University of KwaZulu-Natal
College of Agriculture, Engineering and Science

Investigation of Feature Extraction Algorithms and Techniques for Hyperspectral Images

by

Hannah Morenike Adebanjo

212562332

Supervisor:

Prof. Jules-Raymond Tapamo

in fulfillment of the academic requirements for the degree of

Doctor of Philosophy,

School of Engineering,

University of KwaZulu-Natal

Submitted August, 2017

EXAMINER'S COPY

Abstract

Hyperspectral images (HSIs) are remote-sensed images that are characterized by very high spatial and spectral dimensions and find applications, for example, in land cover classification, urban planning and management, security and food processing. Unlike conventional three bands RGB images, their high dimensional data space creates a challenge for traditional image processing techniques which are usually based on the assumption that there exists sufficient training samples in order to increase the likelihood of high classification accuracy. However, the high cost and difficulty of obtaining ground truth of hyperspectral data sets makes this assumption unrealistic and necessitates the introduction of alternative methods for their processing. Several techniques have been developed in the exploration of the rich spectral and spatial information in HSIs. Specifically, feature extraction (FE) techniques are introduced in the processing of HSIs as a necessary step before classification. They are aimed at transforming the high dimensional data of the HSI into one of a lower dimension while retaining as much spatial and/or spectral information as possible. In this research, we develop semi-supervised FE techniques which combine features of supervised and unsupervised techniques into a single framework for the processing of HSIs. Firstly, we developed a feature extraction algorithm known as Semi-Supervised Linear Embedding (SSLE) for the extraction of features in HSI. The algorithm combines supervised Linear Discriminant Analysis (LDA) and unsupervised Local Linear Embedding (LLE) to enhance class discrimination while also preserving the properties of classes of interest. The technique was developed based on the fact that LDA extracts features from HSIs by discriminating between classes of interest and it can only extract $C-1$ features provided there are C classes in the image by extracting features that are equivalent to the number of classes in the HSI. Experiments show that the SSLE algorithm overcomes the limitation of LDA and extracts features that are equivalent to

the number of classes in HSIs. Secondly, a graphical manifold dimension reduction (DR) algorithm known as Graph Clustered Discriminant Analysis (GCDA) is developed. The algorithm is developed to dynamically select labeled samples from the pool of available unlabeled samples in order to complement the few available label samples in HSIs. The selection is achieved by entwining K-means clustering with a semi-supervised manifold discriminant analysis. Using two HSI data sets, experimental results show that GCDA extracts features that are equivalent to the number of classes with high classification accuracy when compared with other state-of-the-art techniques. Furthermore, we develop a window-based partitioning approach to preserve the spatial properties of HSIs when their features are being extracted. In this approach, the HSI is partitioned along its spatial dimension into n windows and the covariance matrices of each window are computed. The covariance matrices of the windows are then merged into a single matrix through using the Kalman filtering approach so that the resulting covariance matrix may be used for dimension reduction. Experiments show that the windowing approach achieves high classification accuracy and preserves the spatial properties of HSIs. For the proposed feature extraction techniques, Support Vector Machine (SVM) and Neural Networks (NN) classification techniques are employed and their performances are compared for these two classifiers. The performances of all proposed FE techniques have also been shown to outperform other state-of-the-art approaches.

Declaration - Supervisor

As the candidate's supervisor, I agree to the submission of this dissertation

---

Prof. Jules-Raymond Tapamo

Declaration - Plaigiarism

I, Hannah Morenike Adebanjo, declare that

1. The research reported in this dissertation, except where otherwise indicated, is my original research.
2. This dissertation has not been submitted for any degree or examination at any other university.
3. This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   (a) Their words have been re-written but the general information attributed to them has been referenced
   (b) Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the References sections.

Hannah Morenike Adebanjo

Declaration - Publications

DETAILS OF CONTRIBUTION TO PUBLICATIONS that form part and/or include research presented in this dissertation

Publication 1:

Adebanjo, Hannah M., and Jules R. Tapamo. "Semi-supervised local feature extraction of hyperspectral images over urban areas." *In 2013 International Conference on Adaptive Science and Technology*, pp. 1-5. IEEE, 2013.

Publication 2:

Adebanjo, Hannah M., and Jules R. Tapamo. "Graph Embedded Discriminant Analysis for Extraction of Features in Hyperspectral Images." *International Journal on Advanced Intelligent Paradigms*, (In Press), 2016.

Publication 3:

Adebanjo, Hannah M., and Jules R. Tapamo. "Windowed linear feature Extraction for Hyperspectral Images Submitted to...

Hannah Morenike Adebanjo

Acknowledgments

To GOD be the glory great things HE hath done. I give thanks to my Creator and Saviour.

I could not have had a better supervisor than Prof Jules-Raymond Tapamo. Your insight, patience and fatherly love have helped in the completion of this research. Thanks for your guidance and for taking time to read through, correct and re-correct every work I submitted. Indeed, I have been able to stand on the shoulders of a giant.

Sometimes I am astray, sometimes I find a path, sometimes it makes no sense and sometimes it does; all these have been my experiences. It could not have been easier without my lovable husband: The best in the world. Thanks for being there for me. And to the most beautiful girl in the world, Victoria, I love you always.

Special thanks also go to my Dad and Mum for their emotional, financial and spiritual support. You are the best definition of what a parent should be. I keep learning from you everyday. To my four soldiers (siblings)- John, Joseph, Joel and Joshua, I love you beyond words. May you all be the best in your chosen field.

Lastly, I appreciate my research colleagues, all of whom we laugh, play, work and discuss research ideas. Special thanks also goes to my office mates, and to my friends in Church, for their kindness and words of encouragement. In actual fact, it is impossible to mention the names of my numerous friends, brothers and sisters who have all made an impact on my life academically, financially and

spiritually. Your unrelenting love have contributed to my continual happiness, progress and the completion of this research. May God bless you all.

Contents

List of Figures

List of Tables

Abbreviations

AA	**A**verage **A**ccuracy
AVIRIS	**A**irborne **V**isible **I**nfrared **I**maging **S**pectrometer
CASI	**C**ompact **A**irborne **S**pectrographic **I**mager
FE	**F**eature **E**xtraction
FLDA	**F**isher **L**inear **D**iscriminant **A**nalysis
GCDA	**G**raph **C**lustered **D**iscriminant **A**nalysis
HSI	**H**yperspectral **I**mage
HYDICE	**H**yperspectral **D**igital **I**magery **C**ollection
ICA	**I**ndependent **C**omponent **A**nalysis
LE	**L**aplacian **E**igenmaps
LE	**L**inear **E**mbedding
LLE	**L**ocal **L**inear **E**mbedding
LDA	**L**inear **D**iscriminant **A**nalysis
NWFE	**N**onparametric **W**eight **F**eature **E**xtraction
OA	**O**verall **A**ccuracy
PCA	**P**rincipal **C**omponent **A**nalysis
RBF	**R**adial **B**asis **F**unction
ROSIS	**R**eflective **O**ptive **S**ystem **I**maging **S**pectrometer
SDA	**S**emi-supervised **D**iscriminant **A**nalysis
SVM	**S**upport **V**ector **M**achine
SSLE	**S**emi-**s**upervised **L**ocal **Embedding**
WinPCA	**W**indowed **P**incipal **C**omponent **A**nalysis

To my Daughter - Victoria ...

Chapter 1

General Introduction

1.1 Introduction

In recent times, advances in remote sensing technology have led to an increased availability of hyperspectral images (HSI). Hyperspectral imaging, otherwise known as imaging spectroscopy, provides image cubes which are generally characterized by high spectral dimension and wide spatial resolution. The spectral dimension of HSI has made it possible to extract useful information from objects using their reflective properties obtained through sensors. Examples of Hyperspectral Airborne Sensors are Airborne Visible Infrared Imaging Spectrometer (AVIRIS), Hyperspectral Digital Imagery Collection (HYDICE) and Compact Airborne Spectrographic Imager (CASI). These sensors acquire each pixel of an object or scene in the form of a spectral vector which helps in differentiating between spectrally similar objects or materials. In Hyperspectral imagery, each pixel of an image is measured in a wide spectrum of narrow and contiguous bands providing a more detailed image for analysis. Figure 1.1 shows the spectral curves of the AVIRIS Indian Pine Image. In the processing of hyperspectral images, two main properties of the images are crucial in its feature extraction process. The first is the spectral property and

the other the spatial property. Hyperspectral images are comprised of hundreds of spectral bands in contrast to the conventional three-band images and multispectral images [1]. In the following section, a distinction between multispectral and hyperspectral images is presented.

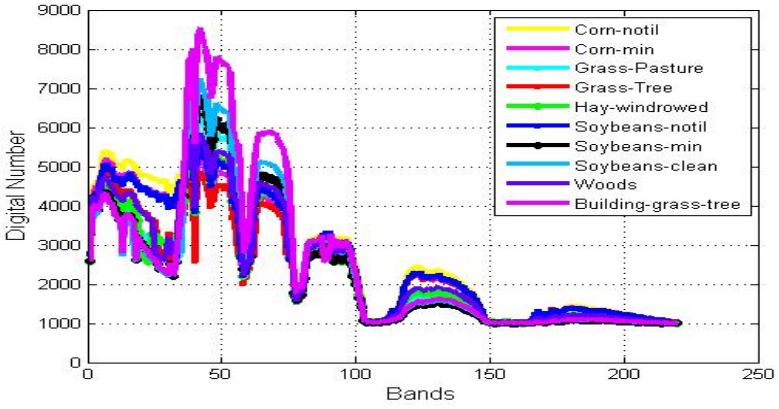

FIGURE 1.1: Spectral Curves of the AVIRIS Indian Pine Image

1.1.1 Categorization of Images Based on their Dimensionality

Remote sensed images can be classified based on the number of bands as either multispectral or hyperspectral. Multispectral images are images with few and separated bands. A typical example is the Landsat images. On the other hand, hyperspectral images contain hundreds of bands, and in addition, these bands are contiguous and narrow. Although, this distinction between multispectral and hyperspectral images is based on the number of spectral bands, they may also be differentiated based on their spectral properties depending on whether the spectral bands in the image are separated or contiguous [2].

1.1.2 Application Areas for Hyperspectral Images

HSIs have found many useful applications in different fields due mainly to the massive amount of information they embed. The enormity of their spatial and

spectral constituents are major attractions which have facilitated their application in agriculture, atmospheric studies, public safety and defence, land management, coastal and forest monitoring, geology and urban planning policy development. Specific applications in the highlighted areas are given in Table 1.1 below.

TABLE 1.1: Application of Hyperspectral Images

General Application Areas	Specific Application
Agriculture	• Estimation of soil parameters [3–5] • Pest and disease monitoring • Characterization of weeds and residues • Monitoring of crop yields
Coastal Monitoring	• Oil spillage monitoring [6–8] • Water quality monitoring [9] • Tidal monitoring [10]
Defence	• Target detection [11, 12] • Landmine detection [13]
Forestry	• Drought impact assessment [14] • Precision forestry [15] • Urban forestry monitoring [16]
Geology	• Mineral Mapping and exploration [17–19] • Characterization of rocks [20, 21]
Public Safety	• Wildfire monitoring [5, 22] • Flood management [23]
Urban Planning and Development	• Urban growth and monitoring [16] • Settlement population • Identification of surfaces and materials [24, 25] • Characterization of urban structures [26]

1.2 Problems in Hyperspectral Image Processing

As highlighted in the previous subsection, HSIs have found applications in many aspects of life. However, their intrinsic characteristics pose fundamental challenges to their analysis and processing. By addressing these challenges, HSIs will find more areas in which they may be applied in solving problems.

The most fundamental challenge in the processing of HSIs is related to their spectral dimensionality. They embed significantly large amounts of information on different land-based objects. These information are very useful in identifying objects or scenes, however there may be redundant information in these spectral bands.

Another challenge posed to the analysis and processing of HSIs is related to their classification. Their large spectral dimension and the availability of usually few labelled samples makes classification very difficult and render traditional classification algorithms incompetent or inapplicable. This challenge is also referred to as the Hughes phenomenon [27]. The result of the limited number of labelled samples and high dimensionality of HSI data has an indirect relationship on the accuracy of the classification process, i.e. the classification accuracy decreases with increase in the dimension of hyperspectral images.

Furthermore, hyperspectral images are unlike the conventional three-band images. Their spectral dimension which spans hundreds of bands also poses great difficulty and complexity in their processing and analysis. This challenge, also referred to as 'the curse of dimensionality' has attracted research efforts towards developing highly efficient and low-complexity feature extraction and classification algorithms for the processing of hyperspectral images of urban areas. Also, the problem of mixed samples in hyperspectral images makes it difficult to classify some images.

1.3 Overview and Motivation

As highlighted in the previous section, the high dimensionality of HSI necessitates the implementation of feature extraction before classification is done. Several methods for feature extraction in hyperspectral image processing have been developed. They may be classified into two categories as either supervised or unsupervised methods. Those methods that do not require the

use of class labels for the transformation of the high-dimensional HSI data space into one of a lower dimension are categorized as unsupervised. These methods are not affected by the limited availability of labelled samples. Common unsupervised feature extraction methods that have been widely used in hyperspectral image processing include Principal Component Analysis (PCA) [28], Minimum Noise Fraction (MNF) [29] and Independent Component Analysis (ICA) [30]. However, it has been observed that the use of these techniques results in the loss of discriminative information which would otherwise, have improved the accuracy of the classification algorithms. Therefore, supervised methods that use class labels or class information during computation are introduced.

Supervised feature extraction algorithms employ class labels for transforming the high-dimensional HSI data space into one of a lower dimension. Commonly implemented supervised algorithms include Non-parametric Weighted Feature Extraction (NPWFE) [31] and Linear Discriminant Analysis (LDA) [32, 33]. Other variants which are improvements on these popular methods have also been developed and implemented for the processing of HSIs [34]. The main advantage of these algorithms lies in their ability to discriminate between classes of interest. But they tend to perform poorly in the presence of few class labels. A new class of feature extraction methods known as semi-supervised method has also been conceived for the processing of HSIs.

A hybrid class of methods for feature extraction known as semi-supervised has also been developed from the consideration of the merits and demerits of supervised and unsupervised feature extraction methods. This class of algorithms thrives with impressive accuracy when the number of labelled samples are few and there is an abundance of unlabelled samples. In fact, they use information available from the few labelled samples as well as the many unlabelled samples to extract features from the HSI and thus, are able to discriminate between classes of interest. Several algorithms for feature extraction have been developed based on the semi-supervised approach. These

are Semi-supervised Local Fisher Discriminant Analysis (SELF) [35] and Semi-supervised Discriminant Analysis (SDA) [36]. However, there still exists the need for novel algorithms that are capable of improving the classification accuracy while taking full advantage of the available unlabeled samples. For instance, in [36], a semi-supervised feature extraction technique was proposed. This algorithm takes on the form of a modified LDA algorithm so that the manifold structure in the unlabeled samples can be preserved. Experiments show that SDA is able to exploit unlabelled samples for highly accurate feature extraction but because it employs tunable variables in the computation of its projection matrix, its complexity is greatly increased. SELF [35] is known for its ability to preserve neighbourhood structure in HSIs. It overcomes the demerit of supervised LDA by being able to extract C features from hyperspectral images unlike LDA, which extracts $C - 1$ features. However, when there is a limited number of training samples, SELF performs poorly. Thus, there is still the need to develop novel semi-supervised algorithms that will address the highlighted shortcomings of existing semi-supervised algorithms while also ensuring that the neighbourhood information of the image is preserved in its low-dimension feature spaces.

1.4 Thesis Objectives

In this thesis, we present an investigation into various feature extraction methods for hyperspectral images with the aim of developing and introducing novel methods. An examination of state-of-the-art algorithms for feature extraction in hyperspectral image processing underscores the necessity for feature extraction before classification. This is crucial in achieving improved classification accuracy, reduced computational complexity and processing time. Although feature extraction has a lot of merits, it is often difficult because of the limited number of training samples available and the large dimension of hyperspectral images. In this work, algorithms that are developed to overcome

the aforementioned limitations of feature extraction methods in hyperspectral image processing are presented. These algorithms are able to :

- fully utilize the abundant unlabelled samples during the feature extraction process,
- preserve the manifold structure of the hyperspectral image while still achieving high class discrimination,
- preserve the locality of the different classes in the reduced space using graph-based clustering, and
- stand in good performance in comparison with existing and related methods in terms of classification accuracy.

1.5 Thesis Contributions

Contributions of this thesis are as follows:

1. A semi-supervised feature extraction method for HSI feature extraction [37] has been developed. In this technique, the supervised Linear Discriminant Analysis (LDA) and the unsupervised Local Linear Embedding were integrated into a new semi-supervised algorithm namely the Semi-supervised Local Linear Embedding (SSLE). This approach enhances the class discriminating property while preserving the neighbourhood information of different classes of interest during the feature extraction process. This is to overcome the problem of few labelled samples in HSI and the curse of dimensionality.
2. Furthermore, a graph-based method for reducing the dimensionality of HSI has been introduced to fully utilize unlabeled samples. Two objectives were highlighted which are firstly to ensure that class separability is maximized

using the unlabelled samples and secondly to ensure that the manifold structure of the image is preserved. LDA which is able to discriminate between classes, makes use of labelled samples. In order to utilize the unlabeled samples in the proposed algorithm, we developed a framework that uses the unsupervised k-means in the semi-supevised approach. The idea was to first group the unlabelled samples into clusters so that labels from the clusters are used to extract the features in a semi-supervised approach.

3. The developed graph-based method was optimized in a way that enhanced the maximum discrimination between classes of interest in the feature extraction process, thereby increasing the expected accuracy when compared to fully supervised approaches.

4. A partitioned approach before the use of feature extraction methods has been introduced. This method uses the approach of dividing images into windows of varying sizes and uses a novel covariance addition method before the computation of projection matrices.

1.6 Thesis Outline

This dissertation is presented in six chapters. In Chapter Two, related works are reviewed and categorized based on their use of labelled samples for computation. The works reviewed therefore qualify as either supervised, unsupervised and semi-supervised methods.

Chapter Three presents the description of a novel semi-supervised method, known as Semi-supervised Local Linear Embedding (SSLE). This SSLE approach is based on the amalgamation of a supervised algorithm, Linear Discriminant Analysis (LDA) and an unsupervised method, Local Linear Embedding (LLE). The algorithm was developed to exploit the merits of LDA

and LLE. The LDA algorithm is capable of maximizing the between-class similarity and also minimize the within-class similarity while LLE, on the other hand, is efficient in preserving the HSI's data local neighbourhood pattern in the low-dimensional feature space. Finally, the experimental results of an investigation into the performance of the SSLE method on real hyperspectral data are presented and a discussion of the results follows.

In Chapter Four, a graph-based approach to dimensionality reduction which dynamically selects unlabelled samples using the k-means method is developed. The method, christened as Graph Clustered Discriminant Analysis (GCDA) has two objectives: the first is to maximize class separability through the use of unlabelled samples and secondly, to ensure that the manifold structure of the hyperspectral image is preserved. In the GCDA approach, the unlabelled samples of the hyperspectral image data are firstly clustered, thereafter, the resultant labels from all clusters are used to obtain the reduced data space in a semi-supervised manner. The classification is done using Neural Networks (NN) and Support Vector Machine (SVM). The results obtained from experiments performed to evaluate the performance of the GCDA algorithm on hyperspectral image of an urban area show that the spatial and spectral properties of the HSI are preserved in the reduced dimension obtained. GCDA's performance is also shown to outperform those of some existing dimensionality reduction methods.

A partitioning approach to dimensionality reduction which highlights the importance of employing varied window sizes is discussed in Chapter Five. The approach considers the partitioning of HSIs into smaller sizes (windows) before the computation of covariance matrices for each window. Experimental results show that the proposed approach is suitable for preserving the spatial property of hyperspectral images in the dimension reduction process.

The thesis concludes with a general discussion of the contributions made in the research in Chapter Six. It presents specific concluding remarks on the research

topics and perspectives on possible future developments of the work are presented.

Chapter 2

Research Background and Review

2.1 Introduction

In this chapter, a detailed review and background of methods and algorithms employed to process and analyze remote-sensed hyperspectral images (HSI) are expounded. The description of related feature extraction techniques that are used in transforming a hyperspectral image data from a high dimensional space into a lower dimensional one, is also presented. The general framework for FE in HSI processing is presented and we further categorize HSI into three main classes based on the usage of training samples in the computation of the FE process. Moreover, a review of various contributions from different works on feature extraction is also done.

2.2 Feature Extraction in Hyperspectral Imaging

As discussed in the previous chapter, hyperspectral images are unique because of their characteristic high spatial and spectral features. Thus, the hyperspectral image data is usually rich in spectrum. In order to analyze them, efficient dimensionality reduction techniques are used to transform the HSI data space from a high-dimensional one to one of a lower dimension. This transformation is achieved by feature reduction techniques. Feature reduction techniques are of two types: these are feature selection [38, 39] and feature extraction [37, 40]. The feature Selection (FS) process involves the search for and selection of an appropriate subset of features of an HSI. On the other hand, the feature extraction (FE) process concerns the utilization of the selected features subspace in transforming the highly-dimensional HSI data space into one with reduced dimensions. Therefore, FE methods reduce the computational complexity involved in the processing of HSI. It also aims to increase the classification accuracy that will be achieved as opposed to a classification process without prior FE.

The General FE transformation process may be reckoned as follows: Given a high-dimensional data $\{p\}_{i=1}^{N}, p_i \in R^d$, FE techniques aim to transform the high-dimensional data, into $\{q\}_{i=1}^{N}$ and $q_i \in R^p$ a low dimensional data : $p \leq d$. The variables d and p are the dimensions of the original data and transformed data respectively. Therefore, the objective of FE is to find a $d \times p$ projection matrix W, which can be mapped one-to-one from its high dimensional data space to the low-dimensional data space.

Several FE algorithms have been developed in the literature; they may be classified into any of three categories as supervised, unsupervised and semi–supervised. These classifications are considered in the following subsections.

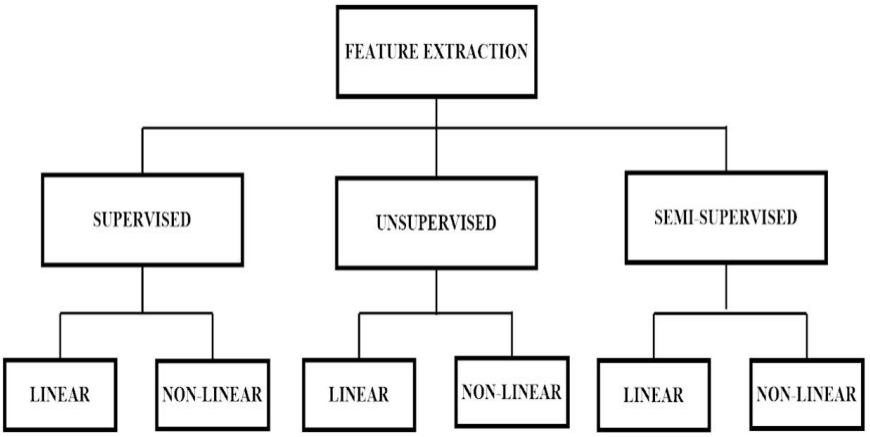

FIGURE 2.1: Feature Extraction in HSI

2.2.1 Supervised Feature Extraction

A feature extraction method is said to be supervised if it uses a-priori knowledge of the structure of data by means of training samples. Examples of supervised feature extraction algorithms are Non-parametric Weighted Feature Extraction (NWFE) [31] and Linear Discriminant Analysis (LDA) [41, 42]. Linear Discriminant Analysis (LDA) is a very popular supervised feature extraction technique and is efficient in discriminating between classes by maximizing the between-class scatter matrix while also minimizing the within-class scatter matrix [42]. However, due to the singularity of the within-class scatter matrix, LDA would fail when the sample size is small. In [43], a non-linear version of LDA known as Generalized Discriminant Analysis (GDA) is proposed. Both versions are capable of extracting maximum $C-1$ features, when the total number of classes is C. On the other hand, NWFE in its calculation of non-parametric scatter matrices, uses the weighted means and is also capable of extracting more than $C-1$ features [31, 44].

The authors in [33] developed and proposed the Regularized Discriminant Analysis (RLDA) in order to address the challenge of limited training samples during the classification of HSIs. By tuning a regularization parameter, the algorithm achieves high classification accuracy in the presence of few labelled samples but is impacted by high computational complexity. In [45], Cui et al.

proposed the Angular Discriminant Analysis (ADA) technique whose goal is to find the best subspace that separates classes in an angular manner that minimizes the ratio of the inner products of the between-class to that of the within-class. Furthermore, the authors introduced the Local Angular Discriminant Analysis (LADA), a technique that uses an affinity matrix for the preservation of data locality in the projected space. Like ADA, the different class samples are separated in an angular manner. However, the ADA and LADA algorithms give best performance when the method employed for classification of the image is one that is based on angular distance. These algorithms are used to improve the performance of classifiers using the cosine angle distance and sparse representation-based classification (SRC). Although these methods were reported to perform better when used with Nearest Neighbour classifier (NN), it was not compared against SVM. Moreover, LADA was intended to address scenarios where class-specific samples are situated across multiple clusters.

2.2.2 Unsupervised Feature Extraction

Unsupervised feature extraction methods on the other hand are mainly considered when the structure of the data is not fully known and there are few or no labelled samples to be used. Examples of unsupervised feature extraction algorithms are Principal Component Analysis (PCA), Independent Component Analysis (ICA), Singular Value Decomposition (SVD) and Projection Pursuit. PCA [46] seem to be the most popular linear unsupervised feature reduction technique and it uses an orthogonal transformation for its eigen value decomposition; it rearranges the bands in order of their variance, with the first band having the largest variance. Since PCA is unsupervised, it does not use class discrimination in its covariance matrix estimation. This is one major limitation in its application to HSI processing. The impact of this is the loss of important spatial information in its lower eigen vectors. Many variants and

extensions of PCA have been developed and applied to hyperspectral imagery to enable PCA preserve class discriminatory properties even in its unsupervised mode. A two-dimensional PCA was introduced in [47] for facial recognition. It computes covariance matrices on a two-dimensional image and has an increased recognition rate over the conventional PCA. Also, Segmented PCA which outperforms two-dimensional PCA when applied to hyperspectral images in [48] was originally introduced by Du et al. [49]. Segmented PCA aims at preserving the spatial properties of a hyperspectral image. It computes covariance matrices on grouped bands instead of on the whole lot of bands in conventional HSI.

2.2.3 Semi-Supervised Feature Extraction

The ability to discriminate between classes of interest is important and a major goal in the feature extraction of hyperspectral images while reducing the dimension. Most often, algorithms for discriminating between classes are majorly supervised. The presence of few or no labeled samples reduces the classification accuracy of these algorithms. This led to the evolution of semi-supervised algorithms.

The basic idea behind the development of semi-supervised algorithms is the calculation of the projection matrix by the use of labelled and unlabelled samples. Deng Cai [36] also introduced the Semi-supervised Discriminant Analysis (SDA) which finds projection with respect to the discriminant structure deduced from the labelled samples as well as the inherent geometrical structures deduced from both labelled and unlabelled samples. SDA uses the labelled samples achieve maximum separability among the different classes. On the other hand, it uses the unlabelled samples to derive the inherent geometric structure of the hyperspectral image data.

Semi-supervised methods which were used in [50] employ preserving the spatial and spectral properties of the data. Morphological profiles were used in [50] for

pre-processing. This approach makes use of manifold algorithms in preserving the spatial structures through the use of graph-based approaches [51]. Zhang et al. also used morphological filtering as preprocessing for watershed segmentation in order to preserve the contextual boundaries of HSI and further formed super-pixels for the objects from watershed segmentation.

A texture-based feature extraction was developed based on 3-D wavelet transform in [52]. The approcah used in the developed algorithm considers each patch of image as a cube which enhance the representation of the image both spectrally and spatially. However, the classification accuracy is dependent on the effective selection of window sizes.

In [53], another spectral and spatial method of the feature extraction of hyperspectral images is introduced, namely the Semi-supervised Discriminative Locally Enhanced Alignment (SDLEA). This method is an hybrid of the Discriminative Locality Alignment (DLA) [54] and is developed so that the spatial and spectral properties of the hyperspectral image are preserved in the reduced space.

2.3 Related Works

As earlier mentioned, algorithms developed for the feature extraction of HSIs have different properties but the overall aim is to preserve the utmost amount of information in the reduced dimensional space. In this section we review works done in the feature extraction of HSIs. Finding a connection between samples of labelled and unlabelled HSI data is crucial towards overcoming the problem of a small training sample size. In [50], a semi-supervised graph was developed which maximizes class discrimination and preserves the local neighbourhood information by combining labelled and unlabelled samples. It connects labelled samples according to their label information and unlabelled samples by their nearest neighbourhood information. Neighbourhood

information preservation has been a major backbone in the development of semi-supervised algorithms. Cai et al. [36] also introduced Semi-supervised Discriminant Analysis (SDA) to mitigate the availability of few training samples in LDA. In SDA, the discrimination is achieved using labelled data points while the locality of points within the classes is preserved by the unlabelled samples. However, like LDA, SDA can extract only $C-1$ features from the image. The authors in [35] attempted to improve the number of extracted features in the processing of HSIs and proposed the Semi-supervised Local Fisher Discriminant Analysis (SELF), a method that combines LDA and LPP. The SELF method shows that the number of features that are extracted from the image is synonymous with the number of classes of interest. However, both SDA and SELF have tunable parameters in the computation of their projection matrices and this may affect the classification accuracy of the extracted features. Liao et al. [55] also proposed a semi-supervised improvement of the LDA method of feature extraction by combining LDA with a number of unsupervised local linear feature extraction methods such as locality preserving projection (LPP), neighbourhood preserving embedding (NPE) and linear local tangent space alignment (LLTSA). In the proposed framework, the data set is initially separated into labelled and unlabelled sets; then, LDA is applied to the labelled sampled while the unsupervised methods are applied to the unlabelled sampled. The proposed method further exploits the benefits of supervised and unsupervised methods through a non-linear combination of both and overcomes the need for the optimization of tuning parameters.

Patch alignment techniques have also been used in major algorithms used in discriminating features in HSI. In [56], a semi-supervised discriminative locally enhance alignment (SDLEA) was used in hyperspectral image processing. SDLEA was built upon the limitation of Discriminative Locality Alignment (DLA) [54] which assumes that a fixed number of neighbouring points are in each patch. It overcomes this by using multi-segmentation in selecting

unlabelled samples. Tensor discriminative locality Alignment (TDLA) [57] also used the patch alignment framework. TDLA uses multilinear algebra with supervised manifold learning algorithms. Although, TDLA preserves the spatial property of the data, its supervised nature makes it difficult for it to be used in the presence of few or no labelled samples. In [58], it was discussed that the spectral and spatial information at each pixel is integrated in which an explicit and nonlinear mapping is done between the unlabelled data and the feature space. All these algorithms aim at preserving the spectral and spatial properties of each pixel in hyperspectral images while their features are being extracted.

In [59], another method known as Neighbourhood Preserving Orthogonal PNMF (NPOPNMF) for feature extraction in the classification of hyperspectral images is proposed. In NPOPNMF, it is assumed that all pixels (data points) of an HSI can be individually represented as a linear association of its neighbouring pixels towards overcoming the Euclidean limitation of PNMF. A unique feature of the NPOPNMF method is its capability of being operated in dual modes either as a supervised or unsupervised method. The modes are determined based on how the adjacency graphs are constructed. The limitation of the traditional Negative Matrix Factorization (NMF) of ignoring the labels of data points was overcome in NPOPNMF. A novel supervised NMF algorithm to improve the discriminative ability of the new representation by using the class labels. Using the class labels, data sample pairs are separated into within-class pairs and between-class pairs. The discriminative ability was further improved by minimizing the maximum distance of the within-class pairs in the new NMF space, and meanwhile maximizing the minimum distance of the between-class pairs.

In [60], the authors propose another method of dimensionality reduction known as Nearest Feature Line Embedding (NFLE) transformation. With focus on the discriminant analysis phase, the NFL measurement is integrated with the transformation phase instead of the matching phase. Furthermore, by giving

simultaneous consideration to the separability of classes, preservation of the neighbourhood structure and NFL measurements, NFLE achieves efficient and discriminating transformation in the eigenspaces for land cover classification. Tan et al. [61] also proposed a semi—supervised feature extraction method based on a block-sparse graph for discriminant analysis in hyperspectral image processing. While it is aimed at overcoming the intrinsic challenge of few label samples in HSIs, the proposed method devises the inclusion of unlabelled samples with labelled samples when graphs are being constructed. But the selection of unlabelled samples is done using sparse and collaborative graph representations. The results show that this semi-supervised block-sparse graph dimensionality reduction method is capable of significantly outperforming a supervised method, given the limited availability of training samples.

As indicated earlier, both the spectral and spatial features of hyperspectral images are important in effectively discriminating classes of interest. An example of such spatial algorithm is morphological profiles [62, 63] which have gained wide usage in the extraction of features for HSIs.

Also, another method being used to discriminate between the spectral and spatial features of HSI is the graph methods. In [64], the graph discrimination method was used which takes into consideration the number of pixels in the neighbourhood. In [65], the authors propose a supervised graphical method based on determinantal point process which uses a fully probabilistic model to select the representative bands and to preserve the relevant information in the original spectral bands. After the band selection process, multiple Laplacian Eigenmaps are performed on the selected bands. These multiple Laplacian Eigenmaps are defined by encoding the spatial-spectral proximity on each band and then collectively on the selected bands. Furthermore, an unsupervised manifold feature extraction was proposed by Gan et al. [66]. This method uses multi-structure based feature learning approach. The feature learning method was constructed using the sparse graph and hypergraph as well as the local linear structure for the dimensionality reduction process. This method was

tested using the Salina and Pavia University hyperspectral images and shows improved accuracies. However, this multi-structure feature learning method uses more running time with increased computational cost when compared with other feature extraction methods. There is always a trade-off between accuracy and computational complexity. To reduce the computational complexity feature extraction is often performed on a PCA transformed space. In [67], PCA was first computed on HSI thereafter Gabor filters was performed on the PCA transformed feature space. Moreover a PCA-Edge Preserving Feature (PCA-EPF) was introduced in [68] to overcome the limitation of EPF based methods which has been found to decrease the spectral differences of similar objects of variable ratios. Also, Matrix Discriminant Analysis (MDA) [69] is posed with the difficulty of optimal choice of scale which led to the development of robust-MDA to overcome the shortcomings of MDA.

2.4 Conclusion

This chapter presents a background to the subject of feature extraction in hyperspectral image (HSI) processing. It highlights the essence of the feature extraction process in the processing of hyperspectral images. The feature extraction process, which precedes classification, is essential in the processing of HSIs and is aimed at achieving a reduction in the dimension of HSI data for easier processing. The chapter begins with a discussion on the two stages involved in the transformation of HSI data from a high-dimensional feature space to one of lower dimension. These include feature selection and feature extraction. However, more emphasis is laid on feature extraction because this research is based on it. Secondly, the chapter also discusses the three classes of feature extraction algorithms namely supervised, unsupervised and semi-supervised algorithms and a review of existing works related to each class. The supervised algorithms use available training samples obtained from the HSI data set for the selection of features. The unsupervised algorithms for

feature extraction are developed to select features when none or very few labeled samples are available. However, the lack or limited number of training samples makes it difficult to discriminate between classes of interest during the feature extraction process. Therefore, semi-supervised algorithms have been developed to exploit the merits of supervised and unsupervised algorithms and to ameliorate challenges related to the implementation of these two classes of algorithms. Finally, other approaches to feature extraction that are developed in the literature were reviewed noting their merits and challenges.

Chapter 3

Semi-Supervised Local Linear Embedding

3.1 Introduction

This chapter presents a semi-supervised method of feature extraction in the processing of hyperspectral images. It introduces a hybrid feature extraction method using the semi-supervised approach. The proposed technique combines the supervised Linear Discriminant Analysis (LDA) with unsupervised Local Linear Embedding (LLE) techniques, and exploits the merits of both techniques to achieve comparatively higher classification accuracy.

LDA and LLE have been widely used for feature extraction in the processing of HSI. While LDA is popular for its class-wise discriminatory capability, the LLE technique is known for its ability to preserve manifold properties during the feature extraction process. However, both techniques have some drawbacks. In LDA, if there are C number of features, then the number of classes that may be extracted is limited to $C - 1$. In addition, it is unable to discover the spatial property of the image when there are only a few labelled samples. Besides, the LLE technique is computationally expensive because it is a nonlinear method.

Therefore, the algorithm proposed in this chapter is developed to exploit the merits of LDA and LLE.

We discuss LDA and LLE and the detailed analysis of the developed semi-supervised local linear embedding is presented in section 3.2. In section 3.3, the algorithm of the proposed feature extraction method is presented. Section 3.4 presents a discussion on the experimental results of the performance of the proposed method on real hyperspectral images. The chapter concludes in section 3.5 with a summary of the highlights of the proposed method.

3.2 Background of Related Feature Extraction Methods

3.2.1 Linear Discriminant Analysis

The Fisher Local Discriminant Analysis proposed in [70] is an algorithm that has found extensive application for feature extraction in hyperspectral image processing. Being a supervised method, this algorithm projects all data points in a hyperspectral image data space into a new space of lower dimension so that the separability between-class is maximized and the variability within-class minimized. The Fisher LDA algorithm may be set up as the minimization of an objective function as follows:

$$W_{LDA} = \underset{w}{\text{argmax}} \frac{W^T S_b W}{W^T S_w W} \tag{3.1}$$

where S_b and S_w respectively denote the scatter matrices of between-classes and the within-classes.

In reducing the dimensionality of the image from a high dimensional space to a lower dimensional one, we denote $\Re^d$ as the d- dimensional vector spaces and $\Re^p$

the p- dimensional vector spaces such that $p \leq d$. Let $\{C_k\}_{k=1}^{c}$ denote C classes of interest where $C_k = \{x_1^k, x_2^k, ..., x_{N_k}^k\}$ and $\{x_j^k\}_{j=1}^{N_k}$ is the k_{th} class and contains N_k patterns and the j_{th} pattern in class C_k denoted by

$x_j^k = \{x_{1j}^k, x_{2j}^k, ...x_{dj}^k\}^T$ is a d- dimensional vector in the space $\Re^d$ Let $N = N_1 + N_2 + \cdots + N_c$ be the total number of training patterns.

From Fisher Local Discriminant Analysis [70], we can form the total, between and within scatter matrix. Let $\mu = (1/N)\sum_{k=1}^{N}\sum_{j=1}^{N_k} x_j^k$ be the global mean and $\mu_k = (1/N_k)\sum_{j=1}^{N_k} x_j^k$ be the mean of class C_k Then

$$S_T = \sum_{k=1}^{C}\sum_{j=1}^{N_k}(x_j^k - \mu)(x_j^k - \mu)^T \tag{3.2}$$

$$S_w = \sum_{k=1}^{C}\sum_{j=1}^{N_k}(x_j^k - \mu_k)(x_j^k - \mu_k)^T \tag{3.3}$$

$$S_b = \sum_{k=1}^{C} N_k(\mu_k - \mu)(\mu_k - \mu)^T \tag{3.4}$$

we can say

$$S_T = S_b + S_w \tag{3.5}$$

Assume that $\Xi = \{x_1, x_2...x_N\} = \{x_j^k\}_{j=1,k=1}^{N_k,C}$ are all data training samples. The Fisher's Discriminant Analysis algorithm finds a weight matrix $W = [w_1, w_2...w_{c-1}]$ of $d \times (C-1)$ dimension, which projects all data samples $x \in \Xi$ in an $\Re^d$ space into y in a low dimensional feature space $\Re^p; y \in \Re^p$ such that all projected data samples y's yield the best possible class separability by

$$y = w^t X \tag{3.6}$$

with

$$y_k = w_k^t X : 1 \leq k \leq C-1 \tag{3.7}$$

where w_k is the k_{th} column vector with dimensionality $d \times 1$ in w and $y = (y_1, y_2...y_N)^t$. Similarly, using equations 3.3 and 3.4, we can define within-class

and between-class scatter matrices for the projected samples y given by $y = w^t x$ as follows:

$$\bar{S}_w = \sum_{k=1}^{C}\sum_{j=1}^{N_k}(x_j^k - \bar{\mu_k})(x_j^k - \bar{\mu_k})^T \tag{3.8}$$

$$\bar{S}_b = \sum_{k=1}^{C} N_k(\bar{\mu_k} - \bar{\mu})(\bar{\mu_k} - \bar{\mu})^T \tag{3.9}$$

where $\bar{\mu} = (1/N)\sum_{k=1}^{N}\sum_{j=1}^{N_k} y_j^k$ and $\bar{\mu_k} = \sum_{j=1}^{N_k} y_j^k$ substituting equations 3.3, 3.4, 3.6, 3.7 into 3.8 and 3.9 results in

$$\bar{S}_w = W^T S_w W \tag{3.10}$$

$$\bar{S}_b = W^T S_b W \tag{3.11}$$

In terms of class separability, an optimal linear transformation matrix, W, may be found using the Fisher's Discriminant Function Ratio, otherwise known as Raleigh's Quotient. The Raleigh's Quotient is found as the ratio of the between-class scatter matrix to the within-class scatter matrix as follows:

$$J(W) = \frac{\mid \bar{S_B} \mid}{\mid \bar{S_W} \mid} = \frac{\mid W^T S_B W \mid}{\mid W^T S_W W \mid} \tag{3.12}$$

where the $\mid . \mid$ operator denotes a matrix' determinant. An optimal solution to $J(W)$ given by

$$W^*_{d\times(C-1)} = [W_1^*, W_2^* ... W_{C-1}^*]_{d\times(C-1)} \tag{3.13}$$

may be obtained through the solution of a generalized eigen value problem defined as follows:

$$S_B W_K^* = \lambda_K S_W W_K^* \tag{3.14}$$

where W_K^* corresponds to the eigen value λ_k. The C-1 eigen vectors, $\{W_k\}_{k=1}^{C-1}$, form a set of Fisher's Linear Discriminant functions which may be applied in equation 3.7 as

$$y_k = (W_k^*)_x^t : 1 \leq k \leq C-1 \tag{3.15}$$

It is worthy of note that although there are C classes, only C-1 eigen values denoted by$\{\lambda_i\}_{i=1}^{C-1}$ are non-zeros. Each eigen value λ_i generates its own eigen vectors W_j^*. Using these eigen vectors, $\{W_i^*\}_{i=1}^{C-1}$ and equations 3.6, 3.13 and 3.15, an optimal linear transformation T^* that is based on Fisher's Discriminant Analysis is derived as

$$y = T^*(x) = (W^*_{d\times(C-1)}) \tag{3.16}$$

3.2.2 Local Linear Embedding

The Local Linear Embedding (LLE) [71] algorithm is an unsupervised dimensional reduction algorithm that seeks to compute a low-dimensional embedding by assuming that nearby points in the high-dimensional space will remain nearby and co-located to each other in the low-dimensional space. The algorithm is described as local because only neighbours can contribute to each reconstruction. On the other hand, it is described as linear because reconstructions are confined to linear sub-spaces in a non-linear dimension. Assuming that $X = [x_1, x_2..., x_n] \in R^{p\times n}$, consists of n data points $x_i, \{i = 1, 2, 3, ...n\}$, each with dimensionality p. Then, each data point x_i, lies with its neighbours, in close proximity to a locally-linear manifold, which dictates the manner in which the weight coefficients $\mathbf{W}_{ij}$ are constructed. The LLE algorithm has been employed in the processing of hyperspectral image [72] and its implementation may be summarized in the following three stages:

1. An adjacency graph of i nodes is constructed such that each data point x_i corresponds to the ith node of the graph. The graph is constructed using the method of K-nearest neighbour in that an edge is constructed between any two nodes i and j, if a data point x_j is one of K-nearest neighbours of x_i.

2. Calculate the weight W_{ij} with which each data point may be reconstructed from its neighbour while the cost is minimized by constrained linear fits. The reconstruction weights are computed as follows:

$$\xi(W) = \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{k} W_{ij} x_j \right\|^2 \tag{3.17}$$

under two constraints; $\sum_{j=1}^{k} W_{ij} = 1$, if x_i and $x_j$are neighbours while $W_{ij} = 0$ if x_i and x_j are not neighbours. Where $\| \ . \ \|_2$ denotes L_2 norm of a vector and W is the weight matrix which describes the local neighbourhood relationship between the data points.

3. Compute the vectors best reconstructed by $[W_{ij}]$ minimizing the quadratic form by its bottom non-zero eigen vectors. Embedded coordinates y_i i = 1,2,...,N are computed by minimizing the following embedding cost function for the fixed weights:

$$\begin{aligned} \Phi(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{n} W_{ij} y_j \right\|^2 \\ = tr(YMY^T) \end{aligned} \tag{3.18}$$

subject to (1/N). $\sum_{i=1}^{N} y_i y_i^T = I_{d\times d}$ and $\sum_{i=1}^{N} y_i = 0$ where y_i is the coordinate of the data point i.e the output matrix. x_i in the low dimensional embedding. $M = (I - W)^T(I - W)$ is a sparse, symmetric, and semi-positive definite matrix. The tr denotes the operation of trace and I = $diag$(1,1,...1).

LLE has a computational complexity of $O(dn^2), O(dnK^3)$ and $O(rn^2)$ for the three steps respectively where d is the input dimensionality, K is the number of nearest neighbours, n is the number of data points and r the output dimensionality.

After a careful survey, it has been established that LDA maximizes the between-class similarity while minimizing the within-class similarity. LLE on the other hand preserves the local neighborhood of the data in the low dimension feature space.

3.3 Semi-supervised Local Embedding

In this section we explain the development of the proposed algorithm which exploits the advantages of the LDA and LLE to yield a new approach known as Semi-supervised Local Embedding (SSLE). It magnifies the advantages of LDA and LLE while compensating for their disadvantages. With LDA being supervised and LLE unsupervised and non-linear, the hybrid algorithm produces a semi-supervised non-linear algorithm.

Let $\{p_i\}_{i=1}^{N}, p_i \in R^d$ denote high-dimensional data, $\{q_i\}_{i=1}^{N}$ and $q_i \in R^r$ the low dimensional data $r \leq d$. In the proposed method, d is the dimensionality of the original data and r is the dimensionality of the extracted features. The goal of this feature extraction method is to find a $d \times r$ projection matrix W, which can be mapped one -to-one to its high dimensional data space. From equation 3.9, by centering the data i.e we assume that $\mu = 0$, we have

$$\begin{aligned} S_B &= \sum_{k=1}^{C} N_k(\mu_k)(\mu_k)^T \\ &= \sum_{k=1}^{C} N_k(\frac{1}{N_k}\sum_{i=1}^{n_k} x_j^k)(\frac{1}{N_k}\sum_{i=1}^{N_k} x_j^k)^T \end{aligned} \tag{3.19}$$

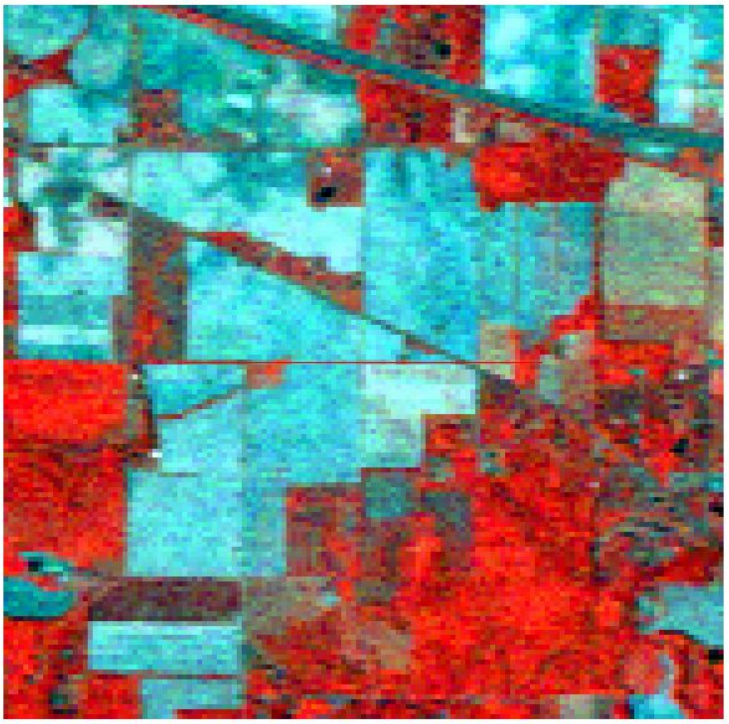

FIGURE 3.1: 3 band Display of the AVIRIS Indian Pines Image

where P^k is a $N_k \times N_k$ matrix with all the elements equal to $\frac{1}{N_k}$. We define a $N \times N$ matrix $P_{N\times N}$ as:

$$P_{N\times N} = \begin{bmatrix} P^{(1)} & 0 & \dots & 0 \\ 0 & P^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P^{(n)} \end{bmatrix} \tag{3.20}$$

we have

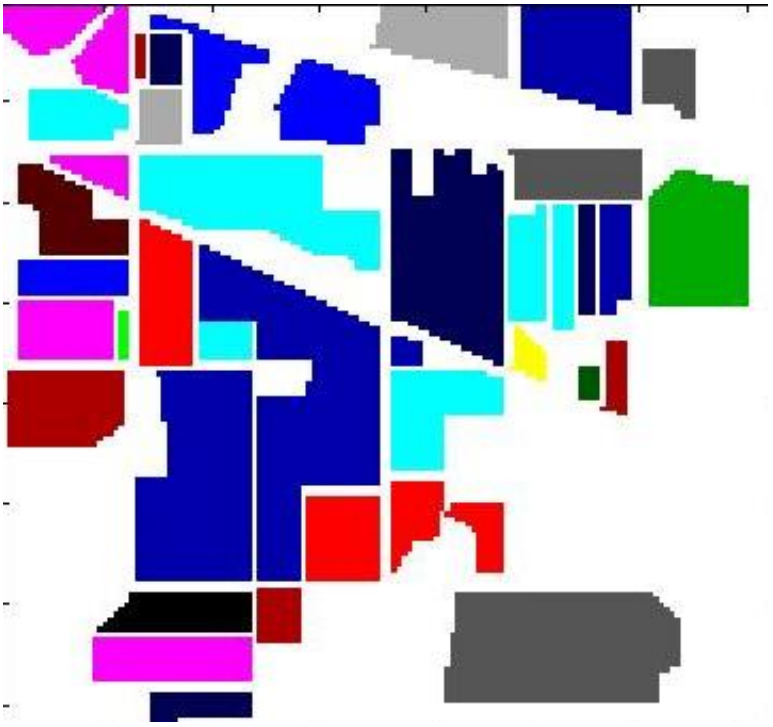

FIGURE 3.2: Ground Truth of the AVIRIS Image showing the Sixteen Classes

$$\begin{aligned} S_B &= \sum_{k=1}^{C} X^{(k)} P^{(k)} (X^{(k)})^T \\ &= X_{labeled} P_{n\times n} (X_{labeled})^T \end{aligned} \tag{3.21}$$

By subtracting the between-class scatter S_b, S_w can be obtained

$$\begin{aligned} S_w &= X_{labeled}(X_{labeled})^T - X_{labeled} P_{n\times n} (X_{labeled})^T \\ &= X_{labeled}(I_{n\times n} - P_{n\times n})(X_{labeled})^T \end{aligned} \tag{3.22}$$

A non-linear dimensionality reduction method, LLE is used in the within-class scatter matrix of LDA. This was in order that class discrimination may be inferred from labelled samples while the local embedding from the labelled and unlabelled samples is preserved. Also, the symmetries of locally linear reconstructions which may be reduced to a sparse eigenvalue problem are used to compute the LLE

mappings. The embedding can be computed from equations 3.18 and 3.21, using the within-class scatter matrix. This gives

$$W^r_{SSLE} = \underset{w}{\operatorname{argmax}} \frac{W^T X_{labeled} P_{n\times n} (X_{labeled})^T W}{W S_w W^T} \tag{3.23}$$

Where

$$S_w = X_{total}(I_{n\times n} - P_{n\times n})(X_{total})^T + (I - W)(I - W)^T \tag{3.24}$$

Noting that hyperspectral images are characterized by few training samples and that LDA, a supervised method uses labelled training samples, over-fitting may occur during the embedding. This is avoided by using the unlabelled training samples in LLE.

Furthermore, the computational complexity of LLE is reduced by introducing a pre-processing step in which PCA is applied to the data in order to reduce its overall dimension. Thereafter, the principal components (PCs) obtained are substituted for the overall 200 data dimensions after the removal of the 20 water absorption bands. The various stages of the proposed feature extraction method are summarized in Algorithm 1

Algorithm 1: Computation of the SSLE matrix

Input : $\{p\}_{i=1}^{n}, p_i \in R^d$ Labeled samples
$\{y\}_{i=1}^{n}, y_i \in C$ Corresponding labels
$\{p\}_{i=n+1}^{N}, p_i \in R^d$ unlabeled samples

Output: $\{p\}_{i=1}^{n}, p_i \in R^d$ Lower dimension projected matrix

1 **Extract** $w_{PCA} \in \Re^q$ from $\{p\}_{i=1}^{N}$ using PCA. **Divide** the samples into training and test samples.

2 **Extract** features (W^r_{SSLE}) using equation 3.23.

3 **Derive** from W^r_{SSLE} the matrix W^s_{SSLE} where $s < r$.

TABLE 3.1: Information classes of selected samples of AVIRIS data

Class Identifier	Number of Samples
C1	1434
C2	834
C3	497
C4	747
C5	489
C6	968
C7	2468
C8	614
C9	1294

TABLE 3.2: Classification results when projection matrix is trained using 20% labelled samples per class

Class Name	PCA(%)	LDA(%)	LLE(%)	SSLE (%)
C1	66.38	72.42	65.48	78.02
C2	68.97	71.37	69.42	72.54
C3	82.81	90.18	85.86	84.91
C4	93.91	99.11	76.89	95.48
C5	98.41	99.55	82.40	97.01
C6	70.76	67.43	69.99	86.34
C7	78.04	79.78	85.40	91.58
C8	53.71	87.16	78.61	96.01
C9	90.73	94.16	79.99	98.67
ACA	78.19	84.57	77.11	88.95

3.4 Experimental Results and Discussion

The performance of the proposed algorithm was investigated using the AVIRIS Indian Pines data set that was collected over an agricultural area of northern Indiana in 1992 [73]. The image is characterized by 145× 145 pixels, 220 spectral bands ranging from 400 to 2500 nm and 16 ground-truth classes. This data set is unique in that it portrays a scenario that is quite challenging for land-cover classification. Also, the land area is mainly cultivated with corn and soybeans which are in the early stages of growth, and has about 5% canopy cover.

A three band display of the image is shown in Figure 3.1 and Figure 3.2 shows the ground truth of the image. Nine classes are considered in evaluating the

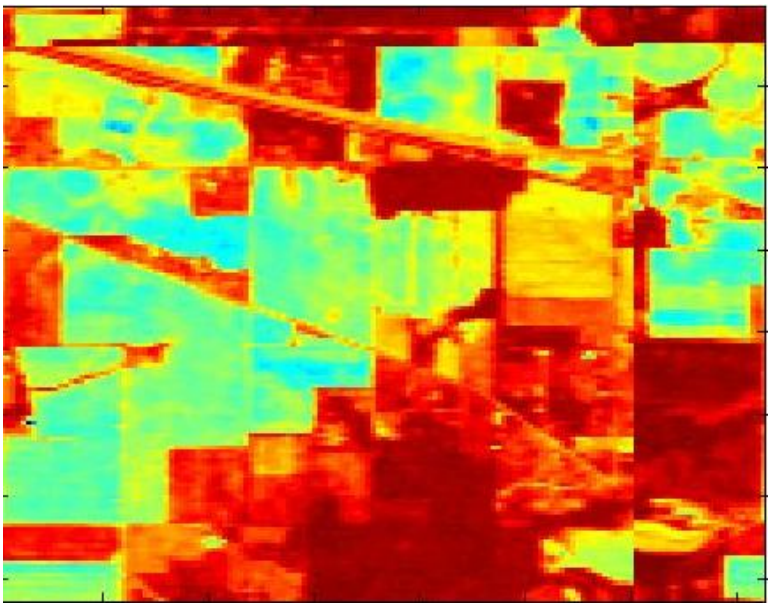

FIGURE 3.3: Spatial Display of the First Extracted Principal Component

percentage accuracies for our proposed method. The classes are Corn-notil, Corn-min, Grass/Pasture, Grass/Tree, Hay-windrowed, Soybeans-min, Soybeans-clean, Soybeans-notil and Woods labelled $C1, C2, \ldots, C9$ respectively. The water absorption bands (104–108,150–163,220) were discarded following Tadjudin and Landgrebe's work [74]. Table 3.1 shows the information classes of selected samples of interest of the AVIRIS Indian pine dataset

We first computed the Principal Component Analysis to extract and represent 99% of the total variance of the image which can be represented in few extracted bands by PCA. Landgrebe in his work shows that the majority of the hyperspectral image bands contain null spaces [75]. Hence, the need for the computation of principal components before the semi-supervised approach to reduce the computational complexity that would be incurred using the whole 200 bands. Figures 3.3 - 3.5 show the first three extracted principal components.

The samples of different ground cover types according to the known ground cover information were collected and then the samples were divided into training samples and test samples. Each of these was done for the three selected principal components. From the training samples, the transform matrices were computed

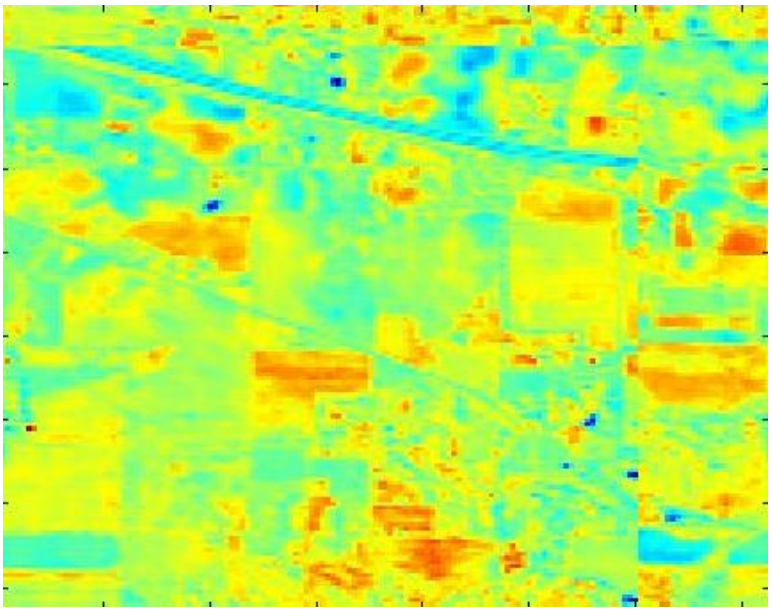

FIGURE 3.4: Spatial Display of the Second Extracted Principal Component

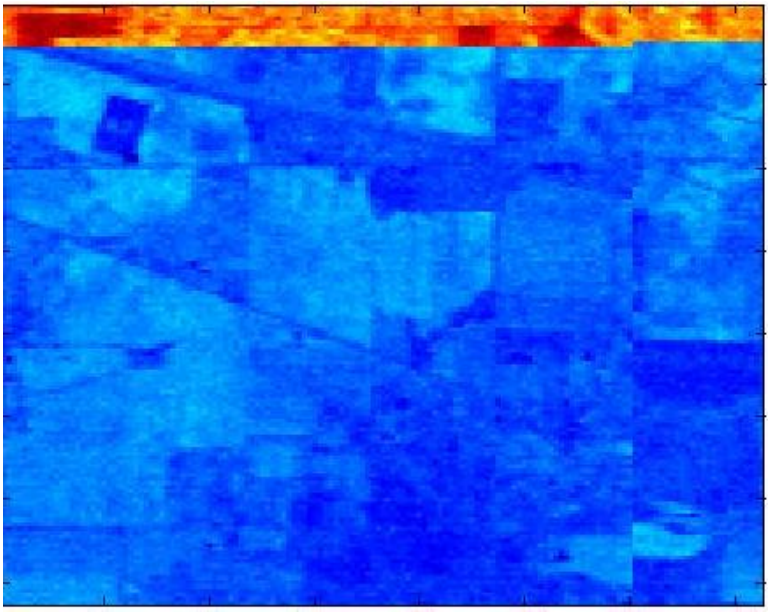

FIGURE 3.5: Spatial Display of the Third Extracted Principal Component

for different feature extraction methods separately, including PCA, LDA, LLE and SSLE. The features were then extracted.

Thereafter, we trained the classifier with the features extracted from the training samples and evaluated the classification result on the test samples. Table 3.2 shows the classification results and the overall accuracies, OA, obtained in the classification. $X = (x_i)_{i=1}^{n}, X \in \Re^{d \times n}$ represents n labelled samples in a d-dimensional feature space and the associated labels $y = (y_i)_{i=1}^{n}, y_i = \{+1, -1\}$ Let the unlabelled dataset $X* = (x_i)_{i=n+1}^{n+m}, X* \in \Re^{d \times m}$ consist of m unlabelled samples. SVM with RBF kernel was used through the MATLAB interface of LIBSVM [76]. The RBF kernel has two parameters: C and γ. We applied a grid search on C and γ using a 10-fold cross validation. 20% of the labelled samples per class of groundtruth was selected for training. The trained classifier was applied to the remaining 80% of the known groundtruth pixel in the scene. The proposed method performs well with the availability of few labelled samples, which is always the case in hyperspectral images over urban areas.

Figure 3.6 shows the relation between the Average classification accuracy under different number of labeled samples per class. The comparison is done using different FE methods. LDA, PCA, LLE were compared with SSLE. PCA performed poorly as it does not take the class property into consideration during the FE process. LDA and LLE performed comparatively closer to each other although LLE had higher accuracies than LDA as a result of the singularity problem in LDA.

3.5 Conclusion

This chapter presents a semi-supervised non-linear method for feature extraction in known as Semi-supervised Local Embedding (SSLE), for the classification of remote-sensed hyperspectral images. The proposed technique aims at extracting features of hyperspectral images in a non-linear manner

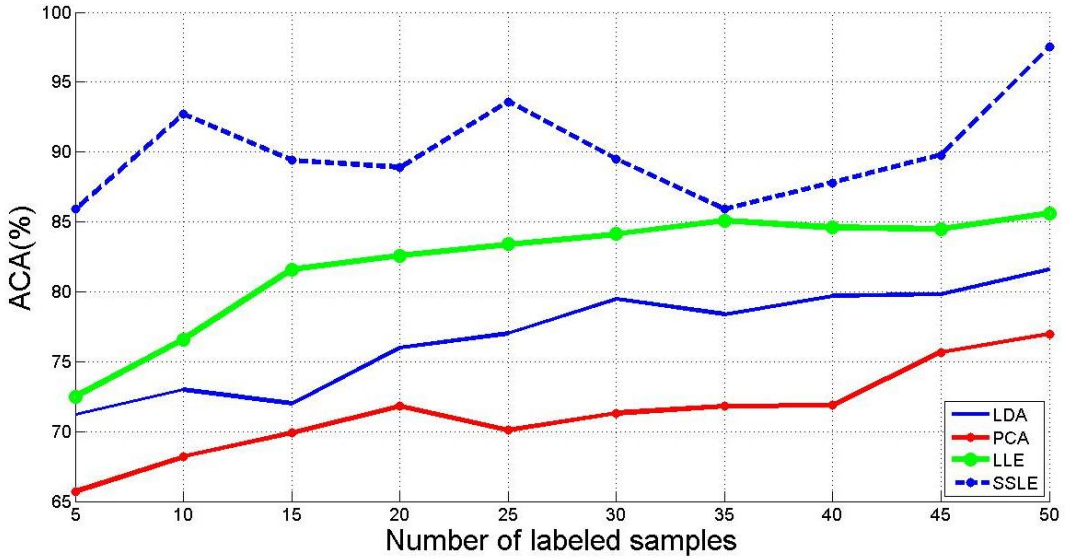

FIGURE 3.6: Average Accuracy with varied number of labelled samples

while ensuring that the extracted features are embedded in the low-dimensional space. It has been developed by the combination of supervised Linear Discriminant Analysis (LDA) and unsupervised Local Linear Embedding (LLE) techniques. Experimental results show that the proposed technique preserves the class property of images in the low-dimensional feature space.

Chapter 4

Graph Discriminative Feature Extraction

4.1 Introduction

In this chapter, a feature extraction method which dynamically selects labelled samples from a pool of unlabelled samples is proposed. Discriminative feature extraction methods perform dimensionality reduction through the decomposition of the eigenvalue of the covariance matrix of the image data. In this process, the eigenvalues and eigenvectors are calculated before the most significant ones are extracted [48]. Local Discriminant Analysis (LDA) [32] and Non-parametric Weight Feature Extraction (NWFE) [44] are common discriminative feature extraction methods that find wide application in hyperspectral image processing due to their class discriminatory properties. LDA discriminates between classes by minimizing the within- and maximizing the between-scatter matrices in order to efficiently classify HSIs. However, it is limited by the fact that the features extracted are dependent on the number of classes in the image. This limitation is addressed in [44] by the development of NWFE which assigns weight to samples and computes the means of their

weight. The distance between weighted means and the samples is used for computing the scatter matrices. Although NWFE overcomes the singularity problem of LDA, it takes longer to complete the feature extraction process, especially when there are only a few labelled samples available. In order to overcome their limitations, several variants and extensions of LDA and NWFE have been developed [33, 34]. In [77], the computation of Principal Component Analysis (PCA) was introduced before LDA by the authors to solve the problem of few labelled samples. The proposed method assumes there is a prior knowledge of the number of classes in the image but this assumption poses difficulty in the implementation of the algorithm. In spite of the development of several variants of LDA, the need to preserve the spatial locality of HSI still needs to be addressed. Recently, the manifold-based learning feature extraction methods [78] in which neighbourhood algorithms are entwined with LDA have been developed for HSI processing. Local Linear Embedding (LLE) [71], Laplacian Eigenmaps (LE) [79], ISOMAP [80] and Local Tangent Space Alignment (LTSA) [81] are typical manifold learning-based techniques that have been successfully implemented in different image processing problems. In addition, nonlinear methods also show great success when implemented in the processing of HSIs because of the nonlinear manifold nature of hyperspectral images [78]. However, their computational complexity and large memory requirement are their major limitations. Hybrid algorithms that combine linear and nonlinear algorithms to obtain new algorithms that eliminate the limitations of linear and nonlinear techniques have been proposed. Hybrid manifold learning-based algorithms seek to preserve the class properties while making the classes as separable as possible. Other hybrid techniques such as Local Fisher Discriminant Analysis (LFDA) [35] combine Linear Discriminant Analysis (LDA) and Local Preserving Projections (LPP) for the reduction of multimodal labelled data. In LFDA, LDA discriminates between the classes of interest while LPP calculates neighbourhood distance by constructing an n-nearest neighbour graph matrix.

Furthermore, manifold learning-based algorithms are sometimes implemented in conjunction with LDA in the processing of HSIs. For instance, the LFDA algorithm which was originally validated with various data sets in [82–84] has been implemented for HSI feature extraction in [51] and classified using the Gaussian-mixture model and support vector machines. In addition, the Nearest Feature Line Embedding (NFLE) algorithm proposed in [60] uses the nearest feature line approach for neighbourhood preservation and shows very interesting performance in dimensionality reduction for HSI classification. The performance of this algorithm can be controlled based on two parameters K1 and K2. In [85], another graph-based supervised feature extraction algorithm is proposed. The algorithm is built on graph-embedded learning with a supervised discriminating method for facial recognition. It uses labelled data points in constructing its within-class graph. However, because labelled data points are always very expensive, processing large-scale unlabelled data would be preferable. Thus, semi-supervised manifold algorithms are been developed. These algorithms utilize labelled and unlabelled data points in the processing of HSIs. One such algorithm is the Semi-supervised Discriminant Analysis (SDA) proposed in [36]. Unlike fully-supervised algorithms such as MMDA, the SDA algorithm has parameters that can be tuned. This is a major advantage of semi-supervised algorithms over fully-supervised ones. Fully-supervised algorithms such as MMDA also suffer a major limitation as a result of an assumption that each class of interest in the image has the same number of samples. But this assumption is not consistent with the processing of HSIs. This major limitation may be addressed by dynamically selecting labelled samples using graph-based clustering techniques.

A detailed review of manifold-learning techniques may be found in [78]. The different manifold-learning methods proposed in the literature may be classified into two categories, namely supervised and non-supervised methods. A method may be classified as supervised if labelled samples are used during computation, and as unsupervised if otherwise. The LDA algorithm [86] is a supervised linear

method; however, it performs poorly when there are limited training samples. Semi-supervised methods such as co-training [87] and graph-based techniques [88–91] are usually preferred and are very popular mainly due to the ease with which unlabelled data may be obtained. In fact, utilizing abundant unlabelled data in conjunction with limited number of labelled data usually presents better classification results. Recently some algorithms have been developed based on discriminant analysis. In [92], the authors propose an algorithm that discriminates the feature space in order to reduce the dimension of hyperspectral image data sets. The proposed algorithm is based on a two–fold mechanism. The first is aimed at increasing the dissimilarities among extracted features by maximizing the between-spectral scatter matrix. The second mechanism aims at minimizing the within–class scatter matrix while concurrently maximizing the between-class scatter matrix. Also in [93], a Sparse Discriminant Manifold Embedding (SDME) algorithm is developed to discriminate features in HSI by boosting the compactness of intra-manifold embedding. Looking closely at these proposed algorithms, it can be deduced that LDA plays a major role in their development. The work presented in this chapter also follows this trend. Therefore, in this chapter we propose the use of graph-based technique for dimensionality reduction in HSI processing which preserves the spectral-spatial property of the image. The proposed algorithm is a semi-supervised manifold transformation algorithm which aims to overcome the limitations of LDA by the use of weighted graphs for computing the within- and between-class scatter matrices. The proposed algorithm, which is referred to as Graph Clustering Discriminant Analysis (GCDA), prevents the merging of classes when their discriminative information is graphically close to one another, a scenario common in HSI processing. GCDA embeds LDA with clustering techniques for dynamic selection of unlabelled samples.

The main contributions of this work are highlighted in the following:

- In order to extract features of the hyperspectral image, GCDA selects

unlabelled samples using graph clustering techniques with manifold learning. The essence is to overcome the problem of few available labelled samples in HSIs and to keep the cluster-derived labels after graph embedding, thereby ensuring that points from the same classes are still close to one another.

- The labelled samples obtained from the clustered segmentation which helps to overcome the problem of few labelled samples, are compared with the labelled samples from the image groundtruth to examine the coherence between both sets of samples.

- We focus on preserving the underlying features of the image after dimensionality reduction by correlating the spatial information from clustered segmentation with spectral information.

The remaining part of the chapter is organized as follows: Related works on graph-based dimensionality reduction are reviewed in section 4.2. The proposed Graph Clustering Discriminant Analysis (GCDA) algorithm is introduced in section 4.3, and the results of experiments conducted to show the effectiveness of the proposed method are discussed in section 5.4. The chapter concludes in section 4.5.

4.2 Related Work

This section introduces three dimensionality reduction methods that are crucial to the formulation of the algorithm developed in this work. These are Linear Discriminant Analysis (LDA), the K–means Clustering method and Laplacian Eigenmaps (LE).

4.2.1 Linear Discriminant Analysis

As earlier mentioned, LDA is a supervised subspace learning method that has found application in many aspects of remote sensing due to its ability to discriminate between classes of interest [33]. Originally introduced in [32], the supervised linear feature extraction method maximizes the ratio of the between-class covariance matrix to the within-class covariance matrix. Let $\{x\}_{i=1}^{n} : x_i \in R^d$ which denotes a high-dimension image data. The labelled samples of each class k where $k = 1 \dots C$ and C represents the number of classes used in the computation of the scatter matrices. The projection matrix W_{LDA} is derived from the following:

$$S_w = \sum_{k=1}^{C} \sum_{j=1}^{N_k} (x_j^k - \mu_k)(x_j^k - \mu_k)^T \tag{4.1}$$

$$S_b = \sum_{k=1}^{C} (\mu_k - \mu)(\mu_k - \mu)^T \tag{4.2}$$

where x_j^k is the j^{th} sample in the k^{th} class, μ is the mean of the total samples, μ_k is the mean vector of class k and N_k is the total number of samples in the given class k. S_w represents the within–class scatter matrix and S_b represents the between–class scatter matrix.

The transformation matrix of LDA, W_{LDA}, is computed by

$$W_{LDA} = \underset{w}{\text{argmax}} \frac{W^T S_b W}{W^T S_w W} \tag{4.3}$$

where $W_{LDA} = (w_1, w_2, w_3, \dots, w_r)$.

4.2.2 K-means Clustering

K-means [94] is an unsupervised clustering method which produces clusters with high intra-class and low inter-class similarities. It has been widely applied to remote-sensed image processing especially in data visualization and image segmentation [95, 96], mainly due to its simplicity, efficiency, ease of implementation and ability to partition data into clusters [97]. Such partitioning algorithms are usually preferred for the processing of remote-sensed images because of the multidimensional characteristics of image data. As an unsupervised method, K–means partitions the image data set into k clusters. The partitioning begins by a random initialization of k centroids, one for each cluster, and then an assignment of input samples to a cluster whose centroid is closest to it. Thus, samples are grouped based on class similarities. The data set is said to have been fully partitioned when the location of the centroids remains unchanged.

In other words, the K-means algorithm partitions a collection of N vectors $x_i : i = 1, \ldots, N$ into K groups $G_k : k = 1, \ldots, K$, and then selects a cluster center c_k for each group. The cost function, J, of the Euclidean distance is a measure of the difference between a vector x_i in the k^{th} group and is defined as follows:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{4.4}$$

4.2.3 Laplacian Eigenmaps

Laplacian Eigenmaps (LE) computes a low-dimensional representation of the data in which the distance between a datapoint and its k-nearest neighbour is minimized. This is done by computing the weights using a cost function. An extension of LE introduced in [98] provides a computationally efficient approach to LE and spectral clustering. This provides a background for

preserving neighbourhood properties of each datapoint while clustering, hence reducing the dimensionality. To represent a low-dimensional data from a high dimensional one, we put an edge between nodes i and j if two datapoints x_i and x_j are close. They are connected if i is among the N nearest neighbour of j or j is among N nearest neighbour of i. For any two datapoints x_i and x_j connected by an edge, we place a weight $W_{ij} = 1$ and the eigenvalues and eigenvectors are computed from

$$Ly = \lambda Dy \tag{4.5}$$

where $L = D - W$ is the laplacian matrix. D is the diagonal weight matrix whose entries are the column sum of W, $D_{ii} = \sum_j W_{ji}$.

Laplacian can be thought of as an operator of functions defined on vertices of G. A reasonable criterion for choosing a good map is to minimize the objective function $\sum_{ij}(y_i - y_j)^2 W_{ij}$ under appropriate constraints.

$$\frac{1}{2}\sum_{ij}(y_i - y_j) = y^T L y \tag{4.6}$$

where $L = D - W$ and W_{ij} being symmetric, $D_{ii} = \sum_j W_{ij}$. It follows that

$$\begin{aligned}\sum_{ij}(y_i - y_j)^2 W_{ij} &= \sum_{ij}(y_i^2 + y_j^2 - 2y_i y_j)W_{ij} \\ &= \sum_i y_i^2 D_{ii} + \sum_j y_j^2 D_{jj} - 2\sum_{ij} y_i y_j W_{ij} \\ &= 2y^T L y\end{aligned} \tag{4.7}$$

Therefore, the minimization problem reduces to

$$\underset{y^T D y = 1}{\operatorname{argmin}} y^T L y \tag{4.8}$$

It is said that the constraint $y^T Dy = 1$ removes an arbitrary scaling factor in the Laplacian Embedding. A natural measure on the vertices of the graph is provided by Matrix D. Laplacian Eigenmaps has been previously used in hyperspectral image processing [99].

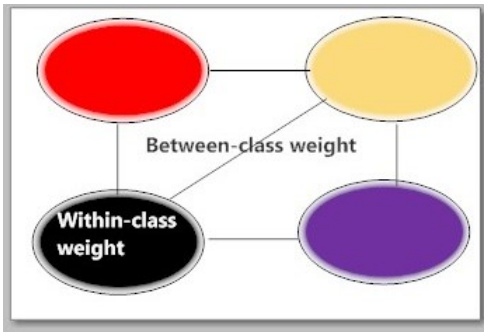

FIGURE 4.1: Illustration of the within-class and between-class weight

4.3 Graph Clustered Discriminant Analysis (GCDA)

The general hyperspectral image feature extraction problem can be described as taking high dimensional data and transforming it into a lower dimensional data set. Let $\{x\}_{i=1}^{N}, x_i \in R^d$ denote the high-dimensional image data and $\{y\}_{i=1}^{N}$, $y_i \in R^p$ the low dimensional image data; where d is the dimension of the original data, p is the dimension of the features that are extracted from it and $p \leq d$. The objective is about finding a projection matrix W of dimension $d \times p$, that can be matched to its high-dimensional data space and allow for better discrimination between classes of interest. The idea is to construct a graph $G = (V, E)$ where V is the set of pixels in the image and E denotes the set of edges that connect them. The edges of the graph are assigned weights based on their spectral similarity, thus weighted graph A is obtained in the process. A is an $n \times n$ matrix which has the weight between any two pixels as entries and n indicates the number of nodes (pixels). In an hyperspectral image, each pixel n is a vector of d dimension –

the number of bands in the hyperspectral image. We assign weight to the graph using the Laplacian Eigenmaps (LE) [100].

In other words, in the extracted low dimensional subspace, points from the same class are meant to be kept as spatially close as possible, and points from different classes are kept as far from each other as possible. Each vertex of the graph is represented in the low-dimensional vector by preserving similarities between the vertex pairs, where the similarity is measured by the LE functions. As shown in Figure 4.1, two types of graphs will be defined; the first graph has a nodes and is referred to as the within-class graph G_{wn} while the second which has b nodes is the between-class graph G_{bn}. For the within-class graph, points from the same class are considered, i.e. points with similar labels. For any two nodes x_i and x_j that have similar labels, an edge is constructed between them. Clusters with lower between-class weight are merged based on similarity. G_{wn} only considers points with the same class labels as shown in Figure 4.2.

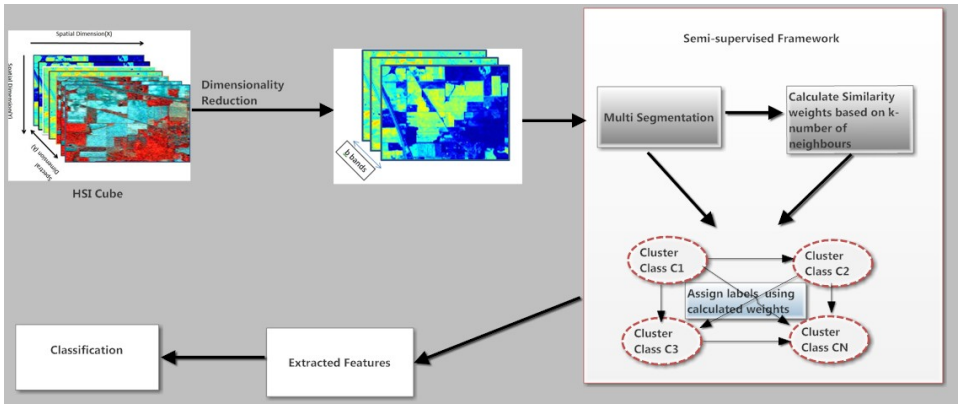

FIGURE 4.2: Flow chart of the proposed approach

In the construction of the GCDA algorithm, we firstly merge the LDA algorithm with K-means algorithms in order to have a single framework, since both algorithms are similar in terms of their classwise discrimination. The main objective behind this merging is to obtain the most discriminative subspace of the image in an unsupervised way while also ensuring that the intrinsic nature of the image is preserved. Unlabelled samples are used in constructing the graphs and in reconstructing the supervised subspace into a semi-supervised

one, using labelled and unlabelled data points. This is achieved by using the hybrid of K-means and LDA and preserves the neighbourhood using LE.

LDA, which is a supervised method, requires the use of class labels. We seek a way to combine LDA and K-means algorithm [101] to provide an unsupervised framework. Recall from equation 4.3:

$$W_{LDA} = \underset{w}{\text{argmax}} \frac{W^T S_b W}{W^T S_w W} \tag{4.9}$$

It is given that

$$S_t = S_b + S_w \tag{4.10}$$

and

$$S_t = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \tag{4.11}$$

From equation 4.10, $S_b = S_t - S_w$

$$\underset{w}{max} \frac{TrW^T S_b W}{TrW^T S_w W} = \frac{TrW^T (S_t - S_w) W}{TrW^T S_w W} \tag{4.12}$$

Therefore, we obtain

$$\frac{TrW^T S_t W}{TrW^T S_w W} - 1 \tag{4.13}$$

In order to normalize the data,

$$\begin{aligned} &x = (x_i)_{i=1,\ldots,n} \text{ is transformed to} \\ &\bar{x} = (\bar{x}_i)_{i=1,\ldots,n} \end{aligned} \tag{4.14}$$

where

$$\begin{aligned} \bar{x}_i &= x_i - \mu \text{ and} \\ \mu &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned} \tag{4.15}$$

$$\begin{aligned} \min_{w} TrW^T S_w W &= Tr \sum_k \sum_{i \in C_k} U^T (x_i - m_k)(x_i - m_k)^T U \\ &= \sum_k \sum_{i \in C_k} \parallel U^T x_i - U^T m_k \parallel^2 \end{aligned} \tag{4.16}$$

This is similar to the K-means clustering in the transformed subspace $W^T X$. We can then construct the within-class graph G_{wn} by constructing, from the same class, an edge between nodes x_i and x_j. The Laplacian Eigenmaps function defines the similarity between these two nodes.

The within-class graph-manifold structure is formulated by incorporating the eigenmaps into unsupervised LDA as

$$\underset{p}{\operatorname{argmin}}\, P^T X L_{wn} X^T P \tag{4.17}$$

where L_{wn} is a Laplacian matrix given by $D_{wn} - A$ and D_{wn} denotes a diagonal matrix with $D_{wii} = \sum_j W_{ij}$ indicating the sum of column (or row) in A.

The matrix A of affinity weight and the diagonal matrix D_{wn} can be respectively expressed as:

$$A = \begin{bmatrix} A_1, & A_2, & \cdots & 0 \\ \vdots & \ddots & & 0 \\ 0 & \cdots & & A_c \end{bmatrix} \tag{4.18}$$

$$D_{wnk} = \begin{bmatrix} D_{wn1}, & D_{wn2} & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & & \cdots & D_{wnc} \end{bmatrix} \tag{4.19}$$

where the affinity weight in individual classes is represented by $A_1 ... A_c$, and $D_{wn1} ... D_{wnc}$ denoting its point importance. According to the matrix D_{wnk}, the weighted center for class k can be computed:

$$mn = \frac{1}{\sum_i D_{kii}} \left(\sum D_{kii} x_{ki} \right) \tag{4.20}$$

The class weighted centers $M = (mn_1, mn_2 ... mn_c)$ which can be more representative than the original mean of each class are then computed. Any two nodes m_i and m_j are connected with an edge which is assigned the weight

$$B_{ij} = \exp^{-\frac{\|m_i - m_j\|^2}{t}} \tag{4.21}$$

In equation 4.21, the weight directly enhances the contribution of each class that has small distances. The condition for between-class graph-penalizing is given as follows:

$$\underset{p}{\operatorname{argmax}} P^T X L_{bn} X^T P \tag{4.22}$$

A unique manifold structure exists for each class, and there also exists the possibility to find non-similar classes residing on different manifolds.

Considering the weight centres of each class in the between-class graph G_{bn}, only points that are linked to each other are used.

According to graph embedding, two optimization criteria should be satisfied:

$$\begin{cases} \underset{p}{\operatorname{argmin}}\, P^T X L_{wn} X^T P \\ \underset{p}{\operatorname{argmax}}\, P^T X L_{bn} X^T P \end{cases} \tag{4.23}$$

which could be further represented as follows:

$$P = \underset{P}{\operatorname{argmax}} \frac{P^T X L_{bn} X^T P}{P^T X L_{wn} X^T P} \tag{4.24}$$

Algorithm 2: Procedure to develop GCDA transformation

Input : $\{p\}_{i=1}^{n}, p_i \in R^d$ Labelled samples
$\{y\}_{i=1}^{n}, y_i \in C$ Corresponding labels
$\{p\}_{i=n+1}^{N}, p_i \in R^d$ Unlabelled samples

1 Extract $w_{PCA} \in \Re^q$ from $\{p\}_{i=1}^{N}$ using PCA.
2 Compute clusters on W_{pca}.
3 Obtain the labels using equations 4.12 - 4.16.
4 Merge similar clusters using LE distance metric in equation 18 to calculate the shortest distance.
5 Recompute the cluster until consistency is reached.
6 Use equations 4.17- 4.23 to compute transformation matrix

Output: The transformation matrix **W**

4.4 Experiments

We validated the proposed algorithm using two HSI data sets and have presented experimental results showing the merits of GCDA. The objective is to compare the performance of GCDA method with other existing feature extraction methods such as LDA, LPP, SDA and FLDA. The first data set is the popular AVIRIS Indian Pines [102] which we also used to validate and benchmark our proposed feature extraction method. The Indian Pines scene contains two-thirds agriculture, and one-third forest or other natural perennial vegetation. The ground truth available is designated into sixteen classes in which ten out of the sixteen available classes are considered from the image

ground truth due to the number of samples available in the ground truth for training the classifiers. These are Corn-notil, Corn-mintil, Grass-pasture, Grass-trees, Hay-windrowed, Soybean-notil, Soybean-mintil, Soybean-clean, Woods, and Buildings-Grass-Trees-Drives. More information about the AVIRIS Indian Pines HSI is available in [102]. The second data set was obtained by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor in northern Italy. The number of bands generated by this sensor is 115 with a spectral coverage ranging from 0.43 - 0.86μm. The Pavia university data set was reduced to 610 $\times$ 340 and 103 spectral bands. The geometric resolution of the image is 1.3 meters and the image's ground truths differentiate 9 classes. Table 4.1 shows the class information of the two data sets used in this work. All data used in the experiment is normalized to a range [0, 1].

TABLE 4.1: Training and Test Samples of Sample Set Used in the Experiment

Indian Pine		Pavia University Area	
Class Name	Samples	Class Name	Samples
Corn-notil (C1)	1434	Asphalt	6631
Corn-min (C2)	834	Meadows	18649
Grass-Pasture (C3)	497	Gravel	2099
Grass-Trees (C4)	747	Trees	3064
Hay-windrowed (C5)	489	Metal Sheets	1345
Soybeans-notil (C6)	968	Soil	5029
Soybeans-min (C7)	1294	Bitumen	1330
Soybeans-clean (C8)	2468	Bricks	3682
Woods (C9)	614	Shadows	947
Bldg-grass-tree (C10)	380		

We evaluated the performance of our algorithm using the overall accuracy (OA), average accuracy (AA) and Kappa coefficient (K) as described in [103]. OA is the ratio of the total correctly classified pixels to the number of pixels in the ground truth. Average accuracy (AA) is defined as the average value of class-specific accuracy obtained for each class. K is a statistical measurement which is the percentage agreement corrected by the level of agreement that could be expected due to chance. These were used to evaluate the performance of our proposed method in comparison with state-of-the-art feature extraction

TABLE 4.2: Class-specific Rates in Percentages for AVIRIS Indian Pine Data Set

DR + Classifier		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	OA	K
Original	SVM	68.50	68.23	82.77	88.05	97.39	77.23	79.68	75.44	91.12	65.18	79.06	0.7674
	NN	64.18	68.05	89.78	83.68	99.18	72.60	81.07	74.60	94.29	62.25	78.52	0.7503
LDA	SVM	78.42	49.92	78.07	90.61	100	75.41	84.83	35.41	97.20	59.06	78.33	0.7451
	NN	63.73	51.52	87.72	97.67	100	42.16	73.44	62.30	96.13	53.97	75.68	0.7203
LPP	SVM	74.22	69.68	96.34	97.39	100	68.63	78.34	49.88	99.42	61.13	80.16	0.7675
	NN	69.07	70.55	92.60	91.79	99.80	80.07	78.31	75.29	93.75	73.24	81.04	0.7805
SDA	SVM	81.61	73.76	97.91	99.30	100	67.45	82.51	64.45	98.55	65.63	83.12	0.8121
	NN	73.40	71.08	88.32	91.48	99.19	74.57	74.48	85.71	94.60	75.94	79.88	0.7732
FLDA	SVM	86.72	79.94	98.69	99.13	99.95	79.09	85.16	91.15	99.52	64.53	88.47	0.8656
	NN	77.17	75.63	93.93	92.63	99.14	83.86	86.30	76.86	94.59	71.02	85.19	0.8286
GCDA	SVM	89.03	81.25	100	100	100	80.25	91.82	93.88	100	70.19	92.10	0.9113
	NN	79.97	75.17	91.50	92.45	100	85.55	86.19	84.68	94.52	79.46	87.22	0.8654

methods. Image classification was done using two classifiers, namely the Support Vector Machine (SVM) and the Nearest Neighbour (NN) classifiers because they achieve high efficiency with high dimensional data. The Library of Support Vector Machine's (LIBSVM) radial basis function (RBF) was used based on a one-against-one classification approach. The two parameters, slack variable C and γ of the RBF kernel, were tested on different range of values; the values of C are bounded in the range [10-1000] with step-size increments of 20 while values for γ are bounded in the range [0.1-2.0] with step-size increment of 0.1. Their optimal values which were obtained using a five-fold cross-validation approach were reported. We compared the proposed algorithm with other state-of-the-art manifold learning algorithms. Our evaluation of the different feature representations was done by average classification rates that were obtained from ten independent experiments. The results obtained for the two data sets are presented in Table 4.2 and Table 4.3. The classification rates are reported using SVM and NN classifiers. The training samples used in each independent experiment were selected from the referenced data randomly and the remaining samples were used as test samples.

The classification rates obtained for the two data sets are shown in Table 4.2 and Table 4.3. Table 4.2 shows the class-specific rates for the ten classes in the AVIRIS Indian Pine data set. The overall accuracy using SVM and NN are reported for each algorithm.

In Table 4.2, class-specific accuracies of the AVIRIS data set were obtained for the proposed DR method as well as four other DR methods, namely LDA, LPP, SDA and FLDA. Similar metrics were also obtained for the image processed without the implementation of DR. For each of the DR methods, the accuracies were obtained for the NN and SVM classifiers. We compared the class accuracies, overall accuracy and Kappa coefficients for the original image without DR as well as for the four DR methods (LDA, LPP, SDA and FLDA) with the proposed GCDA algorithm. In C1, the least accuracy was recorded when LDA is classified with SVM, while the highest accuracy was recorded for GCDA when it is classified with SVM. A similar performance was recorded with C2, albeit with a lower accuracy. The least accuracy of 49.92% was recorded for the LDA+SVM while the highest accuracy recorded was for GCDA+SVM with a value of 81.25%. For C3, GCDA+SVM achieves 100% accuracy; while LPP, SDA and FLDA with SVM and NN classifiers also recorded class accuracies greater than 90.00%; the LDA+SVM recorded the least accuracy of 78.07%. Also, for classes C4, C5 and C9, 100% accuracies were recorded for the GCDA with SVM classifier. The GCDA with NN classifier also achieved 100% accuracy in C5. For all classes (C1–C10), the highest accuracy was recorded for the proposed GCDA algorithm. Furthermore, the efficiency of DR is significantly highlighted by the difference in class accuracy values recorded for the GCDA and original image without DR. The GCDA algorithm achieved significantly higher accuracy than the processing without DR. Considering the overall accuracy (OA), all DR methods achieve OA values greater than 70%. However, the GCDA with SVM classifier achieved the highest OA of 92.10%, while the least accuracy is recorded for the LDA with NN classifier with a value of 75.68%. As for the Kappa coefficient, a value as close to unity is desirable for all DR methods. The GCDA with SVM achieved the highest Kappa value of 0.9113; this is followed by the FLDA with SVM classifier, while the LDA with NN classifier achieved the least Kappa value of 0.7203. This is because SVM is a more robust classifier than NN when dealing with high dimensional data. Generally, the GCDA DR algorithm performed comparatively

TABLE 4.3: Class-specific Rates in Percentage for Pavia University Data Set

DR + Classifier		Asphalt	Meadows	Gravel	Trees	Metal Sheets	Bare Soil	Bitumen	Bricks	Shadows	OA	K
Original	SVM	79.85	84.62	83.80	87.59	99.02	76.07	75.76	66.28	100	81.74	0.7515
	NN	82.12	86.20	64.67	89.23	98.14	58.08	70.08	59.76	94.73	79.87	0.7467
LDA	SVM	84.27	86.26	72.40	89.38	98.95	76.99	83.02	71.90	100	84.18	0.7859
	NN	74.01	79.86	61.32	92.01	99.55	49.48	46.29	52.25	81.45	72.55	0.6256
LPP	SVM	86.54	88.07	65.10	91.31	99.71	78.30	74.55	71.02	100	83.06	0.7981
	NN	85.01	87.84	62.06	86.65	98.97	64.11	69.02	66.08	98.14	80.26	0.7734
SDA	SVM	87.50	89.06	66.74	91.97	99.68	78.56	86.19	73.55	99.98	85.17	0.8347
	NN	73.59	82.47	63.57	93.37	99.61	60.35	49.92	60.65	86.78	79.38	0.7567
FLDA	SVM	88.55	90.01	73.85	92.12	99.45	81.15	85.60	76.58	100	87.35	0.8413
	NN	81.37	85.27	64.49	91.45	97.31	72.76	70.74	65.93	97.94	82.34	0.7931
GCDA	SVM	88.94	90.40	80.67	93.66	99.75	81.14	87.57	80.28	100	89.10	0.8615
	NN	86.75	88.26	68.16	94.95	99.85	74.96	71.56	68.18	99.89	84.12	0.8226

better than all other four DR methods considered.

In Table 4.3, classes are considered in the order Asphalt, Meadows, Gravel, Trees, Metal Sheets, Bare Soil, Bitumen, Bricks and Shadows. For the first and second class, GCDA performed best overall when classified with SVM closely followed by FLDA and SDA both with SVM classification. GCDA performed comparatively well in most of the classes, with SVM classifier showing superior performance over the NN classifier. An exception is found in the fourth class where GCDA with NN classifier outperformed SVM with a difference of 1.29%.

It can be observed that SDA and GCDA performed comparatively well in the presence of few labelled samples. For some classes, using the Pavia University Data in Table 4.3, SDA and FLDA performed comparatively well. Considering the Bare Soil class, using SVM classifier, GCDA and FLDA had similar class-specific accuracy, but considering the overall accuracy, GCDA outperformed both FLDA and SDA. The results show that for each class of interest, GCDA maintained the highest classification rates, OA and Kappa coefficient values with the two classifiers used in our experiment.

A graphical comparison of the average accuracy of the GCDA algorithm and existing methods, namely LDA, LPP, SDA and FLDA is presented in Figures 4.3–4.6 as a function of the number of extracted features. The accuracy is obtained for different numbers of extracted features for the AVIRIS Indian Pines data set. We performed the experiment repeatedly using 16 and 32 training samples and employed NN and SVM as classifiers. In Figure 4.5, the average accuracy of the

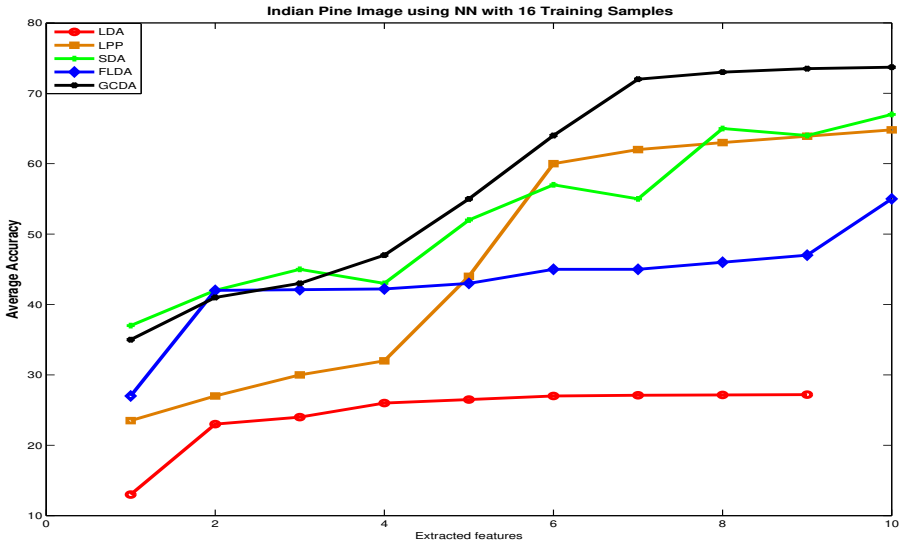

FIGURE 4.3: Indian Pine Image Using NN Classifier with 16 Training Samples

algorithms is compared using 16 training samples with the NN classifier. From the graph, the LDA algorithm had the lowest performance with its accuracy almost constant as the number of extracted features increased. However, the accuracy of other algorithms increased with the number of extracted features; SDA slightly outperformed GCDA with the first two extracted features, but GCDA showed greater accuracy with an increasing number of extracted features. In Figure 4.6, the accuracy of the GCDA algorithm still exceeded those of other algorithms. However, with more training samples, the LDA algorithm showed improved accuracy in its performance. The least accuracy was recorded for the FLDA algorithm.

In Figures 4.7 and 4.8, the average accuracy is presented as a function of the number of extracted features when classified with SVM using 16 and 32 training samples respectively. The graph in Figure 4.7 shows that LDA recorded the least accuracy as in Figure 4.5. However, it performed with higher accuracy when the number of training samples was increased to 32 (see Figure 4.8). Figures 4.7 and 4.8 also show that GCDA achieved the highest accuracy, with the best result recorded with 32 training samples.

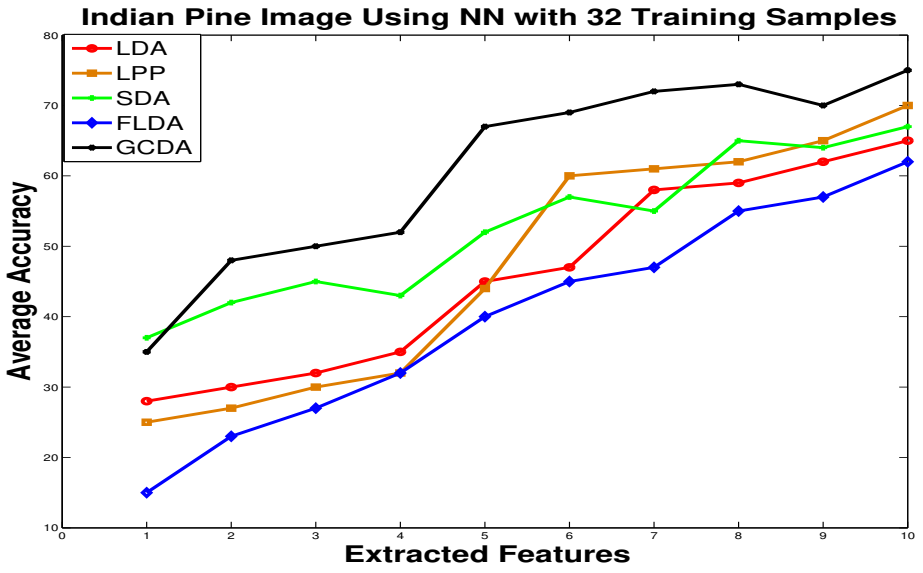

FIGURE 4.4: Indian Pine Image Using NN Classifier with 32 Training Samples

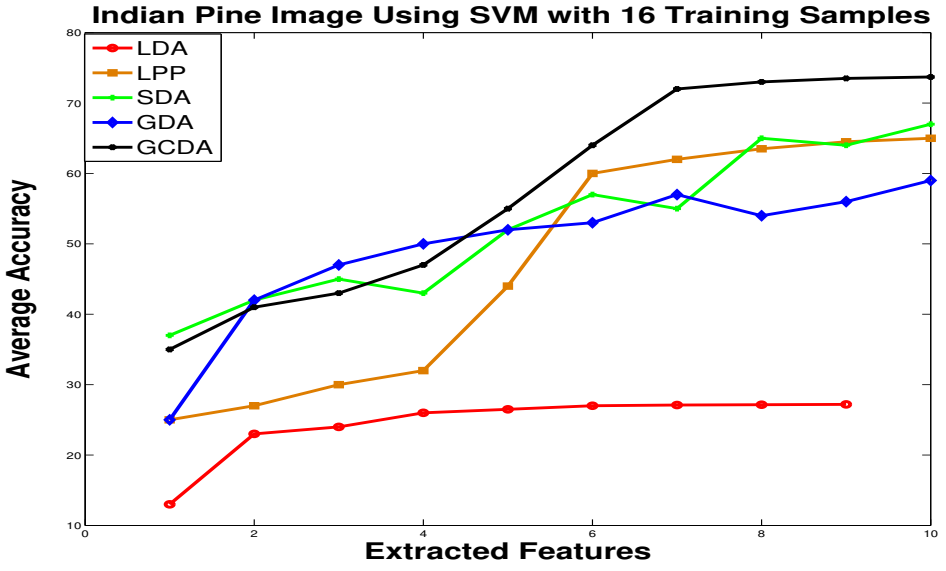

FIGURE 4.5: Indian Pine Image Using SVM Classifier with 16 Training Samples

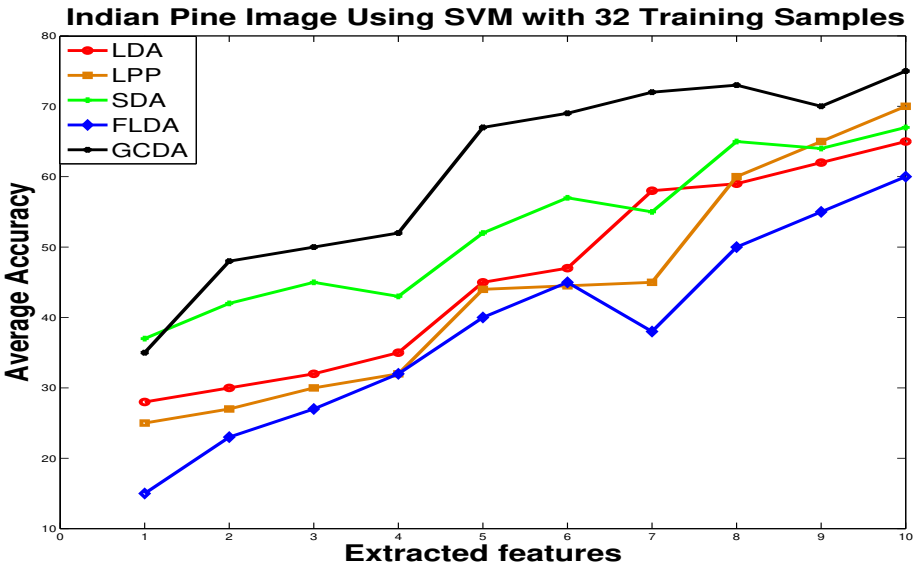

FIGURE 4.6: Indian Pine Image Using SVM Classifier with 32 Training Samples

Furthermore, we investigated the differences in the classification accuracies of the two classifiers (NN and SVM) to determine if such differences were statistically significant. The McNemar's Test [104] for statistical significance was used. The statistical significance of the difference in accuracy of the two classifiers was inferred from the —Z— parameter which may be evaluated as follows:

$$Zc_1c_2 = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \tag{4.25}$$

where f_{ij} denotes the number of samples lying in the confusion matrix at row i and column j. The parameter $|Z_{c_1c_2}|$ indicates that classifier one c_1 performs better than classifier two c_2 if $|Z_{c_1c_2}| > 0$ or the other way round. If $|Z_{c_1c_2}|$ is greater than 1.96 then the difference between the accuracy values of the two classifiers is said to be statistically significant.

The results of the z-score significance test for the Indian Pines AVIRIS data sets using SVM and NN are presented. The results in Tables 4.4 and 4.5 show the z–score significance for the feature extraction algorithms discussed. In Table 4.4, using SVM the difference between GCDA and the other feature

TABLE 4.4: McNemar's Test for Indian Pine Image Using SVM

	GCDA	LDA	LPP	SDA	FLDA
GCDA	0	59.3	52.8	41.6	18.7
LDA	-59.3	0	-6.9	-18.6	-41.6
LPP	-52.8	6.9	0	-11.7	-34.9
SDA	-41.6	18.6	11.7	0	-23.4
FLDA	-18.7	41.6	34.9	23.4	0

TABLE 4.5: McNemar's Test for Indian Pine Image Using NN

	GCDA	LDA	LPP	SDA	FLDA
GCDA	0	45.4	25.8	30.3	9
LDA	-45.4	0	-19.9	-15.4	-36.6
LPP	-25.8	19.9	0	4.5	-16.9
SDA	-30.3	15.4	-4.5	0	-21.4
FLDA	-9	36.6	16.9	21.4	0

extraction methods are statistically significant, that is $|Z| > 1.96$. Also the statistical difference of accuracy $Z = -41.6$ and $Z = -18.7$ shows the benefit of GCDA over SDA and FLDA respectively. Similarly, in Table 4.5 GCDA shows superiority over the other feature extraction algorithm using NN. However, the statistical significance of GCDA over FLDA was greater when using SVM with a z-score value of 18.7, than when using NN with a z-score value of 9.

All experiments were implemented with MATLAB R2014b software installed on an AMD Dual core processor PC with 3GB of RAM. A comparison of the computation times of the implemented algorithms showed that LDA records the least duration of 0.67s. This was due to the presence of few labelled samples in the ground truth. This was closely followed by LPP which ran for 0.71s, then SDA with 0.81s and FLDA with a time of 0.82s. The proposed GCDA algorithm records the highest computation time of 1.07s. However, in spite of its high computation time, its performance in terms of accuracy as well as its ability to preserve the classwise property of images makes GCDA a preferable algorithm for feature extraction in hyperspectral image processing.

4.5 Conclusion

A novel graph-based discriminant analysis for feature extraction in hyperspectral image processing known as Graph Clustered Discriminant Analysis (GCDA) has been developed in this work. GCDA uses a clustered manifold technique for the feature extraction process and has the major advantage of discriminating between classes of interest in the HSI using class labels that are obtained from segmented clusters. In addition, it preserves the neighbourhood regions of the segmented clusters in an HSI. The performance of the proposed method was compared with other state-of-the-art feature extraction methods using two hyperspectral image data sets. The extracted features were classified using SVM and NN classification algorithms. The results from the experiment show that GCDA achieves higher accuracy than the existing methods. However, its computational complexity and running time is slightly higher in comparison to the existing methods. A parallel computing approach to its implementation may be pursued in future work with the goal to reduce the computation time.

Chapter 5

Window Partitioned and Covariance Estimation Feature Extraction

5.1 Introduction

Hyperspectral images (HSIs) are highly dimensional remote-sensed images which are characterized by hundreds of spectral bands and wide spatial dimension. Their rich spectral information is useful in discriminating between classes of interest in the image. However, HSIs are characterized by a limited number of training samples in relation to their spectral dimension, hence, feature extraction is usually introduced as a necessary step before classification [105]. Feature Extraction (FE) methods are developed and implemented in order to transform the high dimension of the image into a low-dimension feature space using the projection matrix. Their implementation also reduces the computational and processing complexities involved in HSI processing. Several feature extraction methods have been proposed for HSI processing in the literature. The proposed methods may be classified as either supervised or unsupervised [70, 72, 76, 106].

Supervised feature extraction methods use class labels for the computation of their projection matrix while the unsupervised ones do not. Recently, Semi-supervised Feature Extraction algorithms which use both labelled and unlabelled samples for computation have been proposed [36].

Algorithms could also be classified as spectral, spatial or spectral-spatial. Related spectral-spatial [107] methods are PCA with morphological profiles [50], and partitional clustering methods [108].

In spectral feature extraction methods pixels are analyzed individually while in spatial methods pixels are grouped based on predefined similarities. Spatial methods are also important in HSI processing as it helps to distinguish the size of an object being classified which sometimes is not detected by spectral methods alone. The composite of both spectral and spatial feature extraction methods yield promising results.

In spite of the existence of other feature extraction techniques, PCA [76] is an unsupervised feature extraction technique that has been widely applied in HSI processing because of its simplicity and ability to represent the total variance of the image in few eigen-vectors (components). It is sometimes used before supervised feature extraction, morphological processing [109] and the application of non-linear methods [37]. However, a common challenge associated with unsupervised linear feature extraction algorithms is the loss of spatial information of certain classes of interest during computation [110]. PCA seeks orthogonal directions that are efficient for representing the data using orthogonal projections. It finds application in data visualization, compression [70] and dimensionality reduction. However, it discards useful discriminating information in its lower eigen vectors, leading to a loss of important spatial information.

In this work, we address this limitation of PCA by introducing a partitioning approach in the covariance computation of PCA. We propose an unsupervised feature extraction method based on Window Partitioning of HSI feature space

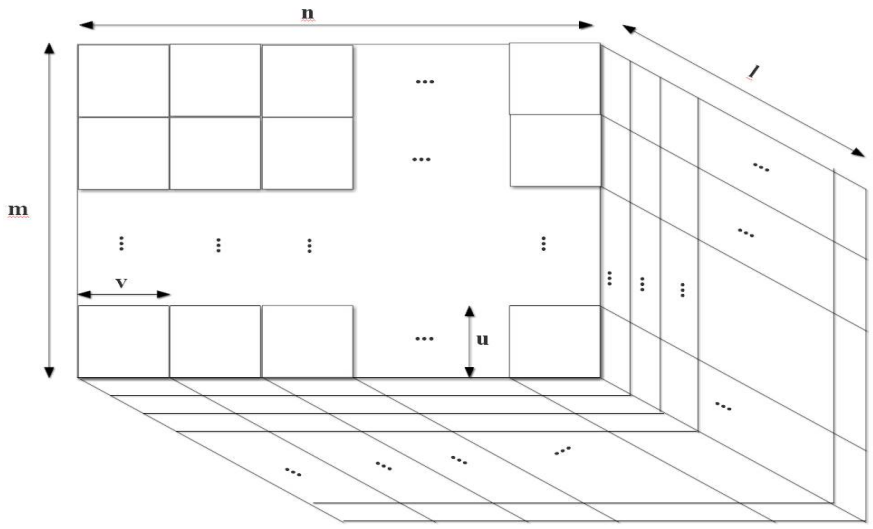

FIGURE 5.1: WinPCA Method

using a static windowing approach named windowed PCA (WinPCA). The idea is to dimension the HSI cube into a specified number of windows and then estimate the covariances for each window in the HSI in order to reduce the loss in spatial information when computing the principal components. The overall covariance of the HSI is then obtained through an iterative merger of the covariance of each window in the HSI based on Kalmer filtering. Experiments with the AVIRIS Indian Pine data set shows that the proposed method preserves more spatial information when compared with the conventional PCA and related approaches.

The remainder of this chapter is structured as follows. Related works are discussed in section 5.2 The proposed method is introduced in section 5.3. The results of experiments are discussed in section 5.4. Section 5.5 concludes the chapter.

5.2 Related Works

In order to address the curse of the dimensionality of hyperspectral images, several feature extraction techniques [111] have been proposed in the literature. The most prominent and widely used method for feature selection

preprocessing of hyperspectral images is the principal component analysis (PCA) which selects the prime bands in the image for classification. A formulation of PCA to extract features of hyperspectral images is done in [46]. To overcome the limitation of PCA the principle of band partitioning before the computation of covariance matrices in HSI has gained prominence. To enhance the display and classification of HSI, Jia and Richards [112] introduced Segmented Principal Component Transformation (PCT). This algorithm was later combined with JPEG2000 in [113]. In recent times, several methods have been proposed to reduce the loss of spatial information in unsupervised linear methods. One of such is the Segmented PCA (Seg-PCA) technique [106] which organizes the whole spectral vectors into a number of sub-vectors and then compute the PCA for each sub-vector. However, this approach is limited by its high computational cost. Another method is the Folded PCA which is proposed in [114]. In its implementation, the spectral vectors are folded into a matrix before the covariance matrix is computed. In terms of computation, Seg-PCA and Folded PCA methods are similar; but they differ in their selection of pixels for the computation of covariance matrix. In Seg-PCA, bands are grouped based on the correlation matrix between each pair of bands as opposed to the folded-PCA method which selects bands based on the common or averaged dimensions of the clusters that can be observed in the main diagonal of the correlation matrix. Partitioning methods based on a training sample selection was introduced in Modified-PCA [115] for covariance matrix computation.

Covariance matrices which are also referred to as second moments play an important role in the formulation of PCA Algorithm. Personen et al. [116] showed that the first- and second-order moments play a major role in the feature extraction of HSI. Manjunath et al. [117] proposed the fusion of covariance matrices of PCA and Fisher Linear Discriminant (FLD) using a product rule to preserve the natures of both covariance matrices and improve performance.

5.3 Windowed Linear Feature Extraction

An hyperspectral image which is represented by a matrix X, of spectral size $m \times n$ in l bands. For a pixel x^{ij} in any location in the $m \times n$ spatial dimension, x^{ij} is a vector with dimension l.

$$x^{ij} = \{x_1^{ij}, x_2^{ij}, x_3^{ij} \ldots, x_l^{ij}\} \tag{5.1}$$

The mean vector of the image for each band b is defined by

$$\bar{x}_b = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_b^{ij}) \tag{5.2}$$

where $\bar{x}_b$ represent the mean vector of band b. For any given p and q bands of the HSI, the term of the covariance matrix is expressed as

$$cov(x_p, x_q) = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_p^{ij} - \bar{x}_p)(x_q^{ij} - \bar{x}_q) \tag{5.3}$$

The covariance matrix then is of the form:

$$cov = \begin{pmatrix} cov(x_1, x_1) & \ldots & cov(x_1, x_l) \\ \vdots & \ddots & \vdots \\ cov(x_l, x_1) & \ldots & cov(x_l, x_l) \end{pmatrix} \tag{5.4}$$

Figure 5.2 shows the covariance image of the AVIRIS Indian Pine using 200 bands. The eigenvalue decomposition of the covariance matrix gives the eigenvalue U and eigenvectors V. The eigenvalue decomposition form of the covariance matrix can be expressed as:

$$cov = VUV^T \tag{5.5}$$

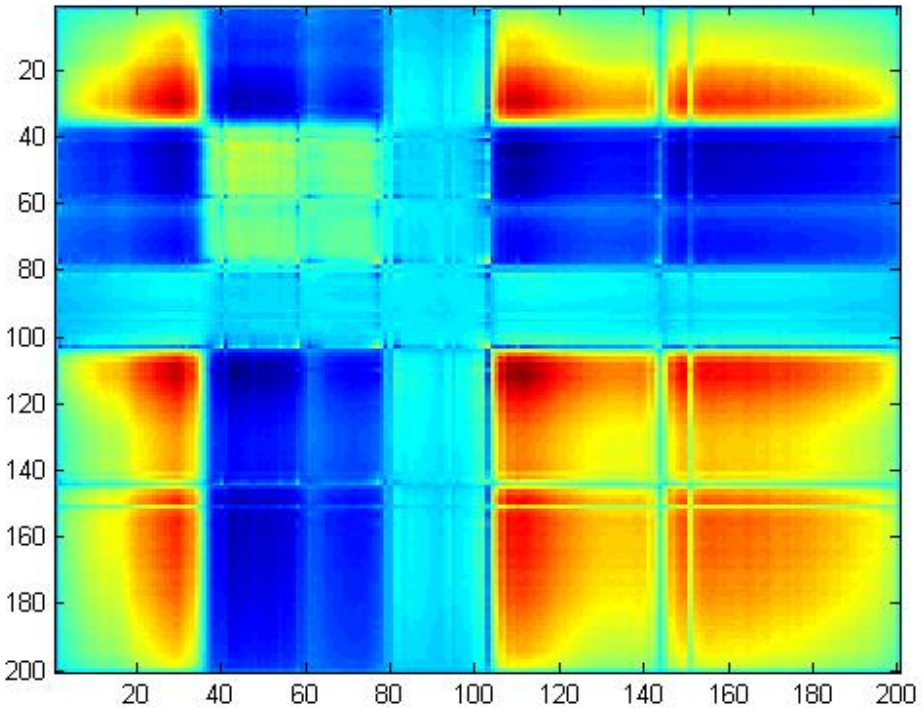

FIGURE 5.2: Covariance for the AVIRIS Indian Pine Data.

U is expressed as a diagonal matrix consisting of the eigenvalue given as:

$$U = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 \\ \vdots & \dots & \dots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \lambda_l \end{pmatrix} \tag{5.6}$$

and V is the orthogonal matrix which contains the corresponding eigenvectors $\{v_1, v_2, v_3 \dots v_l\}$ corresponding to the eigenvalues U. The eigenvalues are arranged in decreasing order [76] and the corresponding eigenvectors are selected for a given number of reduced dimension.

In general, given a data with d-dimensional space, PCA tries to project a data into a low dimensional space r while preserving as much variance as possible

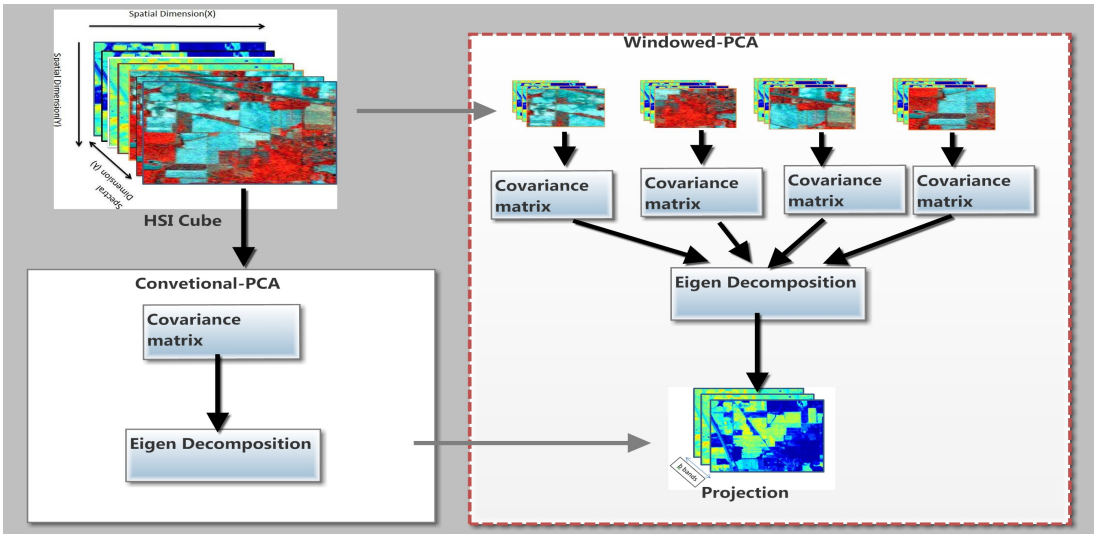

FIGURE 5.3: Windowed-PCA Approach

where $r \ll d$. There are other methods for computing the PCA but the mean-covariance approach described above is the most common.

The proposed approach uses a windowed partitioning technique for feature extraction in hyperspectral image processing. Recalling that the goal of dimension reduction methods is to represent a high dimension data by one of a lower dimension where $r \ll d$ while still retaining as much information as possible in the lower dimension space, the proposed windowed feature extraction technique seeks to achieve this goal. In addition, our approach seeks to preserve the spatial characteristics of the hyperspectral image being processed. The algorithm is described as follows.

Each pixel x_{ij} forms a spectral curve over the spectral bands of the HSI. In Win-PCA, the image X is spatially dimensioned into N windows with each window of size $u \times v$, where u and v indicate the number of pixels on each row and column respectively. The pixel representation for each window is given as

$$c_i = \begin{bmatrix} x_{111} & x_{121} & \dots & x_{1u1} \\ x_{211} & x_{221} & \dots & x_{2u1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{v11} & x_{v21} & \dots & x_{vu1} \end{bmatrix}, \begin{bmatrix} x_{112} & x_{122} & \dots & x_{1u2} \\ x_{212} & x_{222} & \dots & x_{2u2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{v12} & x_{v22} & \dots & x_{vu2} \end{bmatrix} \dots \begin{bmatrix} x_{11l} & x_{12l} & \dots & x_{1ul} \\ x_{21l} & x_{22l} & \dots & x_{2ul} \\ \vdots & \vdots & \ddots & \vdots \\ x_{v1l} & x_{v2l} & \dots & x_{vul} \end{bmatrix} \tag{5.7}$$

The covariance matrix of this small image cuboids is computed independently using equation 5.3. Since covariance matrices are square matrices, the dimension of the covariance matrix gotten from the windowed image is $l \times l$. It then features into groups for covariance matrix estimation. The goal is to preserve local structures while extracting the most significant components. Then the eigen decomposition method can be performed on the total covariance gotten from the covariance of individual windows.

Given that the dimension of the windows is $u \times v \times l$, the mean vector of the j_{th} window is obtained from

$$\bar{C_{ijb}} = \frac{1}{u \times v} \sum_{s=1}^{u} \sum_{t=1}^{v} X_{ijb}^{st} \tag{5.8}$$

where X_{ijb}^{st} is the element on the t^{th} column of the s^{th} row in the $(ij)^{th}$ window of the b^{th} band. C_{ijb} is the mean of the $(ij)^{th}$ window on the b^{th} band. The covariance matrix cov_{ij} for each window ij of the hyperspectral image is expressed as

$$cov_{ij}(x_p, x_q) = \frac{1}{u \times v} \sum_{s=1}^{u} \sum_{t=1}^{v} (x_{ijp}^{st} - \bar{C_{ijp}})(x_{ijq}^{st} - \bar{C_{ijq}})^T \tag{5.9}$$

The process for the merger of the covariance matrices of the $i \times j$ windows of the HSI is described based on a pairwise merger of covariance matrices that was formulated in [118]. Covariance matrix union process has been used in various applications which includes but not limited to information fusion or robotics for example.

In order to compute the combined mean and covariance of any two windows whose mean vectors and covariance matrices are respectively denoted by (M_j, C_j) and (M_k, C_k), Kalman gain, K, is first obtained as follows:

$$K = C_j(C_j + C_k)^{-1} \tag{5.10}$$

Then the combined mean, M, and combined covariance C are respectively estimated as

$$M = M_j + K(M_k - M_j) \tag{5.11}$$

and

$$C = C_j - KC_j \tag{5.12}$$

Let an image be divided into n windows; after the calculation of means and covariances $(M_1, C_1), (M_2, C_2), \ldots, (M_n, C_n))$ of all n windows, the computation of the total mean vector and total covariance matrix (M, C) of the image in done as indicated by algorithm 3.

The eigenvalues and eigen vectors for the whole image are computed from the combined covariance matrices.

Algorithm 3: Combined Mean and Covariance of multiples windows in an images

input : $(M_1, C_1), (M_2, C_2), \ldots, (M_n, C_n))$ // Means and Covariances of n windows

output: (M, C) // Combined Mean and covariance

$\mathbf{K} \leftarrow \mathbf{C_1}(\mathbf{C_1} + \mathbf{C_2})^{-1}$: Kalman gain calculation;
$\mathbf{C} \leftarrow \mathbf{C_1} + \mathbf{KC_1}$;
$\mathbf{M} \leftarrow \mathbf{M_1} + \mathbf{K}(\mathbf{M_2} - \mathbf{M_1})$;
for $i \leftarrow 3$ **to** n **do**
    $\mathbf{K} \leftarrow \mathbf{C}(\mathbf{C} + \mathbf{C_i})^{-1}$;
    $\mathbf{C} \leftarrow \mathbf{C} + \mathbf{KC}$;
    $\mathbf{M} \leftarrow \mathbf{M} + \mathbf{K}(\mathbf{M_i} - \mathbf{M})$;

Since the HSI is dimensioned into J windows with C_j denoting the estimated covariance of its j_{th} window where (j = 1,2, ..., J). It is required to obtain the total covariance of the HSI by combining the covariances of all the windows of the

HSI. The process for the combination of the covariance matrices begins with the combination of the covariances of any two windows, then the obtained covariance is combined iteratively with that of other windows.

The eigenvalues and eigen vectors for the whole image are computed from the combined covariance matrices.

The total covariance can be expressed as

$$C = VUV^T \tag{5.13}$$

U is expressed as a diagonal matrix consisting of the eigen-value given as:

$$U = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 \\ \vdots & \dots & \dots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \lambda_l \end{pmatrix} \tag{5.14}$$

and V is the orthogonal matrix which contains the corresponding eigen-vectors $\{v_1, v_2, v_3 \dots v_l\}$ corresponding to the matrix of eigenvalues U. Thereafter, the first l eigen vectors from the total covariance are selected with the highest corresponding eigen-values from the projection vectors where V is the eigen-vector and U is the matrix of eigen-values.

$$W_{pca} = [v_1, v_2, \dots, v_m] \tag{5.15}$$

Algorithm 4 shows the description of our proposed WinPCA method for extracting features in hyperspectral images.

Algorithm 4: WPCA

Input : Hyperspectral image X

Output: The projection matrix Y_{W-PCA} which is made up of k eigenvectors associated with the largest n eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$

1 **For** $n \times m$ spatial domain of image X, we divide X into windows of $\frac{n}{q} \times \frac{m}{p}$ and bands of each window into H groups along its spectral dimension to give smaller hyperspectral cuboids c_i
2 **Compute** the covariance matrix of each c_i using equation 5.3.
3 **Combine** the covariance matrix of each group using Algorithm 3.
4 **Find** the eigen vectors and values of the covariance matrix.
5 **Sort** eigenvectors in decreasing order of eigenvalues.
6 **Projecting** the original dataset on the eigen vectors which gives the feature vector (Principal Components).

5.4 Experiments

The proposed approach was validated using the AVIRIS Indian Pines [119] data set and results from experiments to demonstrate the benefits of WinPCA for feature extraction are presented in this section. Classification was done using Support Vector Machine (SVM). The AVIRIS Indian Pine data consist of 220 spectral bands. Table 5.1 gives the information on the data set. Following the work from Tadjudin and Landgrebe [120], bands 104-108,150-163 and 220 are discarded because they were water absorption bands. In this study, all sixteen classes of the AVIRIS HSI image were used for evaluation.Pixels were represented in vectors which correspond to the spectral wavelength of the reflected object. Figure 5.4 shows the spectral curve for AVIRIS Indian Pine [102] hyperspectral image. The performance of the proposed method was measured by three metrics, namely, overall accuracy (OA), average accuracy (AA) and Kappa Coefficient. Also, the Radial Basis Function (RBF) kernel of LibSVM[74] was used for Support Vector Machine (SVM) classification. Using the segmented approach in [75], the results of PCA and Seg-PCA were compared with the proposed WinPCA method.

In the experiment, the training samples were constituted by a random selection of 10, 20 and 30 labelled samples per class and the remaining samples left for

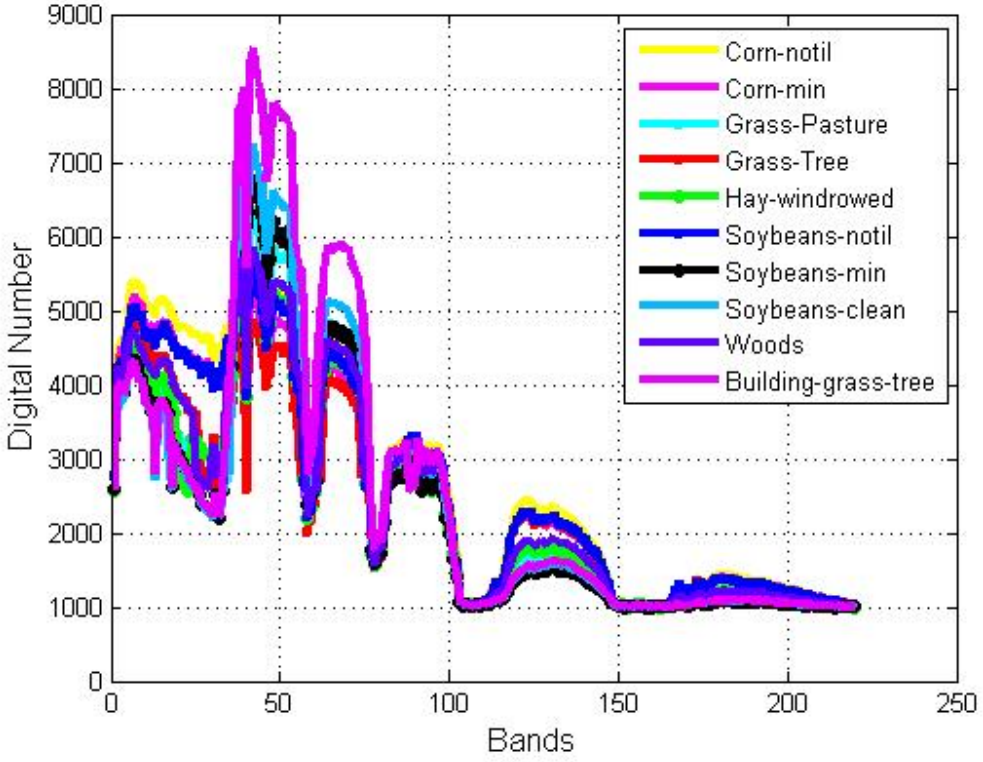

FIGURE 5.4: Spectral Curves for the AVIRIS Indian Pine Data.

TABLE 5.1: Number of Samples in the Data Set Used in the Experiment

Class Name	Samples	Class Name	Samples
Corn-notil(C1)	1434	Soybeans-notil(C6)	968
Corn-min(C2)	834	Soybeans-min(C7)	1294
Grass/Pasture(C3)	497	Soybeans-clean(C8)	2468
Grass/Trees(C4)	747	Woods(C9)	614
Hay-windrowed(C5)	489	Blg-grass-tree(C10)	380

validation. The penalty (C) was tested in the range 10 to 1000 in step-wise increments of 20 and the value of gamma(γ) was selected in the range 0.1 to 2.0 in step-wise increments of 0.1. A five-fold cross-validation approach [121] was employed to obtain the best values. In order to guarantee the accuracy and consistency of the results, each experiment was run 10 times and the average reported. The spatial display of the first twelve principal components for both PCA and WinPCA are shown in Figure 5.6 and 5.7 respectively. The first twelve principal component were chosen for display as it contained over 98.00% of the

total variance after the feature extraction process. The spatial display shows that WinPCA preserved more spatial features. WinPCA was able to preserve more information in its extracted principal components.

Three conventional ways were used for the evaluation of the overall performance of the algorithms, the overall accuracy, average accuracy and kappa coefficient. Details of the metrics used for evaluating the accuracies can be found in Appendix A. For WinPCA, a window size of 29 × 29 divided it into 25 windows and each of these windows was further divided along the bands grouping. Throughout this experiment for Seg-PCA, we set H = 10 and W = 20. The experimental results are summarized in Table 5.2.

In addition, the performances of WinPCA for a different number of windows were compared. Results were obtained for WinPCA for 4, 9, 25, 36, 49 and 64 window sizes. Considering their overall accuracies, best results were seen at 25 windows and the performance decreased slowly as the number of partition increases. This result is presented in Figure 5.5.

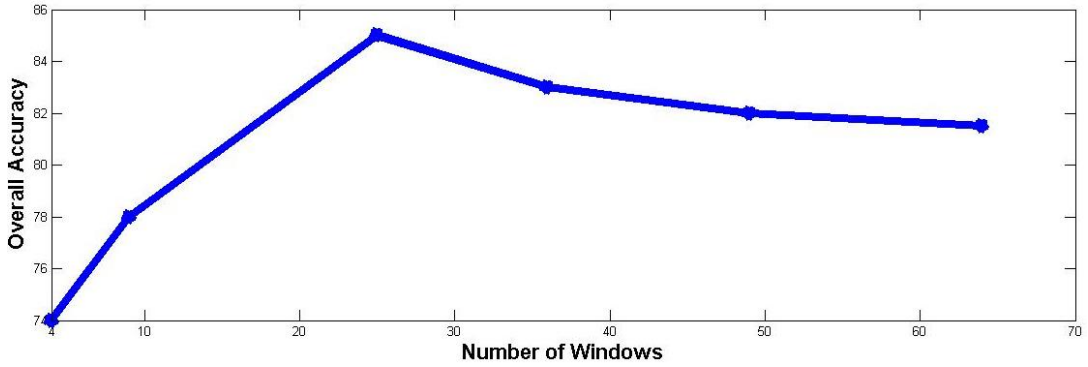

FIGURE 5.5: Classification Accuracy for Extracted Principal Components

Table 5.2 displays the performance of PCA, WinPCA and Seg-PCA when classified with SVM using 10, 20 and 30 training samples per class. The data was scaled between 0 and 1.The training samples were selected randomly while the remaining samples were used for testing. From Table 5.2, it was shown that with the increase in training samples the accuracies of PCA, Seg-PCA and

WinPCA increased but at a varying number of training samples WinPCA outperforms Seg-PCA and PCA feature extraction methods.

With 10 training samples,WinPCA outperformed both PCA and Seg-PCA. The overall accuracy of WinPCA was 66.58% and average accuracy 73.45%. PCA and Seg-PCA performed in the same way with 10 training samples. However, when 20 training samples were used, Seg-PCA showed improved performance over PCA. Also the performance of WinPCA with 30 training samples was better than that of PCA and Seg-PCA. This showed the effectiveness of WinPCA over the conventional PCA and related partitioning approaches.

TABLE 5.2: Overall and Average Accuracy with Kappa Coefficient for the various extraction methods using the first ten principal components

	10 Training Samples			20 Training Samples			30 Training Samples		
	Accuracies (%)			Accuracies (%)			Accuracies (%)		
DR	OA	AA	K	OA	AA	K	OA	AA	K
PCA	56.16	67.06	0.51	64.46	72.91	0.60	67.71	73.21	0.65
Seg-PCA	56.55	66.83	0.52	65.03	73.20	0.60	68.36	72.96	0.64
WinPCA	66.58	73.45	0.72	69.50	74.82	0.79	72.45	78.57	0.81

Figure 5.8 shows the overall accuracy for different principal components. For each class, 20 training samples were selected for PCA, Seg-PCA and Win-PCA. With the first few extracted features, Seg-PCA and WinPCA were comparatively close in terms of their overall accuracies but as the number of features increased, the accuracy of Seg-PCA declined and that of WinPCA increased.

In comparing the computational cost, for the covariance matrix estimation, for PCA the computational cost is $O(R \times C \times S)^2$ and for WinPCA, the computational cost is $O((R/p)(C/q)S^2)$. Where R, C and S denote the number of rows, number of columns and the spectral length of the HSI respectively. The computational complexity of WinPCA could only be reflected in the covariance combination. The experiments were run on MATLAB R2014b software installed on an AMD Dual core processor PC with 3GB of RAM. The computation time for PCA was 2.98s and that of Seg-PCA, 3.01s while that of WinPCA was 3.09s. Moreover, the least computation time was recorded for the

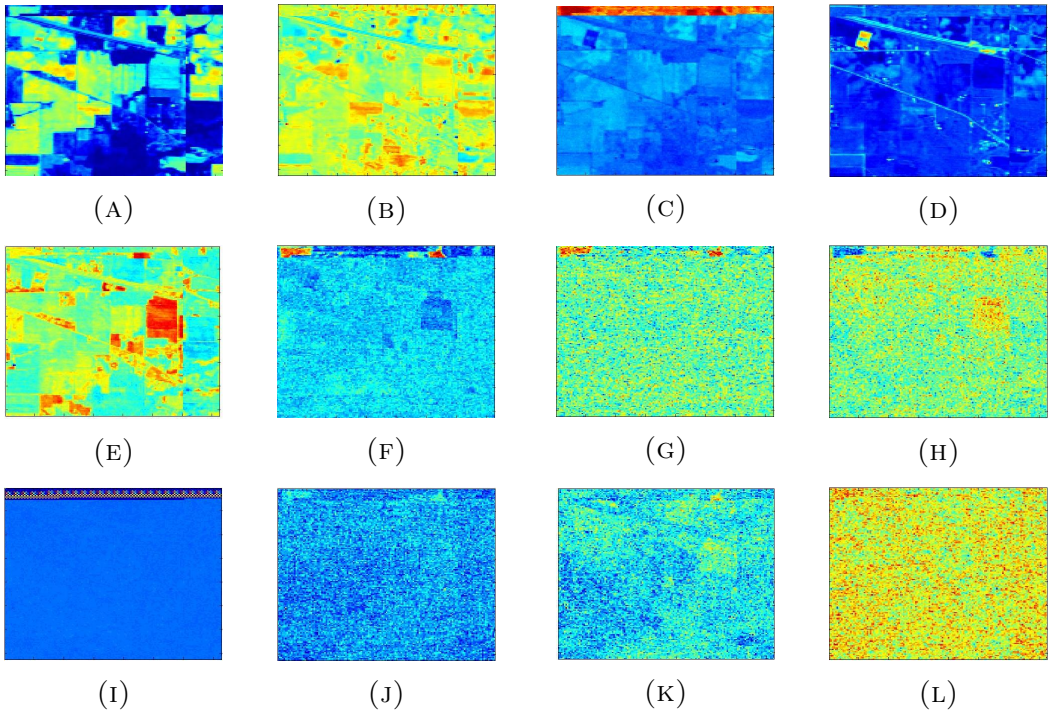

FIGURE 5.6: The First 12 Components Extracted by PCA

PCA algorithm. This is closely followed by Seg-PCA, then WinPCA in increasing order. The proposed WinPCA algorithm records the highest computation time of 3.09s. In spite of its high computation time, its performance in terms of accuracy as well as its ability to preserve spatial properties of the HSI after feature extraction makes WinPCA a preferable algorithm for feature extraction in hyperspectral image processing.

TABLE 5.3: Overall accuracy and average accuracy with varied number of training samples

DR	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	OA	AA	K
Original	0.77	0.65	0.48	0.75	0.68	0.69	0.58	0.79	0.65	0.71	0.67	0.69	0.58
PCA	0.64	0.75	0.74	0.78	0.85	0.80	0.72	0.87	0.81	0.85	0.71	0.65	0.61
Seg-PCA	0.63	0.82	0.77	0.88	0.81	0.87	0.89	0.90	0.87	0.85	0.87	0.71	0.69
Win-PCA	0.81	0.89	0.85	0.93	0.97	0.87	0.85	0.87	0.96	0.96	0.89	0.91	0.88

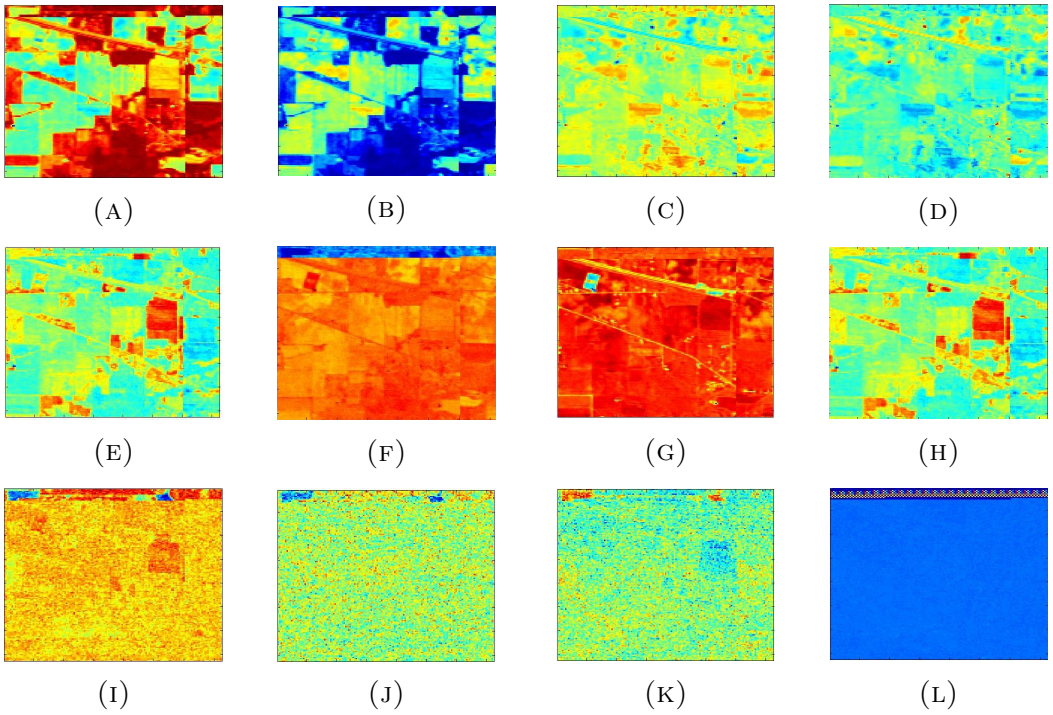

FIGURE 5.7: The First 12 Components Extracted by WinPCA

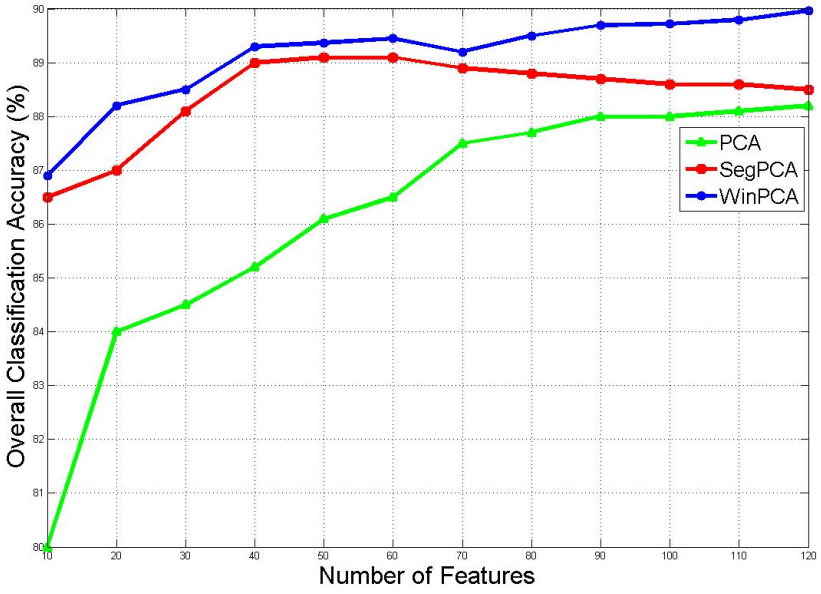

FIGURE 5.8: Classification Accuracy for Extracted Principal Components

5.5 Conclusion

We have presented a windowed approach to PCA for the extraction of features from hyperspectral images. The algorithm, christened Windowed PCA (WinPCA), divides the spatial dimension of an HSI into smaller windows in order to preserve its spatial properties. Then, the overall covariance matrix of the HSI is estimated by merger of the individual covariance matrix of each window. The merger is done by the method of Kalman filtering. The results from the experiments show that the proposed method improves the accuracy of classification. Furthermore, a comparison of the future work will include using Seg-PCA and WinPCA in semi-supervised manifold method for hyperspectral image processing. Comparison between windows of different sizes will also be taken into consideration.

Chapter 6

Summary, Conclusions and Recommendations

This research has investigated the performance of a number of feature extraction techniques that were developed for the processing of hyperspectral images. The proposed techniques adopt the unsupervised and semi-supervised feature extraction methods for their development. In particular, three feature extraction methods are developed, namely, Semi-Supervised Local Linear Embedding (SSLE), Graph Clustered Discriminant Analysis (GCDA) and the Windowed Principal Component Analysis (Win-PCA).

Chapter 2 presents a detailed background of feature extraction for Hyperspectral image processing as well as a review of related works in the literature. Existing feature extraction techniques are classified into three categories namely unsupervised, supervised and semi-supervised, depending on the usage of labelled samples in the computation of their projection matrix. Each of these classes of feature extraction methods have distinct merits and demerits that makes them suitable for different applications. However, as done in this research, the strengths of two or more categories may be exploited in

developing feature extraction techniques that are more efficient and have increased accuracy.

In Chapter 3, Semi-supervised Local linear Embedding (SSLE) feature extraction was developed. The development of the SSLE technique was aimed at exploiting the strengths of Linear Discriminant Analysis (LDA), a supervised method and Local Linear Embedding (LLE) which is an unsupervised feature extraction method. With an understanding that LDA fails to discover the spatial patterns when only a limited number of training samples are available and that LLE is computationally expensive despite its capability to preserve an image's local structure, the SSLE technique merges the strengths of these two techniques to maximize between-class separabilities, minimize within-class similarities and preserve the local neighbourhood of the HSI data. An investigation into the performance of the SSLE technique was done using data sets of the AVIRIS Indian Pines. The results obtained by employing Support Vector Machines (SVM) classifier show that SSLE achieves better performance than PCA, LDA and LLE in terms of overall accuracy when few labelled samples are available.

While still considering that HSIs are characterized by a limited number of training samples in comparison to their dimensions, a graph-based feature extraction technique known as Graph Clustered Discriminant Analysis (GCDA) was developed in Chapter 4. The GCDA technique is built from the local discriminant analysis (LDA), K-means clustering and Laplacian Eigenmap (LE) methods and makes full use of the unlabelled samples in a semi-supervised framework. After careful consideration that LDA and K-means are similar class-wise discriminations, both algorithms were unified in an unsupervised manner before incorporating the graph-preserving property of LE to form a semi-supervised algorithm. The uniqueness of GCDA is in its method of selecting unlabelled samples, which is based on graph clustering and manifold learning. It selects unlabelled samples using the graph clustering technique with manifold learning. This helps in overcoming the problem of few labelled

samples by keeping the cluster-derived labels after graph-embedding and ensures that points in the image that originate from the same class remain close to each other after the feature extraction process is concluded.

Chapter 5 introduces the partitioned method. This method is able to preserve the spatial property of hyperspectral images using the Windowed Principal Component (Win-PCA) method. This was done by dividing the images into subgroups before performing the covariance estimation method. The covariance matrices were then merged using the Kalman filter method. The eigen decomposition was then performed on the merged covariance matrices.

In summary, three FE methods are proposed to explore the rich spectral and spatial complexities of HSI. This methods have been discussed in details in the previous chapters. Two of these, namely Semi-supervised Local Linear Embedding (SSLE)and Graph Clustered Discriminant Analysis (GCDA) are semi-supervised methods while the third, Windowed Principal Component Analysis (WIN-PCA) is an unsupervised method. The importance of these methods are highlighted in the following:

- The semi-supervised approaches are capable of preserving the neighbourhood information while discriminating between classes of interest in the reduced dimension of the feature space.
- The utilization of unlabelled samples in the GCDA approach is another added novelty. This approach makes the developed semi-supervised approach to benefit from the rich abundance unlabelled samples while still making use of the limited labelled samples
- The discriminative feature of both semi-supervised algorithms increases the class accuracy in both algorithms.
- Moreover, with the McNemar test the GCDA approach shows better statistical significance over related methods.

- The Spatial property of the hyperspectral images are preserved using the Graph-based GCDA and the Win-PCA partitioning method.

Based on the research carried out, we are able to propose the following future works and directions:

- Morphological profiles could be investigated and compared with the partitioning approach.
- The partitioning approach in this work was used in an unsupervised manner. With the success shown from this approach, it could be extended to supervised methods.
- Co-training methods could be compared with the use of unlabelled samples.
- Nonlinear PCA could be used instead of PCA in the Win-PCA method and the result used for morphological approaches.

Appendix A

Accuracy Computation

The following provides a brief description of the various metrics used in computing the accuracy and expressing the efficiency of the methods described in this research.

A.1 Confusion Matrix

The confusion matrix is a square matrix used for the computation of other accuracies. It is from the confusion matrix that all other accuracies can be computed

	Classified Samples				
Labelled Samples	C1	C2	C3	C4	Total
C1	C_{11}	C_{12}	C_{13}	C_{14}	$\sum_1^c C_{1j}$
C2	C_{21}	C_{22}	C_{23}	C_{24}	$\sum_1^c C_{2j}$
C3	C_{31}	C_{32}	C_{33}	C_{34}	$\sum_1^c C_{3j}$
C4	C_{41}	C_{42}	C_{43}	C_{44}	$\sum_1^c C_{4j}$
Total	$\sum_1^c C_{i1}$	$\sum_1^c C_{i2}$	$\sum_1^c C_{i3}$	$\sum_1^c C_{i4}$	

TABLE A.1: Sample of a Confusion Matrix

A.2 Class Accuracy

Class Accuracy (CA) is computed as the ratio in percentage of pixels of a given class that have been correctly classified.

$$CA_i = \frac{C_{ii}}{\sum_{j=1}^{c} C_{ij}} \times 100 \tag{A.1}$$

Where C_{ii} denotes correctly classified samples for Class i and C_{ij} denotes the misclassified samples when $i \neq j$

A.3 Overall Accuracy

Overall Accuracy (OA) is computed as the percentage of correctly classified pixels for all the classes that are considered in the image expressed as a percentage.

$$OA = \frac{\sum_{i}^{c} C_{ii}}{\sum_{ij}^{c} C_{ij}} \times 100 \tag{A.2}$$

A.4 Average Accuracy

Average Accuracy (AA) denotes the mean of all class accuracy calculated from the confusion matrix

$$AA = \frac{\sum_{i=1}^{c} CA_i}{c} \tag{A.3}$$

A.5 Kappa Coefficient

Kappa coefficient is measured as a level of agreement between different classes. A value of 1 denotes full agreement and a value of 0 denotes no agreement.

A.6 McNemar Test

The statistical significance of the difference between two classifiers is inferred from the —Z— parameter which is estimated as follows:

$$Zc_1c_2 = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \tag{A.4}$$

where f_{ij} denotes the frequency of points lying in the confusion matrix of element i and element j. The parameter $|Z_{c_1c_2}|$ indicates that classifier one c_1 performs better than classifier two c_2 if $|Z_{c_1c_2}| > 0$ or the other way round. Furthermore,

if $|Z_{c_1c_2}|$ is greater than 1.96, then it can be said that the difference obtained in the classification accuracy of the two classifiers is statistically significant.

Bibliography

[1] José M Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine*, 1(2):6–36, 2013.

[2] Peg Shippert. Introduction to hyperspectral image analysis. *Online Journal of Space Communication*, 3, 2003.

[3] Thomas Jarmer, Joachim Hill, and Sebastian Mader. The use of hyperspectral remote sensing data for the assessment of chemical properties of dryland soils in se-spain. In *Proceedings 5th EARSeL Workshop on Imaging Spectroscopy. Bruges*, volume 1, 2007.

[4] Kun WANG et al. Application of hyperspectral remote sensing in research on ecological boundary in north farming-pasturing transition in china. *Spectroscopy and Spectral Analysis*, 29(6):1636–1639, 2009.

[5] Sarah A Lewis, Peter R Robichaud, William J Elliot, Bruce E Frazier, and Joan Q Wu. Hyperspectral remote sensing of postfire soil properties. In *Tenth Forest Service Remote Sensing Applications Conference: Salt Lake City, Utah*, page 9, 2004.

[6] Foudan Salem and Menas Kafatos. Hyperspectral image analysis for oil spill mitigation. In *Paper presented at the 22nd Asian Conference on Remote Sensing*, volume 5, page 9, 2001.

[7] Ira Leifer, William J Lehr, Debra Simecek-Beatty, Eliza Bradley, Roger Clark, Philip Dennison, Yongxiang Hu, Scott Matheson, Cathleen E Jones, Benjamin Holt, et al. State of the art satellite and airborne marine oil spill remote sensing: Application to the bp deepwater horizon oil spill. *Remote Sensing of Environment*, 124:185–209, 2012.

[8] G Andreoli, B Bulgarelli, B Hosgood, and D Tarchi. Hyperspectral analysis of oil and oil-impacted soils for remote sensing purposes. *EUR*, 22739:1–30, 2007.

[9] AM Bhatti, D Rundquist, J Schalles, and L Ramirez. Application of hyperspectral remotely sensed data for water quality monitoring: accuracy and limitation. In *Proceedings of the Accuracy 2010 Symposium, Leicester, UK, July*, pages 20–23, 2010.

[10] Benjamin D Hennig, BC Cogan, and Inka Bartsch. Hyperspectral remote sensing and analysis of intertidal zones: A contribution to monitor coastal biodiversity. 2007.

[11] Bing Zhang, Wei Yang, Lianru Gao, and Dongmei Chen. Real-time target detection in hyperspectral images based on spatial-spectral information extraction. *EURASIP journal on advances in signal processing*, 2012(1): 1–15, 2012.

[12] Fengchen Huang, Lizhong Xu, Min Li, and Min Tang. High-resolution remotely sensed small target detection by imitating fly visual perception mechanism. *Computational and mathematical methods in medicine*, 2012, 2012.

[13] Nicola Playle. Detection of landmines using hyperspectral imaging. In *Defense and Security Symposium*, pages 62170A–62170A. International Society for Optics and Photonics, 2006.

[14] Gregory P Asner, Daniel Nepstad, Gina Cardinot, and David Ray. Drought stress and carbon uptake in an amazon forest measured with spaceborne

imaging spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6039–6044, 2004.

[15] Kamaruzaman Jusoff. Precision forestry using airborne hyperspectral imaging sensor. *Journal of Agricultural Science*, 1(1):142, 2009.

[16] Hao Zhang, Heng-jia Song, and Bo-chun Yu. Application of hyper spectral remote sensing for urban forestry monitoring in natural disaster zones. In *Computer and Management (CAMAN), 2011 International Conference on*, pages 1–4. IEEE, 2011.

[17] Alvaro Penteado Crósta and Carlos Roberto de Souza Filho. Hyperspectral remote sensing for mineral mapping: a case-study at alto paraíso de goías, central brazil. *Brazilian Journal of Geology*, 30(3):551–554, 2000.

[18] Floyd F Sabins. Remote sensing for mineral exploration. *Ore Geology Reviews*, 14(3):157–183, 1999.

[19] Fred A Kruse, Sandra L Perry, and Alejandro Caballero. District-level mineral survey using airborne hyperspectral data, los menucos, argentina. 2006.

[20] Julia Beckert, Veerle Vandeginste, Cédric M John, Dominique Guérillot, Jurgen Foeken, et al. Hyperspectral remote sensing for the characterization of dolomite bodies: A case study in the central oman mountains-lower khuff analogue. In *International Petroleum Technology Conference*. International Petroleum Technology Conference, 2015.

[21] John B Adams, Milton O Smith, and Paul E Johnson. Spectral mixture modeling: A new analysis of rock and soil types at the viking lander 1 site. *Journal of Geophysical Research: Solid Earth*, 91(B8):8098–8112, 1986.

[22] Philip E Dennison, Kraivut Charoensiri, Dar A Roberts, Seth H Peterson, and Robert O Green. Wildfire temperature and land cover modeling using hyperspectral data. *Remote Sensing of Environment*, 100(2):212–222, 2006.

[23] D NAGESH KUMAR. Remote sensing applications to water resources. In *Research Perspectives in Hydraulics and Water Resources Engineering*, pages 287–316. 2002.

[24] Qihao Weng. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sensing of Environment*, 117: 34–49, 2012.

[25] Uta Heiden, Karl Segl, Sigrid Roessner, and Hermann Kaufmann. Determination of robust spectral features for identification of urban surface materials in hyperspectral remote sensing data. *Remote Sensing of Environment*, 111(4):537–552, 2007.

[26] Uta Heiden, Wieke Heldens, Sigrid Roessner, Karl Segl, Thomas Esch, and Andreas Mueller. Urban structure type characterization using hyperspectral remote sensing and height information. *Landscape and urban Planning*, 105(4):361–375, 2012.

[27] Behzad M Shahshahani and David A Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and remote sensing*, 32(5):1087–1095, 1994.

[28] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[29] JR Harris, D Rogge, R Hitchcock, O Ijewliw, and D Wright. Mapping lithology in canada's arctic: application of hyperspectral data using the minimum noise fraction transformation and matched filtering. *Canadian Journal of Earth Sciences*, 42(12):2173–2193, 2005.

[30] Jing Wang and Chein-I Chang. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. *IEEE transactions on geoscience and remote sensing*, 44(6):1586–1600, 2006.

[31] David Landgrebe. Hyperspectral image data analysis. *Signal Processing Magazine, IEEE*, 19(1):17–28, 2002.

[32] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons,, 1999.

[33] Tatyana V Bandos, Lorenzo Bruzzone, and Gustavo Camps-Valls. Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 47(3): 862–873, 2009.

[34] Qian Du. Modified fisher's linear discriminant analysis for hyperspectral imagery. *IEEE geoscience and remote sensing letters*, 4(4):503, 2007.

[35] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8:1027–1061, 2007.

[36] Deng Cai, Xiaofei He, and Jiawei Han. Semi-supervised discriminant analysis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7. IEEE, 2007.

[37] Hannah M Adebanjo and Jules R Tapamo. Semi-supervised local feature extraction of hyperspectral images over urban areas. In *Adaptive Science and Technology (ICAST), 2013 International Conference on*, pages 1–5. IEEE, 2013.

[38] Sebastiano B Serpico and Lorenzo Bruzzone. A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 39(7):1360–1367, 2001.

[39] Claudio Persello and Lorenzo Bruzzone. Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5):2615–2626, 2016.

[40] Mark Nixon. *Feature extraction & image processing*. Academic Press, 2008.

[41] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1990.

[42] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[43] Gaston Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

[44] Bor-Chen Kuo and David A Landgrebe. Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geoscience and Remote Sensing*, 42(5):1096–1105, 2004.

[45] Minshan Cui and Saurabh Prasad. Angular discriminant analysis for hyperspectral image classification. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1003–1015, 2015.

[46] Craig Rodarmel and Jie Shan. Principal component analysis for hyperspectral image classification. *Surveying and Land Information Science*, 62(2):115, 2002.

[47] Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(1):131–137, 2004.

[48] Jinchang Ren, Jaime Zabalza, Simon Marshall, and Jiangbin Zheng. Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. *Signal Processing Magazine, IEEE*, 31(4):149–154, 2014.

[49] Qian Du and James E Fowler. Hyperspectral image compression using jpeg2000 and principal component analysis. *IEEE Geoscience and Remote sensing letters*, 4(2):201–205, 2007.

[50] Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Wilfried Philips, and Youguo Pi. Classification of hyperspectral data over urban areas using directional morphological profiles and semi-supervised feature extraction. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(4):1177–1190, 2012.

[51] Wei Li, Saurabh Prasad, James E Fowler, and Lori Mann Bruce. Locality-preserving dimensionality reduction and classification for hyperspectral image analysis. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(4):1185–1198, 2012.

[52] Xian Guo, Xin Huang, and Liangpei Zhang. Three-dimensional wavelet texture feature extraction and classification for multi/hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 11(12):2183–2187, 2014.

[53] Qian Shi, Liangpei Zhang, and Bo Du. Semi-supervised discriminative locally enhanced alignment for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens*, 51(9):4800–4815, 2013.

[54] Tianhao Zhang, Dacheng Tao, and Jie Yang. Discriminative locality alignment. *Computer Vision–ECCV 2008*, pages 725–738, 2008.

[55] Wenzhi Liao, Aleksandra Pizurica, Paul Scheunders, Wilfried Philips, and Youguo Pi. Semisupervised local discriminant analysis for feature extraction in hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(1):184–198, 2013.

[56] Qian Shi, Liangpei Zhang, and Bo Du. Semisupervised discriminative locally enhanced alignment for hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(9):4800–4815, 2013.

[57] Liangpei Zhang, Lefei Zhang, Dacheng Tao, and Xin Huang. Tensor discriminative locality alignment for hyperspectral image spectral–spatial

feature extraction. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(1):242–256, 2013.

[58] Peng Zhang, Haixia He, Zhou Sun, and Chunbo Fan. Supervised feature extraction of hyperspectral image by preserving spatial-spectral and local topology. In *Intelligent Computing Theories and Methodologies*, pages 682–692. Springer, 2015.

[59] Jinhuan Wen, Zheng Tian, Xiangzeng Liu, and Wei Lin. Neighborhood preserving orthogonal pnmf feature extraction for hyperspectral image classification. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(2):759–768, 2013.

[60] Yang-Lang Chang, Jin-Nan Liu, Chin-Chuan Han, and Ying-Nong Chen. Hyperspectral image classification using nearest feature line embedding approach. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(1): 278–287, 2014.

[61] Kun Tan, Songyang Zhou, and Qian Du. Semisupervised discriminant analysis for hyperspectral imagery with block-sparse graph. *Geoscience and Remote Sensing Letters, IEEE*, 12(8):1765–1769, 2015.

[62] Pierre Soille. *Morphological image analysis: principles and applications*. Springer Science & Business Media, 2013.

[63] Jón Atli Benediktsson, Jón Aevar Palmason, and Johannes R Sveinsson. Classification of hyperspectral data from urban areas based on extended morphological profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(3):480–491, 2005.

[64] Fulin Luo, Hong Huang, Zezhong Ma, and Jiamin Liu. Semisupervised sparse manifold discriminative analysis for feature extraction of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6197–6211, 2016.

[65] Xiangtao Zheng, Yuan Yuan, and Xiaoqiang Lu. Dimensionality reduction by spatial–spectral preservation in selected bands. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5185–5197, 2017.

[66] Yuhang Gan, Fulin Luo, Juhua Liu, Bing Lei, Tao Zhang, and Ke Liu. Feature extraction based multi-structure manifold embedding for hyperspectral remote sensing image classification. *IEEE Access*, 5:25069–25080, 2017.

[67] Zhen Ye, Lian Tan, and Lin Bai. Hyperspectral image classification based on spectral-spatial feature extraction. In *Remote Sensing with Intelligent Processing (RSIP), 2017 International Workshop on*, pages 1–4. IEEE, 2017.

[68] Xudong Kang, Xuanlin Xiang, Shutao Li, and Jón Atli Benediktsson. Pca-based edge-preserving features for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

[69] Renlong Hang, Qingshan Liu, Yubao Sun, Xiaotong Yuan, Hucheng Pei, Javier Plaza, and Antonio Plaza. Robust matrix discriminative analysis for feature extraction from hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5):2002–2011, 2017.

[70] Qian Du and James E Fowler. Hyperspectral image compression using jpeg2000 and principal component analysis. *Geoscience and Remote Sensing Letters, IEEE*, 4(2):201–205, 2007.

[71] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[72] David H Kim and Leif H Finkel. Hyperspectral image processing using locally linear embedding. In *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*, pages 316–319. IEEE, 2003.

[73] M Baumgardner, L Biehl, and D Landgrebe. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3. *Purdue University Research Repository*, 2015.

[74] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[75] Jinchang Ren, Jaime Zabalza, Stephen Marshall, and Jiangbin Zheng. Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. *Signal Processing Magazine, IEEE*, 31(4):149–154, 2014.

[76] IT Jollife. Principle component analysis. *SpringerVerlag, New York*, 1986.

[77] Qian Du and Nicolas H Younan. Dimensionality reduction and linear discriminant analysis for hyperspectral image classification. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 392–399. Springer, 2008.

[78] Dalton Lunga, Saurabh Prasad, M Crawford, and Okan Ersoy. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *Signal Processing Magazine, IEEE*, 31(1):55–66, 2014.

[79] X Niyogi. Locality preserving projections. In *Neural information processing systems*, volume 16, page 153, 2004.

[80] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

[81] Zhen-yue Zhang and Hong-yuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University (English Edition)*, 8(4):406–424, 2004.

[82] Hui-Ling Chen, Da-You Liu, Bo Yang, Jie Liu, and Gang Wang. A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis. *Expert Systems with Applications*, 38(9):11796–11803, 2011.

[83] Yogachandran Rahulamathavan, Raphael C-W Phan, Jonathon Chambers, David J Parish, et al. Facial expression recognition in the encrypted domain based on local fisher discriminant analysis. *Affective Computing, IEEE Transactions on*, 4(1):83–92, 2013.

[84] Hong Huang, Jiamin Liu, Hailiang Feng, and Tongdi He. Ear recognition based on uncorrelated local fisher discriminant analysis. *Neurocomputing*, 74(17):3103–3113, 2011.

[85] Wankou Yang, Changyin Sun, and Lei Zhang. A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognition*, 44(8):1649–1657, 2011.

[86] Dijun Luo, Chris HQ Ding, and Heng Huang. Linear discriminant analysis: New formulations and overfit analysis. In *AAAI*, 2011.

[87] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[88] Gustavo Camps-Valls, Tatyana V Bandos Marsheva, and Dengyong Zhou. Semi-supervised graph-based hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(10):3044–3054, 2007.

[89] Santiago Velasco-Forero and Vidya Manian. Improving hyperspectral image classification using spatial preprocessing. *IEEE Geoscience and Remote Sensing Letters*, 2(6):297–301, 2009.

[90] Jun Bai, Shiming Xiang, and Chunhong Pan. A graph-based classification method for hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(2):803–817, 2013.

[91] Nam Hoai Ly, Qian Du, and James E Fowler. Sparse graph-based discriminant analysis for hyperspectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(7):3872–3884, 2014.

[92] Maryam Imani and Hassan Ghassemian. Feature space discriminant analysis for hyperspectral data feature reduction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102:1–13, 2015.

[93] Hong Huang, Fulin Luo, Jiamin Liu, and Yaqiong Yang. Dimensionality reduction of hyperspectral images based on sparse discriminant manifold embedding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 106: 42–54, 2015.

[94] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[95] Saman Ghaffarian and Salar Ghaffarian. Automatic histogram-based fuzzy c-means clustering for remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 97:46–57, 2014.

[96] Dominique Lavenier. Fpga implementation of the k-means clustering algorithm for hyperspectral images. In *Los Alamos National Laboratory LAUR*. Citeseer, 2000.

[97] James P Theiler and Galen Gisler. Contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation. In *Optical Science, Engineering and Instrumentation'97*, pages 108–118. International Society for Optics and Photonics, 1997.

[98] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[99] Biao Hou, Xiangrong Zhang, Qiang Ye, and Yaoguo Zheng. A novel method for hyperspectral image classification based on laplacian eigenmap pixels distribution-flow. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(3):1602–1618, 2013.

[100] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[101] Chris Ding and Tao Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th international conference on Machine learning*, pages 521–528. ACM, 2007.

[102] NW Aviris. Indiana's indian pines 1992 data set. *URL: ftp://ftp. ecn. purdue. edu/biehl/MultiSpec/92AV3C (original files) and ftp://ftp. ecn. purdue. edu/biehl/PC_MultiSpec/ThyFiles. zip (ground truth)(last date accessed: 16 June 2008)*, 2012.

[103] Russell G Congalton. A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1):35–46, 1991.

[104] Giles M Foody. Thematic map comparison. *Photogrammetric Engineering & Remote Sensing*, 70(5):627–633, 2004.

[105] David Landgrebe. Hyperspectral image data analysis. *Signal Processing Magazine, IEEE*, 19(1):17–28, 2002.

[106] Xiuping Jia and John A Richards. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 37 (1):538–542, 1999.

[107] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, Jocelyn Chanussot, and James C Tilton. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101(3):652–675, 2013.

[108] Yuliya Tarabalka, Jón Atli Benediktsson, and Jocelyn Chanussot. Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2973–2987, 2009.

[109] Jon Aevar Palmason, Jon Atli Benediktsson, Johannes R Sveinsson, and Jocelyn Chanussot. Classification of hyperspectral data from urban areas using morphological preprocessing and independent component analysis. In *Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. Proceedings. 2005 IEEE International*, volume 1, pages 4–pp. IEEE, 2005.

[110] Anil Cheriyadat and Lori Mann Bruce. Why principal component analysis is not an appropriate feature extraction method for hyperspectral data. In *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International*, volume 6, pages 3420–3422. IEEE, 2003.

[111] Antonio Plaza, Jon Atli Benediktsson, Joseph W Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, et al. Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, 113:S110–S122, 2009.

[112] Xiuping Jia and John A Richards. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 37 (1):538–542, 1999.

[113] Qian Du, Wei Zhu, He Yang, and James E Fowler. Segmented principal component analysis for parallel compression of hyperspectral imagery. *Geoscience and Remote Sensing Letters, IEEE*, 6(4):713–717, 2009.

[114] Jaime Zabalza, Jinchang Ren, Mingqiang Yang, Yi Zhang, Jun Wang, Stephen Marshall, and Junwei Han. Novel folded-pca for improved feature extraction and data reduction with hyperspectral imaging and sar in remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93:112–122, 2014.

[115] Cheng Wang, M Menenti, and Zhao-Liang Li. Modified principal component analysis (mpca) for feature selection of hyperspectral imagery. In *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International*, volume 6, pages 3781–3783. IEEE, 2003.

[116] Maiju Pesonen, Henri Pesonen, and Jaakko Nevalainen. Covariance matrix estimation for left-censored data. *Computational Statistics & Data Analysis*, 92:13–25, 2015.

[117] DS Guru, MG Suraj, and S Manjunath. Fusion of covariance matrices of pca and fld. *Pattern Recognition Letters*, 32(3):432–440, 2011.

[118] Randall C Smith and Peter Cheeseman. On the representation and estimation of spatial uncertainty. *The international journal of Robotics Research*, 5(4):56–68, 1986.

[119] D Landgrebe. Aviris nw indiana's indian pines 1992 data set, 1992.

[120] Saldju Tadjudin and David A Landgrebe. Covariance estimation for limited training samples. In *Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS'98. 1998 IEEE International*, volume 5, pages 2688–2690. IEEE, 1998.

[121] Cheng-Hsuan Li, Chin-Teng Lin, Bor-Chen Kuo, and H-H Ho. An automatic method for selecting the parameter of the normalized kernel function to support vector machines. In *Technologies and Applications of Artificial Intelligence (TAAI), 2010 International Conference on*, pages 226–232. IEEE, 2010.