Computers and Electronics in Agriculture 109 (2014) 23–31

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Hyperspectral band selection for detecting different blueberry fruit maturity stages

Ce Yang [a,1], Won Suk Lee [b,*], Paul Gader [c,2]

[a] Bioproducts and Biosystems Engineering Department, University of Minnesota, St. Paul, MN 55108, United States
[b] Agricultural and Biological Engineering Department, University of Florida, Gainesville, FL 32611, United States
[c] Computer Information Science and Engineering Department, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

Article history:
Received 10 March 2014
Received in revised form 19 August 2014
Accepted 20 August 2014

Keywords:
Band selection
Blueberry
Hyperspectral imagery
Kullback–Leibler divergence
Precision agriculture
Yield mapping

ABSTRACT

Hyperspectral imagery divides spectrum into many bands with very narrow bandwidth. It is more capable to detect or classify objects, where visible information is not sufficient for the task. However, hyperspectral image contains a large amount of redundant information, which eliminates its discriminability. Band selection is used to both reduce the dimensionality of hyperspectral images and save useful bands for further application. This study explores the feasibility of hyperspectral imaging for the task of classifying blueberry fruit growth stages and background. Three information theory based band selection methods using Kullback–Leibler divergence: pair-wise class discriminability, hierarchical dimensionality reduction and non-Gaussianity measures were applied. Three classifiers, *K*-nearest neighbor, support vector machine and AdaBoost were used to test the performance of the selected bands by the three methods. The selected bands achieved classification accuracies of 88% and higher. Therefore, the band selection methods are very useful in reducing the volume of the hyperspectral data, and constructing a multispectral imaging system for detecting blueberry fruit maturity stages.

1. Introduction

Labor expense of handpicked blueberries for fresh markets is increasing due to severe shortages of available farm workers. Management cost of Florida's commercial blueberry field excluding harvesting labor is approximately $9884 ha (Williamson et al., 2012). The average blueberry yield in Florida is 6310 kg/ha (USDA, 2012). Morgan et al. (2011) estimated that the hand harvest cost is $1.59 kg. Therefore, the cost of harvesting labor takes higher than $10,000 ha, which is more than half of the total management cost of the blueberry production. Efficient harvesting labor assignment in a large blueberry field can reduce much of the harvesting cost. Furthermore, yield estimation prior to harvest helps growers to find problems in their fields as early as possible. It is useful for growers to make further decisions such as irrigation, pest control, and weed control. Therefore, yield estimation of blueberry field prior to harvesting is beneficial for the growers. During the harvest season, individual blueberries in a fruit cluster usually mature at different times. A cluster may contain all growth stages including young fruit (green color), intermediate fruit (red color) and mature fruit (dark blue/purple) at the same time. Fig. 1 is an example picture taken from a blueberry field during the blueberry harvest season in 2013.

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 the study contained only mature fruit, which was easily distinguishable because of its significant color contrast in the blue band. However, as shown in Fig. 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. Hyperspectral imaging has been used in detecting fruit and vegetable quality such as maturity, firmness, starch content, soluble solid contents for over a decade. Lu and Peng (2006) investigated peach fruit firmness using hyperspectral

* Corresponding author. Tel.: +1 352 392 1864x227.
E-mail addresses: ceyang@umn.edu (C. Yang), wslee@ufl.edu (W.S. Lee), pgader@ufl.edu (P. Gader).
[1] Tel.: +1 612 626 6419.
[2] Tel.: +1 352 392 1527.

http://dx.doi.org/10.1016/j.compag.2014.08.009
0168-1699/

Fig. 1. A blueberry fruit bunch that contains all three growth stages: young, intermediate and mature.

scattering. They selected 10 or 11 wavelengths with r^2 of 0.77 and 0.58 for two peach cultivars. Nagata et al. (2004) estimated strawberry maturity by measuring the soluble solids content of 'Akihime' strawberries and had a correlation coefficient of 0.784 using five-predictor firmness model. However, all the five predictors they chose are in the visible range. Rajkumar et al. (2012) studied banana maturity stages at different temperatures using hyperspectral imaging and obtained coefficient of determinant of 0.85, 0.87 and 0.91 for total soluble solids, moisture and firmness, respectively. There are also studies on maturity estimation of mango, peanut pod, etc. (Sivakumar et al., 2011; Carley, 2006). However, all of these research were carried out as a post-harvest step in the packing house or lab, which have more ideal condition compared to the outdoor environments. The samples for these experiments were placed in the lighting house and images were taken with even illumination, and without wind or shadow factors. These methods are not applicable for on-site early crop maturity detection.

In a previous study, 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 a 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 stable indoor light source. In addition, the samples were well prepared without much noisy background. However, field measurement uses the sunlight as its illumination source, and the background contains not only leaves, but also soil, sky, and man-made objects such as PVC irrigation pipes. A portable spectrometer can only measure either a spot or a small area as one spectrum, which cannot provide sufficient information. A color image is not easy to detect all the fruit maturity stages because of the similar colors of young fruit and leaves. On the other hand, hyperspectral images obtained from field conditions 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 objects.

Due to the high spectral resolution, hyperspectral images contain considerable amount of redundancy. The images usually have several hundred bands, but some bands are useless or even hinder the discriminability of useful bands. Adjacent bands in the spectrum tend to be highly correlated (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 have physical information. Some selected original bands can be used for yield estimation using a multispectral camera system. A multispectral camera is of lower cost and higher processing speed compared to a hyperspectral camera system. Therefore, a multispectral imaging system 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 calculation complexity. They did not focus on what specific selected bands were.

The objectives of this study were to explore the feasibility of hyperspectral imagery in classifying different blueberry growth stages, and to select useful bands that are suitable for a multispectral imaging system, 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 the Kullback–Leibler divergence (KLD) was proposed, which measures pair-wise discriminability of spectral bands. The proposed method was compared with two other band selection methods: hierarchical dimensionality reduction and non-Gaussianity measures. The band selection methods were tested by *K*-nearest neighbor (KNN), support vector machine (SVM) (Martinez-Uso et al., 2007; Chang and Wang, 2006) and AdaBoost (Freund and Schapire, 1995) by their performance in classifying the blueberry growth stages and background.

2. Materials and methods

2.1. Hyperspectral image acquisition

Hyperspectral images were obtained from a blueberry research and demonstration farm at the University of Georgia cooperative extension in Alma, GA, United States (31.534°N, 82.510°W) 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 15.2 × 15.2 cm^2 of the view was acquired. A hyperspectral imaging system was used for image acquisition, consisting of 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 data acquisition 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, the encoder generated pulses, which was sent to the program for generating a trigger signal. The camera acquired one line

Fig. 2. Hyperspectral camera used in the blueberry field.

image once it received a trigger signal from the program. The camera body that is mounted on a tilting head and the encoder are mounted on a tripod, which are shown in Fig. 2.

The original image contained a total of 776 bands with a highest spectral resolution of 0.79 nm, which would make the image size very large. Therefore, binning was used to reduce the spectral resolution by a half. After binning, there were 388 hyperspectral bands with spectral resolution of 1.59 nm, which was sufficient for our study. The spectral range was 398–1010 nm, and 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 was n (number of lines) $\times$ 1312 (pixels/line) $\times$ 388 (bands). Reflectance images were created using a universal white standard (Spectralon, Labshpere Inc., North Sutton, NH, USA). Fig. 3 shows the RGB bands of an example 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.

A training/testing pixel set was collected by manually collecting 600 pixels together with labeling information. A half of the pixels of each class (mature fruit, intermediate fruit, young fruit and leaf) were selected randomly and put in the training set and the other half were in the testing set.

Matlab R2012a (The MathWorks Inc., Natick, MA, USA) was used to implement three hyperspectral band selection methods in this study. KLD was used in the three methods as a criterion of determining variation between distributions. Among the three methods, pair-wise class discriminability (PWCD) measure was proposed as a supervised band selection method. Hierarchical dimensionality reduction (HDR) and non-Gaussianity (NG) measure were also applied which were unsupervised band selection methods. In this study, these two band selection methods were used based on the training set, which was labeled manually by three colleagues who did not have prior spectral knowledge. The labeling results were obtained according to a majority vote of the three opinions.

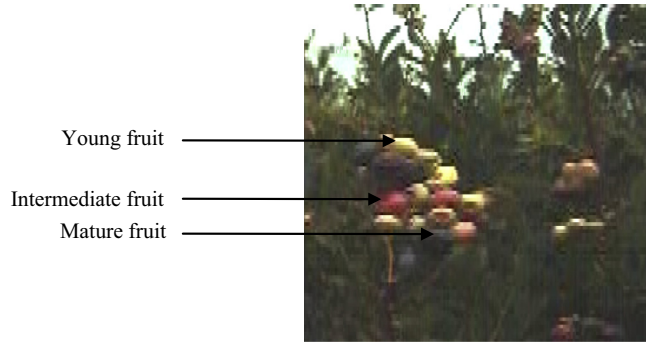

Fig. 3. RGB bands of a hyperspectral image with all blueberry fruit growth stages.

2.2. Band selection methodology

2.2.1. Information theory and Kullback–Leibler divergence. In information theory, entropy is a measure of the amount of information of a random variable. The entropy of a random variable X with a probability density function $p(x)$ is:

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

where Ω includes all possible events, and x is the value of X.

If X is a discrete random variable, then $p(x)$ is the probability mass function of all possible events. Entropy $H(X)$ for a discrete random variable is defined as

$$H(X) = -\sum_{x\in\Omega} p(x)\log p(x) \tag{2}$$

KLD is an information divergence measure, which shows the dissimilarity distance between two probability distributions. The original KLD is non-symmetric, and 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 as:

$$D_{kl}(X_i, X_j) = \sum_{x\in\Omega}\left[p_i(x)\log\frac{p_i(x)}{p_j(x)} + p_j(x)\log\frac{p_j(x)}{p_i(x)}\right] \tag{3}$$

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

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

2.2.2. Pair-wise class discriminability

The proposed PWCD method 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 a 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 can be normalized (Jia and Richards, 2002) so that the values in each histogram would sum up to one. Our goal is to find the band that has the most discrepancy between two classes. It is expressed as the following equation:

$$\mathrm{argmax}_B(D_{kl,B}(C_i, C_j)) = \mathrm{argmax}_B\sum_{c\in\Omega}\left[hist_{i,B}(c)\log\frac{hist_{i,B(c)}}{hist_{j,B}(c)} + hist_{j,B}(c)\log\frac{hist_{j,B}(c)}{hist_{i,B}(c)}\right] \tag{4}$$

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. In this study, the bands that maximized the KLD of class pairs were chosen. Since there were four classes

(mature fruit, intermediate fruit, young fruit and background), which would make six pairs of classes, six bands were selected in the end.

2.2.3. Hierarchical dimensionality reduction (HDR)

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:

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

where $D_{kl}(B_i, B_j)$ is the KLD of band i (B_i) and band j (B_j) in an image. $hist_i(b)$ and $hist_j(b)$ are the normalized histograms of the two bands.

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

2.2.4. Non-Gaussianity measures

The non-Gaussianity measures were 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 similar to a Gaussian distribution. In contrast, the more the histogram differs from a Gaussian distribution, the better it is. The difference between them can be expressed as:

$$D_{kl}(B_i, B_{ig}) = \sum_{b\in\Omega}\left[hist_i(b)\log\frac{hist_i(b)}{P_{ig}(b)} + P_{ig}(b)\log\frac{P_{ig}(b)}{hist_i(b)}\right] \quad (6)$$

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

2.3. Supervised classification

In order to compare the performance of the band selection methods, three supervised classifiers were applied to the testing data set: K-nearest neighbor (KNN), support vector machine (SVM), and Adaptive Boosting (AdaBoost). 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 is taken into consideration. If $K = 1$, the test sample is assigned only to the class of the nearest neighbor. Larger K reduces the effect of noise and outliers in the classification. However, the boundaries among classes are less clear. It does not require a training step because all the distance calculations are in the testing step.

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 is constructed with a 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, kernel functions are introduced so that 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. Widely used kernels include polynomial kernel, Gaussian radial basis function (RBF) kernel, etc.

AdaBoost 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. Classifiers such as best-first tree (BFTree), functional tree (FT), logistic model tree (LMT), and classification and regression tree (CART) can be applied in this method (Hall et al., 2009). The method sequentially runs the classifiers, and the weight of each training sample is modified during the application of the classifiers. The incorrectly 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. Matlab R2012a was used to apply KNN. Weka software from the University of Waikato was used to apply SVM and AdaBoost.

3. Results and discussion

3.1. Blueberry spectra

In order to show that the four observed classes have different spectra, ten pixels of each class were randomly selected from the training set and their spectra are shown in Fig. 4. Leaves occupy most of the background. Therefore, leaf pixels are used to represent the background spectrum. Leaves contain the highest chlorophyll content, which results in high reflectance in the NIR range. Mature fruit is very dark and therefore, has a very low reflectance in the visible range. It also has relatively lower value in the near-infrared (NIR) range, as shown in the figure. Intermediate fruit appears red in the color image and has a 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.

3.2. Principal component analysis

Principal component analysis (Yang et al., 2012) 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) were extracted to show the distribution of the classes in Fig. 5. 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 diamonds 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 image acquisition under direct sunlight. Another reason is that the images were acquired from many different conditions, considering the depth of view, the influence of water evaporation, etc., and thus the pixels are widely scattered. All in all, although various conditions in the field have strong impacts on spectral properties, the four classes are separable by the three PCs.

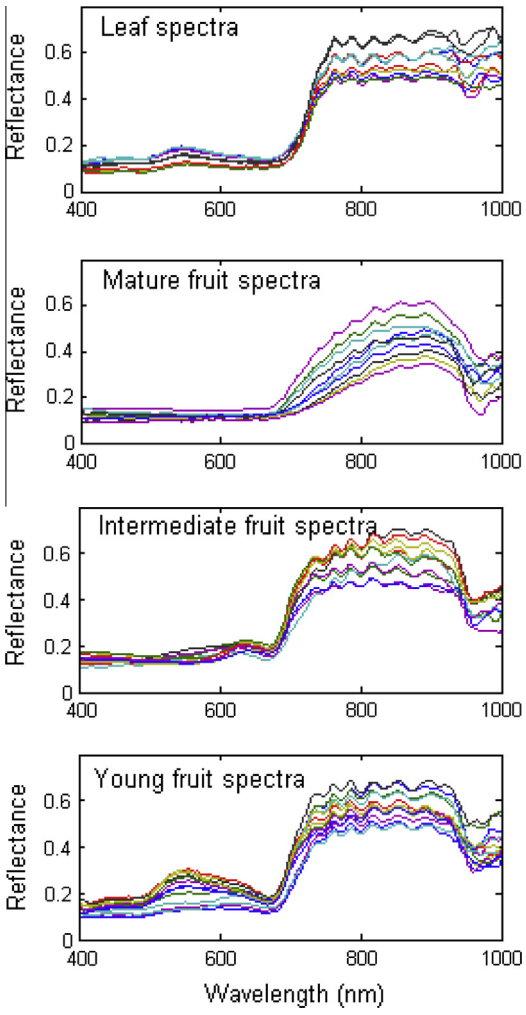

Fig. 4. Spectra of ten pixels for each class: background (leaf), mature fruit, intermediate fruit, and young fruit.

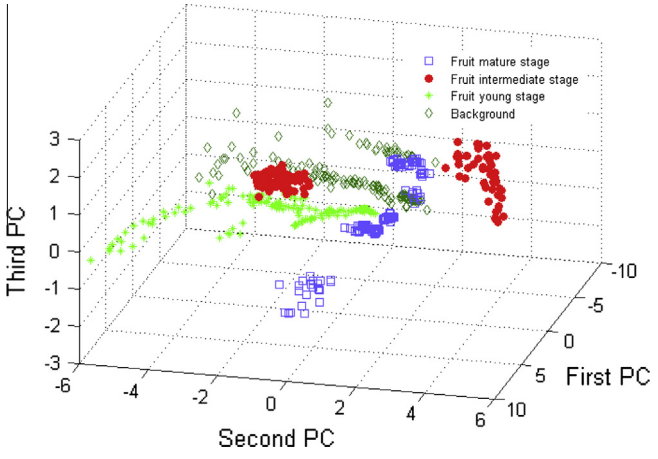

Fig. 5. Principal component transform of the four classes: mature fruit, intermediate fruit, young fruit and background.

3.3. Band selection results

Hierarchical dimension reduction (i.e., HDR) and non-Gaussianity measures (i.e., NG) are unsupervised methods. However, in this study, they were used as supervised methods. The pixels were labeled based on both spatial and spectral information of the hyperspectral images and divided into training and validation pixel sets. HDR, NG measures, and PWCD were applied to the training set to select bands. Then the selected bands by the three methods were tested using the validation set.

3.3.1. Pair-wise class discriminability

By calculating the KLD between probability distributions, PWCD method selected six bands for separating class pairs. Fig. 6 shows the normalized histograms of the bands for class pairs that were selected by this method. Band 41 (457 nm) was used to discriminate mature and intermediate fruits. Band 303 (870.5 nm) had the highest discriminability for mature fruit and young fruits. Band 68 (498.4 nm) was the best for separating mature fruit from background. Band 176 (666.7 nm) separated intermediate fruit from young fruit with the best result. Band 145 (617.9 nm) separated intermediate fruit from background with the best result. Band 164 (647.8 nm) achieved the best separation result for young fruit from background. Some of the histograms in this figure show multi-modal distributions, which is mostly because of the shadows caused by the direct sunlight.

3.3.2. Hierarchical dimensionality reduction

Hierarchical dimensionality reduction was applied on the labeled training and testing pixel sets. This method aggregates the bands that have very similar normalized histograms from the training set. The bands are then grouped into clusters. The mean of each cluster is calculated. The band with the highest correlation with the cluster mean is chosen to represent the cluster. In the end, six bands were selected from six clusters. As expected, the clusters mainly aggregate neighboring bands. Fig. 7 shows the band clustering result and the selected bands. The Y-axis is the band cluster numbers from 1 to 6. The selected bands are: 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). One cluster 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.

3.3.3. Non-Gaussianity measure

The NG measure method directly sorts the bands by their Gaussianity. The top bands chosen are those with the highest KLD values, which are the NG measures, between the original distribution and the simulated Gaussian distribution. The result of this method is shown in Table 1, listing the top 25 bands. 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 nearby bands have higher correlations. In addition, the band selection results will be applied to a multispectral camera system, in which the band width is usually at least 10 nm. Therefore, some top-ranked bands which were close to each other were categorized into a group with a spectral range of 10 nm. The first column of Table 1 is the ranking, the second column is their NG measures, and the group numbers are in the last column. The NG measure decreases quickly from the first ranked band to the second, however it decreases much slower after that. The top ranked band in every group was chosen as a representative band for that group. The selected bands are underlined in Table 1. Thus, the final selected bands are: 192 (692.1 nm), 246 (778.3 nm), 175 (665.1 nm), 181 (674.6 nm), 162 (644.6 nm) and 142 (613.2 nm).

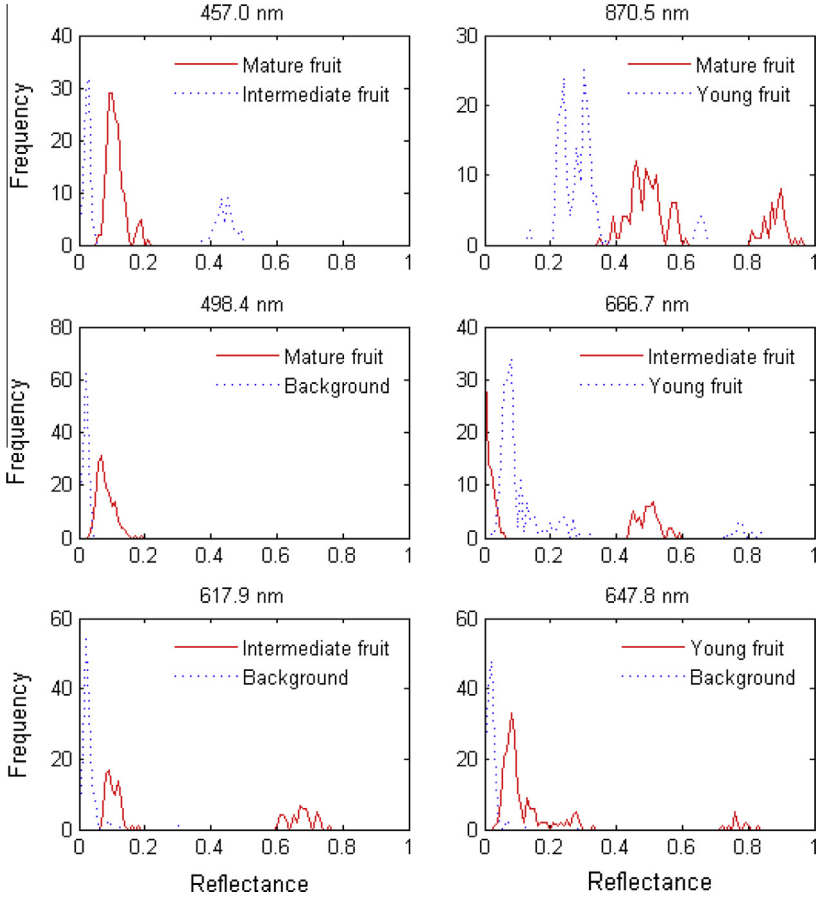

Fig. 6. Separation ability of the selected bands by PWCD.

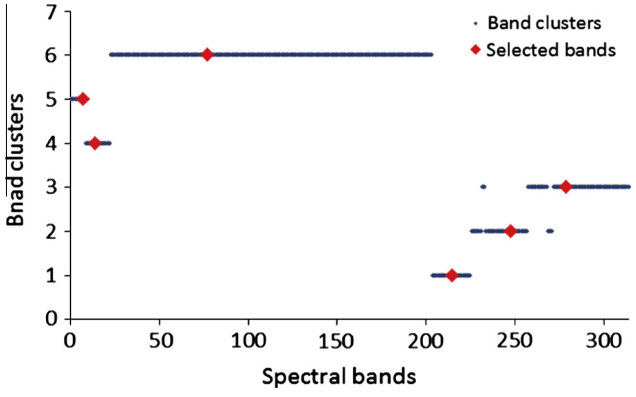

Fig. 7. Band clustering result and the selected bands by calculating correlations between cluster mean and individual bands.

Table 1
Sorted bands using the 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
21	1694.6	142	613.2	VI
22	1691.3	255	792.8	VII
23	1690.3	188	685.7	IV
24	1687.8	248	781.5	II
25	1685.4	170	657.3	III

3.4. Classification using the selected bands

KNN classifier, SVM and AdaBoost were applied to test the performance of the bands selected by the three methods. The classification results using the bands selected by PWCD are shown in Table 2. The overall correct detection was calculated by dividing the amount of correctly detected samples by the total amount of pixels. The overall false positive was calculated in the same way. 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 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 a model. Given a dataset that is much larger,

Table 2
Classification results of three classifiers using the 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 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

the calculation time can be a problem. However, it is worth mentioning that the long processing time will be for building the classifier rather than testing new pixels.

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 limitation of quantity of the training pixels. The overall accuracy is 96.8% and the false positive rate is 3.2%, which are shown in Table 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, average prediction accuracies with up to $K = 10$ (data not shown) were 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 Pearson VII Function-Based Universal Kernel (PUK) (Trivedi and Dey, 2013). Polynomial, RBF kernel, and other parameters achieved much lower accuracy.

Table 3 summarizes the classification results of the three classifiers using the bands selected by HDR. They achieved 97.8% overall classification accuracy using the KNN classifier when $K = 1$. The best classification result using SVM is 95.8% with $c = 5$ and a PUK kernel. AdaBoost method obtained 92.3% of an overall accuracy when using NBTree classifiers.

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

3.5. Discussion

The selected wavelengths and classification results are listed in Table 5 for comparison. Bands selected by PWCD are well scattered across the spectral range. However, PWCD did not consider the correlations between bands. Since HDR groups bands based on their discrepancy on the training set, it is logical to use the band clustering by HDR to analyze the selected bands from other methods. Among the selected bands, wavelength 415.9 nm is near the carbohydrate absorption band, which is 424 nm (Yang et al., 2012). 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 reflectance peak, which is very high for leaf and young fruit. Although HDR did not always achieve the highest prediction accuracy, it kept a relatively stable prediction accuracy using all three classification methods. However, five bands are from one HDR cluster (numbered 6 in Fig. 7). It is possible that HDR cluster 6 lost a lot of information since its range is too wide, covering all the three visible bands and the red edge. This might be the reason of the lowest prediction result of AdaBoost using the HDR 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 (Yang et al., 2012), 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

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

Band selection methods	Six selected wavelengths (nm)	KNN (%)	SVM (%)	AdaBoost (%)
PWCD	457.0, 498.4, 617.9, 647.8, 666.7, 870.5	96.8	90.6	97.5
HDR	405.3, 415.9, 512.2, 728.6, 781.5, 831.5	97.8	95.8	92.3
NG measure	613.2, 644.6, 665.1, 674.6, 692.1, 778.3	98.7	88.2	98.4

613 nm to 776 nm, which are the red band and red edge, a very narrow range compared to the spectral range of the image. This caused lower prediction accuracy using SVM, compared with 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 red band and red edge are crucial for the classification task in this research.

SVM transforms the original features into infinite dimensions where the samples are linearly classifiable. 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 a model. Given a large dataset, AdaBoost might be computationally intensive.

4. Conclusion

In this study, three information theory based band selection methods, pair-wise class discriminability (PWCD), hierarchical dimensionality reduction (HDR) and non-Gaussianity (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. Kullback–Leibler divergence (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 were 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 discriminating 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 study. The bands that have the highest correlations with the centers of the band clusters were chosen. 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).
- 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 were 192 (692.1 nm), 246 (778.3 nm), 175 (665.1 nm), 181 (674.6 nm), 162 (644.6 nm) 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 detection 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 hyperspectral imagery. NG measure method selected bands from only the red band and the red edge, which yielded the highest prediction accuracy using KNN with $K = 1$ and AdaBoost. Therefore, the red band and red edge are very important for distinguishing the fruit growth stages and leaf. Overall, the selected bands achieved 88% and higher accuracies using all three band selection methods, which means that the proposed band selection methods are capable of classifying the blueberry maturity stages without using all the spectral bands. The band selection methods can be adapted in the maturity stage detection of other fruits and vegetables using hyperspectral imaging as well.

Acknowledgements

The authors would like to thank Dr. Changying Li from University of Georgia for offering the blueberry demonstration field for data collection. Many thanks go to Mr. John Ed Smith from University of Georgia, Ms. Han Li, Mr. James Park, and Mr. Hao Ma, from the Precision Agriculture Laboratory, University of Florida, who helped collect the images in the field. This study was funded by the Graduate School Fellowship at the University of Florida.

References

Cai, S., Du, Q., Moorhead, R.J., 2007. Hyperspectral imagery visualization using double layers. IEEE Trans. Geosci. Remote 45 (10), 3028–3036.
Carley, S.D., 2006. Potential Use of Hyperspectral and Multispectral Remote Sensing Imagery to Enhance Management of Peanut (*Arachis hypogaea L.*). Ph.D. Dissertation. North Carolina State University, Raleigh, North Carolina.
Chang, C.-I., Wang, S., 2006. Constrained band selection for hyperspectral imagery. IEEE Trans. Geosci. Remote 44 (6), 1575–1585.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
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.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. 11 (1), 10–18.
Jia, S., Qian, Y., Ji, Z., 2008. Band selection for hyperspectral imagery using affinity propagation. In: Proceedings of the 2008 Digital Image Computing: Techniques and Applications, pp. 137–141.
Jia, X., Richards, J.A., 2002. Cluster-space representation for hyperspectral data classification. IEEE Trans. Geosci. Remote Sens. 40 (3), 593–598.
Lu, R., Peng, Y., 2006. Hyperspectral scattering for assessing peach fruit firmness. Biosyst. Eng. 93 (2), 161–171.
Martinez-Uso, A., Pla, F., Sotoca, J.M., Garcia-Sevilla, P., 2007. Clustering-based hyperspectral band selection using information measures. IEEE Trans. Geosci. Remote 45 (12), 4158–4171.
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.
Nagata, M., Tallada, J. G., Kobayashi, T., Cui, Y., Gejima, Y., 2004. Predicting maturity quality parameters of strawberries using hyperspectral imaging. In: ASAE/CSAE Annual International Meeting, Ottawa, Ontario, Canada, Paper No. 043033.
Rajkumar, P., Wang, N., EImasry, G., Raghavan, G.S.V., Gariepy, Y., 2012. Studies on banana fruit quality and maturity stages using hyperspectral imaging. J. Food Eng. 108 (1), 194–200.
Rifkin, R., Klautau, A., 2004. In defense of one-v.s.-all classification. J. Mach. Learn. Res 5, 101–141.
Sivakumar, D., Jiang, Y., Yahia, E.M., 2011. Maintaining mango (*Mangifera indica*) fruit quality during the export chain. Food Res. Int. 44 (5), 1254–1263.
Trivedi, S.K., Dey, S., 2013. Effect of various kernels and feature selection methods on SVM performance for detecting email spams. Int. J. Comput. Appl. 66 (21), 18–23.
USDA, 2012. 2012 Fruit and Tree Nuts Yearbook Table-D2, Blueberries: Acres, Yield, Production, Price, by State, 80/81-to Date.
Webb, A., 2002. Statistical Pattern Recognition, second ed. Wiley, Hoboken, NJ.

Williamson, J., 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.
Yang, H., Du, Q., Su, H., Sheng, Y., 2011. An efficient method for supervised hyperspectral band selection. IEEE Trans. Geosci. Remote 8 (1), 138–142.
Yang, C., Lee, W.S., Williamson, J., 2012. Classification of blueberry fruit and leaves based on spectral signatures. Biosys. Eng. 113 (4), 351–362.
Zaman, Q.U., Schumann, A.W., Percival, D.C., Gordon, J., 2008. Estimation of wild blueberry fruit yield using digital color photography. Trans. ASABE 51 (5), 1539–1544.