the main inflection point. Alternatively they proposed a collection of several different features obtained from high resolution spectra which jointly considered can describe small differences in the chemical and morphological status of plants. One of the requirements for reliable algorithms of pigment analysis is their low sensibility to morphological-anatomical traits of plant tissues. Remarkably, many approaches for nondestructive pigment assessment are based on invariant spectral signatures that require knowledge of reflectance only at few certain wavelengths. This has been the basis for the development of *spectral indices* using reflectances corresponding to wavelengths with maximum and minimum sensitivity to variation in pigment concentration (Gitelson and Merzlyak, 1996). These indices can serve as indicators of stress, disease and senescence in different plants and crops.

Stress-induced changes (including dehydration, flooding, freezing, ozone, herbicides, competition, disease, insects, deficiencies and fertilization) affect the reflectance spectra of plants. Multiple non-intrusive remote sensing techniques at plant leaf-level have been described in the bibliography for detecting stress factors (Smith et al., 2004; Vogelmann, 1993; Carter, 1991; Gitelson and Merzlyak, 1996; Zarco-Tejada et al., 2004). They describe how changes on physiological properties of the leaf can alter the interaction of light with the foliar medium. The most common and widespread change occurs in the proportion of light absorbing pigments mentioned, most notably in the green peak (525-605nm) and along the red edge (750nm) .

Non-intrusive techniques are essential for capturing data in the continuous manner necessary for monitoring vegetative production systems. A number of hyperspectral sensing techniques have been studied to monitor contamination (Lu and Chen, 1998; Kim et al., 2001; Kim et al., 2002; Mahl et al., 2004), detect defects (Nagata et al., 2006; Ariana et al., 2006) and nutritional stress or diseases (Keulemans et al., 2007; Liu et al., 2007; Lee et al., 2008; Qin et al. 2009; Mishra et al., 2007). A key question to consider in these remote sensing applications is the analysis of how leaf spectra signatures are preserved at the canopy level or when the plant reflected spectra is sensed distorted with other background radiance (Borel and Gerstl, 1994; Zarco-Tejada, 2004).

In the particular case of citrus crops and *HLB* disease, Mishra et al., 2007 performed an analysis of the spectral characteristics of citrus greening showing the potential of hyperspectral spectroscopy to detect this disease.

Numerous studies demonstrate that hyperspectral reflectance and its correlation with the plant biochemical and biological status were affected by the autocorrelation and multicollinearity of the data due to the continuous wavebands. Many authors have reported the use of feature selection techniques to reduce the dimensionality of the dataset and tackle these problems. Liu et al., 2007 used different techniques involving model dimensionality reduction (stepwise regression, principal component regression and partial least squares regression) to the analysis of rice brown spot disease using hyperspectral reflectance. Renzullo et al., 2006 applied recent advances in regularized regression techniques to improve the results of discriminant analysis applied to hyperspectral data. In their article, they describe the use of Penalized Discriminant Analysis and a technique based on the Lasso regularization (OSLASSO) reporting better results than previous approaches.

DATA COLLECTION

This section describes the dataset used in the experiments presented in this paper and the data collection procedure.

The leaf samples were captured in a citrus grove in south-west Florida which contained trees infected with HLB. Visual surveys of the grove were conducted by field scouting crews with experience in recognizing HLB symptoms. Trees suspected of HLB infection were flagged with marking tape and were later re-examined by a highly qualified head scout who could confirm or reject the diagnosis.

The samples were chosen according to the following procedure:

- We selected 100 trees that were considered healthy during the inspections and collected a sample from each one of them.
- Similarly, we selected another 100 trees labeled as diseased by the human scouts and acquired 2 samples from each of them. One sample was picked from a spot where the HLB symptoms were visually identifiable. This spot was previously selected and flagged by the scouts since their diagnosis was based on the visual detection of these particular leaves in the canopy. The other sample was collected from a spot where the tree was asymptomatic.

Consequently, the data collection involved the selection of 300 samples in total. For each of these samples we captured the VNIR spectra and collected multiple leaves.

The spectral reflectance curves of the samples were captured using a SVC HR-1024 portable spectroradiometer (Spectra Vista Corporation, Poughkeepsie, New York). The spectral range of the SVC is 350-2500 nm with spectral resolution of 3.5 nm (350-1000nm), 9.5 nm (1000-1850nm) and 6.5 nm (1850-

2500nm). The field of view of the instrument is 4^{o} covering a rectangular area and it has a minimum integration time of 1ms. Five different spectra were captured per sample. The instrument was calibrated for dark current at the beginning of the session and for white reference normalization before capturing each sample.

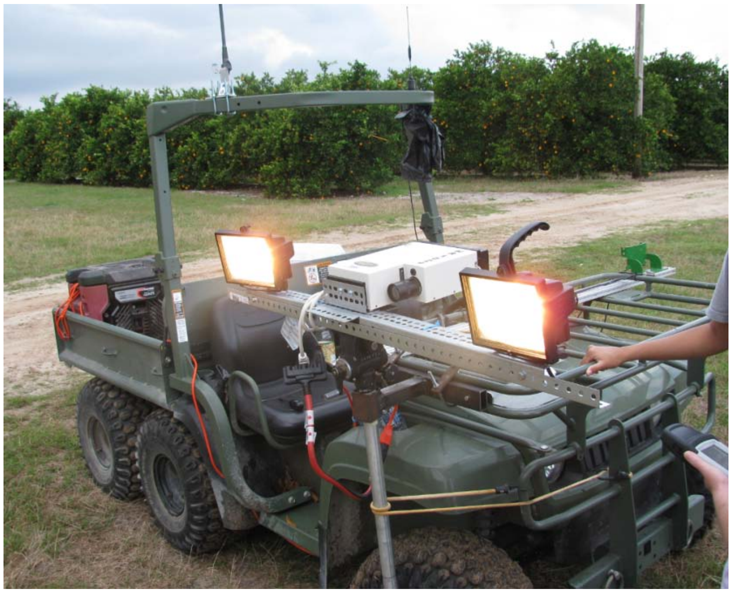


Fig. 2. Data capturing system including hyperspectral radiometer and artificial illumination

The SVC data was captured in the field under natural illumination. We intended to have a realistic “in-field” configuration with varying angles in the incident light and different relative orientation between the leaves and the sensor. In many cases the most representative symptomatic leaves were located on inaccessible branches, and in those situations the samples were removed and temporally placed in more accessible locations in the canopy. The data was collected during the day, at different times ranging from 8am to 5pm. The distance between the sensor and the leaves was set to approximately 0.75m. Different noise factors were involved in the data capturing process: changing ambient lighting conditions, background reflectance that is collected in the sensor FOV due to the irregular contour of the leaves and movement of the branches leading to misalignments between the sensor and the target leaves.

To minimize some of the noise factors, we tried to increment the signal to noise ratio (SNR) in the captured reflectance by complementing the natural illumination with artificial light sources (two 500W halogen flood lights). We attempted to prevent misalignments during the spectra acquisition by using a rigid support for the radiometer which was attached to a vehicle for easy transportation between different trees (see Fig. 2).

The 300 samples were later analyzed for HLB infection in the laboratory using real-time PCR (Li et al., 2006), and the results are presented in Table 1. A sample was considered to test positive for HLB if it produced a FAM CT value of 30 or less (Irey et al., 2006). It is important to indicate that due to scheduling constraints the data was collected in late February, a time of the year which is known to be suboptimal for PCR analysis for citrus greening. As a result, the chances of obtaining false negative results from PCR are much higher than during more favorable months (starting in August).

	Symptomatic	**Asymptomatic**	**Healthy**
Total count	100	100	100
Count with PCR <30 (HLB positive)	89	24	0
Count with PCR >30 (HLB negative)	11	76	100

Table 1: The three columns correspond to leaves that presented visual greening symptoms and were in an HLB-tagged tree (symptomatic), leaves that were asymptomatic but were in an HLB-tagged tree (asymptomatic), and leaves that looked healthy and were not in an HLB-tagged tree (healthy). For each column we indicate how many examples were identified as HLB-positive by PCR.

This bias towards negative PCR HLB results is noticeable in the symptomatic column (where 11% of the leaves are not confirmed as infected although several scouts confirmed the diagnosis). Unfortunately, this limits the conclusions what could be drawn from any analysis of the asymptomatic samples, as we will see in our experimental results section.

SPECTRA CLASSIFICATION AND WAVELENGTH SELECTION

In this section we analyze different classification and feature selection techniques for the identification of leaves from HLB infected trees using hyperspectral radiometry captures.

We first consider *Logistic Regression* classification as an initial approach for the analysis of the discriminative capabilities of hyperspectral reflectance when discerning healthy and diseased leaves. This is a simple classification algorithm based on a generalized linear model that is frequently used to provide a benchmark for more sophisticated methods. Because of its simplicity, it rarely overfits the training data (assuming that the training dataset has enough samples) and it has the advantage of being fast to train.

If we consider that reflectance samples *x* are distributed in the "healthy" and "diseased" classes denoted by C_h and C_d respectively, the posterior probability for the "disease class" C_d can be written as:

$$p(C_d|x) = \frac{p(x|C_d)p(C_d)}{p(x|C_d)p(C_d) + p(x|C_h)p(C_h)} = \frac{1}{1 + exp(-a)} = \sigma(a)$$

where we have defined

$$a = ln\frac{p(x|C_d)p(C_d)}{p(x|C_h)p(C_h)}$$

and $\sigma(a)$ is the *logistic sigmoid function* defined by

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

When the class conditional probabilities $p(x|C_d)$ and $p(x|C_h)$ are Gaussian with common covariance matrix Σ and means μ_d and μ_h, we have

$$a = w^T\, x + w_0$$

where w and w_0 are the coefficients and the bias of the linear model and are defined as

$$w = \Sigma^{-1}(\mu_d - \mu_h)$$

$$w_0 = -\frac{1}{2}\mu_d{}^T\Sigma^{-1}\mu_d + \frac{1}{2}\mu_h{}^T\Sigma^{-1}\mu_h + \ln\frac{p(C_d)}{p(C_h)}$$

Although the real distribution for the reflectance spectra will not obey this Gaussian form, we can consider it a good simplified model from where to start drawing preliminary conclusions. In the general case, defining the optimal classifier will involve determining a maximum likelihood estimate for the parameters w and w_0 by minimizing the following cost function:

$$E(w,w_0) = -\ln\big(p(t|w,w_0)\big)$$

$$= -\sum_{n=1}^{N}\Big(t_n \ln\big(p(C_d|x_n)\big) + (1-t_n)ln(1-p(C_d|x_n))\Big)$$

where t_n are the disease labels (i.e. $t_n=\{0,1\}$, 0 denoting healthy and 1 diseased) associated to each of the reflectance samples x_n for $n=1\ldots N$.

If we use all the spectral information available for the reflectance samples (i.e. 989 values for the different wavelength bands between 350 and 2500 nm) the resulting classifier trained from this dataset offers poor generalization. This is a consequence of the small ratio between number of samples and number of features (the “curse of dimensionality”) that has been mentioned in previous sections (Baum and Haussler, 1989). Numerous studies have shown that hyperspectral reflectance and its accuracy in the detection of plant biochemicals were affected by the multicollinearity and autocorrelation of the data due to the continuous wavebands (Liu et al., 2007). Since collecting a large number of training samples is difficult in practice, a commonly used approach is to reduce the number of features used for classifying each sample. Two possible approaches can be used:

- *Feature extraction* algorithms perform either a linear or a nonlinear mapping of the original features into a space with lower dimensionality. The new features can each be a function of all the original features.
- *Feature selection* algorithms select a subset of the original features that can be used to discriminate between the classes of interest.

In our disease detection application, feature selection algorithms have special relevance. A fast and inexpensive multispectral inspection system can be fabricated using optical filters if our algorithm uses only a reduced number of spectral bands. Although several reviews of feature selection methods have been published, no algorithm is clearly superior for high dimensional data (Doak, 1992;

Jain and Zongker, 1997).

The problem of obtaining the optimal features for a linear classifier is a NP complete problem and is computationally unsolvable in most large applications (Cover and Campenhout, 1977; Wenton et al., 2003). A popular strategy is to use a continuous, convex relaxation of the non-convex feature selection problem using a regularizer that encourages the sparsity in the model. In particular, in this paper we make use of a *L1 regularization* term in a Logistic Regression model to incorporate feature selection in the parameters estimation process (Weston et al., 2003; Schmidt et al., 2007). Incorporating the regularization term, the cost function is transformed into:

$$L(w, w_0) = E(w, w_0) + \lambda |w_0| + \lambda \|w\|_1$$

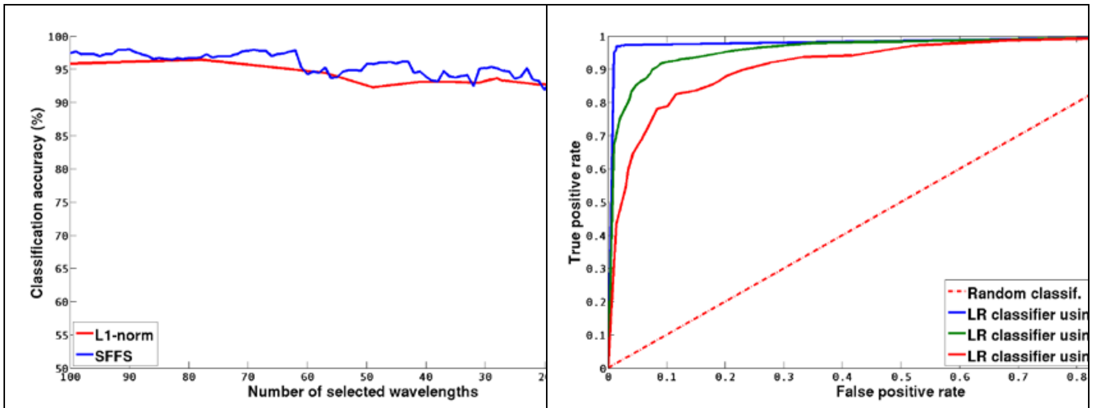
This loss function minimization with a *L1*-penalty term yields a sparse solution with most of the coefficients in *w* being zero. The remaining non-null coefficients will correspond to the selected features.

EXPERIMENTAL RESULTS

This section presents experimental results obtained with the described techniques for HLB detection in citrus. The primary goals of our experiments were the following:

- Compare the accuracy of hyperspectral based methods to the labels produced by human scouts
- Compare the accuracy of hyperspectral based methods to the labels produced by PCR analysis.

The first experiment was performed using visually distinguishable health conditions, as identified by human scouts. This should normally be the most favorable detection scenario, since visual symptoms appear at later stages of the disease. Figure 3(a) presents the accuracy obtained by using a logistic regression classifier when using different numbers of hyperspectral bands. We use the feature selection method based on regularization logistic regression presented in the previous section to choose between 1 and 100 wavelengths. For comparison purposes we also show the classification accuracies obtained when the wavelengths are selected using the *Sequential Floating Forward Selection* (SFFS) method measuring the Mahalanobis distance (Jain and Zongker, 1997). We use n-fold cross validation (*n=50*) to determine the classification accuracy in mutually exclusive training and testing sets. The figure illustrates how the classifier accuracy for the testing set is similar in both feature selection methods. This is a useful sanity check since the two techniques are based on the supposition of having a Gaussian distribution for the considered classes. Figure 3(a) indicates that the number of bands can be reduced to approximately 10 while maintaining a classification accuracy rate of approximately 90%.



(a) (b)

Fig. 3. Classification using symptomatic vs. healthy leaves (labeling based on scouting decisions). (a) Classifier accuracy as a function of the number of selected wavelengths. (b) ROC curves corresponding to classifiers trained with 6, 10 and 100 different bands.

By varying the threshold used on the outputs of a logistic regression classifier one can choose the true positive/false negative ratio according to the desired performance of the system. The Receiver Operating Characteristic (ROC) curves (Fawcett, 2004) represented in Figure 3(b) illustrate this mechanism: it is possible to detect a high number of diseased samples if we are willing to accept an increase in the number of healthy leaves that are incorrectly recognized as diseased. We can see how in the case of using 100 wavelengths the classifier performance approximates an ideal curve in which it appears to be possible to detect nearly all of the diseased cases with a very low false positive rate. Although this might look like a very positive result, we believe that this behavior is likely to indicate an overfitting problem: we are modeling a high dimensional space (100 wavelengths) using a reduced number of samples (100 per class). In order to achieve good generalization performance, a commonly accepted rule of thumb is that the number of training sample should be at least around 10 times the number of the feature dimensions. Notice that overfitting seems to occur although we are using cross-validation to estimate the generalization performance.

Figure 4 illustrates the classification performance obtained when we consider the PCR labels to be the ground truth value for the health status of a leaf. In this case we use all the leaf samples available (healthy, symptomatic and non-symptomatic leaves, as described in Table 1) to evaluate the classification results considering a limited number of spectral bands. The classification accuracy is lower than in the previous experiment. In this case the ROC curve for 100 wavelengths seems to point to a more realistic scenario in which a tradeoff between the rate of true positives (correct detections) and the false positive rate is required. The two possible explanations for this result are that the classification problem is intrinsically more challenging (since non-symptomatic leaves are harder to classify correctly), and the additional number of samples considered (the 100 non-symptomatic leaves) make overfitting slightly less likely.

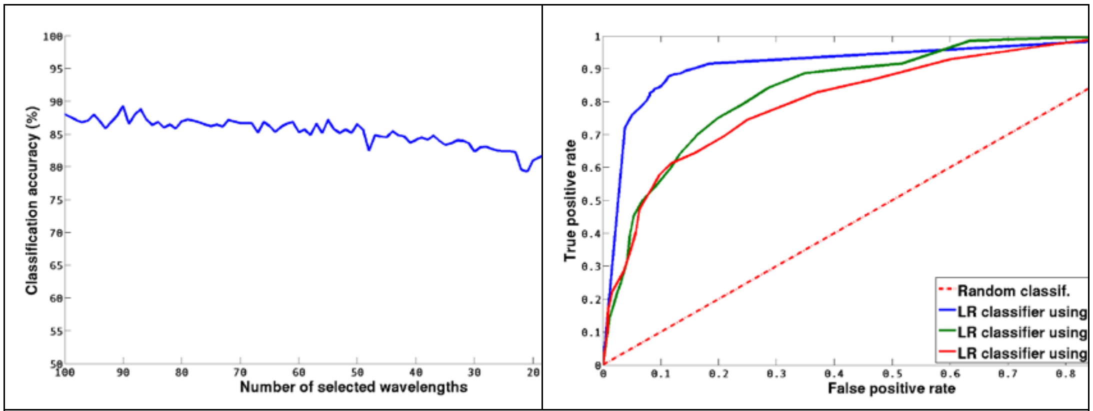


Fig. 4. Classification using PCR labels. Samples with CT values lower than 30 were considered HLB-positives while values higher than 30 were considered negative. (a) Classifier accuracy as a function of the number of selected wavelengths. (b) ROC curves corresponding to classifiers trained with 6, 10 and 100 different bands.

CONCLUSSIONS AND FUTURE WORK

In this paper we have presented preliminary results of using hyperspectral radiometry for the detection of HLB. We evaluate the logistic regression performance in determining the health condition of leaf samples. We have proposed the use of feature selection methods that can be translated into simple implementations based on multispectral technologies. We have illustrated how the use of L1-norm regularization makes it possible to formulate the feature selection process in terms of a criterion of sparseness that can be easily incorporated into the cost function, and we have observed that its performance is similar to the SFFS algorithm.

We have attempted to use data captured in somewhat realistic field conditions. Although the leaves had to be removed from their original locations in the tree so that they can be brought in the field of view of the sensor, the data was still affected by the different levels and directions of illumination encountered during a day. Our ROC curves indicate that a significant amount of discriminative power resides in the hyperspectral signatures we collected, even when using only 10 out of the 989 wavelengths produced by the SVC.

The most severe limitation of our results comes from the unfavorable timing of our hyperspectral data collection and PCR tests. Although our current results validate the feature selection approaches we considered and some of the classification results are promising, only very limited claims can be made given that the PCR results which are considered our ground-truth were obtain outside the time frame when PCR testing is considered reliable. The data set of three hundred samples we have collected is a bare minimum for the type of problem under consideration. In order to make strong claims about performance generalization potential across different data collection days, across different ways to sample leaves from the tree and the different stages of the disease a much larger data set needs to be collected. We intend to collect this additional data starting in August 2010, when both scouting and PCR methods will be able to produce much more reliable ground truth data.

With respect to the envisioned final application for the disease detection technologies we described –having an autonomous system that can scan trees--, it is important to notice that two challenges remain: accessing leaves with detectable traces of the disease, and bringing the classification performance to a level where the false-positive rates are at a manageable level despite the repeated testing that takes place. The current procedure of removing symptomatic leaves from the tree for placing them in front of the sensor simply bypasses the question of how one can devise a practical system for scanning a large enough percentage of the tree canopy. We believe that obtaining good classification results on asymptomatic leaves would confirm that our approach could work even if the leaves that we scan in an HLB infected tree are not the ones presenting visual symptoms. Designing a system concept that is focused on identifying a subset of high-risk trees that need to be inspected by human experts as opposed to generating a definitive diagnosis about which trees are infected is an approach that can lead to a useful application even if the false positive rates are too high for leaving the entire diagnosis up to an automated system.

On the positive side, the results we presented here were only based on a linear classifier (logistic regression) and on relatively simple raw hyperspectral features corrected for the ambient illumination by the SVC device. Having a larger dataset will allow us to consider more powerful classification algorithms.

One of the key lessons learned from our efforts in the area of hyperspectral sensing for disease detection was that obtaining good datasets that include both good sensor data and correct ground truth labels is time consuming and expensive. We believe research in this important area of precision agriculture is hampered by the lack of large publicly available datasets that can be shared among researchers for both comparing and reproducing technical results and for allowing research on data processing algorithms without the requirement of collecting large amounts of data in the field. Our modest contribution to the field in this area consists of making all the hyperspectral data we have collected freely available on our project's website[1].

ACKNOWLEDGEMENTS

The authors would like to acknowledge the generous advice and support provided by Tim Gast and Michael Irey from Southen Gardens Citrus regarding PCR testing and HLB scouting. The horticultural advice provided by Dr. Gene Albrigo and the field support coming from Dr. Reza Ehsani's research group at University of Florida's Citrus Research and Education Center (CREC) were greatly appreciated. This material is based upon research supported by USDA/CSREES under Award Number 2008-51180-04890.

REFERENCES

[1] Currently, our project's website is located at http://www.rec.ri.cmu.edu/usda

Ariana D.P., R. Lu, and D.E. Guyer. 2006. Near-infrared hyperspectral reflectance imaging for detection of bruises on pickling cucumbers. Comput. Electron. Agri. 53(1):60–70

Baum E. and D. Haussler. 1989. What size net gives valid generalization? Neural Comp., 1:151–160

Boochs F., G. Kupfer, K. Dockter, and W. Kuhbauch. 1990. Shape of the red edge as vitality indicator for trees. Int. J. Rem. Sens., 11:1741–1753.

Borel C.C. and S. A. W. Gerstl. 1994. Are leaf chemistry signatures preserved at the canopy level? Geoscience and Remote Sensing Symposium IGARSS ’94, p. 996–998

Carter G.A. 1991. Primary and secondary effects of water content in the spectral reflectance of leaves. Am. J. Bot., 78:916–924

Carter G.A. and A. K. Knapp. 2001. Leaf optical properties in higher trees: Linking spectral characteristics to stress and chlorophyll concentration. Am. J. Bot., 88(4):677–684

Cover T. and J. Campenhout. 1977. On the possible orderings in the measurement selection problem. IEEE Trans. System Man. and Cybernet., 7(9):657–661

Da Graca J. and L. Korsten. 2004. Citrus huanglongbing: Review, present status and future strategies. Dis. of Fruits and Veg., 1:229–245

Doak J. 1992. An evaluation of feature selection methods and their application to computer security. Technical Report CSE-92-18, University of California, Davis, Department of Computer Science

Floyd J. and C. Krass. 2006. New pest response guidelines: citrus greening disease. Technical report, USDA APHIS PPQ Emergency and Domestic Programs, Riverdale, Maryland. Avail. at http://www.aphis.usda.gov/import_export/plants/ppq_manuals.shtml

Fawcett T. 2004. ROC graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories, Palo Alto, CA, March 2004. Avail. at http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

Gitelson A. and M. Merzlyak. 1996. Signature analysis of leaf reflectance spectra: algorithm development for remote sensing of chlorophyll. J. of Plant Physiol., 148:495–500

Irey M.S., T. Gast, and T.R. Gottwald. 2006. Comparison of visual assessment and polymerase chain reaction assay testing to estimate the incidence of the huanglongbing pathogen in commercial Florida citrus. In Proc. Florida State Horti. Soc., 119:89–93

Jain A. and D. Zongker. 1997. Feature selection: evaluation, application and small sample performance. IEEE Trans PAMI, 19(2):153–158

Keulemans W., S. Delalieux, J. Aardt and P. Coppin. 2007. Detection of biotic stress (venturia inaequalis) in apple trees using hyperspectral analysis: non-parametric statistical approaches and physiological implications. Eur. J. of Ag., 27(1):130–143

Kim M.S., Y.R. Chen, and P.M. Mehl. 2001. Hyperspectral re?ectance and ?uorescence imaging system for food quality and safety. In Trans. of the ASAE, 44(3):721–729

Kim M.S., A.M. Lefcourt, K. Chao, Y.R. Chen, I.Kim, and D.E.Chan. 2002. Multispectral detection of fecal contamination on apples based on hyperspectral imagery part I: application of visible and near infrared reflectance imaging. In Trans. of the ASAE, 45(6):2027–2037

Lee W.S., R. Ehsani, and L. G. Albrigo. 2008. Citrus greening (huanglongbing) detection using aerial hyperspectral imaging. In 9th Int. Conf. on Prec. Ag.

Li W., J.S. Hartung, and L. Levy. 2006. Quantitative real-time PCR for detection and identi?cation of candidatus liberibecter species associated with citrus huanglongbing. J. Microbiology Methods

Liu Z., J. Huang, J. Shu, R. Tao, W. Zhou, and L. Zhang. 2007. Characterizing and estimating rice brown spot disease severity using stepwise regression, principal component regression and partial least-square regression. J. Zheijiang Univ. SCIENCE B, 8(10):738–744

Lu R. and Y.R. Chen. 1998. Hyperspectral imaging for safety inspection of food and agricultural products. SPIE, pages 121–133

Mahl P.M., Y. Chen, M. S. Kim, and D. E. Chan. 2004. Development of hyperspectral imaging technique for the detection of apple surface defects and contaminations. J. of Food Eng., 61:67–81

McClean A.P.D. 1970. Greening disease of sweet orange: its transmission in propagative parts and distribution in partially disease trees. Phytophylactica, 2:263–268

Merzlyak M., A. Gitelson, A. Chivkunova, and S. Pogosyan. 2003. Application of re?ectance spectroscopy for analysis of higher plant pigments. Rus. J. of pPlant Phys., 50:704–710

Mishra A.R., R. Ehsani, G. Abrigo, and W.S. Lee. 2007. Spectral characteristics of citrus greening (huangongbing). Trans. of ASABE. p. 1–11

Nagata M., J.G. Tallada and T. Kobayashi. 2006. Detection of bruises in strawberries by hyperspectral imaging. In Trans. of ASABE, 2006.

Qin J., T.F. Burks, M. A. Ritenour, and W.G. Bonn. 2009. Detection of citrus canker using hyperspectral re?ectance imaging with spectral information divergence. J. of Food Eng., 93:183–191

Renzullo L., A.L. Blanch?eld, and K.S. Powell. 2006. A method of wavelength selection and spectral discrimination of hyperspectral re?ectance spectrometry. In IEEE Trans. on Geo. and Rem. Sens., 44(7):1986–1994

Schmidt M., G. Flung, and R. Rosales. 2007. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In ECML Proc., 2007.

Smith K.L., M.D. Steven, and J.J. Colls. 2004. Use of hyperspectral derivative ratios in the red-edge region to identify plant stress responses to gas leaks. Rem. Sens. of Env., 92(2):207–217

Vogelmann T. 1993. Plant tissue optics. Ann. Rev. Plant Physiol and Mol. Bio., 44:916–924

Weston J., A. Elissee?, B. Scholkopf, and M. Tipping. 2003. Use of the zero norm with linear models and kernel methods. J Mach. Learning Res., 3:1439–1461

Zarco-Tejada P. , J. Miller, A. Morales, A. Berjon, and J Aguera. 2004. Hyperspectral indices and model simulation for chlorophyll estimation in open canopy tree crops. Rem. Sens. of Env., 90:463–476

Zhao X. 1981. Citrus yellow shoot disease(huanglongbing)-a review. Int. Soc. of Citricult., pages 466–469