A Genetic-Algorithm-Based Selective Principal Component Analysis (GA-SPCA) Method for High-Dimensional Data Feature Extraction

Haibo Yao, Student Member, IEEE, and Lei Tian, Member, IEEE

Abstract—In this paper, a genetic-algorithm-based selective principal component analysis (GA-SPCA) method is proposed and tested using hyperspectral remote sensing data and ground reference data collected within an agricultural field. The proposed method uses a global optimizer, the genetic algorithms, to select a subset of the original image bands, which first reduces the data dimension. A principal component transformation is subsequently applied to the selected bands. By extracting features from the resulting eigenimage, the remote sensing data, originally high in dimension, will be further reduced to a feature space with one to several principal component bands. Subsequent image processing on the reduced feature space can thus be performed with improved accuracy. Experiments were conducted using three sets of ground reference data: corn chlorophyll content, corn plant population, and various corn hybrids. The results showed that with GA-SPCA, the number of original bands used for principal component analysis (PCA) could be reduced to 17, 26, and 25 from a 60-band hyperspectral image, respectively. In all cases, the correlation coefficients between image and ground reference data were greater when using GA-SPCA than that for PCA results with all original bands. This indicates that bands with no contribution to a specific application were removed prior to PCA. The variance related to a specific application within the image was transformed with more emphasis by using bands sensitive to that application. The selected bands can also provide useful information for future imaging sensor development.

Index Terms—Feature extraction, genetic algorithm, hyperspectral image, selective principal component analysis, supervised dimension reduction.

I. INTRODUCTION

MANY FACTORS contribute to plant health and affect yield. Identification of such factors is important for crop production. Remote sensing can potentially provide a fast and economic way for this task. The spectral response of a plant varies with species, and crops and weeds can be distinguished using spectral information [1]. Recently, aerial hyperspectral images have been employed in agricultural applications. These applications include calculating vegetation indices for the estimation of green leaf area index and canopy chlorophyll density [2] and describing soil variability [3]. It is expected that information extracted from hyperspectral images will be a viable source of data for in-season crop management and precision farming applications, such as plant growth condition modeling, canopy and weed mapping, crop stress detection, yield estimation, and soil property identification.

Some ranges of wavelength are significant for agricultural applications and these wavelength ranges generally vary for different applications. One example is that with broadband images, crop plants have the greatest reflectance in green (G) and near infrared (NIR) regions and the largest absorbance in red (R) and blue (B) regions [4]. Field information extraction would be improved if image processing focused on the significant wavelength ranges (or image bands). However, in the past, it has been difficult to correctly locate specific wavelength ranges due to sensor limitations. The emergence of hyperspectral imagery provides opportunities for identifying and studying specific wavelength ranges. A hyperspectral image has a very high spectral resolution (one to several nanometers) with a large number of bands (tens to hundreds) [5]. High spectral resolution images can be used to study either the physical characteristics at each pixel location by looking at the shape of the spectral reflectance curves or the spectral/spatial relationships of different classes using pattern recognition and image processing methods. The reflectance curve shape, such as the shape in the red edge region, is closely associated with leave chlorophyll content and has been used for crop nitrogen status estimation [6]. Rather than using the spectral shape, this paper takes a different approach, which is similar to the latter method with certain data transformations in the spectral domain.

One important feature of hyperspectral imagery is that the neighboring image bands are highly correlated (Fig. 1). This correlation is mainly due to material spectral correlation, topography, and sensor band overlap. The band-to-band correlation creates redundant information in the hyperspectral image. It is preferable to remove the redundant information before image interpretation. Furthermore, even with a subset of the hyperspectral image bands (such as in the aforementioned significant wavelength range), the number of bands is still sometimes large, normally in the order of tens. When processing high-dimensional data such as the hyperspectral imagery, the Hughes phenomenon [7] can be observed due to the relatively small amount of training data compared with the large number of bands. The Hughes phenomenon shows how the classification accuracy decreases as data dimension increases. To reduce computational complications and to increase image interpretation accuracy, it is desired that the thematic classification or modeling process be implemented in a reduced feature space rather than the original image space. This requires an image dimension reduction and a feature extraction process before image classifica-
II. BACKGROUND

Feature extraction and optimal band selection are the methods most commonly used for finding useful features in high-dimensional data. Both methods are employed to filter out noise or unwanted image bands while retaining important information in the image. The feature extraction method transforms the original image bands to a new feature space from which new features can be extracted [9]–[13]. The optimal band selection approach is sometimes referred to as feature selection [14], [15]. It involves identification of an optimal subset of the original image bands that best describes a specific application. To properly implement feature selection, two aspects must be considered. One is the search algorithm necessary to identify the important features, and the second is a target function for result evaluation. The search algorithm needs to provide an efficient way to find the most informative features with regard to certain target functions. The target functions generally depend on the application. It is thus not surprising to select different features from the same dataset for different applications.

A. Feature Extraction and Principal Component Analysis

There are many different approaches for hyperspectral image feature extraction. One commonly used feature extraction method is image arithmetic such as band ratio [vegetation index (VI)] [5] or band combination [9]. Decision boundary methods, projection pursuit (PP) [11], and PCA [12] are other approaches for feature extraction. The PCA approach can be viewed as a special case of PP as the projection index represents the data variance. Studies [13] have also been carried out to automate the feature extraction process by using a collection of predefined image processing operators.

PCA is a feature extraction process that first transforms the original image into a principal component (PC) image [5] through principal component transformation (PCT), and then extracts informative features from the principal component bands. For each pixel, every band of the PC image is a linear combination of the original bands from that same pixel. The transformation uses the global covariance or correlation matrix from the original image, whose image bands normally correlate with each other. After the transformation, every dimension is orthogonal to each other with no correlation among the PC bands. The total variance of the PC image equals the total variance in the original image, thus preserving the original data information after transformation. However, the first several bands in the PC image contain the majority of the variance in the original image. Most of the total variance from the original image is mapped to the first component with decreasing variance in the following bands. Image analysis can be implemented using features extracted from the variance ranked PC bands rather than from all image bands.

A standard principal component transformation (STD-PCT) uses all spectral bands for transformation. Generally, it is expected that variance related to a specific application be transformed as much as possible to the principal component band(s). However, given that STD-PCT preserves all the variance of the original bands, if variance from some bands is not associated with a certain application, the final results can be quite mis-
leading. It is more beneficial if only the bands with unique information are used rather than all the bands that have the common information. This suggests that PCT can be implemented over a different subset of the original bands, depending on the application. This alternative of doing PCA was defined as selective principal component analysis (SPCA) [16]–[19]. Had the band subset used in SPCA contained most of the original image variance for a specific application, SPCA would enable maximum information preservation for that application. Further, SPCA also provides a way to investigate both the PC image bands and the band subset for transformation. The band subset for SPCA would be useful for studying significant wavelength ranges for different applications.

In one study using the SPCA approach [16], pairs of the Landsat Thematic Mapper (TM) bands with high correlation from the original six bands were used for dimension reduction, while pairs of bands with low correlation were used for spectral contrast mapping. Low correlation bands were used for spectral contrast mapping because the degree of correlation between two bands is inversely related to the contrast amount. For example, high correlation bands have low contrast and vice versa. Results from the SPCA process were easier to visually interpret than results from the standard principal component analysis (STD-PCA) process. Four TM bands were used for a feature-oriented principal components selection (FPSC) method [17] and this approach was suggested for alteration mapping. In an effort to differentiate four geomorphological units from TM images [18], a pairwise SPCA approach indicated that the least correlated bands led to the most useful combinations. The two band combinations were TM 4 and TM 7, and TM 1 and TM 4. Another study [19] found that SPCA using TM bands 5 and 7 was more appropriate for alteration mapping and the second PC image band from this transformation also discriminated vegetated areas in drainage routes. An alternative method to SPCA could be viewed as block SPCA [12] for hyperspectral image processing. In this method, rather than selecting a subset of the original bands, the original bands were first grouped into several blocks with highly correlated bands together in each block. Each block of bands was then transformed using PCT separately. The eigenimage from each block was analyzed and feature extraction was carried out at this stage. The extracted features from each block were then compressed to form a new image for further processing. This process could be repeated until the desired dimension reduction ratio was reached.

B. Optimal Image Band Search With GA

SPCA can be viewed as a two-step dimensionality reduction and feature extraction method for high-dimensional data analysis. The first step is optimal band selection and the second step is PCA based on the selected bands. Previous studies showed the SPCA method has been used in applications utilizing TM imagery due to its small number of bands. The small number of bands made it acceptable to do an exhaustive search on all the band combinations. However, when the number of bands in an image is significantly large, as is the case with hyperspectral imagery, this exhaustive selection approach becomes impractical. Considering a hyperspectral image with 60 bands, the total number of combinations is $2^{60} - 1$. Implementing the above SPCA method also requires significant computational resources. For every band combination, the process involves calculation of the covariance matrix of the selected bands, PCT, and result evaluation. Obviously, an optimal search strategy is desired for this task.

GA is a collection of adaptive search methods based on the principles of the genetic evolutionary process. By using implicit parallelism, the GA optimization process can efficiently explore the search space without being trapped into local optima. GA is suitable for hyperspectral imagery, where multiple local optima are likely due to similarity and high correlation among neighboring image bands. The classic (or simple) GA [8] starts from an initial generation pool and evolves from one generation to the next. The generation pool stores many possible solutions, normally encoded into binary format, for the optimization problem. In each generation, each individual in the generation pool is evaluated by the evaluation functions (called fitness). During the reproduction process, two individuals are selected based on the fitness values to produce the next generation. In this process, feature crossover and mutation generally occur to simulate the natural evolution process. This procedure keeps running until the predefined number of generations is achieved or the stop criterion is met. The series of operations, population initialization, individual selection, crossover, and mutation are collectively called GA operators.

GA has previously been used for image segmentation [20] and pixel classification [21]. Many studies have also focused on using GA for hyperspectral image optimal band selection. A multilevel model [22] was used based on GA and rough set theory to do optimal band selection for hyperspectral image classification. In this approach, most of the redundant and irrelevant bands were removed from the original bands after the dimension reduction and the GA-based filter step. Entropy has been used in a GA optimal band search process [23]. By definition, entropy is a measure of the amount of information associated with the n-dimensional spectral vectors. Band search was first implemented in a material reflectance database and then evaluated in a Hyperspectral Digital Imagery Collection Experiment (HYDICE) image. Another study [24] used classification accuracy as the fitness function based on minimum distance classification. However, the above GA-based band selection methods only selected and used a subset of the image bands for image processing. No data transformation has been applied to the selected bands, preventing further exploration on the transformed image feature space. A related work [13] has used GA for both band and related image processor selection. Upon selecting one of the candidate image processors, such as spectral, spatial, spatio-spectral, logic, or thresholding operators, the selected bands were transformed into a new image space or scratch plane(s) for further investigation. One band was selected for most of the operators, but as many as 16 could be selected for some spectral angle operators.

III. SPCA

Mathematically, the SPCA procedure can be described as in the following.

$$PC_S = H_S \times DN_S$$ (1)
PC\(_S\) is the PC image of the selected bands; \(H_S\) is the eigenvectors for SPCA; and DN\(_S\) is the selected original bands. Theoretically, either the data covariance matrix or the correlation matrix can be considered for calculating the eigenvectors depending on the data measurement scale and unit. Because the image data scale did not have widely different range, the covariance matrix of the selected original bands was used.

IV. IMPLEMENTATION OF CLASSIC GA

This research used GA as the search algorithm for the above SPCA process. The overall implementation is thus collectively called a GA-based SPCA, or GA-SPCA. Fig. 2 is a schematic diagram on how to implement the GA-SPCA algorithm. To use GA in SPCA, each band combination is represented by a binary string, or “chromosome.” Each bit in the chromosome represents one band and is set to either one or zero. When doing SPCA, if the bit is set to one, this band is used in the transformation. If the bit is set to zero, this band is not used. The process that maps band combinations into chromosomes is called encoding. Each band combination is encoded into a chromosome sequentially based on the order of band number so that neighboring bands are also neighboring bits within the chromosome. This encoding schema assures that successful combinations are not easily broken up during the crossover operations for certain cases. For example, when searching significant wavelength regions, the representing bands are generally close to each other. The chromosome length is thus the total number of hyperspectral image bands based on the encoding schema. Each chromosome is also called an individual. Four GA operators (namely, initialization, selection, crossover, and mutation) were used in the study. Following is a brief description of these GA operators, with details in [8].

A. Population Initialization

GA needs an initial individual population to carry out parallel multidirectional searches in each generation. In the beginning of the algorithm, each chromosome bit is randomly set to either zero or one to represent a random start. The population of all individuals normally equals the chromosome length based on Goldberg’s rule of thumb [25].

B. Individual Selection

In every generation, each individual will be evaluated by a fitness function. In this research, each individual was first decoded into selected bands. The selected bands were transformed into their principal components after calculating the corresponded statistics. Individual fitness would be computed in the PC image space (usually over a subset of the PC bands—features to be extracted). The probability of selecting a specific individual can be calculated by using the individual’s fitness and the sum of the population fitness. The cumulative probability for each individual can be subsequently computed based on the individual’s probability. A roulette wheel selection approach will first generate a random number in the range of [0,1]. An individual will be picked based on the cumulative probability distribution and the random number. Individual selection is the most computationally expensive operator among the GA operators, as it needs to compute the fitness of all the individuals first.

C. Crossover

The crossover process defines how genes (chromosomes) from the parents have been passed to the offspring. The process is as follows. In each generation, once two individuals are selected as the parents, a gene from each parent is broken into several segments and recombined with gene segments from the other parent based on a predefined crossover probability. In this research, a two-point crossover method, which breaks the parent gene into three segments, was used and generally needed a high crossover probability [26]. After the crossover operation, every two parents will produce two children. The above selection and crossover process will continue to run in each generation until the number of children equals the population size. At the end of each generation, an elitism policy, wherein the best individual from the current generation is copied directly to the next generation, was also used for fast convergence.

D. Mutation

The mutation process simulates the natural disturbance during crossover. It is a bit-by-bit operation based on the mutation probability—called mutation rate. Mutation rate is generally selected based on the population size and other factors, such as selection method and with or without an elitism policy. The mutation operation follows immediately after the crossover operation.

The target function describes the relationship between images and ground reference. It can be viewed as the fitness function for GA. Here the “ground reference” is a simplification for “ground based measurement,” which represents the real field data. The fitness function varies based on different applications. For example, it can be classification accuracy for supervised classification, statistical distance for class statistics estimation, and correlation coefficient for modeling, etc. It is important to select an appropriate fitness function, as it will guide the search process based on how the PC bands are related to the ground reference data. When the data reduction and
feature extraction processes are completed, the original high-dimensional image will be reduced to an eigenimage with one or several principal component bands. Information regarding certain fitness functions in the original image bands will be retained in the eigenimage, with unrelated original bands removed. A properly designed fitness function would enable the algorithm to reduce the data dimension and simultaneously classify the image.

V. EXPERIMENT

The objectives of the experiment were to use hyperspectral images to separately detect corn nitrogen stress, estimate corn population, and identify corn hybrid, using the GA-SPCA algorithm. The experiment was conducted at the Agricultural Engineering Research Farm, Urbana, IL, during the 2001 growing season. The field was square with an area of 40,469 m² (10 acres) (Fig. 3). There is a 5.75-m elevation variation within the field. The field boundary was first surveyed using a dual frequency GPS receiver with centimeter level horizontal accuracy. The field was then divided into 640 small plots, 6.1 m × 8.2 m, with eight rows of corn planted in each plot, and buffer zones. There were three hybrids of corn; two in the plots and the third in the buffer zone. Corn plant population was initially controlled by varying seeding rate during planting. Four levels of population [49.4, 61.8, 74.1, 86.5 (1000 seeds/ha)] were used. Each population was planted in 160 plots. Both preplant and sidedress applications were used for nitrogen input. The application levels were 50, 101, 151, and 202 kg/ha, each in 64 plots. A precision variable nitrogen sprayer was used for variable nitrogen application. All the various inputs were spatially randomized.

There were eighteen ground control points (GCP), each a 40 × 40 cm white board, been set up for remote sensing image georectification. The GCPs were surveyed using the same dual frequency GPS receiver for boundary measurement. For easy collection of the ground reference data, all plots were marked with flags at each corner of the plot. Corn was planted in early May 2001. Corn plant population was manually counted in July and August. When counting the population, only the center two rows of each plot were used. Chlorophyll content, which is an indicator of plant nitrogen levels, was measured using a Minolta 502 SPAD meter (Minolta, Co. Ltd.). On July 16 (nine days after sidedress application), all the plots were sampled with the SPAD meter at the center location of each plot. When collecting the SPAD data, five data points within a one-meter circle at the central location of the plot were collected and averaged. This approach required measuring multiple corn plants at the same location.

Precision Aviation (Rantoul, IL) collected remote sensing imagery using a Cessna (T210M Centurion II) fixed wing aircraft. The hyperspectral image sensor system was the RDACS/H-3, which is a prism-grating pushbroom scanner developed by NASA/ITD Spectral Visions [27]. The image used in this study, taken on July 13, 2001, had 60 bands with a spectral resolution of 6 nm from 473–827 nm and a spatial resolution of 0.5 m. To calibrate the hyperspectral image, two additional sets of images were taken, a dark current image for sensor noise calculation, and a placard image with standard surface reflectance for illumination calibration. The hyperspectral image was preprocessed in a series of steps defined by NASA/ITD Spectral Visions, Inc. [28]. The raw image was first corrected for geometric distortion [29] along the in-track direction due to the characteristics of the pushbroom scanner. Before image georeferencing, the distortion corrected image was clipped to show only the study field and then rotated to align with the GCPs. A first-order polynomial warp function was used for registration, with the average rms error equal to 0.87 pixels over the GCPs.

A forward minimum noise fraction (MNF) and an inverse MNF transformation were used to remove noise in the image caused by the image sensor [30], [31]. The forward MNF transformation, which used the original image and the dark current image, transformed the original image into data space with one part had large eigenvalues and coherent eigenimages, and a complementary part had near-unity eigenvalues and noise-dominated images. The dark current image was used for noise covariance matrix calculation. To avoid the potential to remove signal when too few bands were used in the inverse MNF transformation, the eigenimages and eigenvalues were checked to determine the best spectral subset for removing noise and minimizing signal loss. The first ten eigenimages were used in the inverse MNF transformation.

Lastly, an illumination calibration was performed via an empirical line approach [31], utilizing the placard image. After image preprocessing, the hyperspectral image was in the Universal Transverse Mercator (UTM) coordinate system using North American Datum 1983 (NAD83). The final image size was 481 × 384 pixels with 60 bands. This final calibrated image was then used in the algorithm.

The three sets of ground reference data, namely SPAD readings for crop chlorophyll content level, corn plant population, and corn hybrid were used to evaluate the algorithm for corn nitrogen stress detection, corn population estimation, and corn hybrid identification, respectively. When computing fitness for each individual, the PC image was generated first. Reflectance data was averaged over each plot for each PC band. The reflectance data was then evaluated with two fitness functions for fitness computation as discussed below.

![Fig. 3. Experiment field overlay with plots (image was taken on 07/13/01).](Image)
A. Feature Extraction With GA-SPCA

The chromosome length and individual population were set to 60, based on the total band number of the hyperspectral image as discussed previously. Crossover probability and mutation rate was chosen to be 0.8 and 0.03, respectively, based on the previous explanation. The number of generations was set to 100 in order to see the overall performance of this algorithm.

A total of two fitness functions were studied in the experiment. Both functions used correlation coefficient between ground reference and the extracted features. Higher correlation coefficient would indicate better nitrogen stress detection, population estimation, and hybrid identification. The first fitness function used the maximum absolute correlation (linear Pearson correlation coefficient) between the PC image and ground reference data. To test the algorithm’s general ability for dimension reduction and feature extraction, all data from each ground reference dataset was used. No attempt was made to generate a separate training and validation dataset. In order to study the algorithm’s performance, all initial chromosome bits were set to one rather than a random initialization. Band reduction was thus started from using all original bands.

The second fitness function, which utilized a modeling approach, was designed to test the algorithm’s generalization ability to deal with the training-validation problem. For each dataset, the ground reference data was randomly and evenly split into training and validation data. During the training phase, a model was built using the training data. The model was later applied to the validation data during the validation phase for result evaluation. The model, which was a multiple linear regression model, used five PC image bands. The correlation between the ground reference and regression model was used as the fitness function. Details of the implementation were: after PCT for each band combination, the linear Pearson correlation coefficient between each PC band and ground reference data was computed; five PC bands with the largest absolute correlations were then picked to build a multiple regression model (2)

\[ Y = \text{const} + w_i \text{PC}_i \]  

where \( Y \) is the ground reference data; const is the constant in the regression model; \( w_i \) is the coefficient for each PC band; and \( \text{PC}_i \) is the ranked PC band with its absolute correlation decreased. This model was dynamic during the training process as it changed from one band combination to another. Both the training and validation \( R^2 \) for the regression model were recorded for result evaluation.

B. Comparison With STD-PCA

To compare with the GA-SPCA results, the above two fitness functions were also tested using the STD-PCA approach, which is available in a commercial software package [31]. For the second fitness function, the same training and validation data, which were generated in the above SPCA process, were used.

VI. RESULTS AND DISCUSSIONS

Fig. 4 describes some statistics of the original image bands and the STD-PCA results. Correlation coefficients between three ground reference data and the original image bands [Fig. 4(a)] and the PC image bands [Fig. 4(b)] are plotted. The STD-PCA process can increase correlations for some applications such as corn chlorophyll content measurement (for corn nitrogen stress detection) and corn hybrid identification. The maximum absolute correlation increased from 0.78 to 0.81 for chlorophyll content. The maximum absolute correlation for hybrid also increased from 0.22 to 0.37. However, corn
TABLE I

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>PCA Mode</th>
<th>Band Num.</th>
<th>PC Band Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPAD STD</td>
<td>60</td>
<td>0.29 -0.57 -0.07 0.74 0.81 0.53 -0.53 0.11 0.54 -0.06</td>
<td></td>
</tr>
<tr>
<td>SPAD SPCA</td>
<td>17</td>
<td>0.37 0.53 0.20 0.68 0.86 0.38 -0.26 -0.51 0.57 0.28</td>
<td></td>
</tr>
<tr>
<td>Population STD</td>
<td>60</td>
<td>0.47 0.29 0.29 -0.17 0.52 0.06 0.19 0.27 -0.55 -0.02</td>
<td></td>
</tr>
<tr>
<td>Population SPCA</td>
<td>29</td>
<td>0.52 -0.15 0.00 -0.24 0.75 -0.02 0.22 0.21 -0.46 -0.04</td>
<td></td>
</tr>
<tr>
<td>Hybrid PCA</td>
<td>60</td>
<td>0.06 -0.04 -0.20 -0.35 0.32 -0.07 -0.05 -0.03 -0.21 -0.37</td>
<td></td>
</tr>
<tr>
<td>Hybrid SPCA</td>
<td>38</td>
<td>0.07 -0.09 -0.22 -0.36 0.26 -0.08 -0.04 -0.15 -0.17 0.43</td>
<td></td>
</tr>
</tbody>
</table>

population estimation has the opposite effect. The maximum absolute correlation for the population decreased from 0.65 to 0.55. This indicates that some band variances that are transformed into the eigenimage are not accounted for in the population estimation. For the other two applications whose correlations increased, unwanted bands are likely included in the eigenimages. These bands need to be removed from the original bands for better analysis.

The first implementation of GA-SPCA used all ground reference data and the maximum absolute correlation fitness function in order to evaluate the algorithm’s performance. The total number of original bands (Fig. 5) used for PC reduced steadily and the corresponding maximum absolute correlation converged asymptotically at generations 50, 60, and 30 for corn hybrid identification, population estimation, and chlorophyll content measurement, respectively. The smallest numbers of bands were at generations 61, 60, and 87 with corresponding maximum absolute correlations of 0.43, 0.75, and 0.86, respectively. The reduction ratios were 36.7%, 51.7%, and 71.7%.

Results for both the STD-PCA and GA-SPCA approaches are given in Table I, with the correlation coefficients for the first ten PC bands listed. The relatively low correlations for hybrid identification with STD-PCA and SPCA indicate that the image bands are not sensitive to hybrid variations based on maximum absolute correlations. The other two ground reference data showed strong correlations with the image bands. With GA-SPCA’s two-step data reduction process, namely a band selection process that selected a subset of the original bands and a feature reduction process that transformed the selected bands into an orthogonal data space, data dimension was reduced to one with the maximum absolute correlation evaluation method. Correlations were increased for all three ground reference data using the GA-SPCA approach.

The following descriptions explain what accounts the increasing correlations. To find the relationship between the PC bands and any ground reference data (such as corn chlorophyll content measured by the SPAD meter for corn nitrogen stress detection), it is expected that most of the (nitrogen) variance from all the image bands has been transformed to one or several PC bands (here assume one PC band will have the most information of nitrogen). It would be expected that bands that do not contribute to nitrogen information would have coefficients close to zero. However, this is not true in a standard PC transformation (STD-PCT) using all the image bands. The reason is because all variance from the original image is transformed to the PC image, and most of the original image variance is due to factors other rather than nitrogen. Actually in this study, PC band 5 (with STD-PCT) has the highest correlation (0.81) with ground reference. When computing coefficients for transforming PC band 5, coefficients for unwanted bands will not be zero automatically because the solution for diagonalizing the original image covariance matrix is unique. In other words, transformation of nitrogen variance from the original image (with STD-PCT) is mixed with other
TABLE II
GA-SPCA RESULTS FOR MULTIPLE REGRESSION CORRELATION FITNESS FUNCTION, MODELS AND $R^2$ OF THE PC BANDS BASED ON SELECTED ORIGINAL IMAGE BANDS. PCI MEANS THE ith PC BAND

<table>
<thead>
<tr>
<th>Ground reference</th>
<th>PCA mode</th>
<th>Band Num.</th>
<th>Regression Model</th>
<th>Training $R^2$</th>
<th>Validation $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPAD</td>
<td>STD</td>
<td>60</td>
<td>$Y=46.3 - 2.5PC5 + 1.6PC4 - 0.3PC2 + 0.3PC9 - 0.3PC6$</td>
<td>0.75</td>
<td>0.70</td>
</tr>
<tr>
<td>SPAD</td>
<td>SPCA</td>
<td>25</td>
<td>$Y=46.0 - 0.12PC5 - 25.6PC8 - 0.7PC - 0.1PC3 + 0.1PC1$</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>Population</td>
<td>STD</td>
<td>60</td>
<td>$Y=21.4 + 3.6PC5 - 4.5PC9 + 0.3PC1 - 0.3PC3 + 2.6PC8$</td>
<td>0.64</td>
<td>0.58</td>
</tr>
<tr>
<td>Population</td>
<td>SPCA</td>
<td>22</td>
<td>$Y=15.8 + 0.8PC1 - 2.2PC3 + 1.46PC8 + 3.1PC5 + 11.7PC9$</td>
<td>0.72</td>
<td>0.67</td>
</tr>
<tr>
<td>Hybrid</td>
<td>PCA</td>
<td>60</td>
<td>$Y=1.8 - 0.1PC4 + 0.2PC5 - 2.3PC10 + 0.3PC9 - 0.01PC3$</td>
<td>0.24</td>
<td>0.19</td>
</tr>
<tr>
<td>Hybrid</td>
<td>SPCA</td>
<td>25</td>
<td>$Y=2.6 + 0.9PC5 - 4.5PC8 - 0.3PC3 - 4.2PC7 - 0.2PC2$</td>
<td>0.37</td>
<td>0.32</td>
</tr>
</tbody>
</table>

variances from the unwanted bands, and transformation of the nitrogen variance is not emphasized.

The use of selective principal component transformation (SPCT) can solve the above problem. For a specific application such as nitrogen stress detection, it is desirable to transform variance due to nitrogen to the PC image as much as possible. By using image-bands sensitive to nitrogen, transformation of nitrogen variance can be emphasized. The unwanted band coefficient is zero because it is not used in the transformation. In the final result, PC 5 (with 17-band SPCT) has the highest correlation with ground reference (0.86). A comparison of the eigenvector (transformation coefficients) of PC 5 for both STD-PCT and SPCT is plotted in Fig. 6. In SPCT, the selected image bands have larger coefficients (absolute value) than the corresponding image bands using the STD-PCT method, meaning those bands are more emphasized. The final results from both approaches are in Fig. 7, which is a scatter plot of the ground reference data and PC 5. Results from the STD-PCT method obviously have less relation with the ground reference data than from the SPCT method. This indicates an accuracy improvement in data analysis using GA-SPCA. The GA-SPCA method can be used to search the image bands sensitive to nitrogen.

The maximum correlation fitness function showed encouraging results for the GA-SPCA algorithm. However, it was applied to the whole ground reference dataset to obtain the results. In a more general situation, it is required to get results based on some training data. The second implementation, which used a multiple linear regression function and the training data, was for this purpose. The training dataset was generated randomly from the ground reference. In the training step, a dynamic multiple regression model was built for each band combination and the regression correlation was used as the fitness for the GA search.

Table II shows results of the second GA-SPCA implementation with the corresponding multiple correlation coefficient of determination ($R^2$), for both the training and validation data. Results from the STD-PCA approach are also listed for comparison. By using the GA-SPCA method with the multiple regression evaluation method, the data dimension was finally reduced to five orthogonal PC bands. The reduction ratios in the band selection stage were 58.3%, 63.3%, and 58.3%, respectively. Generally, PC bands used for building the model are all within the first nine PC bands. $R^2$ values, which are the portion of ground reference variability explained by the PC bands, were higher for both the training and validation data when comparing the GA-SPCA approach with the STD-PCA approach. $R^2$ values for hybrids had the largest increases. The results showed the algorithm’s generalization ability based on training data.

The selected bands for nitrogen stress detection along with the original bands are plotted in Fig. 8. Generally, more bands are selected for the multiple regression approach than the maximum absolute correlation approach. When considering the specific domain knowledge, such as the significant wavelength range for nitrogen stress detection, many studies have shown similar results as concluded in this study. It was found that reflectance measurements near 550 nm [32], the range of 630–690 nm, and 760–900 nm [33] were sensitive for detecting nitrogen stress in corn. Except for a cluster of bands across the blue and green wavelength range selected by the regression approach, both approaches selected band clusters in agreement with other studies. However, results from this study selected much more narrow and discrete wavelength ranges than the previously reported broadband results. This also proves that GA provides an efficient way to identify the optimal/suboptimal sensitive bands rather than an exhaustive search over all band combinations.

Considering the unique information that different bands have for different applications, the band combination results can be
used for further imaging sensor development. There are certain tradeoffs of doing band selections between spectral and spatial resolutions [34]. A reduced number of bands in the imagery can both reduce imaging sensor costs and increase image SNRs.

VII. CONCLUSION

The proposed supervised dimension reduction and feature extraction method, genetic-algorithm-based selective principal component analysis (GA-SPCA), is a practical way to extract hyperspectral image features for different remote sensing applications. The original image band number is first reduced in the band selection phase. The image dimension is subsequently reduced and features are extracted through a principal component analysis phase. A genetic algorithm is used to optimize the band selection process. By using this method, the original image bands can be reduced to one or several principal component image bands, while still retaining most of the application data information in the original image. Experimental results from three ground reference datasets were encouraging. By removing image bands that do not contribute to information extraction for a specific application, the result correlation improved. The GA-SPCA method thus can provide a standard approach for hyperspectral image dimension reduction and feature extraction, as well as provide useful information for imaging sensor development.

VIII. FUTURE RESEARCH DIRECTIONS

SPCA is based on the assumption that a subset of the original image bands represents the most important information within the original image space for a specific application. To successfully implement SPCA, both the search strategies and evaluation functions are indispensable. As a first implementation of utilizing an optimal band selection algorithm in SPCA, the GA-SPCA method generated satisfactory results. There are many other search algorithms that have been studied in the past for hyperspectral image optimal band selection. These algorithms can be considered as the immediate candidates for further SPCA investigation. The application-based evaluation function also plays an important role in the algorithm. It would not be surprising that a properly designed evaluation function would help to efficiently extract the most informative features for a specific application. As a more general consideration, the GA-SPCA method is one case of a hybrid method of feature selection and feature extraction. The main concept of this hybrid method is to do band selection and data transformation sequentially. It will consider both the significant wavelength range and the corresponding feature space for better data analysis. More studies are expected using this hybrid approach.

ACKNOWLEDGMENT

The use of trade names is only intended to provide specific information to the reader and does not constitute endorsement by the University of Illinois. The authors also would like to thank ITD Spectral Visions Midwest for coordinating flight and providing images, and Precision Aviation for data collection.

REFERENCES


Haibo Yao (S’03) received the B.S. degree in aeronautical engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1990, the M.S. degree in biomedical engineering from Sichuan University, Sichuan, China, in 1996, and the M.S. degree in agricultural engineering from the University of Georgia, Athens, in 1998. He is currently pursuing the Ph.D. degree at the University of Illinois, Urbana Champaign.

His research interests include remote sensing, pattern recognition, sensor integration for remote sensing and precision farming, and artificial intelligence.

Mr. Yao is a member of Alpha Epsilon Honor Society.

Lei Tian (M’96) received the Ph.D. degree in biological and agricultural engineering from the University of California, Davis.

He is currently the Director of the Illinois Laboratory for Agricultural Remote Sensing (ILARS), University of Illinois, Urbana. He coordinates research in remote sensing for precision farming, variable-rate chemical application equipment, and field sensor development for site-specific management (SSM). His research interests have been in the development of new-generation artificial intelligence equipment for agriculture: production systems, remote/local field sensing systems, advanced high-density data analysis technology, and image analysis and pattern recognition algorithms for outdoor biological objective detection.