SPECTRAL ANALYSIS AND MULTISPECTRAL/HYPERSPECTRAL IMAGING TO DETECT BLUEBERRY FRUIT MATURITY STAGES FOR EARLY BLUEBERRY YIELD ESTIMATION

By

CE YANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2013

© 2013 Ce Yang

To my family and friends

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor and committee chair Dr. Won Suk “Daniel” Lee, Professor of Agricultural and Biological Engineering, University of Florida (UFL) for his consistent mentoring and support throughout my research work at UF. His every conversation with me has been an excellent learning experience for me.

I am thoroughly grateful to my supervisory committee Dr. Thomas F. Burks, Associate Professor of Agricultural and Biological Engineering, UFL, Dr. Paul D. Gader, Professor of Computer and Information Science and Engineering, UFL, Dr. John Schueller, Professor of Mechanical and Aerospace Engineering, UFL and Dr. Jeffrey G. Williamson, Professor of Horticulture Sciences and Extension Horticulturist, UFL, for their valuable suggestions and insights to complete this dissertation. It was a privilege for me to have had them in my committee.

The lab and field experiments were with the help of Dr. Jeffrey G. Williamson, Dr. Alto Straughn, Dr. Changying Li, Ms. Xiuhua Li, Mr. John Simmons, Dr. Lihua Zheng, Mr. Asish Skaria, Ms. Han Li, Mr. James Park, Mr. John Ed Smith, Mr. Hao Ma, Ms. Rebecca C. Lee, Ms. Rebekah Combs and Dr. Yan Zhu. In addition, Mr. Michael Zingaro and Mr. Orlando L. Lanni helped me with the mechanical adjustment of my camera system and the transportation of the vehicle for field experiment. The experiments would not have been successful without their friendship and help.

Many other people who are not explicitly mentioned here have aided me throughout my academic career. Finally, I would also like thank my family and friends for their love, support and care throughout this milestone of my life. This research was supported by the Graduate School Fellowship at the University of Florida.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

SPECTRAL ANALYSIS AND MULTISPECTRAL/HYPERSPECTRAL IMAGING TO DETECT BLUEBERRY FRUIT MATURITY STAGES FOR EARLY BLUEBERRY YIELD ESTIMATION

By

Ce Yang

December 2013

Chair: Won Suk “Daniel” Lee
Major: Agricultural and Biological Engineering

Blueberry industry has been increasingly important to both Florida and United States economically since 1990’s (USDA, 2012). Because of the mild sub-tropical climate, blueberry harvesting window in Florida is uniquely early, yielding high profits in the fresh market. However, it is relatively short, usually lasting only five to six weeks. After that, the blueberry price drops rapidly. Therefore, early estimation of fruit yield is crucial for the market and for labor planning. This dissertation explores methods for detection of blueberry with all maturity stages by their spectral properties as well as spatial information.

Spectral analysis offers necessary wavelengths for blueberry detection. Spectra of blueberry fruit and leaf samples were obtained and analyzed. The samples were divided into leaf, mature fruit, near-mature fruit, near-young fruit and young fruit. Normalized indices were used as the candidate variables for classification. Classification models were built and their performances were compared. Four to six wavelengths were chosen using different methods and accuracies of more than 94% were obtained for the classification task.

However, a spectrophotometer is very expensive and can only be used in a laboratory. In contrast, computer vision enables in-field data acquisition. In 2011, multispectral images with three bands: near infrared (760 – 900 nm), red (630 – 690 nm) and green (520 – 600 nm) were obtained. Different color components were input features for classification. Accuracies of 84% and 73% were obtained for fruit and background classes, respectively. However, the color features did poorly in separating eight classes: mature fruit, intermediate fruit, young fruit, leaf, branch, soil, sky, and reference board.

Hyperspectral imaging was proved to be more capable of detecting visually similar object. Therefore, hyperspectral images were acquired in 2012 and 2013. Band selection was necessary to find the most important bands for further application in the field. Three sets of bands were selected using three band selection methods and obtained prediction accuracies of more than 88%. The results showed that the selected band sets were capable of classifying blueberry maturity stages and background. It is beneficial to use spatial information upon the spectral properties of objects in the view. Therefore, spectral-spatial image analysis was considered for the detection of fruits with different maturity stages. Two spectral-spatial image analysis procedures were carried out and evaluated based on the labeled images and obtained more than 78% pixel detection accuracy. The spectral-spatial detection improved the prediction accuracy by up to 30% compared to spectral detection.

CHAPTER 1
INTRODUCTION

This dissertation is about the detection of blueberry maturity stages using spectral analysis and multispectral/hyperspectral image processing toward the development of an in-field yield mapping system. In order to estimate blueberry yield, different growth stages of fruit need to be identified and the fruit amount of each growth stage should be estimated. Remote sensing techniques do not physically contact objects, which is suitable for estimating blueberry yield of different growth stages. It is easy to distinguish mature blueberries in regular color images because of its dark-blue color. However, color images do not show much difference for young fruit and leaf, especially when leaves are in well-illuminated conditions. Using spectral analysis and multispectral/hyperspectral image processing, the near-infrared range is utilized.

A glance of precision agriculture, Florida blueberry and remote sensing is given. The details of applications are then introduced. After that, remote sensing methods used in blueberry classification and detection are presented. Finally, the results are discussed and concluded. Further improvements are indicated as well.

Precision Agriculture

Precision agriculture integrates traditional agricultural industry with new technologies in the information age. It is well known that crop fields have in-field spatial variability in soil texture and structure, soil moisture, nutrient status, organic matter contents, weeds, pests, disease, etc. All these factors cause the yield variation across the landscape. Precision agriculture helps farmers to manage the field more precisely and site-specifically by monitoring smaller areas within the field. Technology used in precision agriculture contains but is not limited to: variable rate technology (VRT), yield

mapping, sensor technology, remote sensing, global positioning system (GPS), and geographic information system (GIS). A good soil property monitoring system helps farmers to save on fertilizers, pesticides as well as protect the environment from over-application of the chemicals. Yield monitoring system estimates the yield variation spatially and helps farmers to make correct decision for managing the field. The overall concern of a precision agriculture system is the economic and environmental benefits it brings to the farmer and all human being.

Florida Blueberry

Florida is ideal for producing early-season blueberries because of its warm weather across the year. Berries from Florida mainly supply the fresh markets from early April to late May. Most of Florida blueberries are hand-picked by manual labor. Insufficiency and high cost of labor are the major concerns. Florida's commercial blueberry field has management cost approximately $9,884/ha (Williamson et al., 2012) besides harvesting cost. The average Florida blueberry yield is 6,310 kg/ha (USDA, 2012). Hand harvest cost is estimated to be $1.59/kg (Morgan et al., 2011). The cost of harvesting labor takes approximately $10,032/ha. Therefore, half of the total management cost of the blueberry field goes to harvesting labor.

Remote Sensing

Remote sensing technique acquires information of an object without physically contact or damage the object. Two major remote sensing types are passive remote sensing and active remote sensing. Passive remote sensing collects natural radiation or reflectance from objects, while active remote sensing collects reflectance from a light source specifically prepared for the object. For example, object detection using natural

light source is passive remote sensing. Spectra obtained in a laboratory by a spectrophotometer are active remote sensing.

Remote sensing has been applied in many fields. Satellite imagery and machine vision are used in meteorology, forestry, landscape, agriculture, etc. The images can be in the visible range or other ranges such as infrared and ultraviolet. Spectral analysis and hyperspectral imaging produce spectral information with very narrow bandwidth. The techniques are used in various fields including mineralogy, biology, environmental measurement and agriculture. Besides high spectral resolution, hyperspectral imaging also provides high spatial resolution, with each pixel represented by its spectrum.

Spectral Analysis

Spectroscopy deals with the interaction of electromagnetic radiation and matter. Originally, spectral analysis was about a prism dispersing the visible light according to its wavelength, which was later expanded to ranges such as ultraviolet, near infrared and far infrared. The outcome spectrum shows the strength of reflectance or absorption at each spectral band. Spectroscopy is used in physics and chemistry because of the unique spectra of atoms and molecules. The coupling of information technology with spectroscopy for statistical interpretation of the result has enabled the sub-discipline of spectroscopy used in agriculture. Its ease to use, nondestructive nature and fast analysis has led to its broad application in precision agriculture. Spectroscopy in the ultraviolet-visible-NIR range is widely used in laboratory measurement for food and crop analysis.

Multispectral/Hyperspectral Imaging

Spectral imaging divides the spectrum into more than the three visible bands: red, green and blue. Multispectral images usually contain several bands from near-

infrared range besides visible bands. The bands are usually discrete and narrower than those of color images. Hyperspectral images contain hundreds of successive bands across the visible and near-infrared range, or even further. Spectral image are stored as an image cube. Like color image, where the three color bands can be treated as three grey images, multispectral image can be seen as several grey images with the specifics bands. Hyperspectral image is stored as an image cube since it has considerable amount of data in all the three dimensions. Spectral imaging systems are powerful tools in many fields such as surveillance, physics, chemistry, mineralogy, precision agriculture, food processing, and environment because of their high resolution in both spatial and spectral dimensions.

CHAPTER 2
LITERATURE REVIEW

Application of Spectroscopy in Agriculture

Spectroscopy has been studied and used in precision agriculture in the last two decades. Researchers use spectroscopy for quality and quantity analysis in fruits, vegetables, field crops and meat. Spectral analysis for precision agriculture includes disease detection, nutrition level analysis, mechanical-damage detection, yield estimation and food quality analysis. The whole spectrum or vegetation indices were applied for specific tasks.

Food quality analysis by spectroscopy usually has several criteria for making the decision. Huang and Lu (2010a) applied partial least squares discriminant analysis (PLS-DA) to discriminant healthy apples and apples with the symptom of mealiness based on hyperspectral scattering. Lu et al. (2011) determined the total phenolic content (TPC) and total antioxidant capacity (TAC) of onion varieties and shallot by infrared spectroscopy. They measured the Folin-Ciocalteu, 2,2-diphenyl-picrylhydrazyl, Trolox equivalent antioxidant and ferric reducing antioxidant power assays by Fourier transform infrared spectroscopy, and these assays were used to quantify the TPC and TAC. They also used the infrared spectral features to classify the variety of onions. Liu et al. (2011) measured the sugar content in chestnuts by near-infrared (NIR) spectroscopy in 833 – 2500 nm. They used partial least squares regression (PLSR) on the original spectra. The correlation coefficients of their experiment achieved 0.90 and 0.86 for original spectra and spectra from different pretreatments, respectively. Zou et al. (2011) carried out the identification of rapeseed cultivars using visible/NIR spectra. They predicted models using distance discriminant analysis and back propagation

neural network and achieved a 100% accuracy using the first six principal components of spectral data in 350 nm – 2500 nm. Inacio et al. (2011) used NIR reflectance spectroscopy to determine the protein content of milk powder samples from different brands and cities. They obtained a 100% prediction accuracy by soft independent modeling of class analogy (SIMCA) models. Using principal component regression, PLSR and successive projection algorithm models, they also achieved a very high accuracy. Daszykowski et al. (2011) classified the fat types in rapeseed, a mixture of rapeseed and soybean, and lard oils, using an improved SIMCA model based on NIR reflectance. All these applications of spectral analysis worked well in fulfilling the task. However, they used the full range of spectra, which is not cost effective. The measurement of reflectance across a wide range is both time and space consuming. To develop models and systems that are applicable in industry, more work should be done.

In order to solve the problem of large time and space consumption, spectral analysis based on only limited number of wavelengths was carried out by many scholars. Special wavelengths and vegetation indices were usually used to quickly measure or decide the features of agricultural products. Kane and Lee (2007) selected three wavelengths for the leaf/fruit separation of citrus. They used fisher linear discriminant and the histograms of the reflectance difference. Ariana and Lu (2010) chose four wavebands with 20 nm bandwidth for the detection of internal defect of cucumbers. They obtained a prediction accuracy of 94.7%. They also evaluated four wavebands with 40 nm bandwidth for internal damage detection for pickles, and reported an accuracy of 82.9%. Balasundaram et al. (2009) classified citrus canker and other diseases on grapefruit peels by applying discriminant analysis with important

bands chosen from the visible and NIR ranges and the whole spectrum, and obtained a 100% accuracy for the classification. Huang and Lu (2010b) selected 17 wavelengths for the evaluation of apple firmness using hierarchical evolutionary algorithm. They obtained a correlation coefficient of 0.857 with a root mean square error of 6.2%. Yang et al. (2010) estimated the nitrogen content in cucumber leaves using vegetation indices. Nichol and Grace (2010) evaluated several indices that were found by other researchers on leaf pigment content in heather (*Calluna vulgaris*). However, it showed that the capability of these indices was not as good as expected. Muller et al. (2008) developed prediction models for predicting oilseed rape shoot dry matter and nitrogen content based on a vegetation index in the NIR range. A correlation coefficient of 0.82 was obtained based this model. Liu et al. (2007) evaluated the fungal disease severity of rice brown spot using the spectral reflectance of the samples. They introduced a ratio with reflectance in 692 nm and 530 nm for this task.

Rao (2007) obtained a spectral reflectance curve for a blueberry variety ‘Lingon’ in Sweden peat land. However, the spectrum was only compared with those of feather moss and crowberry. There was no spectral analysis for blueberry.

Application of Machine Vision and Multispectral/Hyperspectral Imaging in Precision Agriculture

Machine vision is broadly applied in industry. It can provide automatic inspection and analysis, guidance, process control, quantity measurement and quality control in industry. In precision agriculture, machine vision is also widely used. Fruit detection, plant identification, nutrition measurement, default detection, disease detection, food quality control, agricultural vehicle guidance and yield mapping all adapt machine vision.

Digital color image processing and multispectral/hyperspectral imaging are the two common ways of fulfilling these tasks.

Color Image Processing

In digital color image processing, color, shape and texture are the most frequently used features for detection, classification or segmentation tasks in precision agriculture. Lee et al. (1999) developed a real-time robotic weed control system for tomatoes. The system worked with a color video camera with a stable light source. The indoor test had much better result compared to outdoor test. Tian et al. (2000) identified tomato seedlings for automated weed control using machine vision under natural illumination. Their environmentally adaptive segmentation algorithm was able to identify non-occluded target plant with an accuracy of 65% to 78% and less than 5% false positive. Ling et al. (1996) and Kurata et al. (1996) used machine vision to measure the tomato canopy. Potato and apple inspection using color information was introduced by Tao et al. (1995). Regunathan and Lee (2005) detected citrus fruit from color images in the Hue, Saturation and Intensity (HSI) color space and obtained the fruit size by an ultrasonic sensor. Apple stem and calyx identification was carried out by Yang (1996). Wijethunga et al. (2008) analyzed the use of RGB image in the L*a*b color space for the development of a kiwifruit counting system using active light source. They obtained much higher accuracy for the detection of gold kiwifruit compared to green kiwifruit. Annamalai (2004) developed a citrus yield mapping system. Aggelopoulou et al. (2011) predicted yield of apple orchards using machine vision. Matiacevich et al. (2011) assessed the quality of blueberries by computer vision. The quality indicators are color, presence of epicuticular wax, size, dehydration, etc. They used color measure in the CIE L*a*b space to show the change of the fruits during the storage processing.

Kurtulmus et al. (2011) identified immature green citrus fruit in natural outdoor conditions using ‘eigenfruit’ approach and Gabor texture features. They obtained an accuracy of 75.3%. Chamelat et al. (2006) used RGB color space, HSI color space and Zernike moments as features for detection of grapes. Zaman et al. (2008) used blue pixels for identifying wild blueberries and estimated the yield. Their images were obtained where only mature berries were in the view.

Multispectral Imaging

While color machine vision was used and analyzed widely, near-infrared imaging provides more information by utilizing information in the near-infrared range, which human vision is not able to see. Wen et al. (1999) conducted defect inspection for apple using rule-based near-infrared machine vision method. Lu (2003, 2004) investigated apple bruises and soluble solid content (SSC) estimation by near-infrared multispectral imaging and hyperspectral imaging. They obtained critical wavelengths for the prediction. Immature green citrus fruit and Huanglongbing (HLB, also known as citrus greening) disease were detected using multispectral imaging by Kane and Lee (2007), Okamoto et al. (2007), Lee et al. (2008), Okamoto and Lee (2009) and Kumar et al. (2009).

Hyperspectral Imaging

Hyperspectral imaging provides both high spectral and spatial resolution, which brings more information for detection and classification purpose in precision agriculture. However, hyperspectral images contain a large amount of redundancy. Some spectral bands are not helpful for detection and classification. In addition, neighboring bands have similar information. Therefore, scholars tried to reduce spectral dimensionality by feature extraction and selection.

Cheng et al. (2004) integrated principal component analysis (PCA) with Fisher's liner discriminant (FLD) to inspect cucumber chilling damage. The integrated method outperformed the PCA and FLD methods when they were used separately. Bruce et al. (2002) applied discrete wavelet transform for dimensionality reduction of hyperspectral data. Prasad et al. (2004) assessed the performance of hyperspectral wavebands for vegetation analysis. They extracted 22 optimal bands by PCA, lamba-lamba R^2 models, stepwise discriminant analysis (SDA) and derivative greenness vegetation indices (DGVI), and obtained over 90% of overall accuracy.

Hyperspectral band extraction reduces dimensionality but the projected features do not have physical meaning. Also, projections need all the bands from the original data, although some of the bands do not contribute to the classification or detection task. In contrast, hyperspectral band selection extracts original features, which contains physical information. For yield mapping purpose in precision agriculture, selected original bands are suitable for a low-cost yield estimation system using multi-spectral imaging. Therefore, band selection is preferable to feature projection.

There are different criteria of measuring importance of bands, such as transformed divergence, Bhattacharyya distance, Jeffries-Matusita distance, etc. (Yang et al., 2011). Selected bands are those that have the largest distance with each other based on these criteria. Mutual information and information divergence have also been used for calculating the relationships among different bands. (Martinez-Uso et al., 2007; Guo et al., 2006). Other methods employed a criterion to rank different bands. Bands with the highest rankings have the highest priorities. Band ranking methods include variance, correlation, signal-to-noise (SNR) ratio, etc. (Chang et al., 1999; Bajwa et al.,

2004). Zare and Gader (2008) conducted hyperspectral band selection and endmember detection simultaneously using sparsity-promoting priors. The method is an extension of sparsity promoting iterated endmember (SPICE) algorithm by adding spectral band weights and a sparsity prior to the SPICE algorithm. Chang and Wang (2006) selected bands using constrained energy minimization (CEM). The method linearly constrains a band image while minimizing band dependence to other band images.

While these band selection methods help in constraining the redundancy of the original hyperspectral data, the purpose is to reduce data volume and calculating complexity. They do not focus on what specific bands are selected and why they are more important than other bands. Only some band selection methods for agricultural application paid attention to the selected bands (Bajwa et al., 2004).

CHAPTER 3
CLASSIFICATION OF BLUEBERRY FRUIT AND LEAVES BASED ON SPECTRAL SIGNATURES

Background

Blueberry is well known for its nutrition value and high levels of anti-oxidants. In addition, the high prices of the hand-harvested early season blueberry enable Florida growers to achieve considerable profits. Therefore, the production acreage of blueberry is expanding in Florida, USA. In 1993, there was a total of 1000 acres of blueberry in Florida. By 2010, blueberry acreage was more than 4000 acres (Braswell, 2010). Florida's mild winter climate is well suited for low-chill southern highbush cultivars such as 'Emerald', 'Jewel', 'Farthing', 'Springhigh', 'Sweetcrisp', 'Star', and 'Windsor'. Southern highbush cultivars grown in Florida ripen during April and May before other North American production areas, and usually receive higher prices than other production areas. However, the production window is relatively short lasting from about April 1 until May 15 after which prices usually drop rapidly as berries enter the market from northern regions Florida growers should adequately prepare for the relatively short harvest season. Moreover, crop yield is considered as the most important information for crop management in precision agriculture (Lee et al., 2010). Especially, early yield estimation is crucial for labor planning and reducing the cost for harvesting. Detection of mature berries before the harvest season is essential for early yield estimation which can help farmers gain better control of harvest and earn higher profits.

Spectral analyses are widely used in precision agriculture for crop quality detection, disease detection, nutrition analysis, etc. (Cozzolino, Cynkar, Shah, Smith, 2011; Nicolai et al., 2007; Menesatti et al., 2010 and Jones, Jones, Lee, 2010). Spectral information can be used for a quick detection model of crop quality. Wang, Li and Wang

(2011) analysed the relationship between the diffuse reflections of hyperspectral images of the range of 539 – 902 nm and the onion quality features (firmness, soluble solids content and dry matter content). They also compared the performance of quality estimation of onion from California, Idaho and Georgia. Disease detection based on spectral analysis has been used to aid in disease control and assist farmers with crop management. Liu, Huang and Tao (2007) estimated fungal disease severity of rice brown spot with hyperspectral reflectance data. They used three vegetation ratio indices in the NIR range. The highest coefficient of correlation with the disease severity was obtained by the ratio of reflectance in 692 and 530 nm. Quick nutritional content measurements using spectral reflectance based on vegetation indices have been applied and evaluated by many researchers. Nichol and Grace (2010) tested several previously published vegetation indices on leaf pigment content in *Calluna vulgaris* shoots and found those indices performed poorly. Yang, Li and Sigrimis (2010) estimated nitrogen content in cucumber leaves using vegetation indices. Muller, Böttcher, Meyer-Schatz and Kage (2008) developed prediction models for oilseed rape shoot dry matter and nitrogen content prediction based on a vegetation index selected from the NIR range, and they obtained a correlation coefficient of 0.82 between tested and predicted values. Liu et al. (2011) applied NIR (833 – 2500 nm) spectroscopy in the measurement of sugar content in chestnuts. Partial least squares regression (PLSR) based on the original spectra and the spectra derived from different pre-treatments were used in their modeling, and the correlation coefficients of the optimized models were 0.90 and 0.86. However, there were no studies about spectral analysis of blueberry fruits and leaves. Rao (2007) obtained a spectral reflectance curve for the blueberry

variety ‘Lingon’ in Sweden peat land, and compared it with those of feather moss and crowberry, but no spectral analysis was made.

Classification of food and crops based on spectral information has also been applied by many researchers. Most of this research was based on the full-spectrum of the NIR range; some also combined the information in the visible range. Zou, Fang, Liu, Kong and He (2011) compared the performance of distance discriminant analysis and back propagation neural network for identification of rapeseed cultivars using visible/NIR spectra, and developed prediction models with 100% accuracy using the first six principal components of spectral data across 350 nm to 2500 nm. Balasundaram, Burks, Bulanon, Schubert and Lee (2009) classified citrus canker and other diseases on grapefruit peels by applying discriminant analysis with important bands in the visible and NIR ranges and the whole spectrum, obtaining a 100% accurate classification of the citrus canker disease. Inácio, Moura and Lima (2011) used NIR reflectance spectrometry to classify and determine the total protein in milk powder samples from different brands and cities, and 100% prediction accuracy by soft independent modeling of class analogy (SIMCA) models. They also obtained high accuracy by the principal component regression (PCR), partial least squares (PLS) and successive projection algorithm (SPA) models. Daszykowski, Orzel, Wrobel, Czarnik-Matusewicz and Walczak (2011) improved the SIMCA modeling for the classification of fat types (rapeseed, a mixture of rapeseed and soybean, and lard oils) based on the NIR reflectance. Huang and Lu (2010a) used partial least squares discriminant analysis (PLS-DA) for the classification of apples with the symptom of mealiness based on hyperspectral scattering technique. However, the prediction accuracy was very low for

the classification of three and four severity levels. All of this research was using the whole wavelength range rather than limited wavelengths for classification. However, for the development of cost effective systems, these methods are not ideal. The measurement of reflectance across a large wavelength range with high resolution can be very expensive.

Some crop classification models use only several wavelengths or wavebands. Kane and Lee (2006) obtained the classification model of citrus fruit and leaf. They obtained three wavelengths for the leaf/fruit separation by fisher linear discriminant and the histograms of the reflectance difference at each wavelength. Ariana and Lu (2010) selected four wavebands with 20 nm bandwidth for detecting internal defect of cucumbers and obtained 94.7% correctly classified. They also tested the other four wavebands with 40 nm bandwidth for internal defect detection of pickles with classification accuracy of 84.9%. Their research used a prototype hyperspectral imaging system and images were collected with 5, 10, 20, 40, 60 nm resolution. The classification was binomial (either healthy or defected), and the wavebands were selected by a branch and bound algorithm combined with the k-nearest neighbor classifier. Huang and Lu (2010b) obtained 17 wavelengths for the prediction of apple firmness using hierarchical evolutionary algorithm (HEA) and the correlation coefficient was 0.857 with root mean square error 6.2%. The model with these wavelengths performed better than the full spectrum model.

Anthocyanin is one of the main antioxidants in blueberry (Prior, et al., 2001). Increased maturity increases the content of anthocyanin (Prior, et al., 1998). Besides anthocyanin, blueberry also contains large amount of flavonol, chlorophyll,

carbohydrates, vitamins and water. According to US highbush blueberry council (2010), carbohydrates in highbush blueberry mainly contain fructose and glucose. Vitamin C is the leading vitamins in blueberry. According to Guisti and Wrolstad (2001), anthocyanin's absorption bands are 490 - 550 nm. Besides, anthocyanin contributes to the color change on the surface of blueberry, which shows major difference in the blue band. Flavonol was reported to respond in the range of 210 nm to 230 nm (Harnly, et al., 2006). Chlorophyll absorption bands are 430 nm - 450 nm and 640 nm - 660 nm. Chlorophyll *a* has bands at 430 nm and 640 nm. Chlorophyll *b* has bands at 450 nm and 660 nm (Jensen, 2000). The absorption wavelength for detecting carbohydrates was 1469 nm for fructose, and 1688 nm for glucose (Giangiacomo, et al., 1981). Yang and Irudayaraj (2002) conducted research on the absorption spectra of vitamin C. They found four wavelengths that determine vitamin C. The four wavelengths were 1457 nm, 1926 nm, 2080 nm and 2242nm. Water has four absorption peaks: 970 nm, 1200 nm, 1450 nm and 1950 nm (Williams, P. and Norris, K., 2001).

Spectral signatures for southern highbush blueberry whole fruits and leaves are neither analyzed nor used for classification. Therefore, the objectives of this study were to analyze the difference among the blueberry fruit growth stages and leaves based on their spectral reflectance, and to build a best classification model for separating the classes of blueberry fruit and leaves based on a limited number of wavelengths. The outcome will later be applied to a blueberry yield mapping system based on the selected wavelengths.

Materials and Methods

Fruit and Leaf Collection

Blueberry fruit and leaf samples were collected from a commercial blueberry farm (Straughn Farm) in Waldo, Florida, USA and a blueberry experimental field in Citra, Florida, USA. Seven representative southern highbush varieties ‘Emerald’, ‘Jewel’, ‘Farthing’, ‘Springhigh’, ‘Sweetcrisp’, ‘Star’ and ‘Windsor’ were chosen for lab spectral reflectance measurement. The in-field sampling was during blueberry harvesting season in April and May, 2011. The samples were classified into five classes: leaf, mature fruit, young fruit, near-mature fruit and near-young fruit. For each class of blueberry fruit and leaves, no more than two to three fruits or leaves were picked from the same plant. Thirty - forty samples for each class of each blueberry variety were obtained. There were 190 - 200 samples for each of the seven varieties, 1378 samples in total. Each sample includes four to six fruits. In 2010, preliminary experiments were taken place in Waldo, FL, USA. 188 samples were collected and different experiment designs were used. The five classes for 2010 samples were dark green leaf, light green leaf, mature fruit, intermediate fruit and young fruit. The limitation of sample size was due to sample availability at that time. It was supposed that there was no significant difference between the two years’ data. To combine all the data together, the two leaf classes of 2010 data were merged into the leaf class. The near-mature and near-young fruit classes of 2011 data were merged into an intermediate fruit class. Therefore, two years’ data together was divided into a total of four classes: mature fruit, intermediate fruit, young fruit and leaf.

Spectral Measurement

A UV-Vis-NIR spectrophotometer (CARY 500, Varian Inc., Palo Alto, California, USA) was used for spectral measurement of the samples. One fruit sample was made of approximately 8-12 fruits, and one leaf sample contained two to three leaves. Fruits were cut into half and put in a sample holder as shown in Figure 3-1. A polytetrafluoroethylene (PTFE) disk was used each time before the day of spectral measurement of the samples in order to obtain the optical baseline for the system. Reflectance of each sample was measured between 200 nm and 2500 nm with a 1 nm increment. The preparations of the samples are shown in Figure 3-1.

Moving average was applied to smooth out some minor fluctuations among successive spectral wavelengths. Eq. (3-1) shows how the simple moving average was calculated.

$$R_n = \frac{\sum_{m=n-i}^{n+i} R_m}{(2i+1)\Delta} \tag{3-1}$$

Where: R_x (X = m or n) is the reflectance at the wavelength X, and Δ is the interval between two wavelengths. In our measurement, Δ was 1 nm, and i was chosen to be 1.

Principal Component Analysis

Principal component analysis (PCA) can help find alternative uncorrelated variables, only a few of which explain the whole spectral data. It is an eigenvector-based method of transforming original correlated variables to a set of uncorrelated variables. These new variables are called principal components. And each component as a vector is orthogonal to the others, which means there is no multicollinearity among them. The principal components are arranged in the order of score and loading, which is

obtained by calculating the variance of any projection of all data. The greater the variance is, the more contribution the principal component has. Usually, the first several principal components can contribute most part of the information of the data. In this study, Matlab (Ver. 7) was used for PCA.

Classification Tree

Classification tree is a method mainly used in machine learning and data mining, and works well for both numerical and categorical data. The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to be used in splitting the set of items (Rokach and Maimon, 2005). Leaves represent classes and branches lead to new classification threshold for separating new classes. The decision is made after all the sub-decisions of the nodes of the tree were made. The predictive model is simple to display and interpret since it is in Boolean format (Sutton, 2005). In this study, R (open source software for statistical analysis and software development) was used for the construction of tree models. The candidate variables were in the format of a normalized index. The indices were chosen based on the largest difference between every two classes out of the five classes. Therefore, a total of ten indices were used as the candidate variables for the classification tree model construction.

Multinomial Logistic Regression

A logistic regression model, also called a logistic model, calculates the probability of an event happening according to the logistic function of several independent variables (Agresti, 2007). For a basic binomial logistic regression model, Eq. 2 explains how the value of a sample can be fitted to a curve confined from -1 to 1. Multinomial

logistic regression is an extension of the binomial logistic regression, as shown in Eq. (3-2).

$$f(y) = \frac{1}{1 + e^{-y}} \tag{3-2}$$

Where *y* is a dataset with n predictive variables, as defined in Eq. (3-3).

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{3-3}$$

Where x_i is the value of the i^{th} variable, β_0 is an intercept, and β_i is the coefficient of variable x_i. R language was used for the application of multinomial logistic regression. The variables for multinomial logistic regression were the principal components of the variables obtained from the tree model.

Spectral Data Analysis

There were 276 leaf samples, 280 mature fruit samples, 275 young fruit samples, 275 near-mature fruit samples, 272 young fruit samples collected in the 2011 blueberry harvest season. The average spectral absorbance for each class is shown in Figure 3-2. Data points between 330 nm and 355 nm are removed because of some noises in all sample measurements. Significant differences between the absorbance of leaves and that of fruit classes are observed in the whole wavelength range.

Mature fruits have more anthocyanin and carbohydrates than immature fruits and leaves do. Therefore, the absorbance of mature fruit between 490 – 550 nm was higher than that of other classes. Fruits need chlorophyll for photosynthesis, which produces glucose. As fruits mature, they continue to accumulate sugar. Fruits have higher absorbance in the NIR range than leaves do. Calculation of indices provided more information for classification, which is explained in the next section.

Absorbance of mature fruit for different varieties is shown in Figure 3-3. All these varieties are southern highbush species, which means that they have the same origins. Therefore, their absorption curves were similar. Anthocyanin and chlorophyll content of the mature fruits differentiate significantly. ‘Jewel’ had the lowest absorbance in the UV and visible ranges. This suggests that Jewel has the lowest concentration of chemical compositions such as flavonol, carbohydrates and anthocyanin. ‘Sweetcrisp’ had the lowest chlorophyll content. ‘Springhigh’ had the highest flavonol, anthocyanin, carbohydrates, chlorophyll, water and vitamin C concentration. ‘Springhigh’ and ‘Sweetcrisp’ were expected to be the easiest varieties to detect.

Since the classification result with specific wavelengths will be used for the in-field yield mapping system, normalized indices were considered to be the best way to eliminate the impact of illumination changes in the open field. Differences between index values of each two classes were calculated in order to find the largest ones. The indices with the largest difference among the five classes were selected as candidate variables for the classification models. A total of 10 indices were obtained among the five classes. The list of the variables is shown in Table 3-1.

A subset of 918 samples out of the 1378 samples was selected randomly as the calibration dataset, and the remaining 460 samples were used as a validation dataset. The classification tree was generated based on the ten candidate normalized indices. A second classification tree was produced based on the subset of the variables used in the first tree. Only variables used in the first classification tree were in the subset. It was expected that classification tree modeling using the less but more useful information would perform as well as the first classification tree did. In this way, fewer wavelengths

may be used to obtain the prediction result. If this is true, then the second classification tree would be an optimized model. In the development of yield mapping system, fewer wavelengths mean lower cost.

Variables used in the classification tree were collected as the input variables for the logistic regression model. Since MNR requires the input variables to be independent, the variables should be rearranged by linear combination to obtain independent variables before the modeling. Uncorrelated variables can be obtained by PCA from the candidate variables, which are not strictly independent. However, these uncorrelated variables already could perform well for the logistic regression. Figure 3-4 shows an MNR example of two variables (u_1 and u_2) using the reflectance of wavelengths λ_1, λ_2, λ_3 and λ_4 as the candidate variables.

Results

Classification Tree Models and MNR Models of 2011 Samples

The first classification tree was obtained based on the ten candidate variables from Table 3-1. The decision was made by the values of three variables: u_2, u_8, and u_9. The wavelengths used were: 1373, 699, 691, 554, 551 and 233 nm. The three variables were able to separate the five classes thoroughly, as shown in Figure 3-5. The nodes of the tree are thresholds. The tree was split into branches and terminal nodes based on the thresholds. Only 14 of the 460 samples in the validation set were misclassified, which yielded 97% accuracy. The validation result for the first classification tree is shown in Table 3-2. Both leaf and mature fruit can be easily divided from the other classes using u_2 and u_9. The prediction accuracy of the classification tree model was very high, with 100% separation accuracy for leaf, 99% for mature fruit and young fruit, and 94% for the two middle fruit stages.

The second decision tree was generated using u_2, u_8, and u_9 as candidate variables. The tree model was exactly the same as the first classification tree with the same accuracy. This result reflects that the first classification tree model is already the best that can be obtained and no more variables can be eliminated.

To obtain a better classification result, multinomial logistic regression was applied. Since u_2, u_8, and u_9 performed well in the classification trees, they were selected as the inputs for the construction of a multinomial logistic regression model. The pre-processing of the data was to find the independent principal components of the three variables. Principal component analysis was applied and the principal components were in the format of linear combinations of the original variables, which are shown in Eq. (3-3) to Eq. (3-5).

$$PC1 = 0.37\,u_2 + 0.55\,u_8 - 0.75\,u_9 \tag{3-3}$$

$$PC2 = 0.78\,u_2 + 0.62\,u_8 - 0.06\,u_9 \tag{3-4}$$

$$PC3 = 0.50\,u_2 + 0.56\,u_8 + 0.66\,u_9 \tag{3-5}$$

Where: *PC1*, *PC2*, and *PC3* are the first, second and third principal components, respectively. These three principal components are the linear combination of the three indices u_2, u_8, and u_9. There are no correlations among the principal components.

The probability of being leaf class is chosen randomly as the denominator of logistic regression results. The numerator of the ratio is the probability of the sample belonging to mature fruit, near-mature fruit, near-young fruit or young fruit class. The model is the logarithm of this ratio. The modeling result is shown in the following equations.

$$\log\left(\frac{P(Y = Mature)}{P(Y = Leaf)}\right) = 11.6 - PC1 + 36.1PC2 - 398.7PC3 \quad (3\text{-}6)$$

$$\log\left(\frac{P(Y = Near - mature)}{P(Y = Leaf)}\right) = -3.8 - 65.8PC1 + 9.8PC2 - 404.4PC3 \quad (3\text{-}7)$$

$$\log\left(\frac{P(Y = Near - young)}{P(Y = Leaf)}\right) = -2.5 - 6.4PC1 - 59.6PC2 - 392.3PC3 \quad (3\text{-}8)$$

$$\log\left(\frac{P(Y = Young)}{P(Y = Leaf)}\right) = -1.2 + 40.3PC1 - 53.7PC2 - 384.8PC3 \quad (3\text{-}9)$$

Y is the dependent nominal variable, whose 'value' is one of the five classes. The final 'value' of *Y* is according to the highest score obtained from the calculation of these logistic regression equations. The validation result is shown in Table 3-3. The overall accuracy is 97.8%, which is higher than that of the classification tree model.

Classification Results for Different Varieties

The same variable collection and data classification methods were applied to each variety. Table 3-4 gives the wavelengths of the ten candidate variables for each of the seven varieties. The description of the ten variables was the same as in Table 3-1.

Since the candidate variables used for the model construction were not the same, the classification trees for the seven varieties were different. Two-thirds of the samples for each variety were used for calibration and the other one-third of the samples was used for validation. Two example results of the varieties 'Farthing' and 'Star' are shown in Figure 3-6. For the variety 'Farthing', the second variable (1373 and 700 nm) and the eighth variable (685 and 554 nm) were both used twice in the tree model. Leaf was the first class being separated because leaf is very different from all stages of fruits. However, for the variety 'Star', fruit growth stages tended to be detected first. Color development in fruit was different for 'Farthing' and 'Star'. 'Star' had more

distinct difference for each growth stage of fruit than ‘Farthing’ did. The color change of ‘Star’ fruits was: entirely green → entirely pink → entirely red → entirely dark blue. At each growth stage, the pigments on the fruit were uniform. However, the color change of ‘Farthing’ fruit was: entirely green → partly green and partly light red → partly green and partly dark red → entirely dark blue. The pigments on the ‘Farthing’ fruits in the middle stages were not uniform.

The wavelengths used for the model construction of each variety and the accuracies of classification trees and multinomial logistic regression models are summarized in Table 5. Classification tree models had relatively lower prediction accuracy than that of MNR model. The lowest prediction accuracy was for the classification tree model of ‘Farthing’, which was 92%. The MNR model increased the accuracy to more than 98%.

According to this table, ‘Farthing’ and ‘Jewel’ had lower prediction accuracy than other varieties. Most of the errors were for separating the near-mature and near-young fruit classes. The prediction ability of leaf and mature fruit was much better than that of the other three classes. Leaf and mature fruit classes obtained 100% accuracy for all varieties.

Classification Results for 2010 and 2011 Samples Together

Since there were only four classes for the 2010 and 2011 samples analyzed together, there were eight candidate indices chosen based on the greatest difference between classes. The indices are shown in Table 3-6. The calculation of these indices was the same as that in Table 3-1.

The classification tree result of the two years' data with four classes is shown in Figure 3-7. Two variables, s_2 and s_6, with wavelengths 553, 688, 698 and 1373 nm, were used for the classification.

Logistic regression model were built using the principal components of the two variables PC1 and PC2 as shown in Eq. (3-10) and (3-11).

$$PC1 = 0.22\, s_2 + 0.22\, s_6 \tag{3-10}$$

$$PC2 = 0.16\, s_2 - 0.13\, s_6 \tag{3-11}$$

The logistic regression models are shown in Eq. (3-12) to (3-14).

$$\log\left(\frac{P(X = Intermediate)}{P(X = immature)}\right) = 2.4 + 12\, PC1 + 14\, PC2 \tag{3-12}$$

$$\log\left(\frac{P(X = Leaf)}{P(X = immature)}\right) = 7.1 + 9\, PC1 + 7.8\, PC2 \tag{3-13}$$

$$\log\left(\frac{P(X = Mature)}{P(X = immature)}\right) = -2.8 + 0.4\, PC1 + 8.3\, PC2 \tag{3-14}$$

Classification Results of Different Varieties

In order to apply precisely the classification model for a specific variety, the variety prediction model was built. According to Figure 3-3, there are some differences in the UV and NIR ranges for mature fruit absorbance of different varieties. However, the variation is much smaller than that of different fruit and leaf classes. Some of the varieties overlapped throughout the whole wavelength range. The largest index differences between any two varieties were calculated. Twenty-one candidate variables were obtained, which are listed in Table 3-8. The calculation of indices was the same as in Table 3-1.

In the classification tree, ‘Sweetcrisp’ was separated in the first step using the value of t_4= index(R_{710}, R_{321}). If t_4>0.26, then the sample was classified into ‘Sweetcrisp’. However, for other varieties, the classification was more complicated. For example, to classify a sample into ‘Farthing’, the sample should pass a series of five or six decision nodes. This makes the classification less accurate. The accuracy of the trees, which was shown in Table 3-9, illustrated that it was difficult to classify most of the varieties. ‘Sweetcrisp’ had the highest classification accuracy (90.0%), and the lowest false positives (9.1%). ‘Windsor’ had the lowest classification accuracy (11.1%), and the highest false positives (80.0%). Except for ‘Sweetcrisp’, all varieties had equal or less than 50% prediction accuracy.

Variables used in the classification tree model were linearly combined to obtain the principal components, and these components were fed into the logistic regression algorithm to get a MNR model for the classification of the seven varieties. The validation result is shown in Table 3-10. The model had 100% prediction accuracy for ‘Sweetcrisp’, and the false detection did not change. For varieties such as ‘Jewel’ and ‘Windsor’, the accuracy of the model was increased. The false positive decreased for ‘Windsor’, however, it increased for ‘Jewel’. For ‘Springhigh’ and ‘Star’, the prediction accuracy decreased.

Discussion

Classification Model of Five Classes Based on 2011 Samples

For the classification model of five classes of 2011 samples, six wavelengths were used. One of them was from the NIR range (1373 nm), one in the UV range (233 nm), and four in the visible range (554 and 551 nm, green band; 699 and 691 nm, red band). There was not an obvious difference in the UV range from Figure 3-2. However,

the important wavelength in the UV range (233 nm) may be selected based on the concentration difference of flavonol. Mature fruits had high concentration of anthocyanin. This may be one reason that two wavelengths approximately 550 nm were chosen. The red edge (680 – 730 nm) was chosen in the model, because of its ability to separate the leaf and fruit. Leaf has lower sugar content, and thus had lower absorption after 700 nm. Therefore, the absorption of the leaf class decreased more rapidly at the red edge than the fruit classes. The color of young fruit is green and its nitrogen content is very high (Gardner, Bradford and Hooker, 1922), which helped in distinguishing mature fruits from young fruits.

The classification tree model had lower prediction accuracy for the classification of fruit stages than the MNR model did, according to Tables 3-2 and 3-3. Compared to the classification tree model, false detections from the MNR model were decreased. Especially, all observed mature fruit samples were classified into this class. It shows that the MNR model was better for the classification of mature blueberries, which was the targeted class. The reason that the MNR model always had the same or higher accuracy than the classification tree may be because the MNR model used principal components of the variables. Collinearity existed among the variables when the classification tree model was built. This may have had a negative impact on the model construction.

Classification Models for Each of the Seven Varieties

The details for the classification models of each variety are listed in Table 3-5. The selected wavelengths, variables, and thus the classification models for each variety were not the same. The classification models of two varieties, ‘Farthing’ and ‘Jewel’, had the lowest prediction accuracy. This was because ‘Farthing’ ripened by different

steps rather than green-pink-red-purple for the four fruit stages. ‘Farthing’ berries changed color unevenly during ripening. A portion of the berry developed the dark purple pigmentation first while the other part of the berry stayed green. This made the condition of the two intermediate fruit stages for ‘Farthing’ very different from other varieties. However, for the leaf, mature fruit and young fruit classes, the prediction accuracy was very high. ‘Jewel’ samples were collected from both the commercial field in Waldo and the experimental field in Citra. Although the climate was the same due to the proximity of the locations, the difference in irrigation and soil condition may be responsible for the low classification accuracy. ‘Windsor’ samples were also collected from both fields. Therefore, the classification result was not 100% correct. ‘Springhigh’ is an early-season variety, which ended harvest in early May in 2011. However, to increase sample size, fruit remaining after commercial harvest were collected from the plants. Although the plants kept enough fruit remaining after harvest, which were still in good condition for the last three sample collections, the classification results showed higher errors. Some of the ‘Emerald’ samples were collected from the greenhouse, where the growing condition was slightly different from the open field. The farm kept some varieties including ‘Emerald’ in the greenhouse so that they could produce berries from January to the end of the blueberry season. Difficult in classifying this variety was expected. However, the result showed that the classification model works well for ‘Emerald’. The classification tree model and MNR model of ‘Sweetcrisp’ both achieved 100% accuracy. And the model only used four wavelengths, which was very good for the yield mapping system development purpose since only four channels were required for the detection of each class even though other factors would affect outdoor

measurement. This showed that the greenhouse environment and a longer harvest window were not a problem for ‘Emerald’.

Classification Model of Four Classes Based on 2010 and 2011 Samples

Only four wavelengths (553, 688, 698 and 1373 nm) were used for the classification of two years’ samples with four classes. Two (688 and 698 nm) of the four wavelengths were from the red edge, and 553 nm was from the nitrogen absorption band. The prediction ability of the model was also very high, with 100% accuracy for mature fruit and leaf. 98.5% accuracy of intermediate fruit classification was already the lowest among the four classes. Compared to that of five-class classification, the model of four classes had higher prediction accuracy. The reason may be that the four-class classification used two years’ samples. Larger sample size would give more accurate result. In addition, near-mature and near-young stages for some varieties such as ‘Farthing’ were difficult to classify due to the different ripening step appearances. This was explained before.

Separation of Different Varieties

The classification tree was already pruned to the simplest model. However, it was still intricate. This was obvious by looking at the absorbance curves of mature fruits of different varieties (Figure 3-3). ‘Springhigh’ and ‘Windsor’ were hoped to be easily classified since their spectra were easily distinguished in the sugar absorption band and water absorption wavelengths. The problem may be that the variation among samples of the same variety was so large that the average reflectance was not representative at all. The MNR model had better prediction ability for most varieties, however was still too poor to accomplish the task. It showed that the average reflectance was not the suitable data source for the selection of variables. The other reason for the problem may be that

several varieties had very close origins. Other methods are required for the separation of varieties rather than finding the largest index difference based on the average reflectance of each variety.

Conclusion

Classification tree and multinomial logistic models for the classification of blueberry leaf, mature fruit, near-mature fruit, near-young fruit and young fruit based on the spectral signatures were built and tested. Six wavelengths (233, 551, 554, 691, 699 and 1373 nm) were used in three vegetation indices for the classification tree construction. Principal components of the three indices were obtained for the multinomial logistic model construction. The MNR model had higher accuracy than that of the classification tree model. The accuracy of MNR model was 100% for the leaf and mature fruit class, and the lowest accuracy occurred in the detection of near-mature fruit, however, it was still higher than 94%.

Classification models for each of the seven varieties were built based on different wavelengths selected. The MNR models of ‘Emerald’, ‘Star’, ‘Sweetcrisp’, ‘Farthing’, ‘Jewel’, ‘Springhigh’ and ‘Windsor’ achieved 100%, 100%, 100%, 98.4%, 98.4%, 98.6% and 98.6% prediction accuracy, respectively.

Samples collected from 2010 and 2011 were combined together, and four classes (leaf, mature fruit, intermediate fruit and young fruit) were designed for the construction of classification models. Four wavelengths were used in the model (553, 688, 698 and 1373 nm). The intermediate fruit class had the lowest accuracy; however, it was 97.2% for the classification model and 98.5% for the MNR model. The prediction accuracy of the mature fruit and leaf classes obtained 100% for both classification tree model and MNR model.

The separation of varieties required a complicated classification tree and had very low prediction accuracy for six of the seven varieties. ‘Sweetcrisp’ was the only variety that obtained 90% classification accuracy. The MNR model also had very low separation ability for varieties. The core reason was that the varieties had very close origins.

In summary, there were strong differences among the blueberry fruit growth stages and leaves in reflectance spectroscopy. Significant wavelengths were identified for the classification of leaves and various fruit growth stages. Simple classification models with very high accuracy were developed based on only several wavelengths. Therefore, these models showed great potential for the development of a low-cost and highly accurate blueberry yield estimation system.

Table 3-1. Candidate variables selected by the maximum differences of index values between each two classes of 2011 samples.

variable	Indices	Description (largest difference between)
u1	index(R1371, R320)[a]	Leaf v.s. Mature fruit
u2	index(R1373, R699)	Leaf v.s. Young fruit
u3	index(R1372, R696)	Leaf v.s. Near-mature fruit
u4	index(R1374, R698)	Leaf v.s. Near-young fruit
u5	index(R554, R231)	Mature v.s. Young fruit
u6	index(R709, R319)	Mature v.s. Near-mature fruit
u7	index(R709, R239)	Mature v.s. Near-young fruit
u8	index(R554, R233)	Young v.s. Near-mature fruit
u9	index(R691, R551)	Young v.s. Near-young fruit
u10	index(R603, R235)	Near-mature v.s. Near-young fruit

[a] $index\ (R_x, R_y) = (R_x - R_y)/(R_x + R_y)$

Table 3-2. Validation result for classification tree.

	Mature fruit	Near-mature fruit	Near-young fruit	Young fruit	Leaf	Pred. Total	Correct Prediction (%)
Pred. Mature fruit[a]	83	5	0	0	0	88	98.8
Pred. Near-mature	1	87	0	0	0	88	93.5
Pred. Near-young	0	1	88	1	0	90	93.6
Pred. Young fruit	0	0	6	91	0	97	98.9
Pred. Leaf	0	0	0	0	97	97	100.0
Total	84	93	94	92	97	460	97.0

[a] Pred. Mature fruit means predicted mature fruit

Table 3-3. Validation result for multinomial logistic regression.

	Mature fruit	Near-mature fruit	Near-young fruit	Young fruit	Leaf	Pred. Total	Correct Prediction (%)
Pred. Mature fruit[a]	84	4	0	0	0	88	100
Pred. Near-mature	0	88	0	0	0	88	94.60
Pred. Near-young	0	1	92	2	0	95	97.80
Pred. young	0	0	2	90	0	92	97.90
Pred. Leaf	0	0	0	0	97	97	100
Total	84	93	94	92	97	460	97.80

[a] Pred. Mature fruit means predicted mature fruit

Table 3-4. Selected candidate wavelengths of the ten variables for each of the seven varieties.

Varieties	Wavelengths (nm)									
	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
Farthing	1376	1373	1373	1376	554	707	707	685	692	606
	693	700	696	699	229	327	325	554	550	237
Jewel	1370	1375	1370	1376	713	711	708	555	689	602
	320	698	696	697	313	322	322	230	550	242
Springhigh	1357	1373	1373	1373	715	710	709	556	687	606
	330	698	696	697	236	549	319	232	551	226
Star	1374	1374	1373	1373	556	707	1064	553	552	599
	692	699	697	698	231	319	712	235	235	239
Sweetcrisp	1370	1376	1372	1376	554	707	707	685	691	607
	319	699	697	698	226	317	235	554	553	235
Windsor	1373	1373	1357	1371	715	708	709	553	692	603
	324	699	697	698	234	318	235	232	550	240
Emerald	1371	1373	1372	1372	714	711	711	553	690	560
	321	698	695	697	234	328	241	226	551	237

Table 3-5. Selected wavelengths for different varieties and the related accuracy. Models were built separately for each of the seven varieties.

Varieties	Wavelengths (nm)	Classification tree model (%)	MNR model (%)
Emerald	1371, 711,690, 553, 551, 328, 321, 226	96.9	100
Farthing	1373, 700, 685, 554	92.2	98.4
Jewel	1370, 713, 602, 555, 320, 313, 242, 230	98.4	98.4
Springhigh	1373, 1357, 710, 698, 556, 549, 330, 232	98.6	98.6
Star	1374, 699, 692, 599, 556, 239, 231	97	100
Sweetcrisp	1376, 707, 699, 317	100	100
Windsor	1373, 715, 553, 550, 324, 234, 232	98.6	98.6

Table 3-6. Candidate indices of two years' data with four classes: mature fruit, intermediate fruit, young fruit and leaf.

Variables	Indices	Description (largest difference between)
s1	index(R1371, R323)	Leaf v.s. Mature fruit
s2	index(R1373, R698)	Leaf v.s. Young fruit
s3	index(R1370, R697)	Leaf v.s. Intermediate fruit
s4	index(R554, R200)	Mature v.s. Young fruit
s5	index(R709, R318)	Mature v.s. Intermediate fruit
s6	index(R688, R553)	Young v.s. Intermediate fruit

Table 3-7. Accuracy of classification tree models and MNR models.

Class	Classification tree (%)	MNR model (%)
Mature	100	100
Intermediate	97.2	98.5
Immature	99	99
Leaf	100.0	100

Table 3-8. Twenty-one candidate variables for separating the seven blueberry varieties

Variable	Indices	Description (largest difference between)
t1	index(R1375, R874)	Jewel v.s. Emerald
t2	index(R800, R313)	Jewel v.s. Farthing
t3	index(R2497, R1199)	Jewel v.s. Springhigh
t4	index(R710, R321)	Jewel v.s. Sweetcrisp
t5	index(R728, R321)	Jewel v.s. Star
t6	index(R799, R313)	Jewel v.s. Windsor
t7	index(R721, R357)	Emerald v.s. Farthing
t8	index(R2231, R804)	Emerald v.s. Springhigh
t9	index(R711, R321)	Emerald v.s. Sweetcrisp
t10	index(R721, R322)	Emerald v.s. Star
t11	index(R321, R200)	Emerald v.s. Windsor
t12	index(R2250, R800)	Farthing v.s. Springhigh
t13	index(R707, R325)	Farthing v.s. Sweetcrisp
t14	index(R2466, R920)	Farthing v.s. Star
t15	index(R721, R201)	Farthing v.s. Windsor
t16	index(R715, R360)	Springhigh v.s. Sweetcrisp
t17	index(R2232, R737)	Springhigh v.s. Star
t18	index(R2216, R800)	Springhigh v.s. Windsor
t19	index(R707, R319)	Sweetcrisp v.s. Star
t20	index(R709, R319)	Sweetcrisp v.s. Windsor
t21	index(R714, R200)	Star v.s. Windsor

Table 3-9. Validation results of the classification tree for separating varieties.

`	Sample size	Correctly classified		False positives		Missed	
		Count	%	Count	%	Count	%
Emerald	10	5	50.0	7	58.3	5	50.0
Farthing	14	7	50.0	10	58.8	7	50.0
Jewel	12	4	33.3	2	33.3	8	66.7
Springhigh	15	6	40.0	12	66.7	9	60.0
Star	12	3	25.0	6	66.7	9	75.0
Sweetcrisp	11	10	90.9	1	9.1	1	9.1
Windsor	9	1	11.1	4	80.0	8	88.9

Table 3-10. Validation results of the multinomial logistic model for the classification of seven blueberry varieties.

Varieties	Correctly classified (%)	False positives (%)	Missed (%)
Emerald	54.5	60.0	45.5
Farthing	50.0	61.1	50.0
Jewel	61.5	50.0	38.5
Springhigh	28.6	66.7	71.4
Star	14.3	71.4	85.7
Sweetcrisp	100.0	9.1	0.0
Windsor	23.1	66.7	76.9

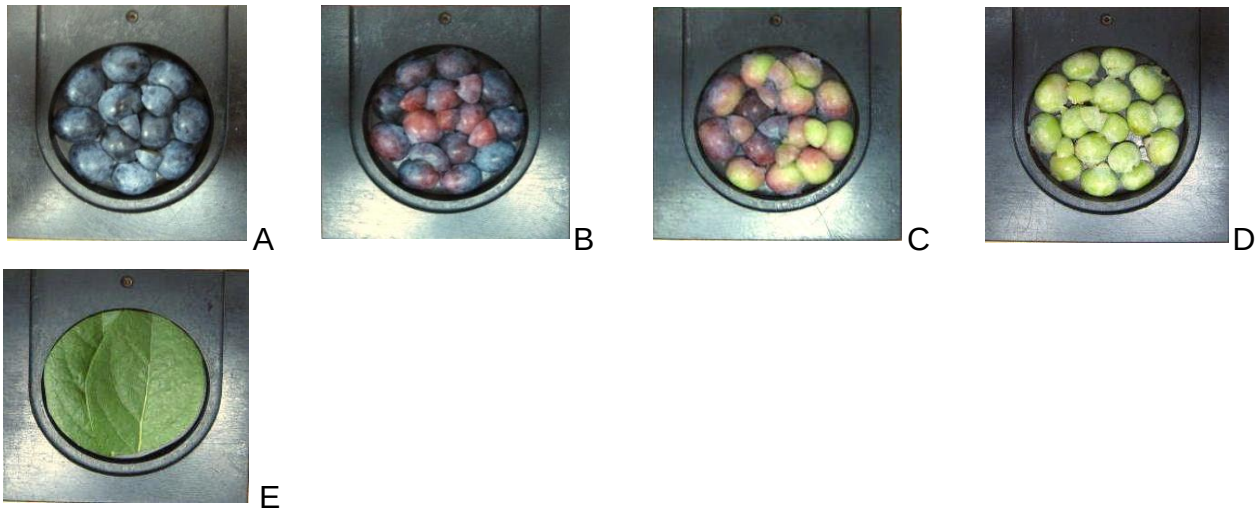


Figure 3-1. Sample preparation for the spectral reflectance measurement. A) mature fruit sample, B) near-mature fruit sample, C) near-young fruit sample, D) young fruit sample, E) leaf sample.

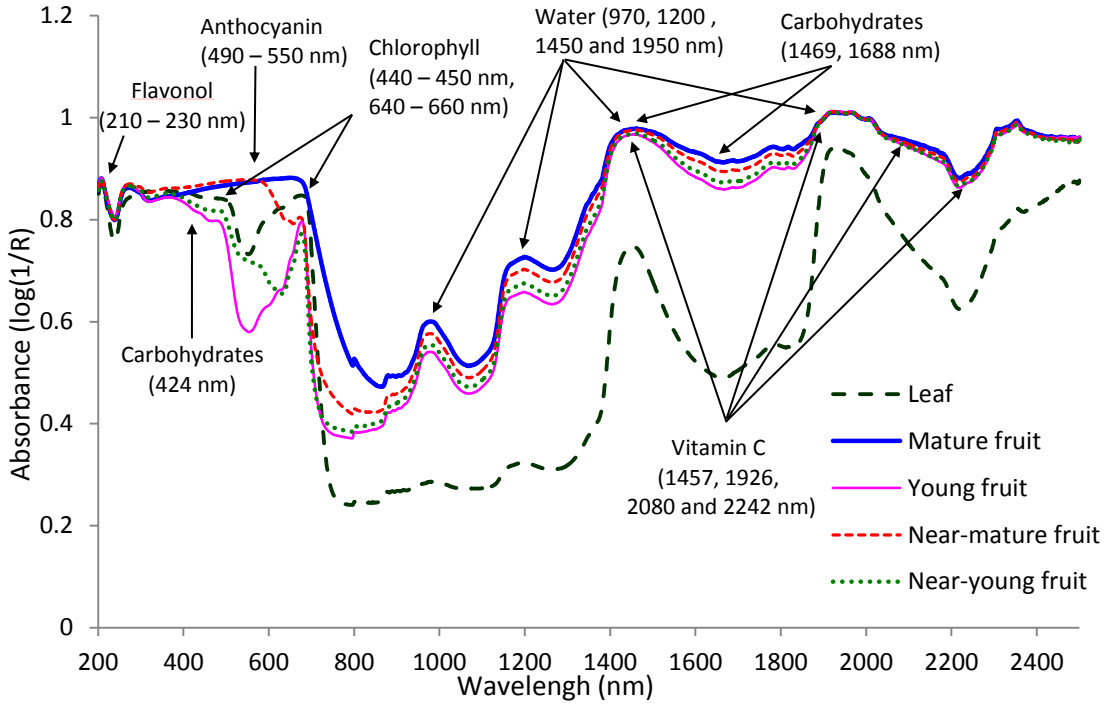


Figure 3-2. Spectral reflectance of blueberry fruit and leaves after moving average.

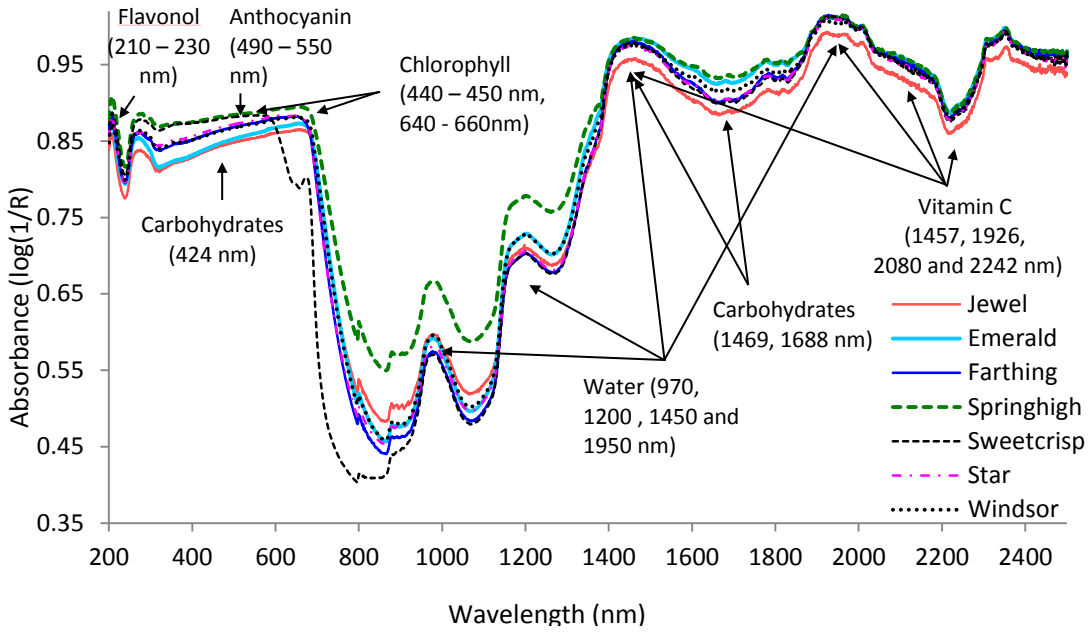


Figure 3-3. Reflectance curves of mature fruit of different varieties.

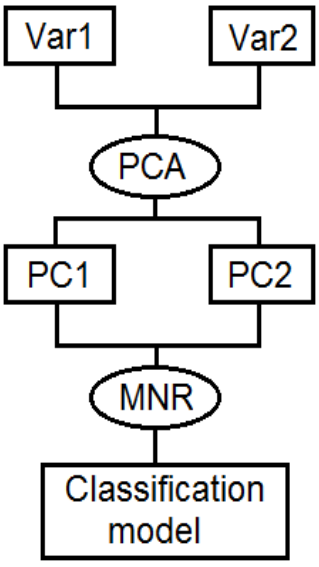


Figure 3-4. An example of logistic regression modeling based on two variables Var1 and Var2 (in the format of indices). PC1 and PC2 are the two principal components obtained from PCA.

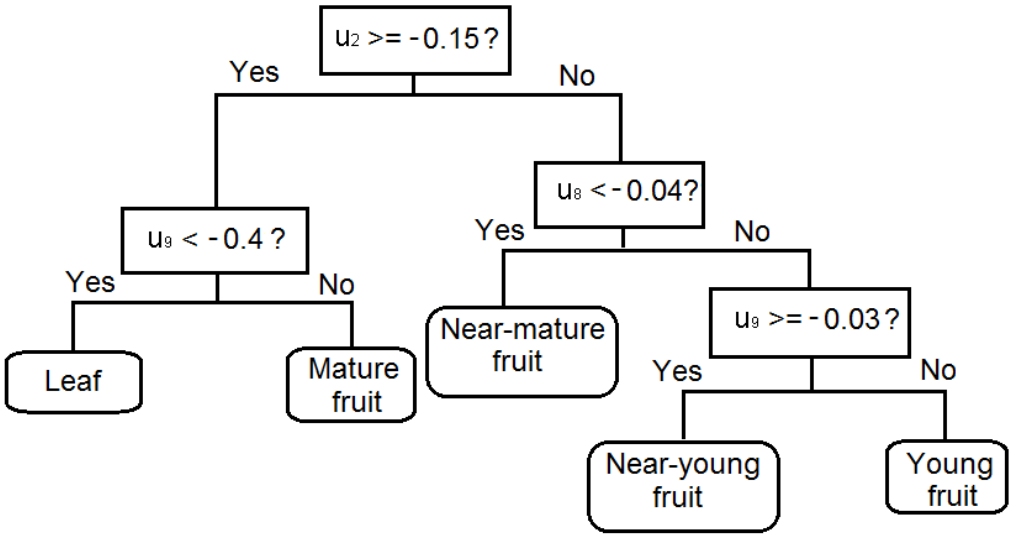
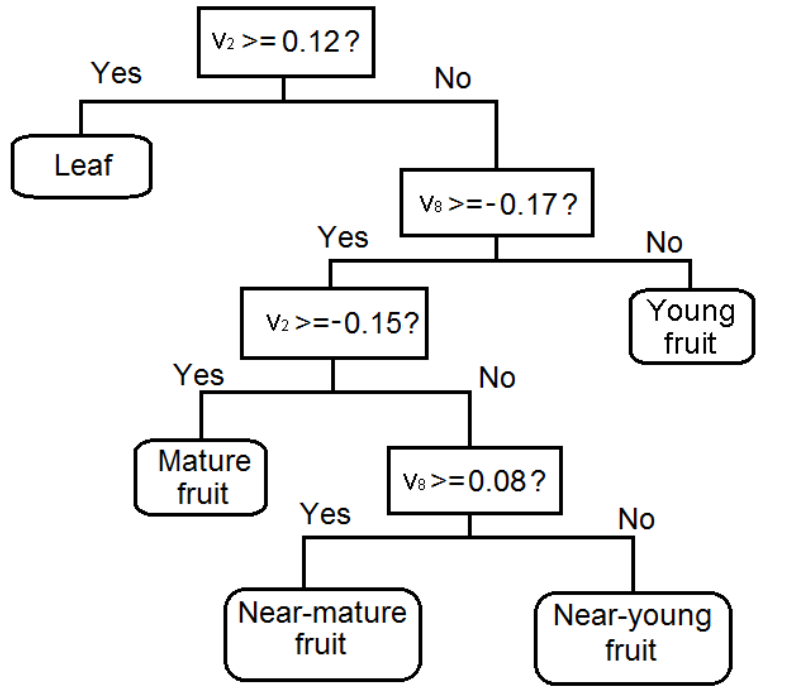
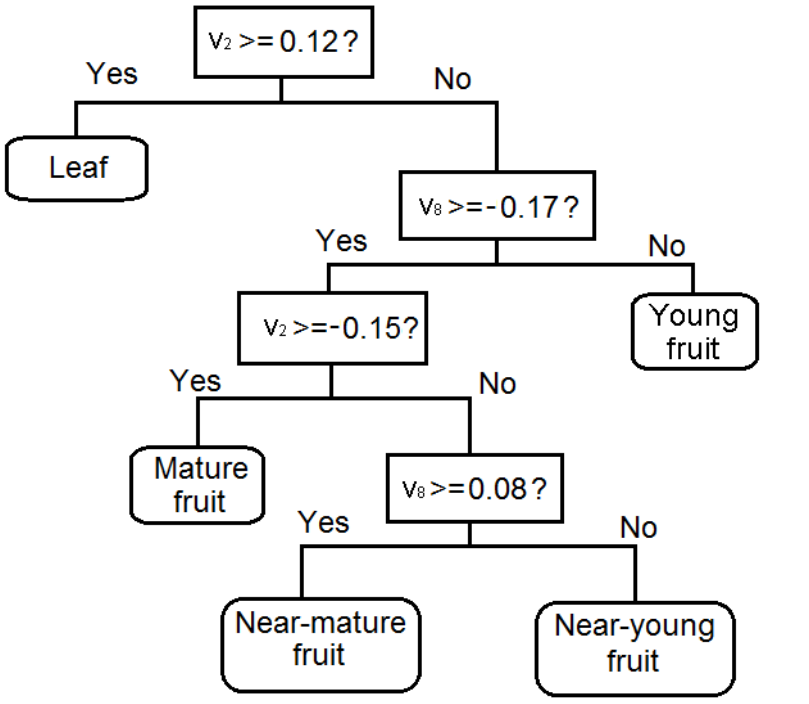


Figure 3-5. Classification tree result for separating five classes of 2011 samples.



A



B

Figure 3-6.Classification tree of two varieties. A) Classification tree of Farthing; B) Classification tree of Star.

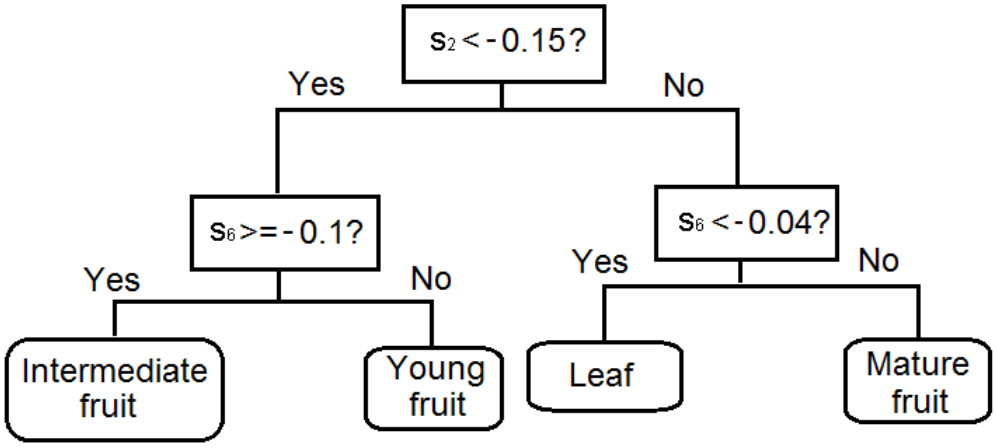


Figure 3-7. Classification tree of two years' data divided into four classes.

CHAPTER 4
BLUEBERRY FRUIT DETECTION BY BAYESIAN CLASSIFIER AND SUPPORT VECTOR MACHINE BASED ON VISIBLE TO NEAR-INFRARED MULTISPECTRAL IMAGING

Background

Florida is a major supplier of early fresh blueberries with high market value. The southern highbush blueberry varieties are the main products for fresh blueberry markets. All the fruits are handpicked, which is very labor-intensive. Therefore, the harvest cost is very high. On the other hand, large-scale commercial blueberry farms tend to be more cost effective and competitive. However, as the blueberry plantings become larger, field conditions such as soil type and irrigation needs will likely vary. These factors can lead to yield variation. Therefore, yield mapping in large scale blueberry farm is very important for the efficient deployment of labor in order to reduce harvest expense and increase profits.

Yield monitors for crops such as wheat and rice have been commercially used for a long time. The monitors measure grain flow, moisture, area covered, and location. Yield estimations are obtained based on crop yield models, and yield maps are generated to show the yield variation in the field. Most fruit varieties for fresh markets are harvested by manual labor. Yield estimation of fruits has different approaches. Schueller et al. (1999) developed a citrus yield measurement system by mounting load cells in open-air vehicles, which were called “goat trucks”. Zaman et al. (2006) estimated the citrus yield by ultrasonically sensing tree size. Machine vision was broadly used by researchers for fruit estimation. Image processing based on RGB image, multispectral and hyperspectral image, thermal image, etc. recently became the main approach for fruit yield estimation.

Fruit detection based on RGB imaging has already been applied to many fruit varieties. Color spaces were considered as useful features for the detection of fruits. Regunathan and Lee (2005) identified citrus fruit from a color camera and obtained fruit size by an ultrasonic sensor, which was used to measure the distance between fruits and camera. They converted RGB color space to Hue, Saturation and Intensity (HSI) as the features for classification. Bayesian, neural network and Fisher's linear discriminant were implemented for fruit detection and basic trigonometry was used for size estimation based on the distance value returned by the ultrasonic sensor. Wijethunga et al. (2008) investigated the use of RGB image under constant light source for the development of a kiwifruit counting system. They converted the image to L*a*b color space and obtained thresholds for the detection of fruit. The accuracy of their methods was 90% for gold kiwifruit images and 60% for green kiwifruit images. Kurtulmus et al. (2011) used the 'eigenfruit' approach and Gabor texture features to detect immature green citrus in natural outdoor condition. 75.3% of the actual fruits were correctly detected for a validation image set. Chamelat et al. (2006) used both RGB and HSI color spaces and Zernike moments as features for detecting grapes. They found that Zernike moments were very helpful and they obtained less than 0.5% errors with small training database. Zaman et al. (2008) estimated wild blueberry (lowbush) fruit yield by counting blue pixels in the center rectangular research region in the images. They obtained accuracies of 98% and 99% for different fields. However, the images were obtained in ideal conditions. There were only mature blueberries, and leaves were background. There were no noises such as ground, sky, etc. in the images.

Besides RGB imaging, some researchers focused on multispectral and hyperspectral imaging for fruit detection, disease detection and fruit quality estimation, etc. Lu (2003, 2004) did research on apple bruises and soluble solids content (SSC) estimation by multispectral imaging and near-infrared hyperspectral imaging and obtained critical wavelengths for the prediction. Peng and Lu (2008) studied on apple firmness by analyzing hyperspectral scattering profiles. Their optimal model obtained correlation coefficient 0.894 for apple firmness estimation and 0.883 for apple SSC estimation. Kane and Lee (2007) detected green fruit pixels using band pass filters based on the wavelength selection in their previous research (Kane and Lee, 2006). Okamoto and Lee (2009) used hyperspectral imaging to detect in-field green citrus and obtained 70 – 85% success rates in pixel identification for different citrus varieties and 80 – 89% success rates in fruit identification. Yang and Lee (2011) investigated multispectral properties of blueberry fruit and leaves using classification tree and logistic regression model and obtained six wavelengths for the prediction of several varieties of southern highbush blueberry fruit. However, at the time of this publication, no study has been conducted on blueberry detection based on multispectral imaging.

Therefore, the objectives of this study were to investigate the feasibility of blueberry detection based on color to near-infrared images and to build a robust classification model, which tolerates outdoor illumination changes and complicated background information. The results of this study will be used for a blueberry yield mapping system for large scale blueberry farms based on multispectral imaging.

Materials and Methods

Image Acquisition

Multispectral images were acquired from a commercial blueberry farm in Waldo, Florida, U.S. from 20 April 2011 to 19 May 2011 using a multispectral camera (Agricultural Digital Camera, Tetracam Inc., Chatsworth, California, U.S.A.). The images included three channels: near infrared (NIR), red (R) and green (G), according to the TM4 (760 nm – 900 nm), TM3 (630 nm – 690 nm) and TM2 (520 nm – 600 nm) bands of the LANDSAT satellite. A Teflon reference board was used in the field to eliminate the outdoor illumination variation by an image-preprocessing step, which was described later in this study. Eighty multispectral images were taken at a distance of 0.5 m – 0.7 m. For each multispectral image, an RGB image was taken at the same distance and direction in order to distinguish fruit stages and background classes in the multispectral image. An example of the multispectral image with referencing RGB image is shown in Figure 4-1. Plants shown in Figure 4-1A) have strong signal in the NIR channel. This represents strong reflectance of plants in the NIR region. According to the RGB image, the globular objects are fruits. In the multispectral image, the mature fruits are dark red, intermediate fruits are light red to yellow, and young fruits are light pink. The background of the image has leaves, branches, soil, sky, and reference board (in the lower left corner of the multispectral image).

Image Preprocessing

The original color components of the multispectral images were the NIR, R, and G. The working principle of the camera was to collect the reflectance of the surface from three filters. Therefore, the images were heavily influenced by the outdoor illumination changes in the open field.

The images were first preprocessed in order to eliminate the influence of outdoor illumination change. Pixel values were adjusted with the mean value of the reference board in each image by Eq. (4-1).

$$V'(x,y) = \frac{V(x,y)}{Mean(reference)} * 256 \quad (4\text{-}1)$$

Where:

V(x, y) is the original pixel value of (x, y) in the image,

Mean (reference) is the mean value of the reference board,

V'(x, y) is the adjusted pixel value of (x, y).

The histograms of the multispectral images were later stretched and equalized evenly to the full range of 0-255. This process enhanced the contrast of the image and corrected images of under- and over-illumination to some extent. A 3x3−pixel moving window average filter was used to reduce noise.

Feature Extraction

The NIR channel was considered to be sensitive to live plants and different textures. Therefore, in order to use the information in the best way, HSI, YIQ and YCbCr color space conversions on the multispectral images were explored and used as features. To distinguish color space conversions for multispectral image from those for RGB image, they were named MHSI, MYIQ and MYCbCr, where 'M' represents multispectral.

The multispectral images were first divided into two regions: fruit and background. Fruit was the region of interest. Later the fruit region was divided into mature fruit, intermediate fruit and young fruit classes according to the fruit growth stages. In the first step of classification, 5000 pure fruit pixels including all stages were

extracted from the image set. The other 5000 pure background pixels containing all possible background objects were extracted as the other class. Classification models based on different classifiers were constructed and tested for separating these two classes. In the multiple-class classification, 1000 pixels for each class (mature fruit, intermediate fruit, young fruit, leaf, branch, soil, sky, and reference board) were collected and classification models were built and tested.

Classifier Application

Two classification techniques were used for the separation of the classes. The first one is the Bayesian classifier. Bayesian classifier is probabilistic based classifier, which requires the features to be in specific distributions, and independent of each other. Eq. (4-2) shows the probabilistic model of the Bayesian classifier.

$$p(C_i|x) = \frac{p(C_i)p(x|C_i)}{p(x)} \quad (4\text{-}2)$$

Where:

$p(C_i|x)$ is the probability of the instance with feature vector x being in class C_i, it is called the posterior,

$p(C_i)$ is the prior, which gives the probability of class C_i,

$p(x|C_i)$ is the likelihood, which gives the probability of instance vector x under the condition that it is in class C_i,

$p(x)$ is the evidence, which is used as a scalar that guarantees the posteriors sum to 1.

Bayesian classifier requires that the features are independent from each other. Therefore, the features were first analyzed using principal component analysis. The uncorrelated principal components were then fed into the Bayesian classifier in Weka,

which is a data mining workbench developed by a machine learning group at the University of Waikato, New Zealand. Bayesian logistic regression, BayesNet and complete naïve Bayesian classifiers were applied with optimized parameters.

The second classification technique applied to the data set is the support vector machine (SVM), which is a supervised classification method. SVM first maps the training sets $\{x_i\}$ into a space H, which may have infinite dimensions. Then an optimal decision hyperplane for separating the training set into different classes would be found. The elements in the training set are already classified into the right classes since they are for supervised learning. The easiest case of using SVM is for separating two classes by the computation of a two dimensional linear margin. Figure 4-2 shows how the margin looks for this case. The four elements on the margin, each two of which are from one class, are called the support vectors. The basic objective is to maximize the margin while reducing the number of outliers, which lie between the margin lines.

The transformation is applied in an implicit manner by applying kernel, and finally the decision function is written as in Eq. (4-3).

$$f(x_i) = w * x_i + b \tag{4-3}$$

Where:

$f(x_i)$ is the projected value of element x_i,

w is the weight vector for element x_i, which maps the feature vector to one dimension,

b is the bias, which enables the separating hyperplane to not necessarily cross the zero point,

'$*$'is kernel application when the kernel can be polynomial function, radial basis function, etc.

If $f(x_i) \geq 1$, then x_i is classified into class 1, and If $f(x_i) \leq -1$, then x_i is classified into class 2. For the outliers that may appear between the margins lines, a parameter C is used as a penalization coefficient. C provides a compromise between outlier counts and width of margin. In order to build the best model for the problem, different values should be given to C and the best validation result would be obtained.

Classification Results

Fruit/Background Classification

As described in the second section, 5000 pure fruit pixels and 5000 background pixels were collected with information of four color spaces. 66% of the pixels were used as a calibration set and the other 34% were used as a validation set. An illustration image of the fruit/background classification is shown in Figure 4-3 B). The white pixels represent the fruit pixels and the black pixels perform like a mask for the background pixels in Figure 4-3 A).

When using the Bayesian classifiers, the prior probabilities of fruit class and background class were both set to be 0.5. The PCA obtained three principal components for the Bayesian classifiers. These principal components are shown in Eq. (4-4) to (4-6).

$$\mathrm{PC1} = 0.35\mathrm{NIR} + 0.35\mathrm{MI_{HSI}} + 0.34\mathrm{MY_{YCC}} + 0.31\mathrm{MI_{YIQ}} \quad (4\text{-}4)$$

$$\mathrm{PC2} = -0.55\mathrm{MS_{HSI}} + 0.43\mathrm{R} - 0.32\mathrm{MI_{YIQ}} - 0.29\mathrm{MQ_{YIQ}} \quad (4\text{-}5)$$

$$\mathrm{PC3} = -0.63\mathrm{MCb_{YCC}} + 0.58\mathrm{MH_{HSI}} - 0.33\mathrm{MQ_{YIQ}} - 0.2\mathrm{MS_{HSI}} \quad (4\text{-}6)$$

Where $\mathrm{PC1}, \mathrm{PC2}, \mathrm{PC3}$ are the three principal components, $\mathrm{MI_{HSI}}$ is the 'I' component of MHSI space, $\mathrm{MY_{YCC}}$ is the 'Y' component of MYCbCr space.

These three uncorrelated new attributes were used as an input for the Bayesian classification. The best model was achieved using the BayesNet classifier. The parameters (Bouckaert, 2004) of the classifier are shown in Table 4-1.

The threshold for the Bayesian model was 0.5. If the probability of an instance being of Class 1 is larger than 0.5, then it is from Class 1, vice versa. The prediction result is shown in Table 4-2.

The accuracy of this method is shown in Table 4-3. The true positive (TP) rate for fruit pixels was 84% and for background pixels was 67%. There were 32% of actual background pixels misclassified into the fruit class (false positive, FP rate). However, only 17% of actual fruit pixels were classified into the background class. The accuracy is the proportion of the examples which truly have class *x* among all those which are classified as class *x*. The result shows that the accuracy of fruit class was lower than that of the background class. It means that the model tended to classify the pixels into the background class rather than the fruit class.

In Weka, there were several methods of obtaining optimization for support vector machine, and sequential minimal optimization (SMO) was used in this study. There was no prerequisite for the feature vector when using the support vector machine. Therefore, the original calibration data was fed into the SMO classifier, which is available under Weka.Classifier.Functions. The parameters (Keerthi et al., 2001) of the classifier are shown in Table 4-4.

The prediction result of the SMO classifier is shown in Table 4-5. Compared to the Bayesian classification result, there were more correctly classified pixels and fewer misclassified pixels.

The accuracy of this method is shown in Table 4-6. The true positive rate for fruit pixels was 84%. For background pixels it was 72%. There were 27% of actual background pixels that were misclassified into the fruit class, however only 16% of actual fruit pixels were classified into the background class. The classification accuracies for fruit and background classes were 76% and 82%, respectively. Both accuracies were higher than those obtained from the BayesNet classifier.

A comparison of the TP rate and FP rate of the two methods are shown in Figure 4-4. Although BayesNet and SMO models obtained equal TP rate for the fruit class, SMO obtained higher TP rate for the background class. The SMO model obtained lower FP rate for both fruit and background classes.

Eight-Class Classification

Similar to the fruit/background classification, the eight-class classification was also applied using the Bayesian and SVM techniques. The eight classes were mature fruit, intermediate fruit, young fruit, leaf, branch, soil, sky, and reference board (the classes were numbered with indices 1, 2, 3, 4, 5, 6, 7, 8, respectively). An illustration image of the classification is shown in Figure 4-5. The left image is the original multispectral image, and the right image is the illustration of the eight classes with indices 1- 8. The classification has the aim of detecting a single fruit stage so that it will help the farmer to either predict the yield in the field, or predict the yield for the next one or two weeks by estimating the amount of intermediate or young fruit. One thousand pure pixels of each class were collected, and 66% of them were in a calibration set, the other 34% were in a validation set.

Before the application of Bayesian classifier, five uncorrelated features were obtained by the PCA from the color features. For the classification of eight classes,

BayesNet classifier still performed the best, with the same parameters except that the initial account alpha became 0.6. The prediction result of the BayesNet classifier is shown in Table 4-6.

The statistics of the BayesNet classifier are shown in Table 4-8. Mature fruit, which is one of the main targets in this classification, obtained 77% of true positive rate, and 13% of false positive rate. However, the intermediate fruit, young fruit, branch and leaf were not well classified. The significantly different classes such as soil, sky and reference board obtained relatively higher prediction accuracy. However, the accuracies for the fruit classes were approximately 50%, which were lower than the fruit/background classification.

The performance of the SMO classifier is shown in Table 4-9. The configuration of the parameters for the SMO was the same as that in the fruit/background classification.

The statistics of the SMO classifier for the eight-class classification is shown in Table 4-10. Mature fruit obtained 70% of TP rate, and 9% of FP rate. These two rates were both lower than those obtained by the BayesNet classifier, which means that the SMO classifier performed worse in the correct classification of mature fruit, but reduced the false detection to some extent. However, the intermediate fruit, young fruit, branch, leaf, soil and sky were all better classified. Therefore, the lower TP rate is the trade-off for the better performance of the other classes. The accuracies of the fruit classes are higher than 50% in this model, which means the SMO classifier performed better than the BayesNet classifier.

The performances of the two classifiers are also compared by the charts of TP rates and FP rates in Figure 4-6 visually. SMO model obtained lower TP rate for the mature fruit class, however higher TP rate for all other classes. FP rate of SMO model for all classes are lower than that of BayesNet model.

Discussion

According to the test results of the classifiers, the support vector machine performed better than the Bayesian classifiers in general. In the fruit/background classification, the TP rate was higher for both classes when using the SMO classifier and the FP rate was lower for both classes in the SMO model. The accuracies also showed that the SMO model performed better because higher accuracies mean less misclassification in the validation data set. The overall accuracy of the SMO model was 79%. The eight-class classification yielded less satisfying result. Still, SMO model performed better in the eight-class classification according to the TP rate, FP rate and accuracy in Tables 8 and 10. The main reason of lower accuracy of Bayesian classifier may be that the variables were not independent, and the distributions of some variables were unknown. In contrast, there was no requirement for the variables when using the support vector machine.

As the target classes, the different growth stages did not show significant differences using the color space features. The TP rate from the BayesNet model and SMO model were both low for the fruit classes, ranging from 38% to 77%. The possible reason for the low classification accuracy for different fruit stages is that the bandwidth of the three channels in the multispectral image was too wide. According to Yang and Lee's (2011) study on the spectral analysis of blueberry fruit and leaves, six

wavelengths (500, 525, 550, 575, 680, and 750 nm) should be used for the classification of different fruit stages from leaves. However, 525, 550 and 575 nm are all in TM2, 680 nm is in TM3, 500 nm and 750 nm are not in any of the TM2, TM3 and TM4 bands of the multispectral image. Therefore, more and narrower bands should be used in order to obtain better features for the classification of different fruit stages.

Conclusion

Multispectral blueberry images were collected and the color spaces NIR-R-G, MHSI, MYIQ and MYCbCr were used as the feature vectors for separating fruit and background classes. Bayesian classifier and support vector machine were applied for fruit/background classification and the eight-class classification. BayesNet classifier and SMO classifier were investigated and proved to have the best performance under specific parameter configuration. In the fruit/background classification, SMO outperformed BayesNet classifier with higher TP rate and lower FP rate for both classes. The support vector machine achieved 84% TP rate and 27% FP rate for the fruit class, and 73% TP rate and 16% FP rate for the background class. The eight-class classification of BayesNet classifier showed strong classification power for mature fruit, which yielded 77% TP rate and 13% FP rate. However, it did not perform well for the intermediate fruit, young fruit, leaf, and branch classes. The SVM classifier obtained lower TP rate for the mature fruit class. However, it performed better for all other classes. Therefore, the overall accuracy of the SMO model was higher than that of the BayesNet model. The importance of new knowledge gained from this research is that color spaces (MHSI, MYIQ, MYCbCr) of multispectral image are useful features for the detection of blueberry fruits using proper classifiers, such as a support vector machine. However, to classify different growth stages of fruits, features according to the spectral

signatures of blueberry should be considered. Cameras with more and narrower bands according to the spectral signature analysis of blueberry may be helpful in the classification of different fruit growth stages.

Table 4-1. Parameters of BayesNet classifier.

Parameter	Value
Estimator	Simple Estimation with alpha 0.5
Search Algorithm	K2
Initial as Bayesian Classifier	True
Random Order	True
Score Type	Bayes

Table 4-2. Prediction result of BayesNet classifier in pixel amounts.

	Predicted fruit pixel	Predicted background pixel
Actual fruit pixel	1424	282
Actual background pixel	553	1141

Table 4-3. Accuracy of BayesNet classifier.

Class	TP Rate (%)	FP Rate (%)	Accuracy (%)
Fruit	84	33	72
Background	67	17	80

Table 4-4. Parameters of SMO classifier.

Parameter	Value
C	8
Epsilon	1.0e-12
Filter type	Normalize training data
Kernel	PUK with Omega=1 and Sigma=1
Tolerance parameter	1.0e-3

Table 4-5. Prediction results of SMO classifier for fruit/background classification in pixel amounts.

	Predicted fruit pixel	Predicted background pixel
Actual fruit pixel	1437	269
Actual background pixel	462	1232

Table 4-6. Accuracy of SMO model for fruit/background classification.

Class	TP Rate (%)	FP Rate (%)	Accuracy (%)
Fruit	84	27	76
Background	73	16	82

Table 4-7. Predicted results of BayesNet classifier for eight classes.

	Predicted Mature fruit	Predicted Intermediate fruit	Predicted Young fruit	Predicted Branch	Predicted Leaf	Predicted Soil	Predicted Sky	Predicted Reference
Mature fruit	271	17	16	9	28	8	0	2
Intermediate fruit	75	130	47	60	13	17	3	1
Young fruit	61	50	126	66	16	4	1	1
Branch	49	30	48	109	58	52	4	0
Leaf	130	5	2	42	171	1	0	0
Soil	1	7	5	7	0	247	36	9
Sky	0	1	2	0	0	34	278	29
Reference board	0	0	0	0	0	0	5	336

Table 4-8. Classification results of BayesNet model for eight classes.

Class	TP Rate (%)	FP Rate (%)	Accuracy (%)
Mature fruit	77	13	46
Intermediate fruit	38	5	54
Young fruit	39	5	51
Branch	31	8	37
Leaf	49	5	60
Soil	79	5	68
Sky	81	2	85
Reference board	99	2	89

Table 4-9. Predicted results of SMO classifier for eight-class classification.

	Predicted Mature fruit	Predicted Intermediate fruit	Predicted Young fruit	Predicted Branch	Predicted Leaf	Predicted Soil	Predicted Sky	Predicted Reference
Mature fruit	247	34	15	11	42	2	0	0
Intermediate fruit	55	159	47	45	33	12	6	0
Young fruit	35	37	166	61	17	7	0	2
Branch	16	36	47	159	59	31	2	0
Leaf	92	1	4	35	217	2	0	0
Soil	3	2	0	9	0	269	19	10
Sky	0	2	1	0	0	19	309	13
Reference	1	0	1	0	0	0	3	336

Table 4-10. Statistics of the SMO classifier for eight-class classification.

Class	TP Rate (%)	FP Rate (%)	Accuracy (%)
Mature fruit	70	9	55
Intermediate fruit	46	5	59
Young fruit	51	5	60
Branch	45	7	50
Leaf	62	6	61
Soil	86	3	79
Sky	90	1	91
Reference board	99	1	93

Figure 4-1. Example multispectral and corresponding RGB images containing fruits with different growth stages, leaves, branches, soil, and sky. A) multispectral image; B) corresponding color image.

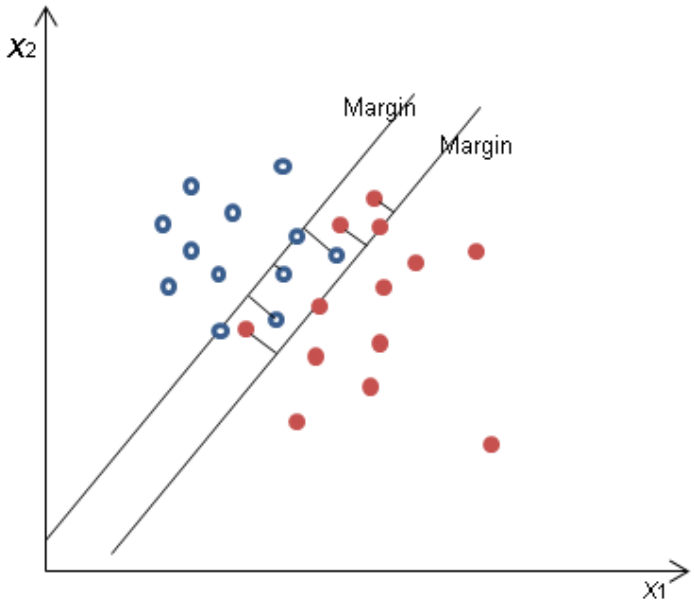
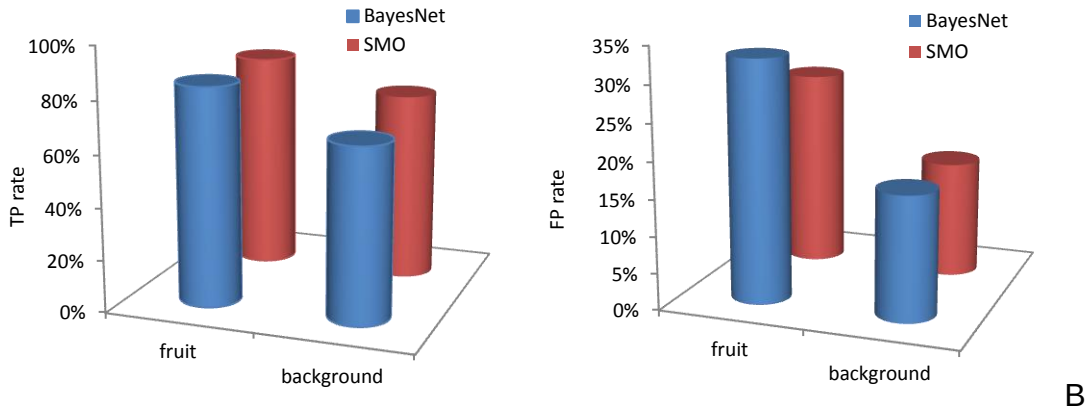


Figure 4-2. SVM for separating two classes by two-dimensional features.

A B

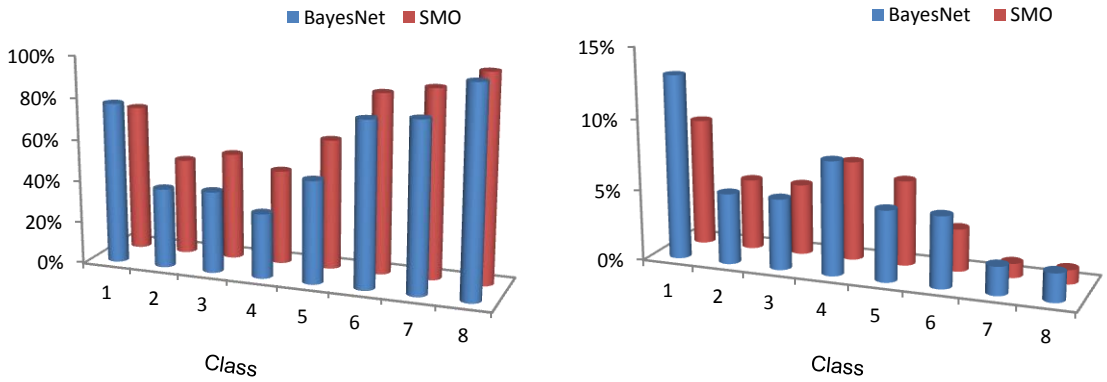
Figure 4-3. Example multispectral image and illustration of fruit/background classification. A) Original multispectral image; B) Binary image.



A B

Figure 4-4. TP rate and FP rate comparison of BayesNet and SMO models for fruit/background classification. A) TP rate comparison; B) FP rate comparison.

Figure 4-5. Illustration of eight-class classification. A) Original multispectral image; B) Illustration image.



A B

Figure 4-6. TP rate and FP rate comparison of BayesNet and SMO models for eight-class classification. A) TP rate comparison; B) FP rate comparison.

CHAPTER 5
HYPERSPECTRAL BAND SELECTION FOR DETECTING DIFFERENT BLUEBERRY FRUIT MATURITY STAGES

Background

Labor expenses for handpicked blueberries for fresh markets are increasing due to severe shortages of available farm workers. Management cost of Florida's commercial blueberry field excluding harvesting labor is approximately $9,884/ha (Williamson et al., 2012). The average blueberry yield in Florida is 6,310 kg/ha (USDA, 2012). Morgan et al. (2011) estimated that the hand harvest cost was $1.59/kg. Therefore, the cost of harvesting labor takes more than $10,000 /ha, which is more than half of the total management cost of the blueberry field. Efficient harvesting labor assignment in large blueberry field can reduce much of the harvesting cost. Furthermore, yield estimation prior to harvest helps grower to find problems in their field as early as possible. It is useful for growers to make further decisions such as irrigation, pest control, weed control, etc. Therefore, yield estimation of blueberry field prior to harvesting is beneficial for the farmers. During the harvest season, individual blueberries in a fruit cluster mature at different times. A cluster may contain all growth stages including young fruit (green color), intermediate fruit (pink/red color) and mature fruit (dark blue/purple) at the same time. Figure 5-1 is an example picture taken from the blueberry field during the blueberry harvest season in 2013, showing different fruit maturity stages and colors.

Efficient labor deployment based on yield monitoring requires that the yield be estimated in advance of berry ripening. Remote sensing is a method of detecting objects without physically touching or breaking them. Therefore, it is logical to use remote sensing for the yield estimation of fruit amount of different growth stages. Wild

blueberry fruit estimation was carried out by digital image processing (Zaman et al., 2008) and high prediction accuracy was obtained. The color images of wild blueberry in this study contained only mature fruit, which was easily distinguishable because of its significant color difference in the blue band. However, as shown in Figure 5-1, a southern highbush blueberry cluster has all growth stages at the same time. It is difficult to distinguish young fruits and intermediate fruits from the noisy background in the visible range. To estimate the blueberry yield in advance of harvesting, all growth stages should be detected so that all fruits on the bushes are considered. Blueberry spectral property was analyzed based on laboratory measured spectral data by Yang et al. (2012). The analysis showed that hyperspectral property would be helpful in classifying different growth stages of blueberry fruit.

While blueberry spectral properties have been analyzed in the laboratory, it cannot be coupled with field measurement directly because of their different measurement conditions. The laboratory is a more ideal environment because of its its uniform and stable indoor light source. In addition, the samples were well prepared without much noisy background. However, field measurement uses sun light as its illumination source, and the background contains not only leaves, but also soil, sky, and man-made objects such as PVC irrigation pipes. Portable spectrometer can only measure either a spot or an area as one spectrum, which would be insufficient for recognition of different fruit maturity stages. Color image is not capable of detecting all the fruit maturity stages because of the similar color of young fruit and leaf. Hyperspectral images obtained from the in-field condition have both high spatial and spectral resolution. Therefore, hyperspectral imagery can be used for the detection of

blueberry of different growth stages in the field with complicated background information.

Because of their high spectral resolution, hyperspectral images contain considerable amounts of redundancy. The images usually have several hundred bands, but some bands are useless or even hinder the discriminability of the useful bands. Adjacent bands in the spectrum tend to have very high correlation (Cai et al., 2007). Band extraction methods such as principal component analysis (PCA), and maximum noise fraction (MNF) reduce dimensionality by projecting the original bands into new dimensions. However, the projected features combine the original information in these methods and do not have physical meaning. In contrast, band selection methods choose original features, which are physical information. The selected original bands can be used for a multispectral camera system for yield estimation. Multispectral camera is of lower cost and higher processing speed compared to a hyperspectral camera system. Therefore, a multispectral camera with selected bands is more suitable for the task of blueberry yield prediction.

During the last decade, many band selection methods have been developed as preprocessing of hyperspectral image analysis. Some methods used different criteria to measure the importance of bands. The separability of bands may be measured with transformed divergence, Bhattacharyya distance, and Jeffries-Matusita distance (Yang et al., 2011). Other methods employed a criterion to prioritize bands, and then bands with the highest rankings in dissimilar band clusters are selected. The band ranking criterion contains variance, correlation, signal-to-noise ratio (SNR), etc. Information measures have also been used for hyperspectral band selection using mutual

information or information divergence (Martinez-Uso et al., 2007). However, the purpose of these band selection methods was to reduce data volume and calculating complexity. They did not focus on what specific bands were selected.

The objectives of this study were to show the feasibility of hyperspectral imagery in blueberry growth stage classification, and to select useful bands that are suitable for multispectral imagery, which is of lower cost and higher processing speed. The selected bands are supposed to yield a high accuracy of classification. A supervised band selection method based on Kullback-Leibler divergence (KLD) was proposed. This method measures the pair-wise class discriminability (PWCD).

Materials and Methods

Hyperspectral Image Acquisition

Hyperspectral images were obtained from the blueberry research and demonstration farm in the University of Georgia cooperative extension in Alma, GA, United States (31.53438°, -82.51019°, WGS84) in July, 2012. There were ten rows with 20 trees per row. In each row, four trees were randomly selected for hyperspectral image acquisition. Therefore, a total of 40 images were obtained. In each image, an area of 15.24 x 15.24 cm^2 of the view was acquired. The camera system was composed of a camera body (a line scanning spectrometer V10E, Specim, Oulu, Finland), a digital CCD camera (MV-D1312, Photonfocus AG, Lachen SZ, Switzerland), a lens (CNG 1.8/4.8-1302, Schneider Optics, North Hollywood, CA, USA), an encoder (Omron-E6B2, OMRON cooperation, Kyoto, Japan), a tilting head (PT785S, ServoCity, Winfield, KS, USA), an image grabber (NI-PCIe 6430, National Instruments Inc. Austin, TX, USA), a DAQ card (NI-6036E, National Instruments Inc. Austin, TX, USA), and a laptop (DELL Latitude E6500) with a control and vision acquisition program written in LabVIEW

(National Instruments Corporation, Austin, TA, USA). The tilting head carried the camera to rotate vertically. When the camera rotated, encoder generated square pulses, which would be sent to the program for generating trigger signal. The camera obtains one line image once it receives a trigger signal from the program.

The highest spectral resolution of the hyperspectral imaging system could be 0.79 nm. However, in that case, there would be a total of 776 bands. The image size would be very large. Therefore, binning was used to reduce the spectral resolution by half. After binning, there were 388 hyperspectral bands with a spectral resolution of 1.59 nm, which was sufficient for our study. The spectral range was 398 – 1010 nm. The spatial resolution was 1 mm. The radiance data was saved in 12-bit binary files. The data was processed to create image cubes of both spectral and spatial data. The size of the image cube is n (number of lines) × 1312 (pixels/line) × 388 (bands). Reflectance images were created using a universal white standard (Spectralon, Labshpere Inc., North Sutton, NH, USA). Figure 5-2 shows the RGB bands of one hyperspectral image. The red band is 690 nm, the green band is 550 nm and the blue band is 450 nm. Dark blue fruits are mature, red fruits are in the intermediate stage, and light green fruits are in the young stage.

Training and testing pixel sets were created by randomly collecting 600 pixels that were labeled manually referring to the relative digital color image and spectra of the pixels. A half of the pixels of each class: mature fruit, intermediate fruit, young fruit and leaf, were put in the training set and the other half were in the testing set.

Hyperspectral Band Selection

Matlab R2012a (The MathWorks Inc., Natick, MA, USA) was used to implement three hyperspectral band selection methods in this study. Kullback-Leibler divergence

was used in the three methods as the criterion of variation between distributions. Among the three methods, pair-wise class discriminability measure was proposed as a supervised band selection method. Ward's Linkage strategy using Information Divergence and non-Gaussianity ranking were originally applied as two unsupervised band selection methods. In this study, these two band selection methods were used based on the training set, which was labeled.

In information theory, Shannon entropy is used to quantify the information contained in a message. The entropy of a random variable X with a probability density function $p(x)$ is shown in Eq. (5-1).

$$H(X) = -\int_{\Omega} p(x) log p(x) dx \quad (5\text{-}1)$$

where Ω includes all possible events. Entropy is a measure of the amount of information of a random variable.

If X is a discrete random variable, then x is the value of X, and $p(x)$ is the probability mass function of all possible events. Entropy $H(X)$ for discrete random variable is defined in Eq. (5-2).

$$H(X) = -\sum_{x \epsilon \Omega} p(x) log p(x) \quad (5\text{-}2)$$

Kullback-Leibler divergence (KLD) is an information divergence measure, which shows the dissimilarity distance between two probability distributions. The original KLD is non-symmetric. Therefore, it is not a real distance. However, its symmetric version is used as a dissimilarity measure in many places (Webb, 2002). The symmetric KLD for discrete random variables is defined in Eq. (5-3).

$$D_{kl}(X_i, X_j) = \sum_{x \epsilon \Omega} [p_i(x) \log \frac{p_i(x)}{p_j(x)} + p_j(x) \log \frac{p_j(x)}{p_i(x)}] \quad (5\text{-}3)$$

Where X_i, X_j are the random variables. $p_i(x)$, $p_j(x)$ are the probability mass function, respectively. The random variables are defined in the finite Ω space.

If two random variables are the same, then the two probability mass functions are identical for every possible value x. The KLD value of this condition is 0.

If the variables are very different, their distributions will be far away from each other and the divergence value will be high. Therefore, it is a way of quantifying the difference of random variables. It can be seen as the cost of substituting one variable with another one. When used in hyperspectral band selection, KLD measures the discrepancy between the probability distribution of a pair of bands in an image or a pair of classes in one band.

The proposed method PWCD calculates the KLD value of pairs of classes in each band. In a specific band, each class is a random variable. Because the pixel values of a class in a specific band can be considered as sample space, the gray-level histograms of class i and class j are analogous to the probability distributions of the two classes. In order to ensure comparability, the histograms were normalized so that the values in each histogram summed up to one. Our goal is to find the band that has the most discrepancy between two classes. It is expressed as in Eq. (5-4).

$$\text{argmax}_B(D_{kl,B}(C_i, C_j)) = \text{argmax}_B \sum_{c\epsilon\Omega}[hist_{i,B}(c) \log \frac{hist_{i,B}(c)}{hist_{j,B}(c)} + hist_{j,B}(c) \log \frac{hist_{j,B}(c)}{hist_{i,B}(c)}] \quad (5\text{-}4)$$

Where B is the band number, $D_{kl,B}(C_i, C_j)$ is the KLD of class i (C_i) and class j (C_j) in band B. $hist_{i,B}(c)$ and $hist_{j,B}(c)$ are the normalized histograms of the two classes. The band that maximized the KLD of C_i and C_j was chosen. Since there are four classes, which make six pairs of classes, six bands were selected in the end.

Hierarchical dimensionality reduction (HDR) is an unsupervised band selection method. It calculates the KLD value of pairs of bands within a hyperspectral image. The normalized histograms of band i and band j are analogous to the probability distributions of the two bands. Therefore, the KLD of the two bands are expressed as in Eq. (5-5).

$$D_{kl}(B_i, B_j) = \sum_{b\epsilon\Omega}[hist_i(b)\log\frac{hist_i(b)}{hist_j(b)} + hist_j(b)\log\frac{hist_j(b)}{hist_i(b)}] \quad (5\text{-}5)$$

Where $\mathrm{D_{kl}(B_i, B_j)}$ is the KLD of band i ($\mathrm{B_i}$) and band j ($\mathrm{B_j}$) in an image. $\mathrm{Hist_i(b)}$ and $\mathrm{hist_j(b)}$ are the normalized histogram of the two bands.

Hierarchical clustering structure using agglomerative strategy (Martinez-Uso et al., 2007) is adapted so as to form the bands with high similarities into clusters. Ward's linkage method merges the clusters repeatedly till the required number clusters are produced. This method minimizes the total variance within each cluster, so that the features that have the least variance are clustered gradually. Bands from different clusters have very low correlation. The mean of each cluster was then obtained, and the representative band was the one that had the highest correlation with cluster mean.

The non-Gaussianity (NG) measure was originally called the information divergence (ID) method because it also utilizes the divergence criterion. However, it assesses the discrepancy of the real distribution with the associated Gaussian probability distribution. If one particular band is good at discriminating classes, its histogram should not be like Gaussian distribution. In contrast, the more the histogram differs from Gaussian distribution, the better. The difference between them can be expressed as in Eq. (5-6).

$$D_{kl}(B_i, B_{ig}) = \sum_{b\epsilon\Omega}[hist_i(b)\log\frac{hist_i(b)}{P_{ig}(b)} + P_{ig}(b)\log\frac{P_{ig}(b)}{hist_i(b)}] \quad (5\text{-}6)$$

Where B_i is band i, and B_{ig} is its associated random variable with Gaussian distribution. The Gaussian distribution $P_{ig}(b)$ was achieved using the mean and variance of the real distribution, which was simulated by normalized histogram $hist_i(b)$. The KLD value of the band is the NG measure. The bands are sorted with their NG measures. The band with greater KLD value has more priority because it has greater deviation from Gaussian distribution.

Supervised Classification

In order to compare the performance of the band selection methods, three supervised classifiers were applied on the testing data set.

K-nearest neighbor (KNN) classifier is one of the most fundamental and widely used classification methods. It is a non-parametric method based on the nearest training samples. Majority vote of the neighbors decides which class the testing sample belongs to. K is the number of the nearest neighbors that are taken into consideration. If K=1, the test sample is assigned only to the class of the single nearest neighbor. Larger K reduce the effect of noise and outliers in the classification, however the boundary between classes are less clear. It does not require a training step because all the distance calculations are in the testing step.

Support vector machine (SVM, Cortes and Vapnik, 1995) is another well-known and widely used classifier. It was originally designed to be a binary linear classifier where an instance was either assigned to one class or the other. The optimal hyperplane will be constructed with the maximum margin and support vectors. When there are more than two classes, different schemes can be used for the classification

task, such as one-against-all (Rifkin and Klautau, 2004). Classification of not linearly separable classes often happens in real problems. Therefore, the original finite-dimensional space is projected into a much-higher dimensional space, which makes the separations appear to be linear in the new space. Kernel function is introduced to replace the original inner product so as to transfer the space. Widely used kernels include polynomial kernel, Gaussian radial basis function (RBF) kernel, etc.

Adaptive Boosting (AdaBoost) (Freund and Schapire, 1995) is a meta-algorithm, which conjuncts multiple learning methods to improve the performance of classifiers. It is one of the most useful learning methods in the history of machine learning (Friedman et al., 2001). AdaBoost combines the outputs of multiple classifiers, which perform just slightly better than guessing. The method sequentially runs the classifiers, and the weight of each training sample is modified during the application of the classifiers. The wrongly classified samples are given higher weight in the next step of classification. The classifiers are also weighted by a majority vote with respect to their contribution. The classifiers that obtained higher accuracy are given higher weight. The weighted classifiers are finally combined to produce the AdaBoost classifier. Weka software from the University of Waikato was used to apply SVM and AdaBoost (Hall et al., 2009).

Results and Discussion

Blueberry Spectra

In order to show that the four observed classes have different spectra, one pixel of each class was selected and their spectra were shown in Figure 5-3. Leaves occupy most of the background. Therefore, a leaf pixel is used to represent the background spectrum. Mature fruit is very dark and therefore, has a very low value in the visible range. It also has relatively lower reflectance in the near-infrared range, as shown in the

Figure 5-3. Intermediate fruit appears red in color image, thus has higher value in the red band. It does not reflect much in the blue and green bands. Young fruit is bright green color. Therefore, it has a high reflectance value in the green band. Leaves have the highest chlorophyll content, which results in high reflectance in the NIR range.

Principal Component Analysis

The principal component analysis of the whole pixel set was carried out in order to check the feasibility of hyperspectral imagery for the separation of different classes of in-field blueberry crops. The first three principal components (PC) are extracted to show the distribution of the classes in Figure 5-4. Purple squares are from the mature fruit class, red dots are from the intermediate fruit class, light green stars are from the young fruit class, and the dark green squares are from the background class, of which leaves comprise the greatest part. Mature fruit, intermediate fruit and background pixels form two clusters each. The young fruit pixels are also scattered. The clustering and scattering of pixels in every class are mainly because of shadows during the daytime under direct sunlight. Another reason is that pixels from all possible conditions were collected, considering the depth of view, the influence of water evaporation, etc. All in all, although various conditions in the field have strong impacts on spectral properties, the four classes are obviously separable by the three PCs.

Band Selection Results

HDR and NG measure are unsupervised methods. However, they were applied to the training pixel set, which was also used in the supervised PWCD method. The pixels were correctly labeled because of the high spatial resolution of the hyperspectral images.

By calculating the KLD between probability distributions, PWCD method selected six bands for separating class pairs. Figure 5-5 shows the normalized histograms of the bands for class pairs that were selected by this method. Band 41 (457 nm) was used for the discriminant of mature fruit and intermediate fruit. Band 303 (870.5 nm) had the highest discriminability of mature fruit and young fruit. Band 68 (498.4 nm) was the best for separating mature fruit and background. Band 176 (666.7 nm) separated intermediate fruit and young fruit the best. Band 145 (617.9 nm) separated intermediate fruit and background with the best result. Band 164 (647.8 nm) achieved the best separation result for young fruit and background. Some of the histograms shown in Figure 5-5 have more than one main value range, which is mostly because of the shadow caused by the strong sunshine.

HDR was applied to the labeled training and testing pixel sets. This method aggregated the bands that had very similar normalized histograms from the training set. The bands were then grouped into clusters. The mean of each cluster was calculated. The band with the highest correlation with the mean of the cluster was chosen to represent the cluster. In the end, six bands were selected from six clusters. As expected, the clusters mainly aggregated neighboring bands. Figure 5-6 shows the band clustering result and the selected bands. The selected bands were: 7 (405.3 nm), 14 (415.9 nm), 77 (512.2 nm), 215 (728.6 nm), 248 (781.5 nm), and 279 (831.5 nm). Cluster 6 contains the most bands, covering from band 23 (429.6 nm) to band 203 (709.5 nm). This cluster goes through the visible range and the red edge, from where only one wavelength should be chosen. Therefore, this might be a loss of useful spectral information.

The NG-measure method directly sorted the bands by Gaussianity. The top bands chosen were those with the highest KLD values between the original distribution and the simulated Gaussian distribution, which are the NG measures. The result of this method is shown in Table 5-1. The top 20 bands are listed in this table. However, the bands are very close to each other. For example, the first and sixth bands are neighbors. The second, third and fifth bands are also neighbors. It is already shown in the HDR method that near-by bands have higher correlations. The first column of Table 5-1 is the ranking, the second column is their NG measures, and the groups of numbers are in the last column. The NG measures decrease quickly from the first ranked band to the second ranked band, however they decrease much slower after that. The top ranked band in every group was chosen as the representative band for that group. The selected bands are underlined in Table 5-1. However, there are only five groups for the top 20 bands. Therefore, more bands were investigated and a sixth band chosen was band 142 (613.2 nm), which was ranked the 21st. The final selected bands are: 192 (692.1nm), 246 (778.3 nm), 175 (665.1nm), 181 (674.6 nm), 162 (644.6nm) and 142 (613.2 nm).

Classification Using Band Selection Results

KNN classifier, SVM and AdaBoost were applied to test the performance of the bands selected by the three methods. The classification results of using bands selected by PWCD are shown in Table 5-2. Intermediate fruit and young fruit are relatively easier to distinguish than mature fruit and background. AdaBoost obtained the best accuracy and the lowest false positive rate when using the functional trees (FT) classifiers. AdaBoost is an advanced machine learning method and usually achieves better results than simple classifiers. However, the tradeoff of combining multiple classifiers is that it

takes much longer to build the model. Given a dataset that is much larger, the calculation time can be a problem. However, it is worth mentioning that most of the processing time is for building the classifier. KNN is a simple and fundamental classifier, which also shows good classification result using the bands from PWCD when K = 1. KNN with K = 1 obtained the highest accuracy compared to other K values. The possible reason is the limited quantity of training pixels. The overall accuracy is 96.8% and the false positive rate is 3.2%. However, K = 1 means that the training samples are classified only based on their nearest training sample. In order to make the classification model represent all possible conditions, the average prediction accuracy up to K = 10 was calculated (Jia et al., 2008) and the comparison with the other classifiers are discussed later in the discussion section. SVM mainly has two parameters to be considered: c (cost) and kernel. When using SVM, the selected bands of the proposed PWCD method obtained 90.6% classification accuracy as the best result. The parameters were set to be c = 5 and a PUK kernel. Polynomial, RBF kernel, and others achieved much lower accuracy. The main incorrectness is the misclassification of mature fruit and background, which is probably because the dark background has a very similar spectrum compared with shadowed mature fruit. The training and testing sets included all kinds of field conditions.

The band set from HDR achieved 97.8% classification accuracy KNN classifier when K = 1. The best classification result using SVM is 95.8% with c = 5 and PUK kernel. AdaBoost classifier obtained 92.3% of overall accuracy when using naïve Bayesian tree (NBTree) classifiers. Table 5-3 shows the detailed results.

The band set from the NG measure achieved the highest classification accuracy using both KNN and AdaBoost. KNN with K = 1 obtained 98.7% of overall accuracy, and AdaBoost with functional trees (FT) classifier obtained 98.4% of overall accuracy. They also have very low false detection rate. Although SVM did not obtain very good classification result, it is still interesting because of the narrow range of the selected bands by NG measure (613.2 – 778.3 nm). Table 5-4 shows the detailed results.

Discussion

The selected wavelengths and classification results are listed in Table 5-5 for comparison. PWCD and NG measure did not consider the correlations between bands. Although HDR did not always achieve the highest prediction accuracy, it kept a relatively stable prediction accuracy using all three classification methods. HDR was originally designed as an unsupervised method. It was directly applied to a whole hyperspectral image, and the bands were clustered based on either information divergence or mutual information. Since HDR groups bands based on their discrepancy on the training set, it is logical to use the band clustering by HDR to see the selected bands from other methods. Wavelength 415.9 nm is near the carbohydrate spectral absorption band, which is near 424 nm. It is crucial for distinguishing the growth stages of the fruits because the berries accumulate more sugar as they mature. Wavelength 512.2 nm is near the chlorophyll absorption band, which is very high for leaf and young fruit.

Bands selected by PWCD are well scattered across the spectral range. However, five bands are from the HDR cluster 6. It is possible that HDR cluster 6 lost much information since its range is too wide, covering all the three visible bands and the red edge. This might be the reason that the prediction result of AdaBoost using the HDR is

the lowest among the three band selection methods. Wavelengths 457 nm, 498.4 nm and 647.8 nm are related to anthocyanin and chlorophyll content in vegetation, which are critical in distinguishing fruit from leaf.

NG measure was also designed as an unsupervised method. It directly sorts the bands by their non-Gaussianity. It selected bands that are from 613 nm to 776 nm, which is the visible red and red edge, a very narrow range compared to the spectral range of the image. This caused lower prediction accuracy using KNN and SVM compared the other two band selection methods. However, it achieved the highest prediction accuracy using the AdaBoost classifier and KNN with K = 1. It shows that the visible red range and red edge is crucial for the classification task in this research.

KNN classification might be weakened because it is more sensitive to over-fitting caused by redundant features, which bring more noise to the system. As a fundamental classifying method, its average accuracy was much higher than SVM. SVM transforms the original features into infinite dimensions where the samples are classifiable linearly. Therefore, the more information it can use, the better result can probably be obtained. NG measure limited the feature to a much narrower range, which is huge information loss for using SVM. Therefore, its prediction ability is very low. AdaBoost iterates many classifiers and adjusts the parameters during the training. Therefore, it achieved much higher prediction accuracy compared to the other two lower level classifiers. Its downside, however, is that it takes much longer to build the model. Given a large dataset, AdaBoost might be computationally intensive.

Conclusion

In this chapter, three information theory based band selection methods PWCD, HDR and NG-measure were applied to the in-field blueberry hyperspectral image. The following are the major band selection results using the three methods.

- PWCD is based on the discriminability of bands for separating every class pair. KLD was used for calculating the discrepancy of the distribution of two classes in each band. The bands with the highest KLD values were chosen. The selected bands are Band 41 (457 nm), Band 68 (498.4 nm), Band 145 (617.9 nm), Band 164 (647.8 nm), Band 176 (666.7 nm) and Band 303 (870.5 nm).

- The second method HDR is based on the assumption that close bands have similar performance for discriminant of objects. KLD was used for calculating the discrepancy of two bands. This method was applied to the labeled training set. Therefore, it is a semi-supervised band selection method in this paper. The bands that have the highest correlations with the centers of the band clusters were chosen. The selected bands are 7 (405.3nm), 14 (415.9nm), 77 (512.2 nm), 215 (728.6nm), 248 (781.5nm), 279 (831.5nm).

- NG measure calculates the difference between the real distribution of each band and its simulated Gaussian distribution. The bands were sorted and grouped since some bands are very close to each other. The selected bands are 192 (692.1nm), 246 (778.3 nm), 175 (665.1nm), 181 (674.6 nm), 162 (644.6nm) and 142 (613.2 nm).

KNN, SVM and AdaBoost classifiers were used to evaluate the performance of the selected bands from the three methods. Although AdaBoost obtained higher accurate rates, it might be too complicated when the data amount is large. HDR had the most stable performance using all classifiers. PWCD achieved the highest average accuracy when using KNN, indicating that PWCD is a promising method for band selection of blueberry hypersepctral imagery. NG-measure method selected bands from only the visible red range and the red edge, which obtained the highest prediction accuracy using KNN with K = 1 and AdaBoost. Therefore, the visible red range and red edge are very important for distinguishing the fruit growth stages and leaf.

Table 5-1. Sorted bands using non-Gaussianity measure. The bands in the same group are very close to each other.

Rank	NG measure	Band	Wavelength (nm)	Group
1	1905.5	192	692.1	I
2	1808.1	246	778.3	II
3	1802.8	245	776.7	II
4	1799.8	175	665.1	III
5	1761.9	244	775.1	II
6	1754.6	193	693.6	I
7	1751.0	173	662.0	III
8	1749.5	171	658.8	III
9	1749.3	243	773.5	II
10	1746.3	247	779.9	II
11	1743.6	174	663.6	III
12	1737.6	194	695.2	I
13	1720.2	181	674.6	IV
14	1719.7	185	681.0	IV
15	1717.7	162	644.6	V
16	1716.1	186	682.5	IV
17	1712.9	191	690.5	I
18	1705.5	176	666.7	III
19	1702.4	163	646.2	V
20	1697.1	172	660.4	III

Table 5-2. Classification results of three classifiers using bands selected by PWCD.

	KNN		SVM		AdaBoost	
	Correct detection (%)	False positive (%)	Correct detection (%)	False positive (%)	Correct detection (%)	False positive (%)
Mature fruit	93.8	5.0	94.3	30.0	95.7	7.1
Intermediate fruit	100.0	0.0	98.4	3.3	100.0	0.0
Young fruit	98.9	4.3	98.9	4.3	96.8	0.0
Background	94.7	2.6	72.9	2.4	97.6	3.6
Overall	96.8	3.2	90.6	9.4	97.5	2.5

Table 5-3. Classification results of three classifiers using bands selected by HDR.

	KNN		SVM		AdaBoost	
	Correct detection (%)	False positive (%)	Correct detection (%)	False positive (%)	Correct detection (%)	False positive (%)
Mature fruit	100.0	1.3	100.0	8.6	97.1	10.0
Intermediate fruit	96.8	0.0	98.4	4.9	95.1	11.5
Young fruit	100.0	6.5	98.9	4.3	96.8	9.6
Background	93.4	0.0	87.1	0.0	81.2	1.2
Overall	97.8	2.2	95.8	4.2	92.3	7.7

Table 5-4. Classification results of three classifiers using bands selected by NG measure.

	KNN		SVM		AdaBoost	
	Correct detection (%)	False positive (%)	Correct detection (%)	False positive (%)	Correct detection (%)	False positive (%)
Mature fruit	97.5	2.5	84.8	39.2	100.0	4.3
Intermediate fruit	100.0	0.0	100.0	1.5	100.0	1.6
Young fruit	98.9	0.0	81.5	2.8	98.9	1.1
Background	98.7	2.6	79.7	15.9	95.3	0.0
Overall	98.7	1.3	88.2	11.8	98.4	1.6

Table 5-5. Comparison of selected wavelengths using different band selection methods and classification methods.

Band selection methods	Six selected wavelengths (nm)	KNN (%)	SVM (%)	AdaBoost (%)
PWCD	457.0, 498.4, 617.9, 647.8, 666.7, 870.5	95.4	90.6	97.5
HDR	405.3, 415.9, 512.2, 728.6, 781.5, 831.5	94.5	95.8	92.3
NG measure	613.2, 644.6, 665.1, 674.6, 692.1, 778.3	93.6	88.2	98.4

Figure 5-1. A blueberry fruit bunch that shows all three growth stages: young, intermediate and mature.

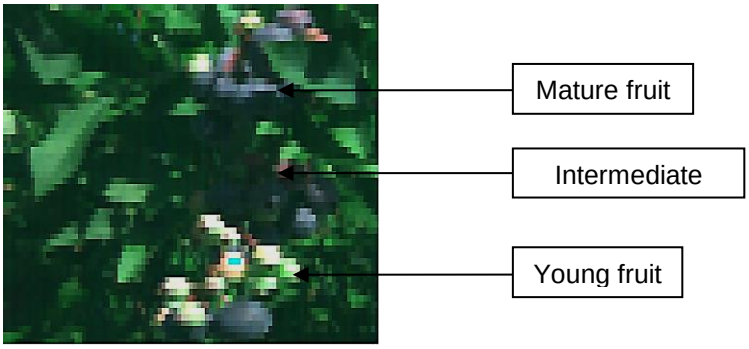


Figure 5-2. RGB bands of a hyperspectral image with all blueberry fruit growth stages.

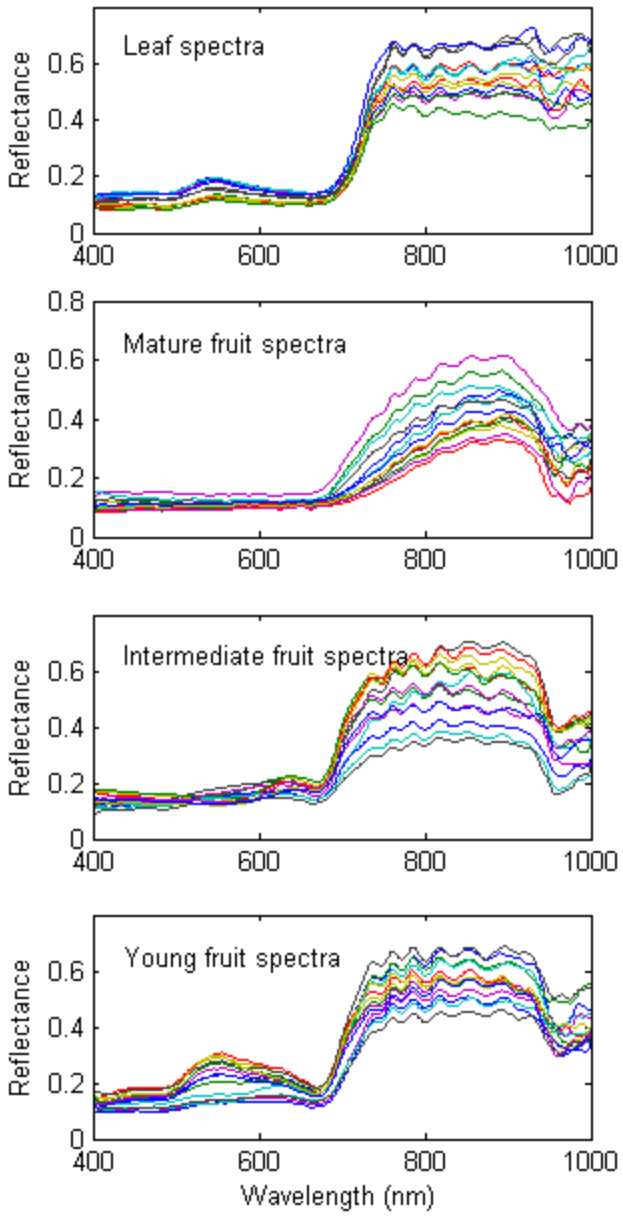


Figure 5-3. Spectra of ten pixels for each class: mature fruit, intermediate fruit, young fruit and background (leaf).

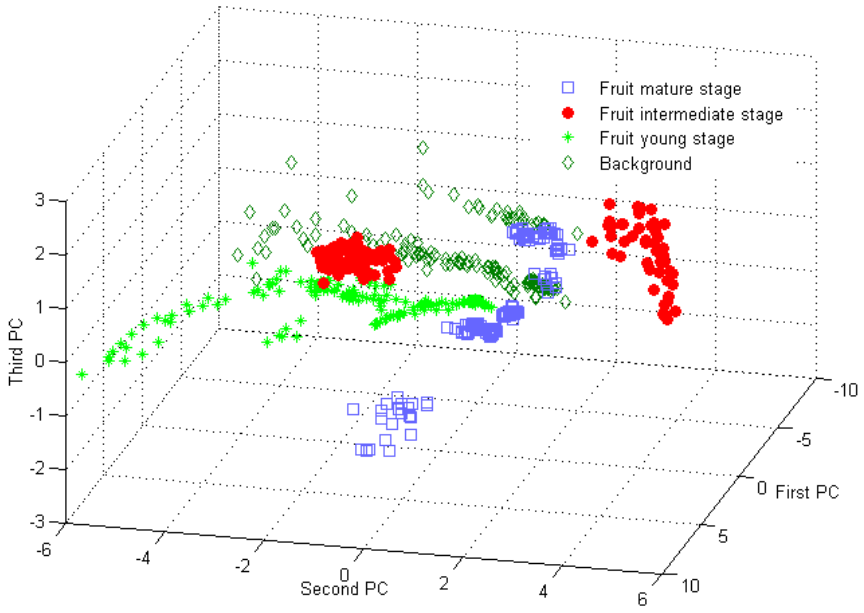


Figure 5-4. Principal component transform of the four classes: mature fruit, intermediate fruit, young fruit and background.

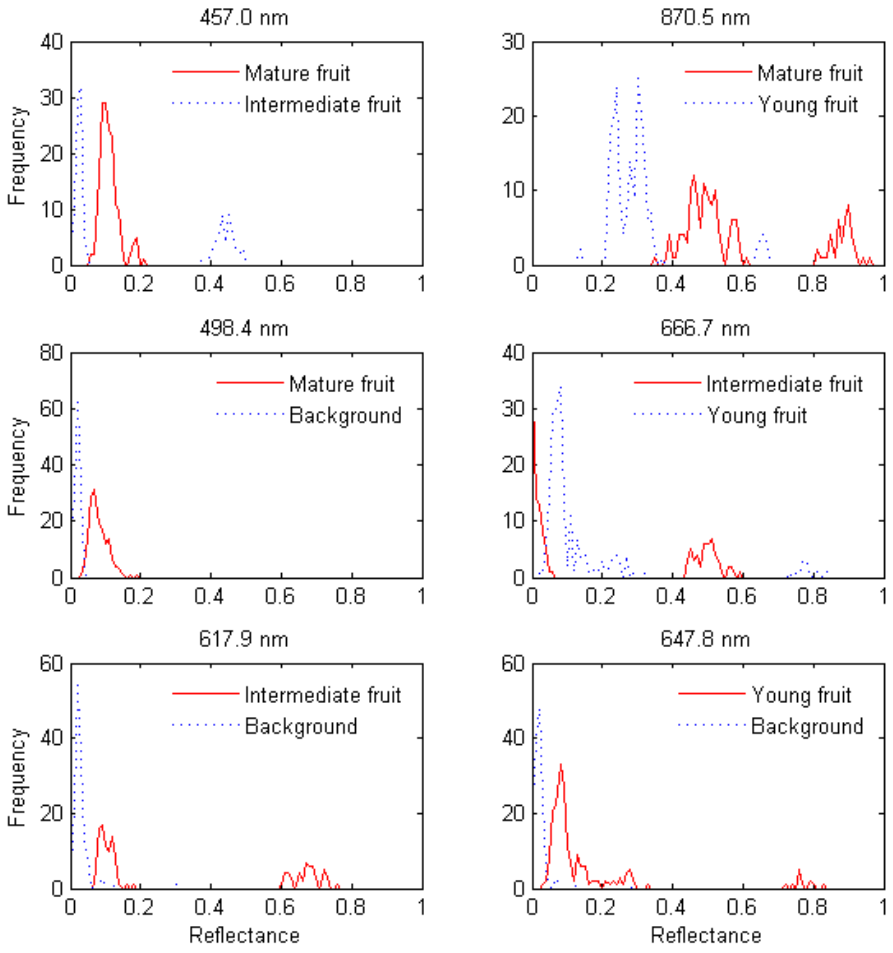


Figure 5-5. Separation ability of selected bands by PWCD.

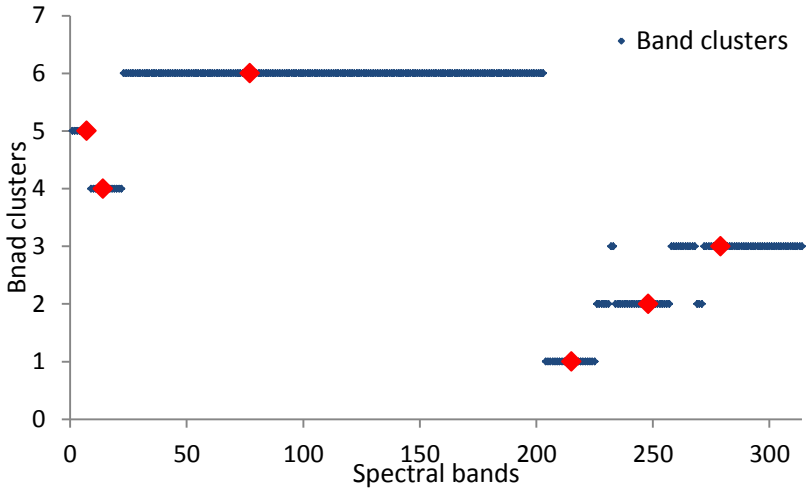


Figure 5-6. Band clustering result and selected bands by calculating correlations between cluster average and individual bands.

CHAPTER 6
BLUEBERRY MATURITY STAGE DETECTION BASED ON SPECTRAL-SPATIAL DETECTION OF HYPERSPECTRAL IMAGE USING SELECTED BANDS

Background

Blueberry farms that supply fresh markets require that the berries be mainly handpicked. Blueberries mature gradually within a single fruit bunch. Therefore, the harvesting is very labor intensive. Labor expense is the major concern of the farmers. It was estimated that more than half of the management cost is for harvest labor (Williamson et al., 2012; USDA, 2012; Morgan et al., 2011). Early yield estimation is beneficial, since it helps farmers to arrange the harvesting labor efficiently based on the estimated yield variation in the field and decrease harvest cost.

Color machine vision has been used for fruit detection in agriculture. Grapes were detected by Chamelat et al. (2006) using RGB color space, HSI color space and Zernike moments as features. However, grapes were harvested by bunch rather than single fruit, which made the detection easier. Blue pixels in the view were used by Zaman et al. (2008) for identifying wild blueberries. However, their images were obtained where only mature berries were in the view. In Florida blueberry fields, all maturity stages exist in a single fruit bunch. Therefore, it is necessary to detect the fruits by their growth stages, which is difficult using color machine vision.

Hyperspectral imagery records a large amount of spectral and spatial information. Compared to color images, hyperspectral images are more helpful for differentiating objects in view. Pixel-scale processing has been widely used in hyperspectral imagery. Much literature describes about applying pixel-scale hyperspectral image processing in agricultural applications, such as food safety inspection, food quality control, nutrition stress detection, crop characterization, meat

inspection, etc. Lu et al. (1999) introduced the implementation of hardware and software of a hyperspectral imaging system. They showed that the hyperspectral imaging system was an effective tool for safety inspection of poultry carcasses. Lelong et al. (1998) extracted information of hyperspectral image for wheat crops using principal component analysis (PCA). They detected water deficiency in the field and estimated the crop vitality according to the stress presence with limited amount of spectral channels. Hyperspectral imagery has been shown to be helpful for blueberry maturity stage detection (Yang et al., 2013). They achieved more than 94% accuracy for the classification of three blueberry maturity stages and background using only spectral information of training and testing pixel sets. However, when applying to whole hyperspectral images, the accuracy decreased because of the variations in the field and in each image. Therefore, accurate detection of blueberry maturity stages is crucial for hyperspectral imaging of this application. To achieve better results, spectral-spatial detection/classification becomes more and more popular since spatial information is also available in hyperspectral images. Benediktsson et al. (2003) proposed the morphological profile originated from the granulometry principle (Serra, 1982). The profile contained opening and closing profiles, which were reconstructed by connected opening and closing operators. The spectral features and morphological profile performed well in terms of classification accuracies, and relatively fewer features were needed. Van der Meer et al. (2005) proposed a spatial-spectral contextual image analysis named the template matching algorithm. The algorithm was used to characterize hydrothermal alteration in epithermal gold deposits. Li et al. (2012) applied supervised spectral-spatial hyperspectral image segmentation. They integrated the

spectral and spatial information in the Bayesian framework by subspace multinomial logistic regression and Markov random fields. Their approach showed accurate characterization for both simulated and real hyperspectral data sets. Tarabalka et al. (2009) proposed the spectral-spatial classification method based on two partitional clustering techniques: ISODATA and Gaussian mixture resolving algorithm. The proposed methods improved classification accuracies and provided decision maps with more homogeneous regions.

The objective of this study was to carry out spectral-spatial detection methods to improve the detection of blueberry maturity stages toward development of an early yield mapping system. Two spectral-spatial detection schemes were applied, one was to combine segmentation of nested clustering results with spectral detection results, and the other was to combine the spectral detection results with morphological operations.

Materials and Methods

Hyperspectral Image Data Set

Hyperspectral images were acquired in a blueberry research and demonstration farm in the University of Georgia cooperative extension in Alma, GA, United States (31.53438°, -82.51019°, WGS84) in July, 2012. The camera system contained a digital CCD camera (MV-D1312, Photonfocus AG, Lachen SZ, Switzerland), a camera body (a line scanning spectrometer V10E, Specim, Oulu, Finland), a lens (CNG 1.8/4.8-1302, Schneider Optics, North Hollywood, CA, USA), an image grabber (NI-PCIe 6430, National Instruments Inc. Austin, TX, USA), a DAQ card (NI-6036E, National Instruments Inc. Austin, TX, USA), an encoder (Omron-E6B2, OMRON cooperation, Kyoto, Japan), a tilting head (PT785S, ServoCity, Winfield, KS, USA), and a laptop (DELL Latitude E6500) with an image acquisition and control program written in

LabVIEW (National Instruments Corporation, Austin, TX, USA). Camera was designed to rotate vertically with the help of the tilting head. The encoder sent a pulse to the program for generating trigger signals for the camera to take a new picture. The images had 388 bands each with spectral resolution of 1.59 nm. The spectral range of the images was 398 – 1010 nm. The spatial resolution was approximately 1 mm. There were three fruit classes according to its maturity stage: mature, intermediate and young fruit. The background included branch, soil, sky and man-made objects such as polyvinyl chloride pipes, ribbons, etc.

The original hyperspectral images had 388 bands, which were time and space consuming for image processing. The six selected bands from Yang et al. (2013) performed well in classifying the three fruit classes and background. Therefore, the bands were utilized in this study instead of the original bands. The selected bands were: 543.1 – 572.6 nm, 627.4 – 658.8 nm, 663.6 – 695.2 nm, 725.4 – 757.4 nm, 773.5 – 805.6 nm and 838 – 870.5 nm. The methods in the following sections are for multispectral image processing with the specific wavebands in the future for an in-field yield estimation system.

Spectral-spatial Processing Based on Nested Clustering Techniques

This spectral-spatial processing based on partitional clustering techniques is adopted from Tarabalka et al. (2009) with several changes. Figure 6-1 shows the flowchart of the steps specifically used in this study. The approach was carried out using Matlab R2012a (The MathWorks Inc., Natick, MA, USA).

There are three stages of the clustering technique: band selection, similarity measure and grouping. The first stage was to selected the six bands, which has been done in Yang et al. (2013). The second stage is to measure the spectral similarity of the

pixels. Distance measures such as Euclidean distance, Mahalanobis distance, cityblock and cosine. After trial and error, Euclidean distance was shown to be the most suitable for the specific blueberry detection task. The third stage is to group the pixels with the most spectral similarity into the same clusters. Besides two partitional clustering algorithms considered in Tarabalka et al. (2009), i.e., iterative self-organizing data analysis (ISODATA) and expectation–maximization algorithm (EM), this study introduced a third algorithm, which was a nested clustering algorithm using agglomerative clustering from linkage (Griffiths et al., 1978). The algorithm started with singleton clusters and successively links clusters to generate a hierarchy of nested clusters. It arranged the clusters and sub-clusters in a tree-structured fashion.

After clustering, every pixel had a unique label, yet the assignment was only based on spectral information. No spatial information was added to the decision yet. Pixels from the same cluster were scattered in the spatial domain with noises and outliers. However, the image plane was to be segmented with unique labels within every single object in the view. For example, normally a mature fruit takes a single connected region in the image. Therefore, union-find data structure based on connected-component-labeling algorithm was used to label the connected components from the same cluster.

Pixel scale detection was parallel to the segmentation based on nested clustering techniques because it did not use any of the segmentation results. There are many pixel scale detection methods, such as spectral angle mapper (SAM), spectral feature filtering analysis (SFFA). SAM is widely used in hyperspectral imagery, which compares the pixel spectra to known spectra by calculating the spectral angle between them. SAM

is insensitive to illumination because it only uses the feature vector direction rather than length. The result of SAM is an image with each pixel labeled to its best matching class. Since the blueberry hyperspectral images taken from the outdoor condition contained a large amount of uneven illumination, SAM was applied for the detection of the pixels.

Until now, two decision images were generated: one from segmentation based on nested clustering strategy and the other from pixel scale detection using SAM. The next step was to combine the two decision images by a majority vote. For every segmentation region, all the pixels were labeled to a most frequent class within that region. After this step, all the segments were assigned according to the pixel scale detection result. A new decision map was generated.

In the end, a post-regularization step aiming at removing noise in the decision map was carried out. The final decision map after the post-regularization would result in more homogeneous regions. Young fruits and some intermediate fruits in the blueberry images often had much smaller size than mature fruits. Some young fruits took only four to six pixels in the image. Therefore, the decision map was filtered by removing salt-and-pepper noise with 8-neighborhoods in case of removing small regions like young fruits.

Spectral-spatial Processing Using Morphological Operations

The spectral-spatial processing for blueberry fruit detection based on morphological operations follows steps shown in Figure 6-2. The main procedure includes pre-processing, fruit detection in the spectral domain, morphological operation in the spatial domain, and post-processing. The method was also carried out by Matlab R2012a (The MathWorks Inc., Natick, MA, USA).

As a preprocessing step, the dark background, man-made objects, soil and sky of the images could be removed by the NIR range difference. These objects have relatively low value in the NIR range comparing to the well-illuminated vegetation. Therefore, the sixth band 838 – 870.5 nm was used as a gray image for performing Otsu's method (1979). Otsu's method was used to automatically perform image thresholding based on the histogram of the gray image. The result was a binary image. Pixels with value higher than the 0.9 times of automatic Otsu gray threshold were given 1 and all other pixels were given 0. The scale 0.9 was used because as many pixels as possible should be saved in case some important pixels are removed. There would be more steps to rule out the non-fruit pixels. However, if important pixels are removed after the Otsu's method, there is no chance to recover them in the later steps.

After removing dark background and other objects of no interest, SAM was applied to the image. Spectral angles of the pixels and the spectra of classes from the library were calculated. Due to the different spectral variations in each class, the spectral angle thresholds varied for different classes.

Since SAM was calculated in the pixel scale with only the spectral information, there were many incorrectly detected pixels that scattered all over the image. In addition, pixels on the fruit edge might be missed because of strong shadow. To eliminate the scattered pixels and noises in the image, spatial information process such as removing salt-and-pepper noise and morphological opening and closing could be used. Opening is the dilation of erosion of a pixel set by a structuring element, such as discs and squares. It helped to remove small objects from the foreground, which were the dark pixels. Morphological closing is the erosion of dilation of a pixel set. It helped to remove

small objects from the background, which were the white pixels. It is logical to close the fruit area so that the fruit pixels in the shadow could be considered as correctly detected pixels if they are missed after SAM. Discs with size 1 to 4 were used in this study by trial-and-error for opening and closing depending on fruit maturity stages. Closing with disc size 1 was used twice and opening with disc size 1 was used once for the mature fruit class. Closing with disc size 3 and 1 were used for the intermediate fruit class. Closing with disc size 2 and 1 and opening with disc size 1 were used for the young fruit class.

After morphological operations, a problem arises: some pixels are labeled as two or more classes. This will happen where fruits of different maturity stages are connected with each other. After morphological closing, both fruits have larger size. This causes the pixels on the edge of the fruits to overlap. Therefore, post-processing is needed to eliminate the overlapping. Majority vote of the pixel's eight neighbors are carried out. The pixel will be labeled to the most frequent class within the 8-neighborhoods window. A final decision map is generated after this step.

Results and Discussions

Spectral-Spatial Detection Result Based on Nested Clustering Techniques

The method was applied to the blueberry images using the selected six bands. In the first step, nested clustering of the images was performed using agglomerative clusters from linkage. Complete linkage was chosen. The pixels were grouped with the nearest Euclidean distance with cutoff value 1.154 by trial and error. The algorithm splits the images into hundreds of clusters.

In the next step, the connectivity within the clusters was further utilized by union-find data structure. The resulting segmentation map contains more segments than the

number of clusters from the previous step. The reason is that many clusters have pixels that are not connected. Therefore, a single cluster can be assigned into several segments. However, some regions are connected well, which represent a whole single object. For instance, the pixels from a mature fruit can be well gathered into one single region. There are still many small segments with only one or two pixels. However, over-segmentation is not a concern (Tarabalka et al., 2009).

SAM detection resulted in another decision map, where spatial information was not considered. Therefore, there were a large amount of pixels that were far away from their reasonable class assignments. Example results of SAM are shown in Figure 6-3 B) with different colors: purple representing mature fruit, red representing intermediate fruit and light green representing young fruit. The optimal thresholds for mature fruit, intermediate fruit and young fruit were 0.15, 0.2 and 0.05, respectively. The different thresholds were because of the different variations of the fruit classes. There are scattered pixels for all the three maturity stages, most of which are false positives. The true positive and false positive detection rates of the SAM detection step are shown in Table 6-1. The highest true positive rate is 75% for young fruit, and the lowest is 52.4% for intermediate fruit. The possible reason was that the intermediate fruit pixels were easier to be classified into the wrong classes because intermediate was the middle stage of the three fruit classes. The false positive rates for all three fruit classes are very high mainly because of the scattered pixels all over the image. Young fruit class obtained the highest false positive rate, which was because the branches in the image had more similar spectra with young fruits.

After the spectral detection, the results were combined with the segmentation decision by majority voting described in Tarabalka et al. (2009). Post regularization was performed on the spectral-spatial based decision map using remove-salt-and-pepper with 8-neighborhoods, since smaller objects such as young fruits would need to be saved. The example final results were shown in Figure 6-3 C) with the same color presentation as in Figure 6-3 B). This step helped to obtain higher intermediate detection accuracy because the segmentation step took majority vote. The regions that seem to be loose in Figure 6-3 B) become more unique in Figure 6-3 C). True positive rate of mature fruit decreased, however, the false positive rates for all three fruit classes significantly decreased after the spectral-spatial operation, as shown in Table 6-1. This is mainly because the false detections and noise pixels were removed by the post regularization step.

Spectral-spatial Detection Result Using Morphological Operations

In the first step, Otsu's method was used on the sixth band in order to remove the dark background and objects other than vegetation. Then SAM detection using only spectral information was carried out with thresholds 0.15, 0.2 and 0.05 for mature fruit, intermediate fruit and young fruit. These thresholds are the same as used in the first spectral-spatial method because they were optimized with several training images under all possible illumination conditions.

Since some dark background pixels that were not removed by the Otsu step had very similar spectra with mature fruit, the mature fruit map after SAM had more scattered pixels than other classes. Morphological close, open, and remove-salt-and-pepper were applied to the mature fruit map. Intermediate fruit was easier to mix up with other classes because it is similar to both mature fruit and young fruit. Morphological

close and remove-salt-and-pepper were applied to the intermediate fruit map. Young fruit has much smaller size than the mature fruit and intermediate fruit. It is very easy to miss. Therefore, two morphological closes were applied. An overview of the detection results using only SAM and combining spectral-spatial operations is shown in Figure 6-4. Purple color represents mature fruit, red represents intermediate fruit and light green represents young fruit. The example mature fruit map, intermediate fruit map and young fruit map before and after the morphological operations of the same hyperspectral image are shown in Figure 6-5. The white pixels are the detected pixels and all the other pixels are shown black. It is shown in Figure 6-4 and Figure 6-5 that results after morphological operations are much more logical than the results of pure SAM detection because the fruits are connected regions and the noises are removed. Table 6-2 shows the detection results after each step of using this spectral-spatial detection method. After removing background and SAM, the true positive and false positive rates are almost the same with the result of the spectral-spatial method based on nested clustering technique. However, the results after morphological operations are much better than the first spectral-spatial detection method. True positive rate of mature fruit is more than 78%, and false positive rate is 13%. Performance of detecting intermediate fruit class increased over 30% from the first step, while false positive decreased to lower than 10%. The main contribution the morphological operations had was to increase the pixel amount on the edge of fruits, which made up for the miss detection of fruit pixels in heavy shadow.

All in all, spectral-spatial detection using morphological operations performed much better than based on nested clustering technique. This shows that the in-field

condition of the blueberry plants was impacted seriously by the heavy shadow. Although SAM is not supposed to be impacted by shadow, when shadow is too strong, it is difficult to classify the pixels under shadow into the correct classes. A possible solution is to consider more classes, such as mature fruit in shadow, intermediate fruit in shadow, and young fruit in shadow. Another problem was caused by the biased opinion of the expert knowledge for labeling the pixels. The variation of decisions was a major concern among experts when labeling some specific pixels on the edge and in the shadow.

Conclusion

Two spectral-spatial detection schemes were carried out, and they both improved the detection of blueberry maturity stages using only spectral information. The first method was to combine segmentation of nested clustering results with spectral detection results, and the second method was to combine the spectral detection results with morphological operations. Remove-salt-and-pepper was also used for noise removal caused by the spectral detection step. The spectral-spatial detection schemes were proved to perform much better than spectral detection. Spectral-spatial detection using morphological operations outperformed the detection based on nested clustering by achieving more than 75% true positive rates for all three fruit classes. The major problem that hinders the performance of the detection schemes are the strong shadows under field conditions and the biased expert opinions for pixels on the edge and in the shadow.

Table 6-1. True positive and false positive rates after each step the spectral-spatial detection based on nested clustering technique.

Class	SAM		SAM+Segmentation	
	TP(%)	FP(%)	TP(%)	FP(%)
Mature fruit	68.7	30.9	65.3	11.1
Intermediate fruit	52.4	43.8	70.5	11.4
Young fruit	75.0	175.0	75.0	25.0

Table 6-2. True positive and false positive rates after each step the spectral-spatial detection using morphological operations.

Class	SAM		SAM+Morphological operation	
	TP(%)	FP(%)	TP(%)	FP(%)
Mature fruit	67.7	25.4	78.3	13.2
Intermediate fruit	52.4	41.9	83.8	9.5
Young fruit	75.0	183.3	75.0	25.0

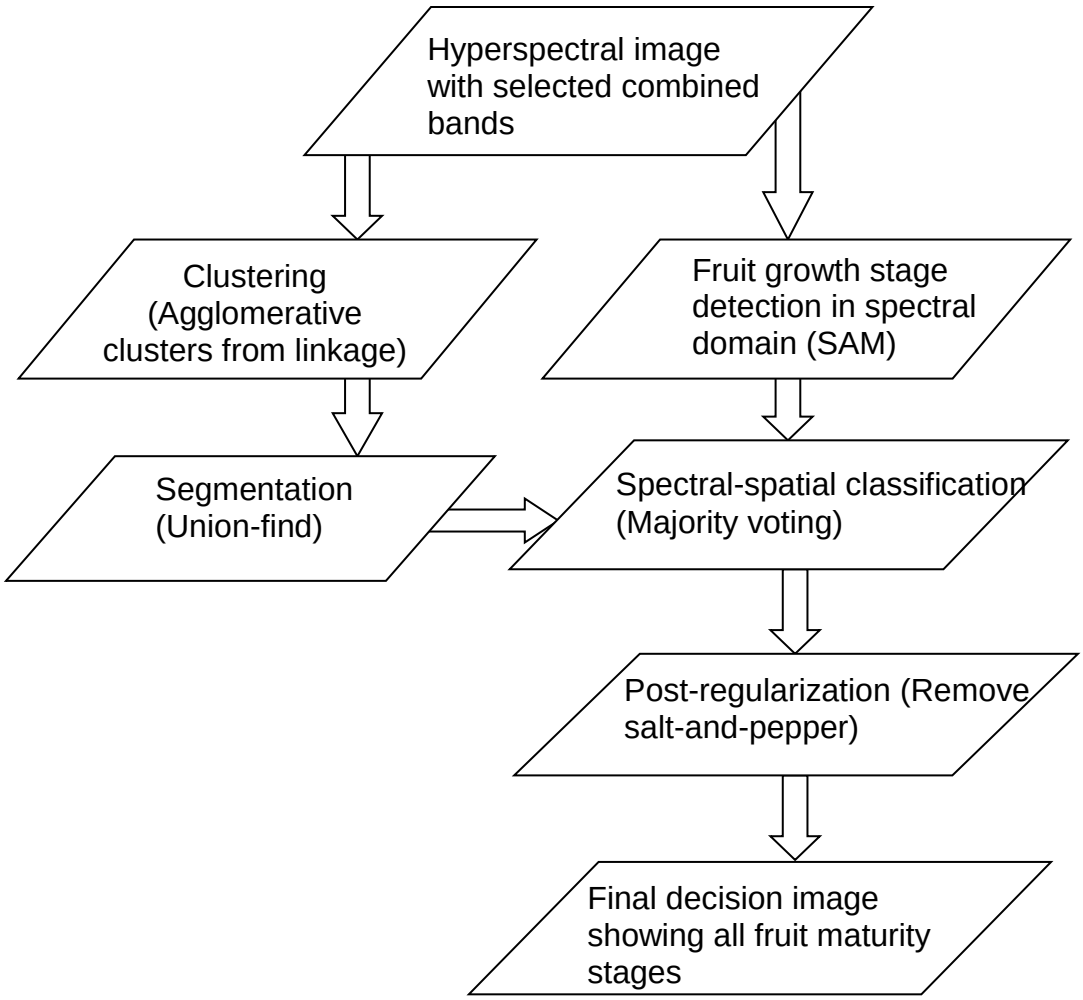


Figure 6-1. Spectral-spatial detection of blueberry fruit maturity stages based on nested clustering.

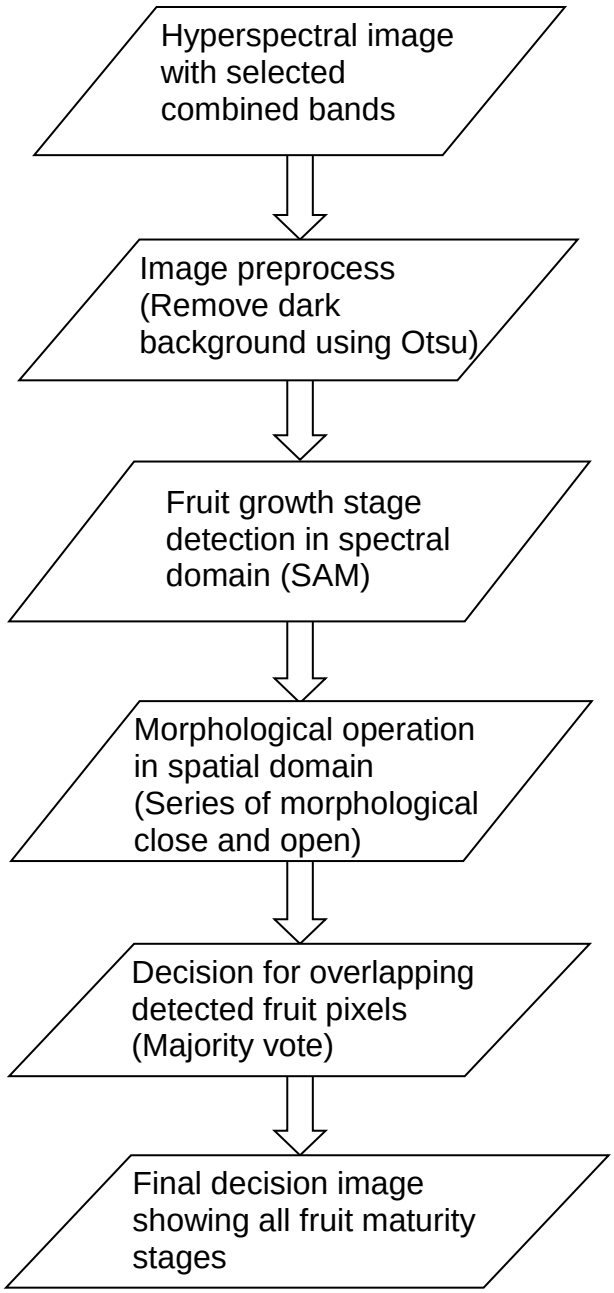


Figure 6-2. Spectral-spatial detection of blueberry fruit maturity stages using morphological operations.

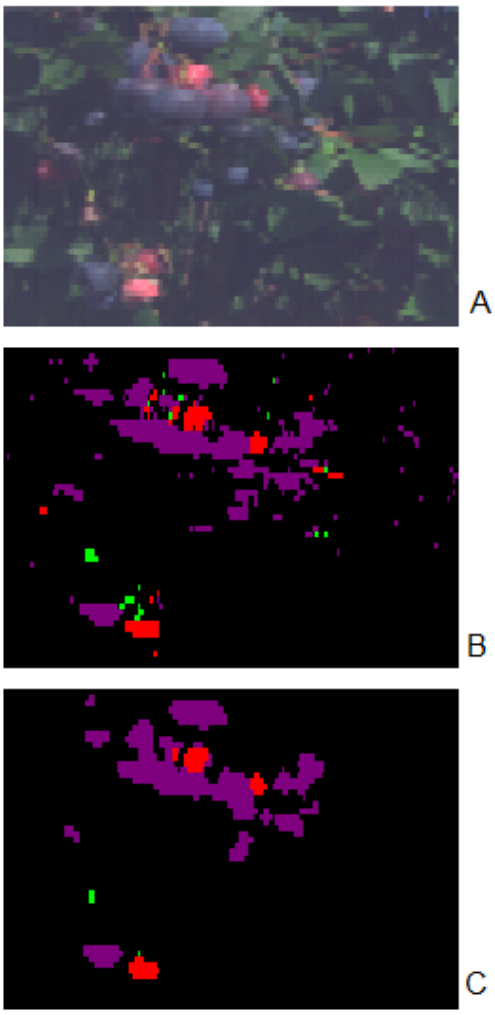


Figure 6-3. Overview of spectral-spatial detection results of a blueberry hyperspectral image based on nested clustering techniques. Purple color = mature fruit, red color = intermediate fruit, green color = young fruit. A) RGB representation of the hyperspectral image, B) SAM detection result, C) SAM combined with segmentation.

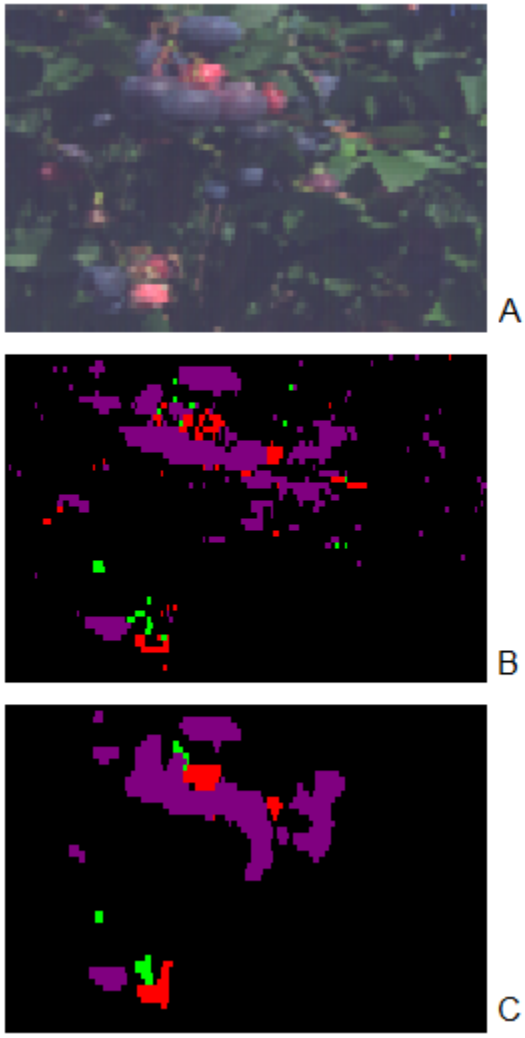


Figure 6-4. Overview of fruit detection results of a testing hyperspectral image based on the selected bands, before and after combining spectral detection and morphological operations. A). RGB representation of the hyperspectral image, B) SAM detection result, C) SAM combined with morphological operations.

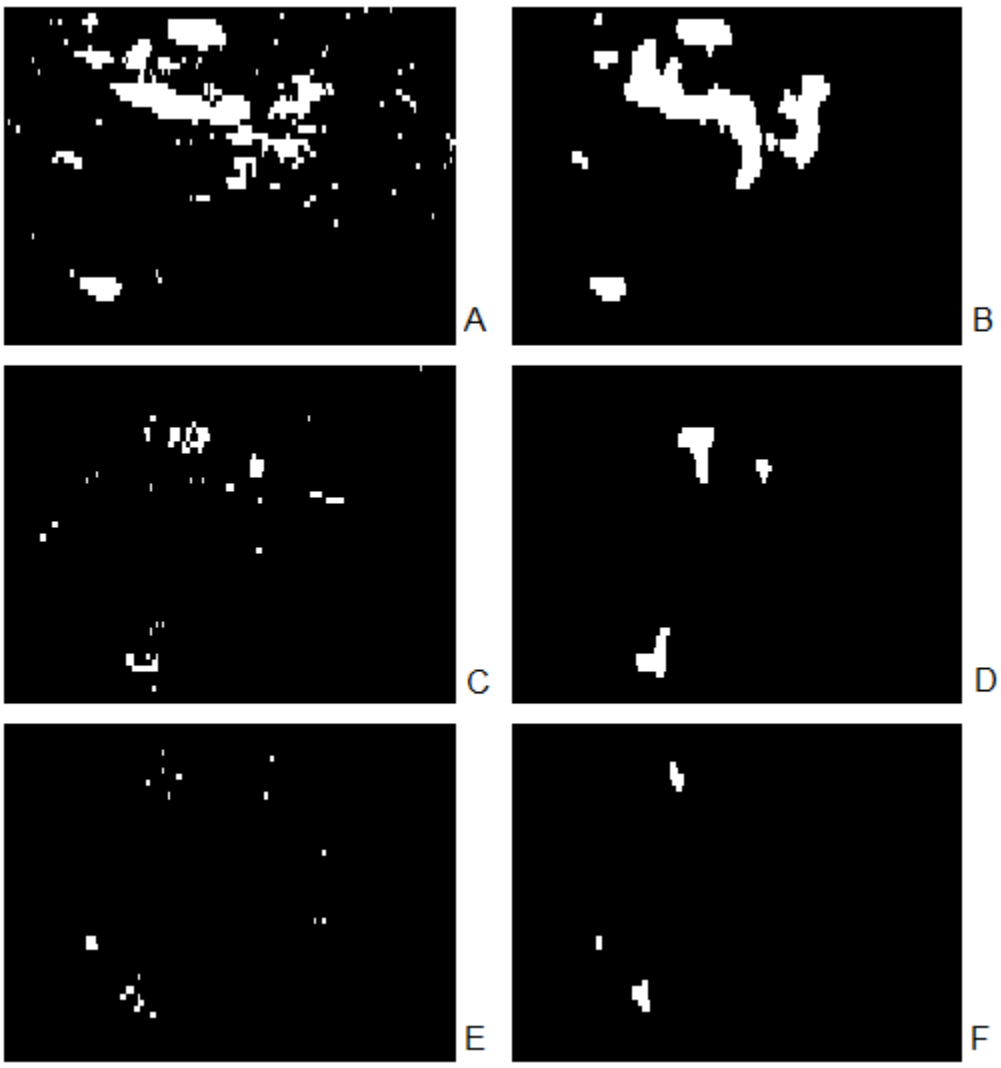


Figure 6-5. Fruit detection results of a testing hyperspectral image based on selected bands, before and after combining spectral detection and morphological operations. A) mature fruit map after SAM, B) mature fruit map after SAM and morphological operations, C) intermediate fruit map after SAM, D) intermediate fruit map after SAM and morphological operations, E) young fruit map after SAM, F) young fruit map after SAM and morphological operations.

CHAPTER 7
SUMMARY AND SYNTHESIS

Spectra of blueberry fruit and leaf samples were obtained in a laboratory and normalized vegetation indices were used as candidate variables for classification of fruit maturity stages and leaf. The selected wavelengths were in the form of indices and achieved accuracies of higher than 94% for the classification task. It showed that spectral analysis of blueberry fruit and leaves is capable of classifying the fruit maturity stages and leaf in a laboratory environment. However, the spectrophotometer was bulky to be used in the field. In addition, spectrophotometer gave a spectrum of a sample with multiple fruits or leaves, usually prepared in a laboratory, which could not be easily suited to yield estimation directly.

In contrast, the advances of computer technology and multispectral/hyperspectral camera enable in-field data acquisition with time and cost efficiency. Multispectral images with three bands: near-infrared (760 – 900 nm), red (630 – 690 nm) and green (520 – 600 nm) were obtained in 2011 in a blueberry field, and different color components were used as input features for classification of the fruit and background. Accuracies of 84% and 73% were obtained for fruit and background classes, respectively. However, the color features did poorly in separating eight classes individually: mature fruit, intermediate fruit, young fruit, leaf, branch, soil, sky, and reference board. It showed that the multispectral camera with only three bands was very limited in fulfilling the task of classifying fruit maturity stages and other classes in the view.

Hyperspectral imaging was more capable of detecting physically similar objects, since it records whole spectrum of an object at each pixel. In this study, hyperspectral

images were acquired in 2012 and 2013 in a blueberry field in Alma, Georgia, USA and Waldo, Florida, USA. However, hyperspectral images generally contain much redundancy, especially in the spectral dimension. Band selection was necessary to find the most important bands for further application in the field. Three band selection methods were used and the selected bands obtained prediction accuracies of more than 88%. It showed that the selected band sets were capable of classifying blueberry maturity stages and background in the field. Although achieved high prediction accuracy, the selected bands could not do well when applied directly to the hyperspectral images. There were a considerable amount of noise pixels and outliers in the result because it did not use any spatial information of the images. Therefore, spectral-spatial image analysis was considered for the detection of fruits with different maturity stages on the hyperspectral images with selected bands. The two spectral-spatial image analysis procedures both obtained rule images with the desired classes in relatively homogeneous regions and noise removed. The spectral-spatial detection using morphological operations outperformed the detection method based on nested clustering by achieving more than 78% pixel detection accuracy. The results improved by up to 30% compared to pure spectral detection. The result might be improved more by controlled in-field condition, which caused strong shadow that impacted the performance of the detection methods.

LIST OF REFERENCES

Aggelopoulou, A. D., Bochtis, D., Fountas, S., Swain, K. C., Gemtos, T. A., & Nanos, G. D. 2011. Yield prediction in apple orchards based on image processing. *Precision Agriculture*,12(3), 448-456.

Agresti, A. 2007. *An introduction to categorical data analysis*. Hoboken, NJ: Wiley-Interscience.

Ariana, D. P. & Lu, R. 2010. Hyperspectral waveband selection for internal defect detection of pickling cucumbers and whole pickles. *Comput. Electron. Agric. 74(1),*137-144.

Balasundaram, D., Burks, T.F., Bulanon, D.M., Schubert, T. & Lee, W.S. 2009. Spectral reflectance characteristics of citrus canker and other peel conditions of grapefruit. *Postharvest Bio. Technol*. *51(2),* 220-226.

Bajwa, S. G., Bajcsy, P., Groves, P., & Tian, L. F. 2004. Hyperspectral image data mining for band selection in agricultural applications. *Transactions-American Society of Agricultural Engineers,* 47(3), 895-908.

Benediktsson, J. A., Pesaresi, M., & Amason, K., 2003. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Transactions of Geoscience and Remote Sensing,* 41(9), 1940-1949.

Braswell, B. 2010. Blueberry report: there's value in growing together. Available at: http://www.growingproduce.com/uncategorized/theres-value-in-growing-together/. Accessed on 1 November, 2010.

Bruce L M, Koger C H, Li J. 2002. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction*. Geoscience and Remote Sensing, IEEE Transactions,* 40(10), 2331-2338.

Bouckaert, R. R., 2004. Bayesian network classifier in Weka. Working paper series. University of Waikato, Department of Computer Science. No. 14/2004. Hamilton, New Zealand: University of Waikato.

Cai, S., Du, Q. & Moorhead, R.J., 2007. Hyperspectral imagery visualization using double layers. *IEEE T Geosci. Remote*. 45(10) 3028–3036.

Chamelat, R., Rosso, E., Choksuriwong, A., Rosenberger, C., Laurent, H. & Bro, P., 2006. Grape Detection By Image Processing. *32nd Annual Conference on IEEE Industrial Electronics*, 3697–3702.

Chang, C. I., Du, Q., Sun, T. L., & Althouse, M. L. 1999. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image

classification. *Geoscience and Remote Sensing, IEEE Transactions.* 37(6), 2631-2641.

Chang, C.-I. & Wang, S., 2006. Constrained band selection for hyperspectral imagery. *IEEE T Geosci. Remote.* 44 (6), 1575–1585.

Cheng, X., Chen, Y. R., Tao, Y., Wang, C. Y., Kim, M. S. & Lefcourt, A. M. (2004). A novel integrated PCA and FLD method on hyperspectral image feature extraction for cucumber chilling damage inspection. *Transactions-American Society of Agricultural Engineers*, *47*(4), 1313-1320.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine Learning.* 20(3), 273 – 297

Cozzolino, D., Cynkar, W. U., Shah, N. & Smith, P. 2011. Multivariate data analysis applied to spectroscopy: potential application to juice and fruit quality. *Food Res. Int*. 44(7), 1888-1896.

Daszykowski, M., Orzel, J., Wrobel, M. S., Czarnik-Matusewicz, H. & Walczak, B. 2011. Improvement of classification using robust soft classification rules for near-infrared reflectance spectral data, *Chemometr. Intell. Lab. Syst*. DOI:10.1016/j.chemolab.2011.08.004.

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

Freund, Y. & Schapire, R. E., 1995. A decision-theoretic generalization of on-line learning and an application to boosting. *J. comput syst sci.* 55, 119-139.

Friedman, J.H., Hastie, T. & Tibshirani, R., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag: Heidelberg.

Gardner, V. R., Bradford, F. C. & Hooker, H. D. 1922. *The fundamentals of fruit production*. NY: McGraw-Hill Book Company.

Giangiacomo, R., Magee, J. B., Birth, G. S. & Dull, G. G. 1981. Predicting concentrations of individual sugars in dry mixtures by near-Infrared reflectance spectroscopy. *J. Food Sci.* 46, 531–534.

Giusti, M. M. & Wrolstad, R. E. 2011. Characterization & measurement of anthocyanins by UV-visible spectroscopy. In *Curr. Protoc. Food Analyt. Chem. (Unit F1.2).*

Griffiths, A., Robinson, L. A. & Willett, P., 1984. Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation*, *40*(3), 175-205.

Guo, B., Gunn, S. R., Damper, R. I., & Nelson, J. D. B. 2006. Band selection for hyperspectral image classification using mutual information. *Geoscience and Remote Sensing Letters, IEEE,* 3(4), 522-526.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009. The WEKA data mining software: an update. SIGKDD Explorations, 11(1), 10-18.

Harnly, J. M., Doherty, R. F., Beecher, G. R., Holden, J. M., Haytowitz, D. B., Bhagwat, S.& Gebhardt, S. 2006. Flavonoid content of U.S. fruits, vegetables, and nuts. *J. Agric. Food Chem.* 54(26), 9966–9977.

Huang, M. & Lu, R. 2010a. Apple mealiness detection using hyperspectral scattering technique. *Postharvest Biol. Technol*. *58(3),* 168-175.

Huang, M. & Lu, R. 2010b. Optimal wavelength selection for hyperspectral scattering prediction of apple firmness and soluble solids content. *Trans. ASABE 53(4),* 1175-1182.

Inácio, M. R. C., Moura, M. F. V. & Lima, K. M. G. 2011. Classification and determination of total protein in milk powder using near infrared reflectance spectrometry and the successive projections algorithm for variable selection. *Vibr. Spectrosc*. *57(2),* 342-345.

Jensen, J. R. 2000. *Remote sensing of the environment.* Upper Saddle River, NJ: Prentice Hall.

Jia, S., Qian, Y. & Ji, Z., 2008. Band Selection for Hyperspectral Imagery Using Affinity Propagation. *Proceedings of the 2008 Digital Image Computing: Techniques and Applications.* 137–141.

Jones, C.D., Jones, J.B. & Lee, W.S. 2010. Diagnosis of bacterial spot of tomato using spectral signatures. *Comput. Electron. Agric. 74(2),* 329-335.

Kane, K.E. & Lee, W.S. 2006. Spectral sensing of different citrus varieties for precision agriculture. 2006 ASAE Paper No. 061065, ASAE, Portland, Oregon.

Kane, K. E. & Lee, W. S., 2007. Multispectral imaging for in-field green citrus identification. 2007 ASABE Paper No. 073025.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., Murthy, K.R.K., 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*. 13(3):637-649.

Kumar, A., Lee, W. S., Ehsani, R. & Albrigo. L. G., 2009. Airborne hyperspectral imaging for citrus greening disease detection.In the Proceedings of the 3rd Asian Conference on Precision Agriculture, Beijing, China.

Kurtulmus, F., Lee, W. S., Vardar, A., 2011. Green citrus detection using 'Eigenfruit', color and circular Gabor texture features under natural outdoor conditions. *Computers and Electronics in Agriculture*. 78(2): 140–149.

Kurata, K., & Yan, J. 1996. Water stress estimation of tomato canopy based on machine vision. *International Symposium on Plant Production in Closed Ecosystems.* 440, 389-394.

Lee, W.S., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D. & Li, C. 2010. Sensing technologies for precision specialty crop production. *Comput. Electron. Agric*. *74(1),* 2-33.

Lee, W. S., Ehsani, R. & Albrigo, L. G., 2008. Citrus greening (Huanglongbing) detection using aerial hyperspectral imaging. 9thInternational Conference on Precision Agriculture, Hyatt Regency Tech Center, Denver, Colorado.

Lelong, C. C., Pinet, P. C., & Poilvé, H., 1998. Hyperspectral imaging and stress mapping in agriculture: a case study on wheat in Beauce (France).*Remote Sensing of Environment*. 66(2), 179-191.

Li, J., Bioucas-Dias, J. M., & Plaza, A., 2012. Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. *Geoscience and Remote Sensing, IEEE Transactions on*, *50*(3), 809-823.

Ling, P. P. & Ruzhitsky, V. N.1996. Machine vision techniques for measuring the canopy of tomato seedling. *Journal of agricultural engineering research,* 65(2), 85-95.

Liu, J., Li, X., Li, P., Wang, W., Zhang, J., Zhou, W. & Zhou, Z. 2011. Non-destructive measurement of sugar content in chestnuts using near-Infrared spectroscopy. *Comput. Computing Technol. Agri. IV.* 246-254.

Liu, Z., Huang, J. & Tao, R. 2007. Characterizing and estimating fungal disease severity of rice brown spot with hyperspectral reflectance data. *Rice Sci*. *15(3),* 232-242.

Lu, R., 2003. Detection of bruises on apples using near-infrared hyperspectral imaging. *Transactions of American Society of Agricultural Engineers*. 46(2): 523–530.

Lu, R., 2004. Multispectral imaging for predicting firmness and soluble solids content of apple fruit. *Postharvest Biology and Technology*. 31 (2): 147–157.

Lu, R., & Chen, Y. R., 1999. Hyperspectral imaging for safety inspection of food and agricultural products. International Society for Optics and Photonics- *Photonics East (ISAM, VVDC, IEMB).* 121-133.

Martinez-Uso, A., Pla, F., Sotoca, J.M. & Garcia-Sevilla, P., 2007. Clustering-based hyperspectral band selection using information measures. IEEE T Geosci. Remote: 45(12), 4158–4171.

Matiacevich, S., Silva, P., Enrione, J., & Osorio, F. 2011. Quality assessment of blueberries by computer vision. *Procedia Food Science*, 1, 421-425.

Menesatti, P., Antonucci, F., Pallottino, F., Roccuzzo, G., Allegra, M., Stagno, F. & Intrigliolo, F. 2010. Estimation of plant nutritional status by VIS-NIR spectrophotometric analysis on orange leaves. *Biosyst. Eng. 105(4),* 448-454.

Morgan, K., Olmstead, J., Williamson, J., Krewer, G., Takeda, F., MacLean, D., Shewfelt, R., Li, C., Malladi, A & Lyrene, P., 2011. Economics of hand and mechanical harvest of new “Crispy” flesh cultivars from Florida. 2011 Blueberry Educational Session, Savannah, GA.

Muller, K., Böttcher, U., Meyer-Schatz, F. & Kage, H. 2008. Analysis of vegetation indices derived from hyperspectral reflection measurements for estimating crop canopy parameters of oilseed rape (Brassica napus L.). *Biosyst. Eng*. *101(2),* 172-182.

Nichol, C.J. & Grace, J. 2010. Determination of leaf pigment content in Calluna vulgaris shoots from spectral reflectance. *Int. J. Remote Sens*. *31(20),* 5409.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I. & Lammertyn, J. 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biol. Technol. 46(2),* 99-118.

Okamoto, H., & Lee, W. S. 2009. Green citrus detection using hyperspectral imaging. *Computers and Electronics in Agriculture*, *66*(2), 201-208.

Okamoto, H., Lee, W. S. & Kane, K. E., 2007. Hyperspectral imaging for green citrus fruit detection.6thEuropean Conference in Precision Agriculture and 2nd European Conference on Precision Livestock Farming, June 3-6, 2007,Skiathos,Greece.

Olennikov, D. & Tankhaeva, T. 2006. Absorption spectra of carbohydrates and related compounds in H_2SO_4. *Chemistry of Natural Compounds.* 42(3), 262–264.

Peng, Y. & Lu, R., 2006. Improving apple fruit firmness predictions by effective correction of multispectral scattering images. *Postharvest Biology and Technology.* 41(3): 266–274.

Peng, Y., & Lu. R., 2008. Analysis of spatially resolved hyperspectral scattering images for assessing apple fruit firmness and soluble solids content*. Postharvest Biology and Technology.* 48(1): 52–62.

Prior, R. L., Cao, G., Martin, A., Sofic, E., McEwen, J., O'Brien, C., Lischner, N., Ehlenfeldt, M., Kalt, W., Krewer, G. & Mainland, C. M. 1998. Antioxidant capacity as influenced by total phenolic and anthocyanin content, maturity, and variety of vaccinium Species. *J. Agric. Food Chem.* 46, 2686-2693.

Rao P. 2007. Remote sensing for assessing vegetational dynamics and productivity of a peatland in southern Sweden. M.S. Thesis. International Institute for Geo-Information Science and Earth Observation, Enschede, The Netherlands.

Regunathan, M. & Lee, W. S., 2005. Citrus yield mapping and size determination using machine vision and ultrasonic sensors. *2005 ASAE Annual Meeting.* ASAE Paper No. 053017.

Rifkin, R. & Klautau, A., 2004. In defense of one-v.s.-all classification. J. Mach. Learn. 5, 101–141.

Rokach, L. & Maimon, O. 2005. Top-down induction of decision trees classifiers - a survey. *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*. *35(4),* 476- 487.

Schueller, J.K, Whitney, J.D., Wheaton, T.A ., Miller, W.M.& Turner A.E. 1999. Low-cost Automatic Yield Mapping in Hand-harvested Citrus. *Computers and Electronics in Agriculture,* 23(2), 145–153.

Sutton, C. 2005. Classification and regression trees, bagging, and boosting. *Handbook of Statistics*. *24*, 303-329.

Tao, Y., Heinemann, P. H., Varghese, Z., Morrow, C. T., & Sommer, H. J. 1995. Machine vision for color inspection of potatoes and apples.*Transactions of the ASAE.* 38(5), 1555-1561.

Tarabalka, Y., Benediktsson, J. A. & Chanussot, J., 2009. Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. *Geoscience and Remote Sensing, IEEE Transactions on*, 47(8), 2973-2987.

USDA, 2012. 2012 Fruit and tree nuts yearbook table-D2, Blueberries: acres, yield, production, price, by state, 80/81-to date.

US Highbush Blueberry Council. 2010. Health benefits of blueberries - blueberry nutrition. Available at: http://www.blueberrycouncil.com/health-benefits-of-blueberries/blueberry-nutrition/. Accessed on 10 May, 2012.

Van der Meer, F. D., Van der Werff, H. M. A., & Van Ruitenbeek, F. J. A., 2005. Spatial-spectral contextual image analysis of hyperspectral data to aid in the characterisation of hydrothermal alteration in epithermal gold deposits. *Proceedings of the 26th Asian conference on remote sensing* (Vol. 9).

Wang, H. C. & Li, M. Wang. 2011. Onion internal quality prediction using line-scan hyperspectral imaging. 2011 ASABE Paper No. 1110708. Louisville, Kentucky.

Webb, A., 2002, Statistical Pattern Recognition, 2nd ed. Hoboken, NJ: Wiley.

Wen, Z., & Tao, Y. 1999. Building a rule-based machine-vision system for defect inspection on apple sorting and packing lines*. Expert Systems with Applications*, 16(3), 307-313.

Wijethunga, P., Samarasinghe, S., Kulasiri, D., & Woodhead, I., 2008. Digital image analysis based automated kiwifruit counting technique. *2008 IEEE Image and Vision Computing 23rd International Conference*, 1–6.

Williamson, J.G., Olmstead, J. & Lyrene, P., 2012. Florida's commercial blueberry industry. Horticultural Sciences Department, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida, HS742.

Williams, P. & Norris, K. 2001. Near-infrared technology. St. Paul, Minnesota: American Association of Cereal Chemists.

Yang, C. & Lee, W.S., 2011. Spectral Signatures of Blueberry Fruits and Leaves. 2011 ASAE Annual Meeting. Paper number 1110582.

Yang, C., Lee, W. S. & Gader, P., 2013. Band selection of hyperspectral images for the detection of blueberry fruit with different growth stages. 2013 ASABE paper No. 1593276.

Yang, C., Lee W.S. & Williamson J., 2012, Classification of Blueberry Fruit and Leaves based on spectral signatures. *Biosys. Engi..* 113(4), 351–362.

Yang, H., Du, Q., Su, H. & Sheng, Y., 2011. An efficient method for supervised hyperspectral band selection. *IEEE T Geosci. Remote,* 8(1), 138-142.

Yang, H. & Irudayaraj, J. 2002. Rapid determination of vitamin C by NIR, MIR and FT-Raman techniques. *J. Pharmacy and Pharmacology.* 54, 1247-1255.

Yang, Q., 1996. Apple stem and calyx identification with machine vision. *Journal of agricultural engineering research.* 63(3), 229-236.

Yang, W., Li, M. & Sigrimis, N. 2010. Estimating nitrogen content of cucumber leaves based on NIR spectroscopy. *Sens. Lett. 8(1),* 145-150.

Zaman, Q. U., Schumann, A. W., Percival, D. C. & Gordon, R. J., 2008. Estimation of wild blueberry fruit yield using digital color photography. *Transactions of the ASABE.* 51(5): 1539–1544.

Zaman, Q. U., Schumann, A. W., Hostler, H. K., 2006. Estimation of citrus fruit yield using ultrasonically-sensed tree size*. Applied Engineering in Agriculture*. 22(1): 39-44.

Zou, Q., Fang, H., Liu, F., Kong, W. & He, Y. 2011. Comparative study of distance discriminant analysis and BP neural network for identification of rapeseed cultivars using visible/near infrared spectra. *Comput. Computing Technol. Agri.* IV. 124-133.

BIOGRAPHICAL SKETCH

Ce Yang was born and raised in Xinji City, Hebei Province in China. She received her bachelor of engineering in electrical engineering in 2007 at China Agricultural University (CAU), Beijing, China. In 2009, Ce obtained her master of science in agricultural electronics and automation at the Precision Agriculture Lab from CAU. In 2009, she came to University of Florida to pursue her Ph.D. degree in the Department of Agricultural and Biological Engineering.