Universiteit Gent
Faculteit Ingenieurswetenschappen
Vakgroep Telecommunicatie en
Informatieverwerking

Feature Extraction and Classification for Hyperspectral Remote Sensing Images

Wenzhi Liao

Proefschrift tot het bekomen van de graad van
Doctor in de Ingenieurswetenschappen:
Computerwetenschappen
Academiejaar 2011-2012

Universiteit Gent
Faculteit Ingenieurswetenschappen
Vakgroep Telecommunicatie en
Informatieverwerking

Promotoren: Prof. Dr. Ir. Wilfried Philips
Prof. Dr. Ir. Aleksandra Pižurica
Prof. Dr. Ir. Youguo Pi

Universiteit Gent
Faculteit Ingenieurswetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
St-Pietersnieuwstraat 41 B-9000 Gent, België

Tel.: +32-9-264.34.12
Fax.: +32-9-264.42.95

Dit werk kwam tot stand in het kader van een specialisatiebeurs van het FWO-Vlaanderen (Flemish Fund for Scientific Research).

Proefschrift tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen:
Computerwetenschappen
Academiejaar 2011-2012

Acknowledgements

This thesis could not have been produced without the support and assistance of my supervisors, my family, numerous friends and colleagues, for whom I would like to include this acknowledgment.

First and foremost I wish to thank my supervisors, Aleksandra Pižurica, Wilfried Philips and Youguo Pi. Thank you for your excellent guidance and stimulating ideas. Thanks for reading all the papers and this thesis with much care and for giving relevant comments. Thanks to all staff of the Gent University and South China University of Technology for making this work possible. My greatest gratitude to Aleksandra Pižurica. You have been my mentor throughout these years. Thank you for your bright ideas and your hand by hand help with numerous problems I encountered, even such as your advices on how to express sentences correctly in English, and how to answer the comments of reviewers in more comfortable ways.

I also thank my colleague Rik Bellens at Telin, for guiding me in programming and discussing our cooperated paper. For advice on English pronunciation and presentations, many thanks to Filip Rooms and Jan Aelterman.

Special thanks to Prof. Paul Scheunders from the university of Antwerp. The FWO project I have worked on during my PhD was cooperated with your group, thank you for supporting me in pursuing this PhD and for your guidance on our cooperated paper. I also wish to thank prof. Paolo Gamba, for providing the data sets of Pavia. I also take this opportunity to mention my friends: Danny De Vleeschauwer, Andres Frias Velazquez, Danilo Babin, Ivana Despotovic and the crew of the badminton, for having shared the few moments of spare time available to me.

Combining a PhD thesis with a professional career and a two years old child at home requires some personal sacrifices. Therefore, I would like to thank my family. My wife, Haixia Liu, thank you for your support, love and patience, I contribute this thesis first to you. My lovely daughters, Yilin Liao, thank you for your love and patience. For your inspiration and for I cannot accompany you as a father within these two years, I contribute this thesis to you as well. My parents and sister Meihong Liao, for your constant support. I owe everything to you, and for that, I contribute this thesis to you too. My parents in law, Shuiping He and Zuming Liu, thank you for taking care of Yilin Liao for so many hours and for your assistance at our home.

Gent, March 2012
Wenzhi Liao

Table of Contents

List of Figures

List of Tables

List of Acronyms

A

AP	Attribute Profile

C

CCA	Curvilinear Component Analysis
CCIPCA	Candid Covariance-Free Incremental Principle Component Analysis

E

EMP	Extended Morphological Profile
EAP	Extended Attribute Profile
EM	Expectation Maximization
EMPN	Extended Morphological Profile with No Reconstruction
EMPP	Extended Morphological Profile with Partial Reconstruction

F

FE	Feature Extraction
FIKPCA	Fast Iterative Kernel Principal Component Analysis

G

GSELD	Generalized Semi-supervised Local Discriminant Analysis
GDA	Generalized Discriminant Analysis
GHA	Generalized Hebbian Algorithm

I

ICA	Independent Component Analysis
IC	Independent Component

K

KPCA	Kernel Principal Component Analysis
KPC	Kernel Principal Component

L

LDA	Local Area Network
LLE	Locally Linear Embedding
LP	Laplacian Eigenmap
LTSA	Local Tangent Space Alignment
LPP	Locality Preserving Projection
LLTSA	Linear Local Tangent Space Alignment
LLFE	Local Linear Feature Extraction
LapSVM	Laplacian Support Vector Machine
LDC	Linear Discriminant Classifier

M

MP	Morphological Profile
MDS	Globus Monitoring and Discovery Service
MPLS	Multi Protocol Label Switching
MI	Mutual information

N

NPE	Neighborhood Preserving Embedding
NWFE	Nonparametric Weighted Feature Extraction
NN	Nearest Neighbor
MNF	Minimum Noise Fraction

P

PCA	Principle Component Analysis
PC	Principal Component

Q

QDC	Quadratic Discriminant Classifier

R

ROSIS	Reflective Optics System Imaging Spectrometer

S

SDA	Semi-supervised Discriminant Analysis
SELF	Semi-supervised Local Fisher Discriminant Analysis
SELD	Semi-supervised Local Discriminant Analysis
SVM	Support Vector Machine
SSL	Semi-Supervised Learning
SLR	Semi-supervised Logistic Regression
SE	Structuring Elements
SOM	Self Organizing Map

T

TSVM	Transductive Support Vector Machine

Nederlandse samenvatting –Summary in Dutch–

De recente technologische ontwikkelingen op het gebied van camera en andere sensoren hebben er toe geleid dat er steeds meer teledetectie data beschikbaar is en dit aan een steeds hogere spatiale en spectrale resolutie. Reeds vele technieken zijn ontwikkeld en getest om zowel de spectrale als de spatiale informatie die in deze data vervat zit, te verkennen. Zo worden vaak kenmerkextractietechnieken gebruikt om de hoge dimensionaliteit van hyperspectrale beelden te reduceren terwijl tegelijkertijd getracht wordt zoveel mogelijk van de spectrale informatie te behouden. Een populaire methode die gebruikt wordt bij het onderzoeken van de spatiale informatie is dan weer de methode van morfologische profielen.

Automatische classificatie technieken die gebruikt worden bij patroonherkenning gaan er gewoonlijk vanuit dat er voldoende trainingsvoorbeelden voorhanden zijn om een betrouwbaar en voldoende nauwkeurig model op te stellen voor de verschillende klassen. Deze veronderstelling is voor classificatieproblemen met hyperspectrale teledetectiebeelden echter maar zelden geldig. Het verzamelen van grondwaarheid voor dit soort data is namelijk een moeilijk en duur proces. Technieken die in staat zijn een betrouwbare classificatie uit te voeren op basis van slechts een beperkt aantal voorbeelden kunnen dus veel tijd en kosten uitsparen. De beperking van een kleine trainingsset is bijgevolg een heel belangrijk probleem in het veld van de hyperspectrale beeldclassificatie.

In recente jaren zijn er steeds meer teledetectiebeelden van stedelijke omgevingen beschikbaar aan zeer hoge spatiale resoluties. De classificatie van zulke beelden is bijzonder uitdagend. In stedelijke omgevingen worden immers veel verschillende materialen gebruikt (baksteen, asfalt, beton, metaal, vegetatie, . . .), maar vaak worden dezelfde materialen of (spectraal) sterk gelijkende materialen gebruikt voor verschillende functies (daken, wegen, parken, pleinen, . . .) . Er is dus geen één op één mapping tussen spectrale karakteristieken en functionele klassen. Bijgevolg is de spectrale informatie onvoldoende om een duidelijk onderscheid te maken tussen alle functionele klassen. Het is dus belangrijk ook de spatiale informatie mee in rekening te brengen om zo de classificatie nauwkeurigheid te verbeteren. Een van de meest populaire methoden om de spatiale informatie in hoge resolutie teledetectiebeelden te onderzoeken zijn morfologische profielen. Bij het gebruik van morfologische profielen in hyperspectrale data, moet men drie belangrijke punten in rekening brengen. Ten eerste, het gebruik van morfologische reconstructie bij het genereren van de morfologische profielen zorgt voor een aantal onverwachte en ongewenste resultaten. Ten tweede, de gegenereerde profielen leiden tot zeer grote data dimensies. En ten slotte, door het toepassen van lineaire kenmerkextractie methoden voor het reduceren van de dimensionaliteit van de hyperspectrale beelden vóór het construeren van morfologische profielen, gaat heel wat van de spectrale informatie verloren.

Om deze problemen op te lossen en de classificatie resultaten te verbeteren, hebben we effectieve kenmerkextractiealgoritmen ontwikkeld en combineren we morfologische kenmerken voor de classificatie van hyperspectrale teledetectiebeelden. De bijdragen van deze thesis zijn de volgende:

Als eerste bijdrage, wordt een nieuwe half-gesuperviseerde lokale discriminantanalyse methode (*semi-supervised local discriminant analysis,* SELD) voorgesteld voor het extraheren van kenmerken in teledetectiebeelden, waardoor de performantie in moeilijke condities verbetert. De voorgestelde methode combineert een niet-gesuperviseerde methode (*Local Linear Feature Extraction Methods,* LLFE) en een gesuperviseerde methode (*Linear Discriminant Analysis,* LDA) in een nieuw kader zonder enige vrije parameters. Het basisidee is om een optimale projectiematrix te construeren, die de lokale omgeving, afgeleid uit de niet gelabelde voorbeelden, bewaart en tegelijkertijd de discriminatie tussen de klassen, afgeleid uit de gelabelde voorbeelden, maximaliseert.

Onze tweede bijdrage is de toepassing van morfologische profielen met partiële reconstructie om de spatiale informatie in hyperspectrale teledetectiebeelden van stedelijke gebieden te beschrijven. Klassieke morfologische openingen en sluitingen zorgen ervoor dat er vervormingen plaatsvinden aan de randen van objecten. Daarom wordt meestal morfologische reconstructie toegepast, die deze randen herstelt. Dit proces heeft echter een aantal ongewenste neveneffecten. Objecten waarvan wegens hun vorm en grootte verwacht zou worden dat ze verdwijnen in een opening of sluiting met een bepaald structuurelement, blijven echter aanwezig wanneer gebruik gemaakt wordt van morfologische reconstructie. Het al dan niet verdwijnen van een object staat hierdoor niet meer in relatie met de grootte van het object. Morfologische profielen met partiële reconstructie daarentegen verbeteren zowel klassieke morfologische profielen als morfologische profielen met reconstructie. De vorm van objecten worden beter bewaard dan in het klassieke geval, terwijl de informatie over de grootte van de objecten beter gerepresenteerd wordt dan in morfologische profielen met reconstructie.

Een derde bijdrage is een nieuw half-gesuperviseerde kenmerkextractie kader voor het reduceren van de dimensie van de gegenereerde morfologische profielen. De morfologische profielen met structuurelementen van verschillende grootte en vorm produceren zeer hoog dimensionale data. Deze data bevat heel wat redundante informatie en vormen bijgevolg een grote uitdaging voor conventionele classificatie methoden, zeker voor diegenen die niet robuust zijn tegen het Hughes fenomeen. Voor zover wij weten, is dit de eerste keer dat half-gesuperviseerde kenmerkextractie wordt gebruikt voor het analyseren van morfologische profielen. De voorgestelde methode, veralgemeende half-gesuperviseerde lokale discriminant analyse (*generalized semi-supervised local discriminant analysis,* GSELD), is een uitbreiding van de SELD methode met een data gestuurde parameter.

Als vierde bijdrage, stellen we een snelle iteratieve kernel principale componenten analyse (*fast iterative kernel principal component analysis,* FIKPCA) voor om de dimensionaliteit van de hyperspectrale beelden te reduceren. In veel toepassingen, zorgen lineaire methoden voor kenmerkextractie, die gebruik maken van een lineaire projectie, ervoor dat niet-lineaire kenmerken van de data verloren gaan. Traditionele niet-lineaire methoden kunnen problemen veroorzaken op het gebied van opslagcapaciteiten en rekenkracht. De methode die we hier voorstellen is een kernel versie van de *Candid Covariance-Free Incremental Principal Component Analysis,* die de eigenvectoren schat via verschillende iteraties. Door de eigendecompositie van de Gram matrix te vermijden, kan onze methode de vereiste geheugencapaciteit en rekenkracht sterk verminderen.

Onze laatste bijdrage tenslotte construeert morfologische profielen met partiële reconstructie op basis van geëxtraheerde kenmerken verkregen met de niet-lineaire methode. In kenmerken verkregen met lineaire methodes, die traditioneel worden gebruikt, gaat te veel spectrale informatie verloren. De niet-lineaire kenmerken zijn beter geschikt om de hogere orde complexe en niet-lineaire distributies te beschrijven. In het bijzonder, hebben we onder andere de kernel principale componenten kenmerken gebruikt om de morfologische profielen te construeren, wat tot een significante verbetering heeft geleid van de classificatienauwkeurigheid.

De experimentele analyse die werd uitgevoerd met de nieuwe technieken die in deze thesis werden ontwikkeld, tonen een duidelijke verbetering van de classificatienauwkeurigheid in verschillende toepassingsdomeinen in vergelijking met andere state-of-the-art methoden.

Summary

Recent advances in sensor technology have led to an increased availability of hyperspectral remote sensing data at very high both spectral and spatial resolutions. Many techniques are developed to explore the spectral information and the spatial information of these data. In particular, feature extraction (FE) aimed at reducing the dimensionality of hyperspectral data while keeping as much spectral information as possible is one of methods to preserve the spectral information, while morphological profile analysis is the most popular methods used to explore the spatial information.

Hyperspectral sensors collect information as a set of images represented by hundreds of spectral bands. While offering much richer spectral information than regular RGB and multispectral images, the high dimensional hyperspectal data creates also a challenge for traditional spectral data processing techniques. Conventional classification methods perform poorly on hyperspectral data due to the curse of dimensionality (i.e. the Hughes phenomenon: for a limited number of training samples, the classification accuracy decreases as the dimension increases). Classification techniques in pattern recognition typically assume that there are enough training samples available to obtain reasonably accurate class descriptions in quantitative form. However, the assumption that enough training samples are available to accurately estimate the class description is frequently not satisfied for hyperspectral remote sensing data classification, because the cost of collecting ground-truth of observed data can be considerably difficult and expensive. In contrast, techniques making accurate estimation by using only small training samples can save time and cost considerably. The small sample size problem therefore becomes a very important issue for hyperspectral image classification.

Very high-resolution remotely sensed images from urban areas have recently become available. The classification of such images is challenging because urban areas often comprise a large number of different surface materials, and consequently the heterogeneity of urban images is relatively high. Moreover, different information classes can be made up of spectrally similar surface materials. Therefore, it is important to combine spectral and spatial information to improve the classification accuracy. In particular, morphological profile analysis is one of the most popular methods to explore the spatial information of the high resolution remote sensing data. When using morphological profiles (MPs) to explore the spatial information for the classification of hyperspectral data, one should consider three important issues. Firstly, classical morphological openings and closings degrade the object boundaries and deform the object shapes, while the morphological pro-

file by reconstruction leads to some unexpected and undesirable results (e.g. over-reconstruction). Secondly, the generated MPs produce high-dimensional data, which may contain redundant information and create a new challenge for conventional classification methods, especially for the classifiers which are not robust to the Hughes phenomenon. Last but not least, linear features, which are used to construct MPs, lose too much spectral information when extracted from the original hyperspectral data.

In order to overcome these problems and improve the classification results, we develop effective feature extraction algorithms and combine morphological features for the classification of hyperspectral remote sensing data. The contributions of this thesis are as follows.

1. As the first contribution of this thesis, a novel semi-supervised local discriminant analysis (SELD) method is proposed for feature extraction in hyperspectral remote sensing imagery, with improved performance in both ill-posed and poor-posed conditions. The proposed method combines unsupervised methods (Local Linear Feature Extraction Methods (LLFE)) and supervised method (Linear Discriminant Analysis (LDA)) in a novel framework without any free parameters. The underlying idea is to design an optimal projection matrix, which preserves the local neighborhood information inferred from unlabeled samples, while simultaneously maximizing the class discrimination of the data inferred from the labeled samples.

2. Our second contribution is the application of morphological profiles with partial reconstruction to explore the spatial information in hyperspectral remote sensing data from the urban areas. Classical morphological openings and closings degrade the object boundaries and deform the object shapes. Morphological openings and closings by reconstruction can avoid this problem, but this process leads to some undesirable effects. Objects expected to disappear at a certain scale remain present when using morphological openings and closings by reconstruction, which means that object size is often incorrectly represented. Morphological profiles with partial reconstruction improve upon both classical MPs and MPs with reconstruction. The shapes of objects are better preserved than classical MPs and the size information is preserved better than in reconstruction MPs.

3. A novel semi-supervised feature extraction framework for dimension reduction of generated morphological profiles is the third contribution of this thesis. The morphological profiles (MPs) with different structuring elements and a range of increasing sizes of morphological operators produce high-dimensional data. These high-dimensional data may contain redundant information and create a new challenge for conventional classification methods, especially for the classifiers which are not robust to the Hughes phenomenon. To the best of our knowledge the use of semi-supervised feature extraction methods for the generated morphological profiles has not been in-

vestigated yet. The proposed generalized semi-supervised local discriminant analysis (GSELD) is an extension of SELD with a data-driven parameter.

4. In our fourth contribution, we propose a fast iterative kernel principal component analysis (FIKPCA) to extract features from hyperspectral images. In many applications, linear FE methods, which depend on linear projection, can result in loss of nonlinear properties of the original data after reduction of dimensionality. Traditional nonlinear methods will cause some problems on storage resources and computational load. The proposed method is a kernel version of the Candid Covariance-Free Incremental Principal Component Analysis, which estimates the eigenvectors through iteration. Without performing eigen decomposition on the Gram matrix, our approach can reduce the space complexity and time complexity greatly.

5. Our last contribution constructs MPs with partial reconstruction on nonlinear features. Traditional linear features, on which the morphological profiles usually are built, lose too much spectral information. Nonlinear features are more suitable to describe higher order complex and nonlinear distributions. In particular, kernel principal components are among the nonlinear features we used to built MPs with partial reconstruction, which led to significant improvement in terms of classification accuracies.

The experimental analysis performed with the novel techniques developed in this thesis demonstrates an improvement in terms of accuracies in different fields of application when compared to other state of the art methods.

1
Introduction

1.1 Introduction

Hyperspectral [1] imaging collects and processes information from across the electromagnetic spectrum. Much as the human eye sees visible light in three bands (red, green, and blue), spectral imaging divides the spectrum into many more bands. This technique of dividing images into bands can be extended beyond the visible spectrum. With the development of technology, hyperspectral sensors have been widely applied in agriculture, mineralogy and surveillance. Hyperspectral sensors look at objects using a vast portion of the electromagnetic spectrum. Certain objects leave unique 'fingerprints' across the electromagnetic spectrum, see Fig. 1.1. These 'fingerprints' are known as spectral signatures and enable identification of the materials that make up a scanned object. For example, a spectral signature for oil helps mineralogists to find new oil fields.

Recently, with the advancement of sensors, hyperspectral imaging has emerged as a new modality in Earth imaging, leading to the definition of hyperspectral remote sensing. Remote sensing is the acquisition of information about an object or phenomenon, without making physical contact with the object. In modern usage, the term generally refers to the use of aerial sensor technologies to detect and classify objects on Earth (both on the surface, and in the atmosphere and oceans) by means of propagated signals (e.g. electromagnetic radiation emitted from aircraft or satellites) [2]. Hyperspectral remote sensing is a relatively new technology that is currently being investigated by researchers and scientists with regard to the

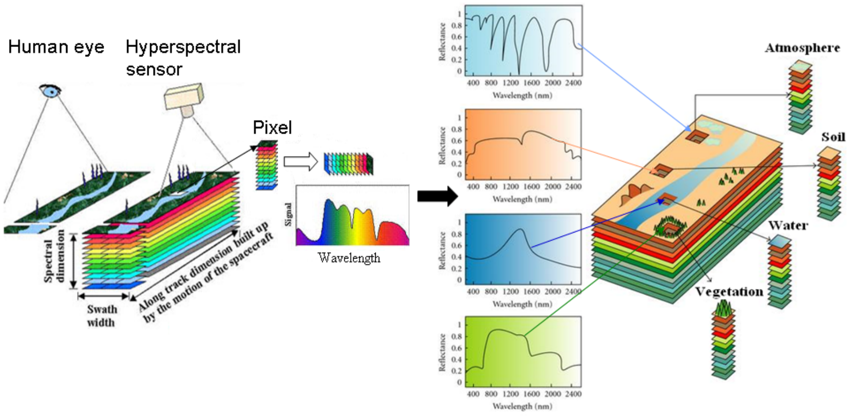


Figure 1.1: Each pixel in a hyperspectral image contains a continuous spectrum that is used to identify the materials present in the pixel.

detection and identification of minerals, terrestial vegetation, and man-made materials and backgrounds.

Hyperspectral sensors collect information as a set of ‘images’. Each image represents a range of the electromagnetic spectrum and is also known as a spectral band. These ‘images’ are then combined into a three-dimensional hyperspectral data cube for processing and analysis, see Fig. 1.2. Hyperspectral cubes are generated from airborne sensors like the NASA’s Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), or from satellites like NASA’s Hyperion [3].

The precision of these sensors is typically measured in spectral resolution, which is the width of each band of the spectrum that is captured. If the scanner detects a large number of fairly narrow frequency bands, it is possible to identify objects even if they are only captured in a handful of pixels. However, spatial resolution is a factor in addition to spectral resolution. If the pixels are too large, then multiple objects are captured in the same pixel and become difficult to identify. If the pixels are too small, then the energy captured by each sensor cell is low, and the decreased signal-to-noise ratio reduces the reliability of measured features.

1.1.1 Differences between hyperspectral and multispectral imaging

Hyperspectral imaging belongs to a class of techniques commonly referred to as spectral imaging or spectral analysis. Hyperspectral imaging is related to multispectral imaging. The distinction between hyper- and multi-spectral imaging is sometimes based on an arbitrary “number of bands” or on the type of measurement, depending on what is appropriate to the purpose.

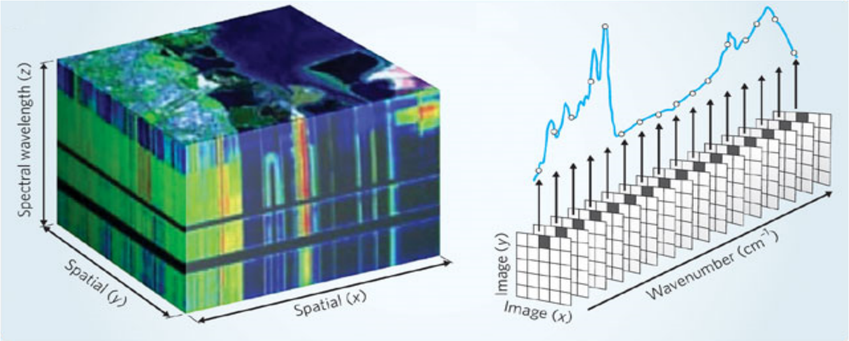


Figure 1.2: Hypercube.

Multispectral imaging deals with several images at discrete and somewhat narrow bands. Being "discrete and somewhat narrow" is what distinguishes multispectral imaging in the visible spectrum from color photography. A multispectral sensor may have many bands covering the spectrum from the visible to the long wave infrared. Multispectral images do not produce the "spectrum" of an object. Landsat is an excellent example.

Hyperspectral deals with imaging narrow spectral bands over a continuous spectral range, and produces the spectra of all pixels in the scene. So, a sensor with only 20 bands can also be hyperspectral when it covers the range from 500 to 700 nm with 20 bands each 10 nm wide. (While a sensor with 20 discrete bands covering the VIS, NIR, SWIR, MWIR, and LWIR would be considered multispectral).

'Ultraspectral imaging' could be reserved for interferometer type imaging sensors with a very fine spectral resolution. These sensors often have (but not necessarily) a low spatial resolution of several pixels only, a restriction imposed by the high data rate.

1.1.2 Applications of hyperspectral remote sensing.

Hyperspectral remote sensing is used in a wide range of applications. Although originally developed for mining and geology (the ability of hyperspectral imaging to identify various minerals makes it ideal for the mining and oil industries, where it can be used to look for ore and oil) [3, 4], it has now spread into fields as widespread as ecology and surveillance, as well as historical manuscript research, such as the imaging of the Archimedes Palimpsest. This technology is continually becoming more available to the public. Organizations such as NASA and the USGS have catalogues of various minerals and their spectral signatures, and have posted them online to make them readily available for researchers.

Hyperspectral remote sensing is increasing by used for monitoring the development and health of crops. For example, hyperspectral images are used to detect grape variety and to develop an early warning system for disease outbreaks [5]. Hyperspectral data can be used to detect the chemical composition of plants [6], and to detect the nutrient and water status of wheat in irrigated systems [7]. Another application in agriculture is the detection of animal proteins in compound feeds to avoid bovine spongiform encephalopathy (BSE), also known as mad-cow disease [8].

Hyperspectral remote sensing of minerals is well developed. Many minerals can be identified from airborne images, and their relation to the presence of valuable minerals, such as gold and diamonds, is well understood. Currently, progress is towards understanding the relationship between oil and gas leakages from pipelines and natural wells, and their effects on the vegetation and the spectral signatures [9, 10].

Hyperspectral imaging is frequently used in military surveillance too. Aerial surveillance with tethered balloons was used by French soldiers to spy on troop movements during the French Revolutionary Wars, and since that time, soldiers have learned not only to hide from the naked eye, but also to mask their heat signatures to blend into the surroundings and avoid infrared scanning. The idea that drives hyperspectral surveillance is that hyperspectral scanning draws information from such a large portion of the light spectrum that any given object should have a unique spectral signature in at least a few out of the many bands that are scanned.

Hyperspectral remote sensing has been used to monitor the environment. The Telops Hyper-Cam, an infrared hyperspectral imager, now offers the possibility of obtaining a complete image of emissions resulting from industrial smokestacks from a remote location, without any need for extractive sampling systems. Emission quantification measurements have been achieved with the Hyper-Cam which can now be used to independently, safely and rapidly identify and quantify polluting emissions from a remote location [11].

1.2 Challenges in hyperspectral data processing

While offering much richer spectral information than regular RGB and multispectral images, hyperspectral data cubes with large number of spectral bands create also a challenge for traditional data processing techniques:

The increasing number of spectral bands causes some problems with storage resources and computational load. Fast computers, sensitive detectors, and large data storage capacities are needed for analyzing hyperspectral data. Significant data storage capacity is necessary since hyperspectral cubes are large datasets, potentially exceeding tens of gigabytes. All of these factors greatly increase the cost of acquiring and processing hyperspectral data. Also, these high-dimensional hyperspectral data may contain redundant information.

The small sample size (SSS) problem [12] is an important issue for high-dimensional data classification. The SSS problem states that the number of available training samples is relatively much smaller than the dimensionality of the sample space. Remotely sensed hyperspectral image data, such as AVIRIS (Airborne Visible InfraRed Imaging Spectrometer) data [13–15] with hundreds of measured features (bands) potentially provide more accurate and detailed information for classification. Some other hyperspectral data from agriculture even have more than thousand spectral bands [16]. However, the cost of collecting ground-truth of remotely sensed hyperspectral image often requires a skilled expert agent to manually classify training examples. The cost associated with the labeling process thus may render a fully labeled training set infeasible.

Very high-resolution remotely sensed images from urban areas have recently become available. The classification of such images is challenging because urban areas often comprise a large number of different surface materials, and consequently the heterogeneity of urban images is relatively high. Moreover, different information classes can be made up of spectrally similar surface materials.

1.3 Overview

Some advanced classifiers, such as neural networks [17], SVM [18,19] and random forest classifiers [19], have been shown to deal efficiently with the problems of the high dimension and small sample size (SSS). The approach of [20] addresses a "K-nearest neighbor classifier based on adaptive nonparametric separability" with a distance metric formed by all the NWFE features. In recent years ensemble learning methods such as bagging [21], boosting [22,23], random subspace method (RSM) [24] and their variants have showed some appealing results for improving the classification performance of "weak classifiers" [25–27].

However, common statistical classifiers are often limited to deal with these cases. The increase in dimensionality of hyperspectral data and the limited number of labeled training samples may create a new challenge for conventional classification methods, especially for the classifiers which are not robust to the Hughes phenomenon [1] (for a limited number of training samples, the classification accuracy decreases as the dimension increases)). Moreover, with the increasing number of spectral bands, this hyperspectral data may contain redundant information. For this reason, feature extraction (FE) or feature selection, aiming at reducing the dimensionality of data, is a desirable preprocessing tool to reduce the dimensionality of hyperspectral data for classification. Relatively few bands can represent most information of the data, making feature extraction (FE) or feature selection very useful for classification of remote sensing data [28–32]. The feature selection method [31, 32] aims to select a suitable subset of the original set of features.

The most important issue relative to feature selection is to find an efficient search strategy for obtaining such a subset for classification. Most of the existing feature selection methods are generally suboptimal [30] because the number of all possible combinations is prohibitive, particularly for high-dimensional data classification. Search strategies to avoid exhaustive search are needed, and the selection of the optimal subset is therefore not guaranteed. Feature extraction uses all the features to construct a transformation that maps the original data to a low-dimensional subspace. The main advantage of feature extraction above feature selection is that no information of the original features needs to be wasted. Furthermore, feature extraction is easier than feature selection in some situations [30].

A number of approaches exist for feature extraction of hyperspectral images [28,33–36], ranging from unsupervised methods to supervised ones. Unsupervised FE methods do not require any prior knowledge or training data, even though they are not directly aimed at optimizing the accuracy in a given classification task [32]. One of the best known unsupervised methods is Principle Component Analysis (PCA) [37], which is widely used for hyperspectral images [33, 38, 39]. Wang and Chang [40] proposed three Independent Component Analysis (ICA) based dimensionality reduction methods for hyperspectral data. Wavelet transforms have been used in hyperspectral data dimensionality reduction [41, 42]. Wavelet transforms can preserve the high and low frequency features during the signal decomposition, hence preserving the spectral signatures. Plaza et al. [39] described sequences of extended morphological transformations for dimensionality reduction and classification of hyperspectral datasets. Harsanyi and Chang [43] investigated hyperspectral image classification and dimensionality reduction by using an orthogonal subspace projection approach. Phillips et al. [44] and He and Mei [45] used singular value decomposition and random projection, respectively, to reduce the dimensions of hyperspectral image data. Lower rank tensor approximation [46] and minimum change rate deviation [47] are proposed for hyperspectral image data by taking into account the spatial relation among neighboring image pixels. Recently, some local methods, which preserve the properties of local neighborhoods were proposed to reduce the dimensionality of hyperspectral images [33, 48–50], such as Locally Linear Embedding [48], Laplacian Eigenmap [51] and Local Tangent Space Alignment [52]. Their linear approximations, such as Neighborhood Preserving Embedding (NPE) [53], Locality Preserving Projection (LPP) [54] and Linear Local Tangent Space Alignment (LLTSA) [55] were recently applied to feature extraction in hyperspectral images [33, 56]. By considering neighborhood information around the data points, these local methods can preserve local neighborhood information and detect the manifold embedded in the high-dimensional feature space.

Supervised methods rely on the existence of labeled samples to infer class separability. Two widely used supervised feature extraction methods for hyperspectral images are the Fisher's Linear discriminant analysis (LDA) [57] and nonparametric weighted feature extraction (NWFE) [35]. Many extensions to both LDA and NWFE have been proposed in recent years, such as modified Fisher's linear discriminant analysis [58], regularized linear discriminant analysis [36], modified nonparametric weight feature extraction using spatial and spectral information [59], and kernel nonparametric weighted feature extraction [60].

In real-world applications, labeled data are usually very limited and labeling large amounts of data may sometimes require considerable human resources or expertise. On the other hand, unlabeled data are available in large quantities at very low cost. For this reason, semi-supervised methods [29, 61–66], which aim at improved classification by utilizing both unlabeled and limited labeled data gained popularity in the machine learning community. Some of the representative semi-supervised learning methods include Co-Training [62] and transductive SVM [63, 64], and Graph-based semi-supervised learning methods [65, 66]. Some semi-supervised feature extraction methods add a regularization term to preserve certain potential properties of the data. For example, semi-supervised discriminant analysis (SDA) [67] adds a regularizer into the objective function of LDA. The resulting method makes use of a limited number of labeled samples to maximize the class discrimination and employs both labeled and unlabeled samples to preserve the local properties of the data. The approach of [68] proposed a general semi-supervised dimensionality reduction framework based on pairwise constraints, which employs regularization with sparse representation. Other semi-supervised feature extraction methods combine supervised methods with unsupervised ones using a trade-off parameter, such as semi-supervised local Fisher discriminant analysis (SELF) [69]. However, it may not be easy to specify the optimal parameter values in these and similar semi-supervised techniques, as mentioned in [68, 69].

Very high-resolution remotely sensed images from urban areas have recently become available. The classification of such images is challenging because urban areas often comprise a large number of different surface materials, and consequently the heterogeneity of urban images is relatively high. Moreover, different information classes can be made up of spectrally similar surface materials. In this case, spatial information is very useful to improve the performances of classification. Many techniques are developed to explore the spatial information of the high resolution remote sensing data. In particular, mathematical morphology [70, 71] is one of the most popular methods. Pesaresi and Benediktsson [72] proposed the use of morphological transformations to build a morphological profile (MP). Bellens et al. [73] further explored this approach by using both disk-shaped and linear structuring elements to improve the classification of very high-resolution panchromatic urban imagery. The approach of [17] extended the method in [70] for hyperspectral data with high spatial resolution. The resulting method built the

MPs on the first principal components (PCs) extracted from a hyperspectral image, leading to the definition of extended MP (EMP). The appoach of [39] performs spectral-based morphology using the full hyperspectral image without dimensionality reduction. In [28], kernel principal components are used to construct the EMP, with significant improvement in terms of classification accuracies compared with the conventional EMP built on PCs. In [74], the attribute profiles (APs) [75] were applied to the first PCs extracted from a hyperspectral image, generating an extended AP (EAP). The approach of [76] improved the classification results by constructing the EAP with the independent component analysis.

However, classical morphological openings and closings degrade the object boundaries and deform the object shapes, which may result in losing some crucial information and introducing fake objects in the image. To avoid this problem, one often uses morphological openings and closings by reconstruction [17, 18, 72, 77, 78], which can reduce some shape noise in the image. However, morphological openings and closings by reconstruction lead to some unexpected results for remote sensing images, such as over-reconstruction, as was discussed in [73]. Objects which are expected to disappear at a certain scale remain present when using morphological openings and closings by reconstruction. The approach of [73] proposed a partial reconstruction for the classification of very high-resolution panchromatic urban imagery. Morphological openings and closings by partial reconstruction can solve the problem of over-reconstruction while preserving the shape of objects as much as possible.

1.4 Objectives and novel contributions of the thesis

The work presented in this thesis aims at investigating and defining novel techniques based on feature extraction for the classification of hyperspectral remote sensing images. State of the art techniques have already proven that the use of feature extraction and morphological features are effective for the classification of hyperspectral data. Nevertheless, several limitations exist (e.g., a very limited labeled samples, very high dimensionality of the data, very high-resolution of the data, high cost on storage and computation, etc.). The work presented in this dissertation attempts to overcoming those limitations.

The novel contributions of this thesis are as follows:

1. *Definition of a novel framework for semi-supervised feature extraction [79, 80].*

   The proposed semi-supervised local discriminant analysis (SELD) method combines unsupervised methods and supervised method without any free parameters. It can find the optimal projection matrix, which preserves the local neighborhood information, while simultaneously maximizing the class discrimination of the data.

2. *Application of morphological profiles with partial reconstruction [81] to hyperspectral remote sensing images.*

   In some applications simultaneous preserving of both size and shape information in the scene is desirable. Therefore, we have applied morphological profiles with partial reconstruction to the classification of very high-resolution hyperspectral data from the urban area.

3. *Pioneering the use of semi-supervised feature extraction to reduce the dimension of generated morphological profiles [81].*

   To the best of our knowledge, the use of semi-supervised feature extraction to reduce the dimension of generated morphological profiles was not yet reported in the remote sensing field before our work of [81].

4. *Application of a nonlinear feature extraction method based on fast iterative kernel principal component analysis to the classification of hyperspectral data [82].*

   High cost of storage and computational time limit the use of nonlinear methods in hyperspectral data. We proposed a fast iterative kernel principal component analysis to extend the limitations of some nonlinear methods by solving the eigenvectors through iteration.

5. *Investigation of extended morphological profiles with partial reconstruction built on kernel principal components [83].*

   In many applications, the preservation of both higher order complex and nonlinear distributions in the extracted features, which will be later used to constructed the extended morphological profiles, is desirable. Thus, we have investigated extended morphological profiles with partial reconstruction built on kernel principal components for the classification of very high-resolution hyperspectral data from the urban area.

1.5 Outline

This dissertation is organized in six chapters.

Some work related to ours is reviewed in Chapter 2, including some unsupervised feature extraction methods, supervised feature extraction methods and semi-supervised methods.

In Chapter 3, a novel semi-supervised feature extraction method, called semi-supervised local discriminant analysis (SELD), is described in detail. Experimental results on both synthetic data and real hyperspectral data are presented to demonstrate its performances.

Morphological profiles with partial reconstruction and proposed semi-supervised feature extraction, are shown in Chapter 4. Experimental results on hyperspectral data from the urban area demonstrates its performance.

The fast iterative kernel principal component analysis and extended morphological profiles with partial reconstruction built on kernel principal components is clearly deduced in Chapter 5. The standard kernel principal component analysis performs eigen decomposition on Gram matrix. Instead, the proposed fast iterative kernel principal component analysis solves the eigenvectors through iteration, which can reduce the space complexity and time complexity greatly. Extended morphological profiles with partial reconstruction built on kernel principal components were investigated with the demonstration of experimental results.

Chapter 6 presents a general discussion of the work described in this thesis reviewing the main contributions of this research. Specific concluding remarks on the research topics treated in the dissertation are also given. Perspectives on possible future developments of the work are presented.

1.6 Publications

1.6.1 Publications in international journals

1. **Wenzhi Liao**, Aleksandra Pižurica, Paul Scheunders, Wilfried Philips, Youguo Pi, "Semi-Supervised Local Discriminant Analysis for Feature Extraction in Hyperspectral Images," *IEEE Transactions on Geoscience and Remote Sensing*, accepted.

2. **Wenzhi Liao**, Rik Bellens, Aleksandra Pižurica, Wilfried Philips, Youguo Pi, "Classification of Hyperspectral Data over Urban Areas Using Directional Morphological Profiles and Semi-supervised Feature Extraction," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 4, 14 pages, 2012.

1.6.2 Publications in international conferences

1. **Wenzhi Liao**, Aleksandra Pižurica, Wilfried Philips, Youguo Pi, "A fast iterative kernel PCA feature extraction for hyperspectral images," *Proceedings of 2010 IEEE 17th International Conference on Image Processing (ICIP2010)*, Hongkong, China, pp. 1317-1320, 2010.

2. **Wenzhi Liao**, Aleksandra Pižurica, Wilfried Philips, Youguo Pi, "Feature extraction for hyperspectral image based on semi-supervised local discrim-

inant analysis," *Joint Urban Remote Sensing Event (JURSE 2011)*, Munich, Germany, pp. 401-402, 2011.

3. **Wenzhi Liao**, Rik Bellens, Aleksandra Pižurica, Wilfried Philips and Youguo Pi, "Classification of hyperspectral data over urban areas based on extended morphological profile with partial reconstruction, " *Proceedings of ACIVS 2012* , submitted.

2
Related work

Hyperspectral sensors collect information as a set of images represented by hundreds of spectral bands. While offering much richer spectral information than regular RGB and multispectral images, this large number of spectral bands creates a challenge for traditional spectral data processing techniques. Conventional classification methods perform poorly on hyperspectral data due to the curse of dimensionality (i.e. the Hughes phenomenon [1]: for a limited number of training samples, the classification accuracy decreases as the dimension increases). Feature extraction aims at reducing the dimensionality of hyperspectral data while keeping as much intrinsic information as possible. Relatively few bands can represent most information of the hyperspectral images [33], making feature extraction very useful for classification, detection and visualization of remote sensing data [29,33,34].

This Chapter presents the background and a brief overview on some related feature extraction methods for the classification of hyperspectral data.

2.1 Introduction

A number of approaches exist for feature extraction of hyperspectral images [28, 33–36], ranging from unsupervised methods to supervised ones. Unsupervised FE methods do not require any prior knowledge or training data, even though they are not directly aimed at optimizing the accuracy in a given classification task [32]. One of the best known unsupervised methods is Principle Component Analysis (PCA) [37], which is widely used for hyperspectral images [33, 38, 39].

The purpose of PCA is to reduce dimensionality according to what percentage of the overall variance can be captured. The kernel-based PCA (KPCA) is to find the directions by performing the PCA in the kernel feature space [84]. Independent component analysis (ICA) is a statistical technique for separating the independent signals from overlapping signals [85]. ICA is related to PCA but is more powerful and capable of finding the underlying factors or sources even when the principal-component approach fails. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-Gaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by [85]. Further techniques, based on image processing approaches, have been proposed in [86] and [87] by combining PCA/ICA and morphological transformations in the context of the classification of hyperspectral images of urban areas. Recently, Wang and Chang [40] proposed three Independent Component Analysis (ICA) based dimensionality reduction methods for hyperspectral data. They have shown better results using their methods than using PCA and MNF.

Wavelet transforms have been used in hyperspectral data dimensionality reduction [41,42]. Wavelet transforms can preserve the high and low frequency features during the signal decomposition, hence preserving the spectral signatures. Plaza et al. [39] described sequences of extended morphological transformations for dimensionality reduction and classification of hyperspectral datasets. Harsanyi and Chang [43] investigated hyperspectral image classification and dimensionality reduction by using an orthogonal subspace projection approach. Phillips et al. [44] and He and Mei [45] used singular value decomposition and random projection, respectively, to reduce the dimensions of hyperspectral image data. Lower rank tensor approximation [46] and minimum change rate deviation [47] are proposed for hyperspectral image data by taking into account the spatial relation among neighboring image pixels. Recently, some local methods, which preserve the properties of local neighborhoods were proposed to reduce the dimensionality of hyperspectral images [33,48–50], such as Locally Linear Embedding [48], Laplacian Eigenmap [51] and Local Tangent Space Alignment [52]. Their linear approximations, such as Neighborhood Preserving Embedding (NPE) [53], Locality Preserving Projection (LPP) [54] and Linear Local Tangent Space Alignment (LLTSA) [55] were recently applied to feature extraction in hyperspectral images [33, 56]. By considering neighborhood information around the data points, these local methods can preserve local neighborhood information and detect the manifold embedded in the high-dimensional feature space.

Supervised methods rely on the existence of labeled samples to infer class separability. Two widely used supervised feature extraction methods for hyperspectral images are the Fisher's Linear discriminant analysis (LDA) [57] and nonparametric weighted feature extraction (NWFE) [35].

Linear discriminant analysis (LDA) [88–92] is a powerful classical supervised feature-extraction method for classification, even if it has been proposed for over 70 years. It is also called the parametric feature extraction method in [90], since LDA uses the mean vector and covariance matrix of each class. Usually, *within-class*, *between-class*, and mixture scatter matrices are used to formulate the criterion of class separability. A kernel-based LDA called generalized discriminant analysis (GDA) was proposed by [93] using a kernel approach. There are three drawbacks of LDA.

1. One is that it works well only if the distributions of classes are normal-like distributions. When the distributions of classes are non normal like or multimodal mixture distributions, the performance of LDA is not satisfactory;

2. The second disadvantage of LDA is that the rank of the between-class scatter matrix is less than or equal to $C-1$, where C is the number of the classes in the image. Hence, assuming sufficient number of observations, the rank of within-class scatter matrix is $r \leq d$, then only r features can be extracted;

3. The third limitation is that, if the within-class covariance is singular, which often occurs in high-dimensional problems, LDA will have a poor performance on classification.

Lee and Landgrebe [94] proposed the Decision-Boundary Feature Extraction (DBFE) that can extract both discriminately informative features and discriminately redundant features from the decision boundary. The approach uses the training samples directly to determine the location of the decision boundary and employs information about the decision hypersurfaces associated with a given classifier to define an intrinsic dimensionality for the classification problem. Then, the corresponding optimal linear mapping can be obtained. NWFE was proposed in [35] to solve the problems of LDA. It also absorbs the idea of DBFE for determining the location of the decision boundary by training samples. The basic ideals of NWFE are asigning different weights on every sample to compute the "weighted means" and compute the distance between samples and their weighted means as their "closeness" to boundary, then defining nonparametric *between-class* and *within-class* scatter matrices which put large weights on the samples close to the boundary and deemphasize those samples far from the boundary. The experimental results of [17] and [87] show that NWFE outperforms LDA and DBFE. In [86] and [95], the authors suggest replacing DBFE by NWFE to obtain more effective features. Other papers show that NWFE outperforms LDA, approximated pairwise accuracy criterion linear dimension reduction, nonparametric discriminant analysis [35], and DBFE [96] in remote-sensing data sets.

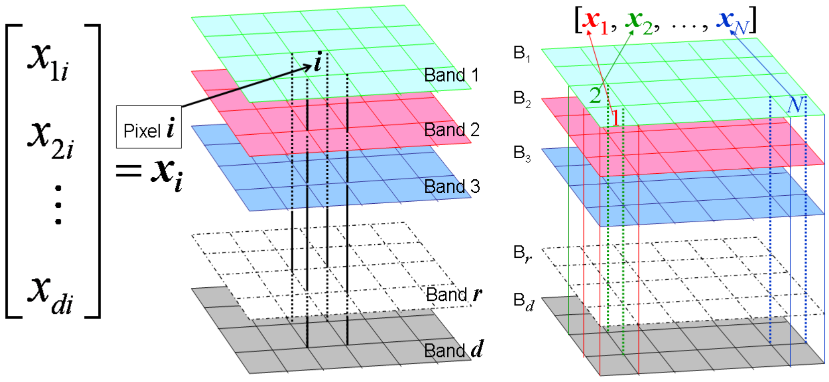
Many extensions to both LDA and NWFE have been proposed in recent years, such as modified Fisher's linear discriminant analysis [58], regularized linear discriminant analysis [36], modified nonparametric weight feature extraction using spatial and spectral information [59], and kernel nonparametric weighted feature extraction [60].

In real-world applications, labeled data are usually very scarce and labeling large amounts of data may sometimes require considerable human resources or expertise. On the other hand, unlabeled data are available in large quantities at very low cost. For this reason, semi-supervised methods [29, 61–66], which aim at improved classification by utilizing both unlabeled and limited labeled data gained popularity in the machine learning community.

Some of the representative semi-supervised learning methods include Co Training [62] and transductive SVM [63, 64], and Graph-based semi-supervised learning methods [65, 66]. Some semi-supervised feature extraction methods add a regularization term to preserve certain potential properties of the data. For example, semi-supervised discriminant analysis (SDA) [67] adds a regularizer into the objective function of LDA. The resulting method makes use of a limited number of labeled samples to maximize the class discrimination and employs both labeled and unlabeled samples to preserve the local properties of the data. The approach of [68] proposed a general semisupervised dimensionality reduction framework based on pairwise constraints, which employs regularization with sparse representation. Other semi-supervised feature extraction methods combine supervised methods with unsupervised ones using a trade-off parameter, such as semi-supervised local Fisher discriminant analysis (SELF) [69]. However, it may not be easy to specify the optimal parameter values in these and similar semi-supervised techniques, as mentioned in [68, 69].

2.2 Feature extraction for hyperspectral images

An image pixel vector $\mathbf{x}_i$ is composed of all pixel values $x_{1i}, x_{1i}, \cdots, x_{1N}$ at one corresponding pixel location of the hyperspectral image data, see Fig. 2.1(a). The dimension of that image vector is equal to the number of hyperspectral bands. For a hyperspectral image with N_R rows and N_C columns there will be $N = N_R \times N_C$ such vectors, namely $i = 1, 2, \cdots, N$, see Fig. 2.1(b). Let $\{\mathbf{x}_i\}_{i=1}^{N}$, $\mathbf{x}_i \in \Re^d$ denote high dimensional data, $\{\mathbf{z}_i\}_{i=1}^{N}$, and $\mathbf{z}_i \in \Re^r$ its low dimensional representations with $r \leq d$. In our application, d is the number of spectral bands of hyperspectral images, and r is the dimensionality of the projected subspace. The assumption is that there exists a mapping function $f : \Re^d \rightarrow \Re^r$, which can map every original data point $\mathbf{x}_i$ to $\mathbf{z}_i = f(\mathbf{x}_i)$ such that most information of the high dimensional data is kept in a much lower dimensional projected subspace. This mapping is usually represented by a $d \times r$ projection matrix $\mathbf{W}$:



(a) (b)

Figure 2.1: Hyperspectral image, (a) a pixel vector; (b) transfer the 3D hypercube into 2D matrix.

$$\mathbf{z}_i = f(\mathbf{x}_i) = \mathbf{W}^T \mathbf{x}_i \tag{2.1}$$

$$\mathbf{z}_i = \begin{bmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{ri} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1r} & \cdots & w_{1d} \\ w_{21} & w_{22} & \cdots & w_{2r} & \cdots & w_{2d} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{r1} & w_{r2} & \cdots & w_{rr} & \cdots & w_{rd} \end{bmatrix} \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ri} \\ \vdots \\ x_{di} \end{bmatrix}$$

where pixel vector $\mathbf{z}_i (i = 1, 2, \cdots, N)$ will form the first r bands of the extracted features.

In many feature extraction methods, the projection matrix $\mathbf{W}$ can be obtained by solving the following optimization problem, where $\mathbf{w}$ denotes one of the columns in the projection matrix $\mathbf{W}$:

$$\mathbf{w}_{opt} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \overline{\mathbf{S}} \mathbf{w}}{\mathbf{w}^T \underline{\mathbf{S}} \mathbf{w}} \tag{2.2}$$

The matrices $\overline{\mathbf{S}}$ and $\underline{\mathbf{S}}$ have specific meaning in different methods as we discuss later in the text. The solution to (2.2) is equivalent to solving the following generalized eigenvalue problem:

$$\overline{\mathbf{S}}\mathbf{w} = \lambda \underline{\mathbf{S}}\mathbf{w} \tag{2.3}$$

Or equivalently:

$$\underline{\mathbf{S}}^{-1}\overline{\mathbf{S}}\mathbf{w} = \lambda \mathbf{w} \tag{2.4}$$

The projection matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ is made up by the *r* eigenvectors of the matrix $\underline{\mathbf{S}}^{-1}\overline{\mathbf{S}}$ associated with the largest *r* eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

2.3 Unsupervised feature extraction methods

Unsupervised feature extraction methods deal with cases where no labeled samples are available, aiming to find another representation of the data in lower dimensional space by satisfying some given criterion. One of the best known unsupervised methods is Principle Component Analysis (PCA) [37], which is widely used for hyperspectral images [33, 38, 39]. Recently, some local methods, which preserve the properties of local neighborhoods were proposed to reduce the dimensionality of hyperspectral images [33, 48–50], such as Locally Linear Embedding [48], Laplacian Eigenmap [51] and Local Tangent Space Alignment [52]. Their linear approximations, such as Neighborhood Preserving Embedding (NPE) [53], Locality Preserving Projection (LPP) [54] and Linear Local Tangent Space Alignment (LLTSA) [55] were recently applied to feature extraction in hyperspectral images [33,56]. By considering neighborhood information around the data points, these local methods can preserve local neighborhood information and detect the manifold embedded in the high-dimensional feature space.

2.3.1 PCA

Principal Component Analysis (PCA) [37] performs feature extraction through analyzing the covariance matrix of the original data. The eigenvalues of the covariance matrix are considered to be an indicator of the information content. Large values suggest more information content and low values indicate the presence of mostly noise. Due to its low complexity and the absence of parameters, PCA has been widely used for feature extraction in hyperspectral images [17]. In mathematical terms, PCA attempts to find a linear mapping **w** that:

$$\begin{aligned} &\max \mathbf{w}^T \mathbf{S}_t \mathbf{w} \\ S.t. \quad &\mathbf{w}^T \mathbf{w} = \mathbf{I} \end{aligned} \tag{2.5}$$

where the covariance matrix is $\mathbf{S}_t = \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^T$, $\mathbf{u} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ is the mean of the vector, **I** and **1** are the identity matrix. The constraint can be enforced by introducing a Lagrange multiplier λ. Therefore, an unconstrained maximization is performed as:

$$\max f(\mathbf{w}) = \mathbf{w}^T \mathbf{S}_t \mathbf{w} + \lambda(\mathbf{1} - \mathbf{w}^T \mathbf{w}) \tag{2.6}$$

Component	Eigenvalue (%)	Cumulative (%)
1	72.48	72.48
2	24.64	97.13
3	1.73	98.86
4	0.37	99.23
5	0.2	99.42

Table 2.1: Eigenvalues and cumulative variance in percentages for AVIRIS Indian Pines with 220 bands.

w can be got by differentiating the above function to **w** and setting the result to 0, which results in:

$$\begin{aligned}\frac{df(\mathbf{w})}{d\mathbf{w}} = \frac{d}{d\mathbf{w}}(\mathbf{w}^T\mathbf{S}_t\mathbf{w} + \lambda(\mathbf{1} - \mathbf{w}^T\mathbf{w})) &= 0\\ \Rightarrow \mathbf{S}_t\mathbf{w} - \lambda\mathbf{w} &= 0\\ \Rightarrow \mathbf{S}_t\mathbf{w} &= \lambda\mathbf{w}\end{aligned}$$

The projection matrix $\mathbf{W}_{PCA} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ can be optimized as follows:

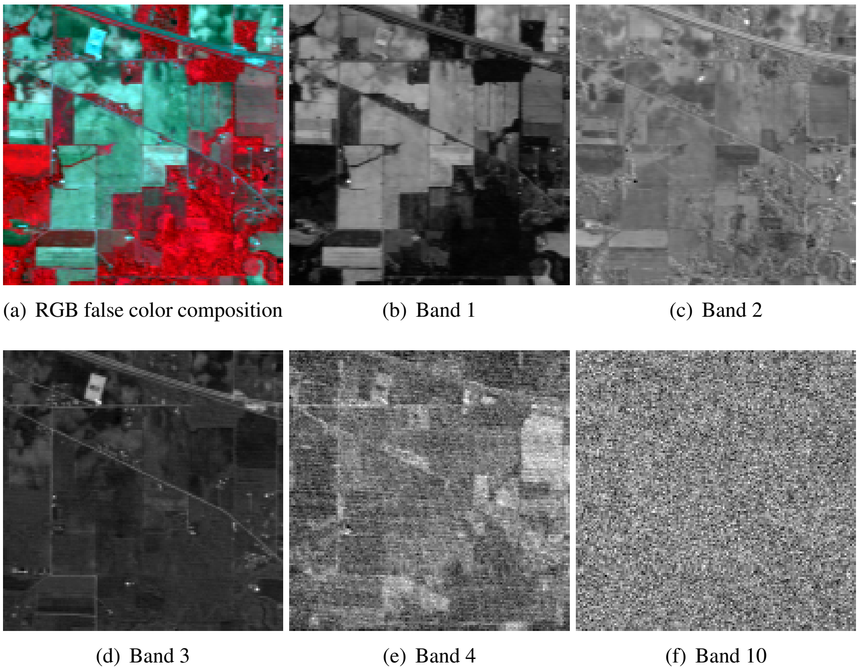
$$\mathbf{w}_{PCA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_t\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \tag{2.7}$$

By setting $\overline{\mathbf{S}} = \mathbf{S}_t$ and $\underline{\mathbf{S}} = \mathbf{I}$, we obtain the projection matrix $\mathbf{W}_{PCA} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ as in (2.2). The features extracted by PCA have the highest contrast or variance in the first band and the lowest contrast or variance in the last band. Therefore, the first r PCA bands often contain the majority of information residing in the original hyperspectral images and can be used for more effective and accurate analyses because the number of image bands and the amount of image noise involved are reduced, see Fig. 2.2 and Table. 2.1.

2.3.2 LLFE

The above subsection presented PCA method for feature extraction which attempts to retain global properties of the data. In contrast, local nonlinear techniques like Isomap [97], Local Linear Embedding (LLE) [98, 99], Laplacian Eigenmaps (LE) [51] and Local Tangent Space Analysis (LTSA) [100], try to find a low-dimensional data representation by preserving local properties of the manifold. They applications to hyperspectral data can be found in [49]. However, when using these nonlinear local methods, one always encounter the following two problems:

1. “Out of sample” problem: this is a phenomenon in such that new samples cannot be projected onto the manifold constructed by training samples;


(a) RGB false color composition (b) Band 1 (c) Band 2

(d) Band 3 (e) Band 4 (f) Band 10

Figure 2.2: Sample PCA bands of the AVIRIS Indian Pines.

2. Huge cost in computation and memory consumption, especially for high resolution hyperspectral images with large samples and lines.

Recently, Chang and Yeung [101] proposed robust locally linear embedding for nonlinear dimensionality reduction, and they demonstrated that the method is better suited for dealing with outliers. In order to speed up this step, the approach of [48] only calculated the distance of the current pixel with those pixels that are within a square neighbourhood window centered at the pixel. However, the "Out of sample" problem cannot be avoided, especially when encountering high resolution hyperspectral images with large samples and lines.

More recently, some of their linear approximations, such as Neighborhood Preserving Embedding (NPE) [53], Locality Preserving Projection (LPP) [54] and Linear Local Tangent Space Alignment (LLTSA) [55] were proposed, and applied to feature extraction in hyperspectral images [33, 56]. These local linear feature extraction (LLFE) [52–54] methods can overcome the "out of sample" problem, as well as inherit the local geometry preserving property.

As a linear approximation to the LLE, Neighborhood Preserving Embedding (NPE) [53] preserved the local properties of the data manifold by writing the high-dimensional datapoints as a linear combination of their nearest neighbors. In the low-dimensional representation of the data, NPE attempts to retain the reconstruction weights in the linear combinations as well as possible. NPE consists of the following three steps:

1. Constructing an adjacency graph: Let G denote a graph with n nodes. The ith node corresponds to the data point $\mathbf{x}_i$. There are two ways to construct the neighborhood graph:

   - K nearest neighbors (KNN): If $\mathbf{x}_j$ is among the K nearest neighbors of $\mathbf{x}_i$, connect nodes i and j;
   - ϵ neighborhood: If $||\mathbf{x}_i - \mathbf{x}_j|| < \epsilon$, connect nodes i and j.

   The graph constructed by KNN nearest neighbors is a directed graph, while the one constructed by the ϵ neighborhood is an undirected graph. In many real world applications, it is difficult to choose a good ϵ. In this work, we adopt the KNN method to construct the graph.

2. Computing the weights : The weights on the edges were computed in this step. Let $\mathbf{Q}$ denote the weight matrix with Q_{ij} having the weight of the edge from node i to node j, and 0 if there is no such edge. Then the reconstruction weights Q_{ij} are calculated by minimizing the reconstruction error, which results from approximating $\mathbf{x}_i$ by its e nearest neighbors:

$$\begin{aligned} \min \sum_i ||\mathbf{x}_i - \sum_{j=1}^{e} Q_{ij}\mathbf{x}_j||^2 \\ S.t. \quad \sum_{j=1}^{e} Q_{ij} = 1 \end{aligned} \tag{2.8}$$

3. Computing the Projections: The extracted features $\mathbf{z}_i$ in the low-dimensional projected subspace that best preserve the local neighborhood information are then obtained as:

$$\begin{aligned} \min \sum_i ||\mathbf{z}_i - \sum_{j=1}^{e} Q_{ij}\mathbf{z}_j||^2 \\ S.t. \quad \mathbf{z}_i^T\mathbf{z}_i = \mathbf{I} \end{aligned} \tag{2.9}$$

   The projection matrix $\mathbf{W}_{NPE} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ can be optimized as follows:

$$\mathbf{w}_{NPE} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{w}} \tag{2.10}$$

where $\mathbf{M} = (\mathbf{I} - \mathbf{Q})^T(\mathbf{I} - \mathbf{Q})$ and **I** represents the identity matrix. Then the extracted feature vector is as follows:

$$\mathbf{z}_i = \mathbf{W}_{NPE}^T \mathbf{x}_i \tag{2.11}$$

Locality Preserving Projection (LPP) [54] is a linear approximation of the LP method. The local properties of LPP method are preserved based on the pairwise distances between near neighbors. LPP computes a low-dimensional representation of the data in which the distances between a datapoint and its e nearest neighbors are minimized. This is done in a weighted manner, i.e., the distance in the low-dimensional data representation between a datapoint and its first nearest neighbor contributes more to the cost function than the distance between the datapoint and its second nearest neighbor. The algorithmic procedure can be summarized below:

1. Constructing an adjacency graph: Let G denote a graph with n nodes. The ith node corresponds to the data point $\mathbf{x}_i$. There are two ways to construct the neighborhood graph:

   - K nearest neighbors (KNN): If $\mathbf{x}_j$ is among the K nearest neighbors of $\mathbf{x}_i$, connect nodes i and j;
   - ϵ neighborhood: If $||\mathbf{x}_i - \mathbf{x}_j|| < \epsilon$, connect nodes i and j.

   The graph constructed by KNN nearest neighbors is a directed graph, while the one constructed by the ϵ neighborhood is an undirected graph. In many real world applications, it is difficult to choose a good ϵ. In this work, we adopt the KNN method to construct the graph.

2. Computing the weights : The weights on the edges were computed in this step. Let **Q** denote the weight matrix with Q_{ij} having the weight of the edge from node i to node j, and 0 if there is no such edge. Then the weights Q_{ij} can be calculated by:

   - Heat kernel. $Q_{ij} = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma}}$, if nodes i and j are connected.
   - Single minded. $Q_{ij} = 1$, if nodes i and j are in the their nearest neighborhood.

3. Computing the projections: In the computation of the low-dimensional representations $\mathbf{z}_i$, the cost function is given as:

$$\min \sum_{ij} ||\mathbf{z}_i - \mathbf{z}_j||^2 Q_{ij} \tag{2.12}$$
$$S.t. \quad \mathbf{z}_i^T \mathbf{D} \mathbf{z}_i = \mathbf{I}$$

Where $\mathbf{D}$ is a diagonal matrix; its entries are the column (or row, since $\mathbf{Q}$ is symmetric) sum of $\mathbf{Q}$, $D_{ii} = \sum_j Q_{ij}$. Large weights Q_{ij} mean small distances between the nodes i and j. Hence, the difference between their low-dimensional representations $\mathbf{z}_i$ and $\mathbf{z}_j$ highly contributes to the cost function. As a consequence, nearby points in the high-dimensional space are projected closer together in the low-dimensional representation. The projection matrix $\mathbf{W}_{LPP} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ can be optimized as follows:

$$\mathbf{w}_{LPP} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}} \tag{2.13}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{Q}$ is the Laplacian matrix [54]. Then the extracted features are the following:

$$\mathbf{z}_i = \mathbf{W}_{LPP}^T\mathbf{x}_i \tag{2.14}$$

In local linear tangent space analysis (LLTSA) [55], the local geometry is described by the local tangent space of each data point. let θ_i of dimensionality d be the local tangent coordinates of $\mathbf{x}_i$. It relates to the global coordinates $\mathbf{z}_i$ by an affine transformation $\mathbf{z}_i\mathbf{H} = \mathbf{L}_i\theta_i + \mathbf{E}_i$, where $\mathbf{L}_i \in \Re^{d\times d}$ is the transformation matrix, $\mathbf{H} = \mathbf{I} - \mathbf{e}\mathbf{e}^T/k$ is a $k \times k$ centering matrix, and $\mathbf{E}_i \in \Re^{d\times r}$ is the reconstruction error matrix. The error is minimized to retain the local geometry in the embedded space via the objective function [55]:

$$\min \sum_i ||\mathbf{E}_i^*||_F^2 = \sum_i ||\mathbf{z}_i\mathbf{U}_i||_F^2 \qquad S.t. \quad \mathbf{z}_i^T\mathbf{z}_i = \mathbf{I} \tag{2.15}$$

where $\mathbf{U}_i = \mathbf{H}(\mathbf{I} - \theta_i^T(\theta_i\theta_i^T)^{-1}\theta_i)$. Minimizing this cost function also becomes the eigenvalue problem, where the $\mathbf{B}$, referred to the alignment matrix, which is constructed with $\mathbf{B}(I_i, I_i) \leftarrow \mathbf{B}(I_i, I_i) + \mathbf{U}_i\mathbf{U}_i^T$ (I_i is the indexes of $\mathbf{x}_i$). The projection matrix $\mathbf{W}_{LLTSA} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ can be optimized as follows:

$$\mathbf{w}_{LLTSA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{w}} \tag{2.16}$$

Then the extracted feature vector is as follows:

$$\mathbf{z}_i = \mathbf{W}_{LLTSA}^T\mathbf{x}_i \tag{2.17}$$

The reasoning behind LLFE is that neighbouring points in the high-dimensional space $\Re^d$ are likely to have similar representation in the low-dimensional projected subspace $\Re^r$ as well, see Fig. 2.3. Therefore, LLFE methods preserve the local neighborhood information of the data in the low-dimensional representation.

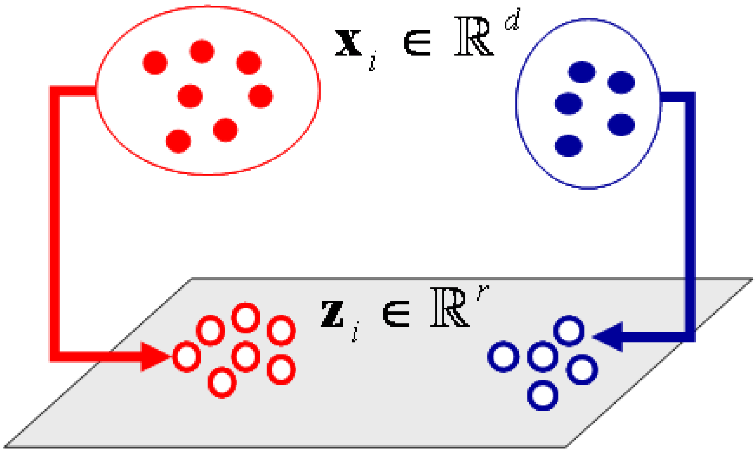

$\mathbf{x}_i \in \mathbb{R}^d$
$\mathbf{z}_i \in \mathbb{R}^r$

Figure 2.3: Basic idea of LLFE.

2.4 Supervised feature extraction methods

Supervised methods rely on the existence of labeled samples to infer class separability. Two widely used supervised feature extraction methods in hyperspectral data are the Fisher Linear discriminant analysis (LDA) [57] and nonparametric weighted feature extraction (NWFE) [35]. Many extensions to these two methods have been proposed in recent years, such as modified Fisher's linear discriminant analysis [58], regularized linear discriminant analysis [36], modified nonparametric weight feature extraction using spatial and spectral information [59], and kernel nonparametric weighted feature extraction [60].

2.4.1 LDA

The best known supervised method is Linear Discriminant Analysis (LDA) [57], which seeks projection directions on which the ratio of the *between-class* covariance to *within-class* covariance is maximized. Taking the label information into account, LDA results in a linear transformation $\mathbf{z}_i = f(\mathbf{x}_i, y_i) = \mathbf{W}^T\mathbf{x}_i$, where y_i is the label of the data point $\mathbf{x}_i$. The corresponding projection matrix $\mathbf{W}_{LDA} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ is optimized as follows:

$$\mathbf{w}_{LDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \tag{2.18}$$

where

$$\mathbf{S}_b = \sum_{k=1}^{C} n_k(\mathbf{u}^{(k)} - \mathbf{u})(\mathbf{u}^{(k)} - \mathbf{u})^T \tag{2.19}$$

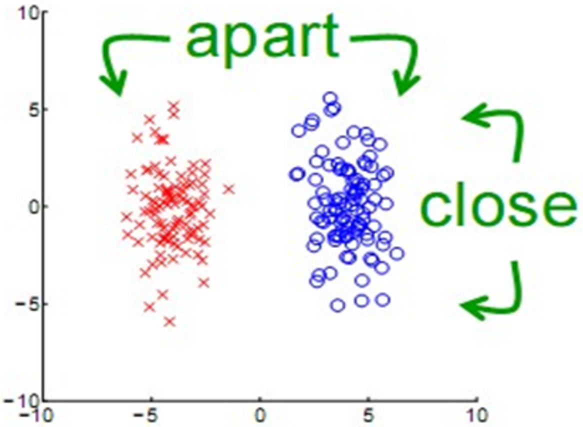


Figure 2.4: Basic idea of LDA.

and

$$\mathbf{S}_w = \sum_{k=1}^{C}(\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \mathbf{u}^{(k)})(\mathbf{x}_i^{(k)} - \mathbf{u}^{(k)})^T) \tag{2.20}$$

where n_k is the number of samples in the kth class, $\mathbf{u}$ is the mean of the entire training set, $\mathbf{u}^{(k)}$ is the mean of the kth class, $\mathbf{x}_i^{(k)}$ is the ith sample in the kth class. $\mathbf{S}_b$ is called the *between-class* scatter matrix and $\mathbf{S}_w$ the *within-classs* scatter matrix. (2.19) is equivalent to

$$\mathbf{w}_{LDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_t\mathbf{w}} \tag{2.21}$$

with

$$\mathbf{S}_t = \sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^T \tag{2.22}$$

form (2.19), (2.20) and (2.22), we have $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$.

By setting $\overline{\mathbf{S}} = \mathbf{S}_b$ and $\underline{\mathbf{S}} = \mathbf{S}_w$ or $\underline{\mathbf{S}} = \mathbf{S}_t$, we obtain the projection matrix $\mathbf{W}_{LDA} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ as in (2.2). LDA seeks projection direction on which the data points within the same class are close while separating all the data points from different classes apart, see Fig. 2.4. However, as the rank of the *between-class* scatter matrix $\mathbf{S}_b$ is $C-1$, LDA can extract at most $C-1$ features, which may not be sufficient to represent essential information of the original data.

2.4.2 NWFE

Similar to LDA, nonparametric weighted feature extraction (NWFE) [35] aims to find the feature space in which *between-class* scatter matrix $\mathbf{S}_b$ is maximized and *within-classs* scatter matrix $\mathbf{S}_w$ is minimized simultaneously. The main idea of NWFE [35] is placing different weights on every sample to compute the ''weighted means" and then applying the distances between samples and their weighted means as their closeness to boundary. Additionally, NWFE addressed a regularized *within-classs* scatter matrix for alleviating the singularity. As a result, NWFE prevents the disadvantages of LDA and obtains satisfactory results. The *between-class* scatter matrix $\mathbf{S}_b^{NW}$ and the *within-classs* scatter matrix $\mathbf{S}_w^{NW}$ of NWFE are defined as:

$$\mathbf{S}_b^{NW} = \sum_{i=1}^{L} P_i \sum_{j=1 j \neq i}^{L} \sum_{k=1}^{n_k} \frac{\eta_k^{i,j}}{n_k} (\mathbf{x}_k^i - M_j(\mathbf{x}_k^i))(\mathbf{x}_k^i - M_j(\mathbf{x}_k^i))^T \tag{2.23}$$

$$\mathbf{S}_w^{NW} = \sum_{i=1}^{L} P_i \sum_{k=1}^{n_k} \frac{\eta_k^{i,j}}{n_k} (\mathbf{x}_k^i - M_i(\mathbf{x}_k^i))(\mathbf{x}_k^i - M_i(\mathbf{x}_k^i))^T \tag{2.24}$$

where the scatter matrix weight $\eta_k^{i,j}$ is defined by:

$$\eta_k^{i,j} = \frac{d(\mathbf{x}_k^i, M_j(\mathbf{x}_k^i))^{-1}}{\sum_{t=1}^{n_k} d((\mathbf{x}_t^i, M_j(\mathbf{x}_t^i))^{-1}} \tag{2.25}$$

and the weight mean is:

$$M_j(\mathbf{x}_k^i) = \sum_{t=1}^{n_j} \gamma_{kt}^{i,j} \mathbf{x}_t^j \tag{2.26}$$

and

$$\gamma_{kt}^{i,j} = \frac{d(\mathbf{x}_k^i, \mathbf{x}_t^j)^{-1}}{\sum_{t=1}^{n_k} d((\mathbf{x}_k^i, \mathbf{x}_t^j)^{-1}} \tag{2.27}$$

The projection matrix of NWFE $\mathbf{W}_{NWFE} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ is optimized as follows:

$$\mathbf{w}_{NWFE} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_b^{NW} \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w^{NW} \mathbf{w}} \tag{2.28}$$

To reduce the effect of the cross products of within-class distances and prevent the singularity, the *within-classs* scatter matrix is regularized by:

$$\mathbf{S}_w^{NW} = 0.5\mathbf{S}_w^{NW} + 0.5 diag(\mathbf{S}_w^{NW}) \tag{2.29}$$

where $diag(\mathbf{A})$ means the diagonal parts of matrix A. The steps of NWFE can be summarized as:

1. Compute the distances between each pair of training samples and form the distance matrix;

2. Calculate the $\gamma_{kt}^{i,j}$ using the distance matrix;
3. Compute the weighted means $M_j(\mathbf{x}_k^i)$ using the $\gamma_{kt}^{i,j}$ calculated in step 2;
4. Compute the scatter matrix weight $\eta_k^{i,j}$;
5. Compute the $\mathbf{S}_b^{NW}$ and regularized $\mathbf{S}_w^{NW}$;
6. Extract features using $\mathbf{z}_i = \mathbf{W}_{NWFE}^T \mathbf{x}_i$.

The NWFE overcomes the limitations of LDA, in which the number of extracted features is depended on the number of classes. However, compared to the fast LDA method, NWFE consumed more computational time as the number of training samples increases. This is because NWFE uses all the training samples to compute $\eta_k^{i,j}$, $\gamma_{kt}^{i,j}$ and $M_j(\mathbf{x}_k^i)$, which results in quite time-consuming for large sample size problem.

When only a small number of labeled samples are available, the performance of supervised feature extraction methods tend to be degraded. Thus, the supervised methods overfit feature spaces to the labeled samples.

2.5 Semi-supervised feature extraction methods

In computer science, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data and a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many applications have shown that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy [63, 64, 67, 69]. The acquisition of labeled data for a learning problem often requires a skilled human agent to manually classify training examples. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is much cheaper. In such situations, semi-supervised learning can be of great practical value.

Recently, semi-supervised feature extraction methods have been proposed and applied to pattern recognition [67, 69]. The idea behind these methods is to infer the class discrimination from labeled samples, as well as the local neighborhood information from both labeled and unlabeled samples. Some semi-supervised feature extraction methods add a regularization term to preserve certain potential properties of the data. For example, semi-supervised discriminant analysis (SDA) [67] adds a regularizer into the objective function of LDA. The resulting method makes use of a limited number of labeled samples to maximize the class discrimination and employs both labeled and unlabeled samples to preserve the local properties of the data. The approach of [68] proposed a general semisupervised dimensionality reduction framework based on pairwise constraints, which employs regularization with sparse representation. Other semi-supervised feature extraction methods combine supervised methods with unsupervised ones using a trade-off parameter, such as semi-supervised local Fisher discriminant analysis (SELF) [69].

2.5.1 SDA

LDA seeks the optimal projections purely on the training (labeled) set. When there are not enough training samples, overfitting may happen. In reality, it is possible to acquire a large set of unlabeled data. In order to prevent overfitting, Semi-supervised Discriminant Analysis (SDA) [67] imposes a regularizer in LDA, and extends LDA to incorporate the manifold structure inferred from the unlabeled data. In particular, the projection matrix is:

$$\mathbf{w}_{SDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_t \mathbf{w} + \alpha J(\mathbf{w})} \tag{2.30}$$

where $J(\mathbf{w})$ is the regularizer, which is the core part in SDA. The parameter α controls the influence of local neighborhood information; for $\alpha = 0$, SDA reduces to LDA.

In SDA, the regularizer $J(\mathbf{w})$ incorporates the manifold structure by constructing a graph in such a way that nearby pixels have similar embeddings (low-dimensional representations), which is similar to classification, namely nearby pixels are likely to have the same label. The weight matrix built on the unlabeled samples is defined as: $Q_{ij} = 1$, if $\mathbf{x}_j$ is in the k nearest neighbors of $\mathbf{x}_i$, otherwise, $Q_{ij} = 0$. Then, the regularizer $J(\mathbf{w})$ can be defined as:

$$J(\mathbf{w}) = \sum_{ij} (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_j)^2 Q_{ij} \tag{2.31}$$

This means that two pixels are likely to be in the same class, if they are linked by an edge. Moreover, their low-dimensional representations are likely to have the same labels.

$$\begin{aligned} J(\mathbf{w}) &= \sum_{ij} (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_j)^2 Q_{ij} \\ &= 2\sum_{i} \mathbf{w}^T \mathbf{x}_i D_{ii} \mathbf{x}_i^T \mathbf{w} - 2\sum_{ij} \mathbf{w}^T \mathbf{x}_i S_{ij} \mathbf{x}_j^T \mathbf{w} \\ &= 2\mathbf{w}^T \mathbf{X}(\mathbf{D} - \mathbf{S})\mathbf{X}^T \mathbf{w} \\ &= 2\mathbf{w}^T \mathbf{X}\mathbf{L}\mathbf{X}^T \mathbf{w} \end{aligned} \tag{2.32}$$

where **D** is a diagonal matrix; its entries are column sum of **Q**, $D_{ii} = \sum_j Q_{ij}$, or row sum of **Q**, since **Q** is symmetric. $\mathbf{L} = \mathbf{D} - \mathbf{Q}$ is the Laplacian matrix [67].

SDA finds a projection which respects the discriminant structure inferred from the labeled data points, as well as the intrinsic geometrical structure inferred from both labeled and unlabeled data points. Specifically, the labeled data points, combined with the unlabeled data points, are used to build a graph incorporating neighborhood information of the data set. The graph provides a discrete approximation to the local geometry of the data manifold. Using the notion of graph Laplacian, a smoothness penalty on the graph can be incorporated into the objective function. In this way, SDA optimally preserves the manifold structure. However, SDA has the same limitation as LDA in the number of extracted features, because the rank of the *between-class* matrix $\mathbf{S}_b$ is $C - 1$.

2.5.2 SELF

Semi-Supervised Local Fisher Discriminant Analysis (SELF) [69] combines linearly PCA and local Fisher discriminant analysis (LFDA) [102]:

$$\mathbf{w}_{SELF} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T[(1-\beta)\mathbf{S}_{lb} + \beta\mathbf{S}_t]\mathbf{w}}{\mathbf{w}^T[(1-\beta)\mathbf{S}_{lw} + \beta\mathbf{I}]\mathbf{w}} \tag{2.33}$$

where $\mathbf{S}_{lb}$ and $\mathbf{S}_{lw}$ are local *between-class* scatter matrix and local *within-class* scatter matrix [102], $\beta(\in [0,1])$ is a trade off parameter, which controls the contribution of the supervised method LFDA and unsupervised method PCA. By setting β to a value between zero and one, SELF can separate samples from different classes while maximizing the variance of the data inferred from both the labeled and unlabeled samples. SELF overcomes some limitations of LDA and SDA (it can extract as much features as the number of the dimensions).

2.6 Conclusion

In this chaper, we briefly reviewed some related feature extraction methods. We explained that unsupervised methods do not rely on the labeled information, but may not discover the class discrimination in the data sets. Supervised methods rely on the labeled information and can separate different classes, but tend to turn to overfit when labeled information is insufficient. Some existing semi-supervised methods can overcome these problems, but usually it is not easy to optimize their parameters.

3 SELD

We propose a novel semi-supervised local discriminant analysis (SELD) method for feature extraction in hyperspectral remote sensing imagery, with improved performance in both ill-posed and poor-posed conditions. The proposed method combines unsupervised methods (Local Linear Feature Extraction Methods (LLFE)) and supervised method (Linear Discriminant Analysis (LDA)) in a novel framework without any free parameters. The underlying idea is to design an optimal projection matrix, which preserves the local neighborhood information inferred from unlabeled samples, while simultaneously maximizing the class discrimination of the data inferred from the labeled samples. Experimental results on synthetic data and real hyperspectral data demonstrate that the proposed method compares favorably with conventional feature extraction methods in the following aspects:

1. No tradeoff parameters to be optimized. The proposed method combines unsupervised methods (LLFE) and supervised method(LDA) in a novel framework without any free parameters;

2. Discrimination maximized and local neighborhood information well preserved;

3. Statistically significant improvement in overall classification accuracy. The McNemar's tests based upon the standardized normal test statistic [103] show the statistical significance of the improvements resulting from the proposed SELD;

4. Less computational cost. The proposed SELD is more efficient than NWFE, SDA and SELF, especially when the number of training samples increases.

3.1 Introduction

In the remote sensing literature, many supervised and unsupervised classifiers have been developed to tackle the multi- and hyperspectral data classification problem [104]. Supervised methods, such as artificial neural networks [105–107] readily revealed inefficient when dealing with a high number of spectral bands, and thus in the recent years, kernel-based methods in general and support vector machines (SVMs) [84, 108] in particular have been successfully used for hyperspectral image classification [109–112]. Certainly, kernel-based classifiers are able to handle large input spaces efficiently, and deal with noisy samples in a robust way [113]. However, the main difficulty with all supervised methods is that the learning process heavily depends on the quality of the training data sets, which is only useful for simultaneous images, or for images with the same classes taken under the same conditions. Even worse, the training set is frequently not available, or in a very reduced number, given the very high cost of true sample labeling. On the other hand, unsupervised methods have demonstrated good results [114–119] in multi- and hyperspectral image classification. Unsupervised methods are not sensitive to the number of labeled samples since they work on the whole image, but the relationship between clusters and classes is not ensured.

In such situations, semi-supervised learning (SSL) [29, 61–66] gained popularity in the machine learning community. In computer science, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data in hyperspectral data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many applications have shown that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy [29, 61–67, 69]. The acquisition of labeled data for a learning problem often requires a skilled human agent to manually classify training examples. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is much cheaper.

The framework of semi-supervised learning is very active in remote sensing and has recently attracted a considerable amount of research [120–122]. Essentially, three different classes of SSL algorithms are encountered in the literature: (1) Generative models, which involve estimating the conditional density $p(x|y)$, such as expectation maximization (EM) algorithms with finite mixture models [123], which have been extensively applied in the context of remotely sensed image classification [124]. (2) Low density separation algorithms, which maximize the margin for labeled and unlabeled samples simultaneously, such as Transductive SVM [63], which have been recently applied to hyperspectral image classification [125]. (3) Graph-based methods [126, 127], in which each sample communicates its label information to its neighbors until a global stable state is achieved on the whole dataset.

Most semi-supervised variants of SVM suffer from a high computational burden and consequently a limited number of unlabeled samples can be used for their training. This gives rise to a poor estimation of the marginal data distribution. Many heuristic approaches have been proposed to reduce the computational cost of the TSVM. In [128], a mixed integer programming was proposed to find the labeling with the lowest objective function. The optimization, however, is intractable for large data sets. In [129], a heuristic that iteratively solves a convex SVM objective function with alternate labeling of unlabeled samples was proposed. Yet, the algorithm is capable of dealing with a few thousand samples only. The improved TSVM still has a cubic cost, and requires storing huge kernel matrices [130]. Several alternative proposals exist for the Laplacian SVM (LapSVM), either by using a sparse manifold regularizer [131] or by using an ℓ_1 penalization term and a regularization path algorithm [132]. A second and important problem with LapSVM is related to the use of a functional form of the Laplacian eigenmaps, which yields a constrained optimization problem that is hard to solve.

On a different note, in most SSL methods, unlabeled data is integrated directly in the dual problem, often in an ad-hoc manner, e.g., via a regularizer, which may lack an intuitive interpretation. Convexity is also a concern for TSVM and related methods. Finally, for most SVM variants the issue of tackling classification problems for a vast number of categories has not been solved entirely. These methods use one-versus-all schemes and majority voting, but this approach is neither natural nor well-motivated. The semi-supervised logistic regression (SLR) algorithm [133], which is founded on information-theoretic principles, was proposed to solve most of the aforementioned problems. SLR allows a natural interpretation of model weights, and has a convex loss function which is a significant advantage. In particular, SLR is based on modifications to the penalty functions of the generalized maximum entropy (MaxEnt) objective in the primal, such that the expectations of similarity features over local regions are consistent. These modifications along with the minimization of the Kullback-Leibler divergence yield the SLR loss. Encoding prior knowledge, e.g., label proportions, is straightforward and scalability is also ensured via sparse similarity features.

Graph-based methods have been lately attracting a lot of attention because of their solid mathematical background, their relationship with kernel methods, sparseness properties, model visualization, and good results in many areas. The algorithms are provided with some available labeled information in addition to the unlabeled information, thus allowing to encode some knowledge about the geometry and the shape of the data set. This idea of exploring the shape of the marginal distribution in the data set can be applied in kernel target detection in order to deform the “measure” of distance in the kernel space according to the geometry of the neighboring pixels. The approach of [134] extended the semi-supervised graph-based method presented in [135] to the classification of hyperspectral image. It preserves the contextual information through the use of composite kernels, which have been recently revealed very useful to improve inductive support vector machines (SVMs) [65, 112, 136]. Semi-supervised kernel Orthogonal Subspace Projection (S^2KOSP) was proposed for target detection applications [137], which introduces an additional regularization term on the geometry of both labeled and unlabeled samples by using the graph Laplacian. The information from unlabeled samples is included in the standard kernel Orthogonal Subspace Projection by means of the graph Laplacian with a contextual unlabeled sample selection mechanism.

Compared to the semi-supervised classifiers, semi-supervised feature extraction methods try to find a projection using very limited number of labeled samples and a large number of unlabeled samples [29, 67–69]. Some semi-supervised feature extraction methods add a regularization term to preserve certain potential properties of the data. For example, semi-supervised discriminant analysis (SDA) [67] adds a regularizer into the objective function of LDA. The regularizer based on graph Laplacian regularization aims to enforce nearby points to have similar representations in the low dimensional feature space. Therefore, the resulting method makes use of a limited number of labeled samples to maximize the class discrimination and employs both labeled and unlabeled samples to preserve the local neighborhood properties of the data.

The approach of [29] proposed a novel semi-supervised feature selection method for the classification of hyperspectral images. It aims at selecting a subset of the original set of features that exhibits at the same time high capability to discriminate among the considered classes and high invariance in the spatial domain of the investigated scene. The feature selection in this method was accomplished by defining a multi-objective criterion function made up of the following two terms: 1) A term that measures the class separability; 2) A term that evaluates the spatial invariance of the selected features. A parameter was used to combine these two terms, which results in the possibility to evaluate in a more flexible way the tradeoffs between discrimination ability among classes and spatial invariance of each feature subset and to identify the subsets of features that simultaneously exhibit both properties.

Some semi-supervised feature extraction methods combine supervised methods with unsupervised ones using a trade-off parameter, such as semi-supervised local Fisher discriminant analysis (SELF) [69], which bridges LFDA and PCA by a parameter so that it can smoothly control the reliance on the global structure of unlabeled samples and class information brought by labeled samples.

The approach of [68] proposed a general semi-supervised dimensionality reduction framework based on pairwise constraints, which employs regularization with sparse representation. It was based on new prior information, i.e., pairwise constraints which specify whether a pair of examples belongs to the same class or not. The resulting methods used a parameter to tradeoff of the following two terms: 1) A discrimination term that assesses the separability between classes; 2) regularization term that characterizes some property of the original data set.

However, it may not be easy to specify the optimal parameter values in these and similar semi-supervised techniques, as mentioned in [68, 69].

In this Chapter, we propose a novel semi-supervised local discriminant analysis (SELD) method to reduce the dimensionality of the hyperspectral images. The proposed SELD method aims to find a projection which can preserve local neighborhood information and maximize the class discrimination of the data. We combine an unsupervised method (from the class of Local Linear Feature Extraction Methods (LLFE), such as NPE, LPP and LLTSA) and a supervised method LDA in a novel framework without any tuning parameters. Contrasting to related semi-supervised methods, such as SELF [69], we do not combine supervised and unsupervised methods linearly. Instead of using both labeled and unlabeled samples together, we first divide the samples into two sets: labeled and unlabeled. Then we employ the labeled samples through the supervised method (LDA) only and the unlabeled ones through an unsupervised, locality preserving method (LLFE) only.

We propose a natural way to combine unsupervised and supervised methods without any free parameters, making fully the use of strengths of both approaches in different scenarios. The supervised component maximizes class discrimination (for the available number of labeled samples) and the local unsupervised component ensures neighborhood information preservation. While we employ the LLFE [53–55] and LDA [57] methods, this novel framework can be applied in combination with other supervised and unsupervised methods too. Another advantage is that our method can extract as many features as the number of spectral bands. This also increases classification accuracy with respect to methods where the number of extracted features is limited by the number of classes (LDA and SDA). The results demonstrate improved classification accuracy when compared to related semi-supervised methods.

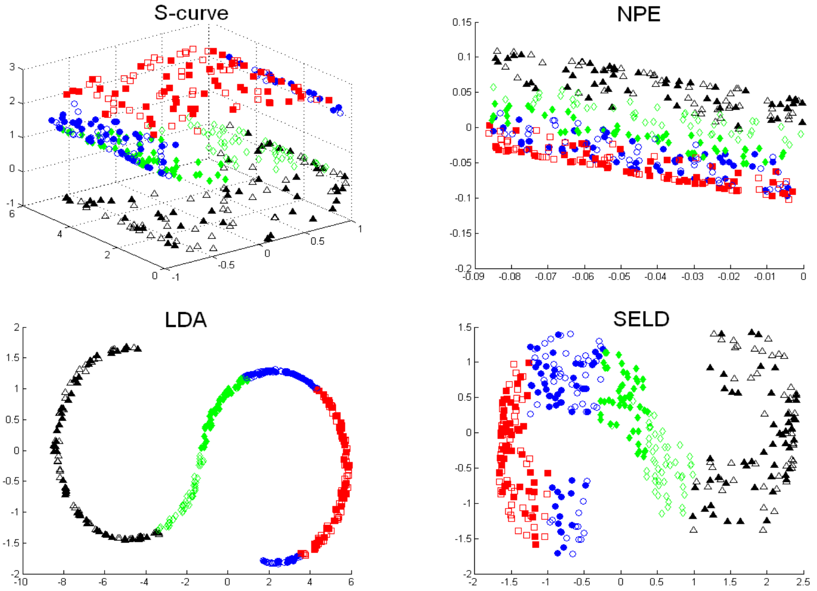


Figure 3.1: Examples of feature extraction by LDA, NPE and the proposed SELD method. Three dimensional S-curve data with four different classes are embedded into a two dimensional subspace. Each class has 100 samples with 50 labeled (filled), 50 unlabeled (unfilled).

3.2 Proposed semi-supervised local discriminant analysis (SELD)

As discussed above, some semi-supervised methods, such as SDA and SELF, can achieve a good class discrimination and preserve the local properties of the data with properly optimized parameters. One important issue is how to optimize tuning parameters, which is common to most of the related semi-supervised methods like [68, 69]. One solution is to employ cross-validation for this purpose. However, except for the computational cost of parameter optimization, cross-validation is not reliable when the number of labeled samples is small [102] (which is sometimes the real case in hyperspectral images). Focusing on class discrimination, LDA is in general well suited to preprocessing for the task of classification, since the transformation improves class separation. However, when only a small number of labeled samples are available, LDA tends to perform poorly due to overfitting. LLFE works directly on the data without any ground truth, and incorporates the local neighborhood information of data points in its feature extraction process.

Motivated by these facts, we propose a novel semi-supervised approach, which combines LLFE and LDA methods in a way that adapts automatically to the frac-

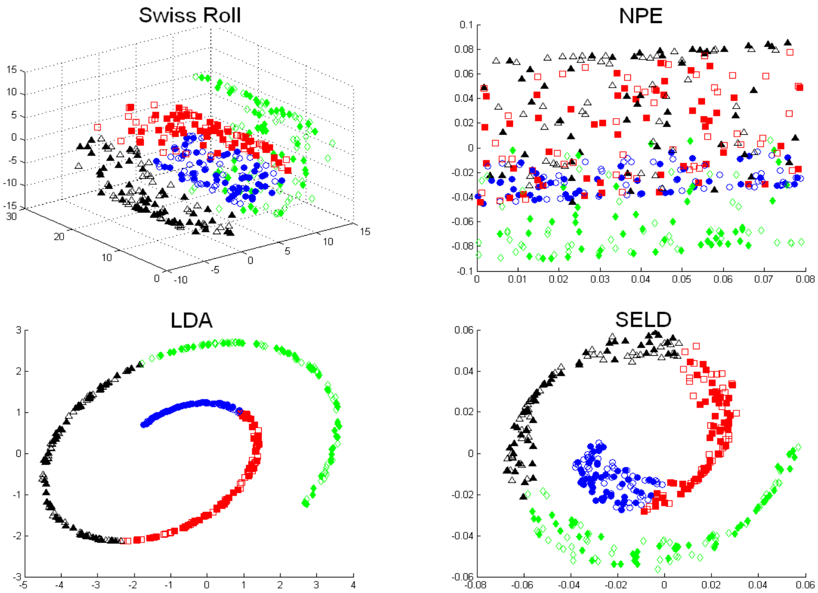


Figure 3.2: Examples of feature extraction by LDA, NPE and the proposed SELD method. Three dimensional Swiss data with four different classes are embedded into a two dimensional subspace. Each class has 100 samples with 50 labeled (filled), 50 unlabeled (unfilled).

tion of the labeled samples without any parameters. The main idea of our approach is to first divide the samples into two sets: labeled and unlabeled. The labeled samples will be used only by LDA (to maximize the class discrimination), and the unlabeled ones only through LLFE (to preserve the local neighborhood information). This will yield a natural way to combine the two as we show next.

3.2.1 Reformulation of supervised LDA and unsupervised LLFE

Suppose a training data set $\mathbf{X}$ is made up of the labeled set $\mathbf{X}_{labeled} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, $y_i \in \{1, 2, \cdots, C\}$, where C is the number of classes, and the unlabeled set $\mathbf{X}_{unlabeled} = \{\mathbf{x}_i\}_{i=n+1}^{N}$ with u unlabeled samples, $N = n + u$, $\mathbf{X} = \{\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}\} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{x}_{n+1}, \cdots, \mathbf{x}_N\}$. The kth class has n_k samples with $\sum_{k=1}^{C} n_k = n$. Without loss of generality, we center the data points by subtracting the mean vector from all the sample vectors, and assume that the labeled samples in $\mathbf{X}_{labeled} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ are ordered according to their labels, with the data matrix of the kth class $\mathbf{X}^{(k)} = \{\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \cdots, \mathbf{x}_{n_k}^{(k)}\}$ where $\mathbf{x}_i^{(k)}$ is the ith sample in the kth class. Then the labeled set can be expressed as $\mathbf{X}_{labeled} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(C)}\}$. We can reformulate the *between-class* scatter matrix as:

$$\begin{aligned}\mathbf{S}_b^{'} = \sum_{k=1}^{C} n_k(\mathbf{u}^{(k)})(\mathbf{u}^{(k)})^T &= \sum_{k=1}^{C} n_k(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i^{(k)})(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i^{(k)})^T \\ &= \sum_{k=1}^{C}\mathbf{X}^{(k)}\mathbf{P}^{(k)}(\mathbf{X}^{(k)})^T\end{aligned}$$

where $\mathbf{P}^{(k)}$ is the $n_k \times n_k$ matrix with all the elements equal to $\frac{1}{n_k}$. If we define a $n \times n$ matrix $\mathbf{P}_{n\times n}$ as:

$$\mathbf{P}_{n\times n} = \begin{bmatrix}\mathbf{P}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}^{(C)}\end{bmatrix}$$

the *between-class* scatter matrix $\mathbf{S}_b^{'}$ can be written as:

$$\mathbf{S}_b^{'} = \sum_{k=1}^{C}\mathbf{X}^{(k)}\mathbf{P}^{(k)}(\mathbf{X}^{(k)})^T = \mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T \tag{3.1}$$

By subtracting the *between-class* scatter matrix from the total scatter matrix $\mathbf{S}_t^{'}$, the *within-class* scatter matrix $\mathbf{S}_w^{'}$ is obtained as:

$$\begin{aligned}\mathbf{S}_w^{'} &= \mathbf{S}_t^{'} - \mathbf{S}_b^{'} \\ &= \mathbf{X}_{labeled}(\mathbf{X}_{labeled})^T - \mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T \\ &= \mathbf{X}_{labeled}(\mathbf{I}_{n\times n} - \mathbf{P}_{n\times n})(\mathbf{X}_{labeled})^T\end{aligned} \tag{3.2}$$

In our approach, the LDA component will use the labeled samples only (to maximize the class discrimination), so we reformulate (2.18) as:

$$\begin{aligned}\mathbf{w}_{LDA}^{'} &= \arg\max_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{S}_b^{'}\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w^{'}\mathbf{w}} \\ &= \arg\max_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}_{labeled}(\mathbf{I}_{n\times n} - \mathbf{P}_{n\times n})(\mathbf{X}_{labeled})^T\mathbf{w}}\end{aligned} \tag{3.3}$$

As unsupervised method, local linear feature extraction (LLFE) [52–54] methods reviewed in [138] seek a projection direction on which neighborhood data points in the high-dimensional feature space $\Re^d$ are kept on neighborhood in the low-dimensional projected subspace $\Re^r$ as well. By considering neighborhood information around the data points, the goal of these methods is to preserve the

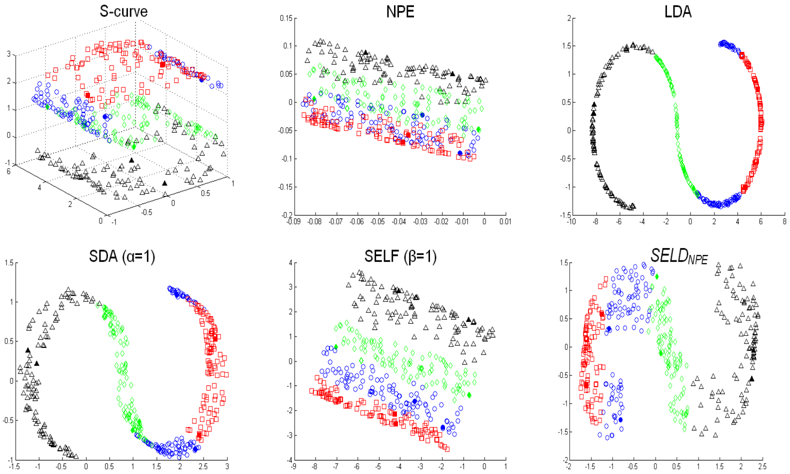


Figure 3.3: Comparison of feature extraction by each method on S-curve. Three dimensional data with four different class are embedded into a two dimensional subspace. Each class has 100 samples, filled/unfiled symbols denote labeled/unlabeled samples, the labeled samples in data sets are 2 for each class, the rest are unlabeled samples. For SDA, the parameter α is optimized in $\{0, 0.1, 0.5, 1, 5$ and $10\}$, and for SELF the parameter β is optimized in $\{0, 0.25, 0.5, 0.75$ and $1\}$.

local properties of the original data. Although the LLFE methods in [53–55] have some characteristic differences [53,55], they are all linear approximations to local nonlinear feature extraction methods and share more or less the same technique of linearization. The optimal solution of all these three methods can be computed by eigen-decomposition. We can express the optimal projection matrix of all LLFE methods from [53–55] in a unified way, so that it only uses the unlabeled samples:

$$\mathbf{w}^{'}_{LLFE} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{X}_{unlabeled} \overline{\mathbf{C}}_{u\times u} (\mathbf{X}_{unlabeled})^T \mathbf{w}}{\mathbf{w}^T \mathbf{X}_{unlabeled} \underline{\mathbf{C}}_{u\times u} (\mathbf{X}_{unlabeled})^T \mathbf{w}} \tag{3.4}$$

For NPE [53], $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{M}$. For LPP [54], $\overline{\mathbf{C}} = \mathbf{D}$ and $\underline{\mathbf{C}} = \mathbf{L}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{L}$ is the Laplacian matrix [54]. For LLTSA [55], $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{B}$, where $\mathbf{B}$ is the alignment matrix [55]. By setting $\overline{\mathbf{S}} = \mathbf{X}_{unlabeled} \overline{\mathbf{C}} \mathbf{X}^T_{unlabeled}$ and $\underline{\mathbf{S}} = \mathbf{X}_{unlabeled} \underline{\mathbf{C}} \mathbf{X}^T_{unlabeled}$, we obtain the projection matrix $\mathbf{W}_{LLFE} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ as in (2.2). The reasoning behind LLFE is that neighbouring points in the high-dimensional space $\Re^d$ are likely to have similar representation in the low-dimensional projected subspace $\Re^r$ as well. Therefore, LLFE methods preserve the local neighborhood information of the data in the low-dimensional representation.

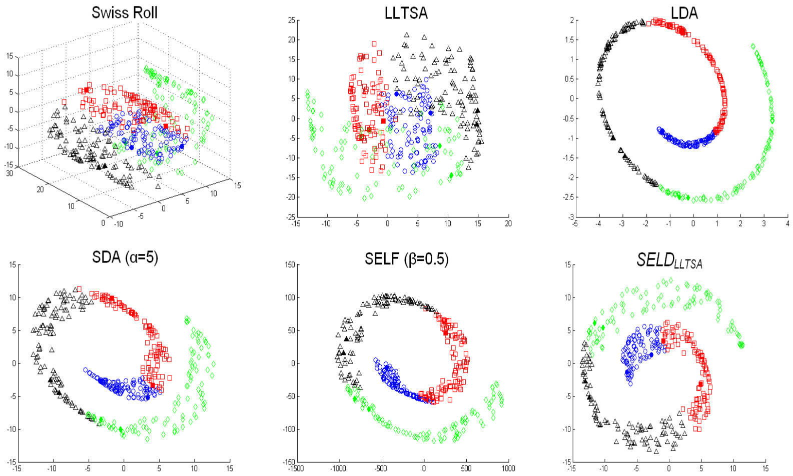


Figure 3.4: Comparison of feature extraction by each method on Swiss Roll data sets. Three dimensional data with four different class are embedded into a two dimensional subspace. Each class has 100 samples, filled/unfiled symbols denote labeled/unlabeled samples, the labeled samples in data sets are 2 for each class, the rest are unlabeled samples. For SDA, the parameter α is optimized in $\{$0, 0.1, 0.5, 1, 5 and 10$\}$, and for SELF the parameter β is optimized in $\{$0, 0.25, 0.5, 0.75 and 1$\}$.

3.2.2 SELD

We define the following matrics:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{n\times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad \bar{\mathbf{I}} = \begin{bmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

$$\underline{\mathbf{C}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \underline{\mathbf{C}}_{u\times u} \end{bmatrix} \quad \overline{\mathbf{C}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{C}}_{u\times u} \end{bmatrix}$$

Now the reformulated optimization problems of LDA and LLFE in (3.3) and (3.4) can be written as follows:

$$\mathbf{w}^{'}_{LDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\mathbf{P}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}(\bar{\mathbf{I}} - \mathbf{P})\mathbf{X}^T\mathbf{w}} \tag{3.5}$$

$$\mathbf{w}^{'}_{LLFE} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\overline{\mathbf{C}}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}\underline{\mathbf{C}}\mathbf{X}^T\mathbf{w}} \tag{3.6}$$

Note that full data vector $\mathbf{X}$ appears in (3.5), (3.6) but due to the structure of the matrices $\mathbf{P}$, $\bar{\mathbf{I}}$, $\underline{\mathbf{C}}$ and $\overline{\mathbf{C}}$, the LDA (3.5) makes use of the labeled samples only and LLFE (3.6) makes use of the unlabeled samples only. In order to make full use of the strengths of both two methods without parameter optimization, we propose a natural way to combine them as:

$$\mathbf{w}_{SELD} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \overline{\mathbf{S}}_{SELD} \mathbf{w}}{\mathbf{w}^T \underline{\mathbf{S}}_{SELD} \mathbf{w}} \tag{3.7}$$

where

$$\begin{aligned}\overline{\mathbf{S}}_{SELD} &= \mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T + \mathbf{X}_{unlabeled}\overline{\mathbf{C}}_{u\times u}(\mathbf{X}_{unlabeled})^T \\ &= [\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}] \begin{bmatrix} \mathbf{P}_{n\times n} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{C}}_{u\times u} \end{bmatrix} [\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}]^T \\ &= \mathbf{X}(\mathbf{P} + \overline{\mathbf{C}})\mathbf{X}^T \end{aligned} \tag{3.8}$$

and

$$\begin{aligned}\underline{\mathbf{S}}_{SELD} &= \mathbf{X}_{labeled}(\mathbf{I}_{n\times n} - \mathbf{P}_{n\times n})(\mathbf{X}_{labeled})^T + \mathbf{X}_{unlabeled}\underline{\mathbf{C}}_{u\times u}(\mathbf{X}_{unlabeled})^T \\ &= [\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}] \begin{bmatrix} \mathbf{I}_{n\times n} - \mathbf{P}_{n\times n} & \mathbf{0} \\ \mathbf{0} & \underline{\mathbf{C}}_{u\times u} \end{bmatrix} [\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}]^T \\ &= \mathbf{X}((\bar{\mathbf{I}} - \mathbf{P}) + \underline{\mathbf{C}})\mathbf{X}^T \end{aligned} \tag{3.9}$$

The resulting method combines supervised and unsupervised components in a nonlinear way, making fully the use of their strengths in different scenarios. In the case when all the samples are labeled, the proposed method reduces to LDA and in the case when all the samples are unlabeled, it reduces to LLFE.

To obtain the projection matrix, we solve the generalized eigenvalue problem of the proposed SELD method, which is equivalent to (2.3):

$$\overline{\mathbf{S}}_{SELD}\mathbf{w} = \lambda \underline{\mathbf{S}}_{SELD}\mathbf{w} \tag{3.10}$$

Through its nonlinear combination of supervised and unsupervised components, the proposed SELD seeks a projection direction on which the local neighborhood information of the data can be best preserved, while simultaneously the class discrimination is maximal, see Fig. 3.1 and Fig. 3.2.

It is important to note that LDA confronts sometimes with the difficulty that the matrix $\mathbf{S}_w$ is singular. The fact is that sometimes the number of labeled training samples n is much smaller than the number of dimensions d. In this situation, the rank of $\mathbf{S}_w$ is at most n as it is evident from (3.2), while the size of the matrix $\mathbf{X}(\bar{\mathbf{I}} - \mathbf{P})\mathbf{X}^T$ in (3.5) is $d \times d$. This implies that the within-class matrix $\mathbf{S}_w$ can become singular. Simultaneously, the *between-class* matrix $\mathbf{S}_b$ in the LDA method uses the labeled samples only. The rank of $\mathbf{S}_b$ is $C - 1$ (as it can be seen from (3.1)), implying that LDA can extract at most $C - 1$ features, which is not always sufficient to represent essential information of the original data.

The proposed SELD method overcomes these problems. The matrices $\overline{\mathbf{S}}_{SELD}$ and $\underline{\mathbf{S}}_{SELD}$ in our approach are both symmetric and positive semi-definite, which makes sure that SELD can extract as much features as the number of the spectral bands and the corresponding eigenvalues are not negative. Since our method can be combined with different LLFE methods, we will use a subscript to identify the particular LLFE methods employed, e.g. $SELD_{NPE}$, $SELD_{LPP}$ or $SELD_{LLTSA}$.

3.2.3 Algorithm

The algorithmic procedure of the proposed SELD is formally stated below:

1. Divide the training set $\mathbf{X}$ into two subsets: $\mathbf{X}_{labeled}$ and $\mathbf{X}_{unlabeled}$, with $\mathbf{X} = \{\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}\} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{x}_{n+1}, \cdots, \mathbf{x}_N\}$. Suppose that the n labeled training samples in $\mathbf{X}_{labeled} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ are ordered according to their labels, with data matrix of the kth class $\mathbf{X}^{(k)} = \{\mathbf{x}_1^{(k)}, \cdots, \mathbf{x}_{n_k}^{(k)}\}$ where $\mathbf{x}_i^{(k)}$ is the ith sample in the kth class, then the labeled subset can be expressed as $\mathbf{X}_{labeled} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(C)}\}$. $u = N - n$ unlabeled samples constitute the unlabeled subset $\mathbf{X}_{unlabeled} = \{\mathbf{x}_i\}_{i=n+1}^{N}$.

2. Construct the labeled weight matrices $\mathbf{P}$ and $\bar{\mathbf{I}}$ from the labeled subset $\mathbf{X}_{labeled}$.

3. Construct the “nearest neighbors” weight matrix $\overline{\mathbf{C}}$ and $\underline{\mathbf{C}}$ from the unlabeled subset $\mathbf{X}_{unlabeled}$. The particular construction depends on the chosen LLFE methods. For NPE: $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{M}$; for LPP: $\overline{\mathbf{C}} = \mathbf{D}$ and $\underline{\mathbf{C}} = \mathbf{L}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{L}$ is the Laplacian matrix [67]; for LLTSA: $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{B}$, where $\mathbf{B}$ is the alignment matrix [52].

4. Compute the eigenvectors and eigenvalues for the generalized eigenvector problem in (3.10). The projection matrix $\mathbf{W}_{SELD} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ is made up by the r eigenvectors of the matrix $\underline{\mathbf{S}}_{SELD}^{-1}\overline{\mathbf{S}}_{SELD}$ associated with the largest r eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

5. SELD embedding: project the original d dimensional data into a lower r dimensional subspace by

$$\mathbf{x} \rightarrow \mathbf{z} = \mathbf{W}_{SELD}^T \mathbf{x}$$

3.3 Experimental results on the synthetic data

In this section, we illustrate low-dimensional representations of the original synthetic data sets, resulting from different approach discussed in this paper. For this purpose, we generated 2 three dimensional data sets: Swissroll and the S-curve, which are well-known synthetic data sets. The data sets are compose of four different classes denoted in Fig. 3.3 and 3.6 by four different colors. Each class has 100 samples. In Fig. 3.3 and 3.6, filled symbols denote labeled samples and unfiled symbols denote unlabeled samples. The number of labeled samples used to

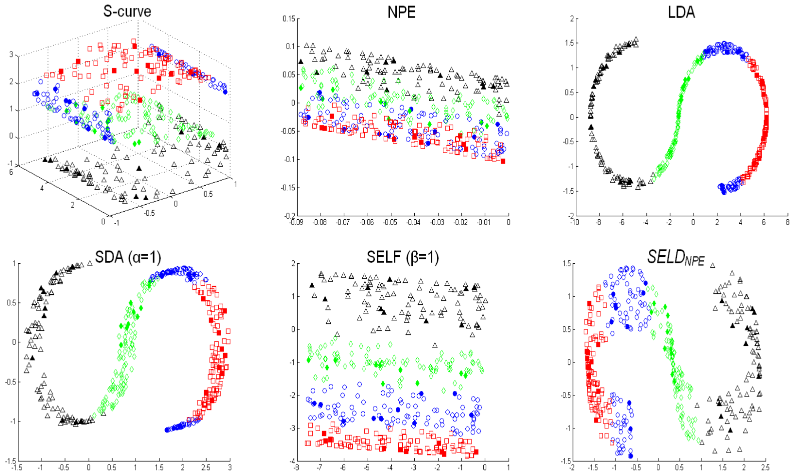


Figure 3.5: Comparison of feature extraction by each method on S-curve. Three dimensional data with four different class are embedded into a two dimensional subspace. Each class has 100 samples, filled/unfiled symbols denote labeled/unlabeled samples, the labeled samples in data sets are 10 for each class, the rest are unlabeled samples. For SDA, the parameter α is optimized in {0, 0.1, 0.5, 1, 5 and 10}, and for SELF the parameter β is optimized in {0, 0.25, 0.5, 0.75 and 1}.

train the projection matrix in one experiment is 2 and in the other experiment is 10. The parameter α in the SDA method was varied in $\{$0, 0.1, 0.5, 1, 5 and 10$\}$, and the parameter β in SELF was varied in $\{$0, 0.25, 0.5, 0.75 and 1$\}$. The proposed SELD combining the LDA method and NPE, LPP and LLTSA methods are respectively recorded as $SELD_{NPE}$, $SELD_{LPP}$ and $SELD_{LLTSA}$. Fig. 3.3 and 3.6 shows the results for the supervised method (LDA), the unsupervised LLFE methods (NPE, LPP and LLTSA) and the proposed SELD, and two other semi-supervised methods SDA [67] and SELF [69] with the best parameters.

Several conclusions can be drawn from these examples. The number of labeled samples does not influence the performance of the unsupervised LLFE methods. However, the projection directions found by LLFE do not take the class discrimination into account, and hence some samples from different classes overlap in the subspace. The supervised LDA method does not consider the local neighborhood information of the data. By optimizing the parameters, SDA and SELF discovered the class discrimination and preserved the local neighborhood information of the data. However, in case where limited labeled samples are available, some unlabeled samples from different classes overlap in the subspace found by LDA, SDA and SELF. This is in particular the case when the number of labeled samples for each class is smaller than the data dimension. The proposed SELD method allows us to extract more informative features even with a very limited labeled samples.

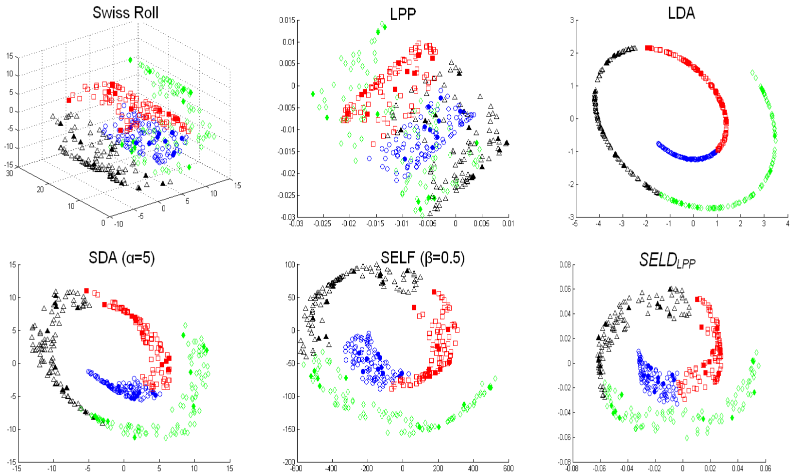


Figure 3.6: Comparison of feature extraction by each method Swiss Roll data sets. Three dimensional data with four different class are embedded into a two dimensional subspace. Each class has 100 samples, filled/unfiled symbols denote labeled/unlabeled samples, the labeled samples in data sets are 10 for each class, the rest are unlabeled samples. For SDA, the parameter α is optimized in $\{$0, 0.1, 0.5, 1, 5 and 10$\}$, and for SELF the parameter β is optimized in $\{$0, 0.25, 0.5, 0.75 and 1$\}$.

By combining the unsupervised LLFE methods and supervised LDA in a novel way, our approach not only preserves local neighborhood information, but also maximizes the class discrimination of the data. Moreover, the proposed SELD does not need to optimize the parameters.

3.4 Experimental results on the real hyperspectral data

3.4.1 Hyperspectral data sets

We use four real hyperspectral data sets in our experiments: the *Indian Pine* (a mixed forest/agricultural site in Indiana [139]), *Kennedy Space Center* (KSC) [140], the *Washington DC Mall* [139] (urban site), and *Okavango Delta, Botswana* [140]. Table 3.1 shows the number of labeled samples in each class for all the data sets. Note that the color in the cell denotes different classes in the classification maps (Fig. 3.8-Fig. 3.11).

Indian Pine data set: was captured by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over northwestern Indiana in June 1992, with 220 spectral bands in the wavelength range from 0.4 to 2.5μm and spatial resolution of 20 me-

No	Indian Pine		KSC		DC		Botswana	
	Class Name	# Samples	Class Name	# Samples	Class Name	# Samples	Class Name	# Samples
1	Corn-notill	1434	Scrub	761	Roof	3834	Water	270
2	Corn-min	834	Willow swamp	243	Street	416	Hippo grass	101
3	Corn	234	Cabbage palm hammock	256	Path	175	Floodplain grasses1	251
4	Grass/Pasture	497	Cabbage palm/oak hammock	252	Grass	1928	Floodplain grasses2	215
5	Grass/Trees	747	Slash pine	161	Trees	405	Reeds1	269
6	Hay-windrowed	489	Oak/broadleaf hammock	229	Water	1224	Riparian	269
7	Soybeans-notill	968	Hardwood swamp	105	Shadow	97	Firescar2	259
8	Soybeans-min	2468	Graminoid marsh	431			Island interior	203
9	Soybeans-clean	614	Spartina marsh	520			Acacia woodlands	314
10	Wheat	212	Cattail marsh	404			Acacia shrublands	248
11	Woods	1294	Salt marsh	419			Acacia grasslands	305
12	Bldg-Grass-Trees	380	Mud flats	503			Short mopane	181
13	Stone-steel towers	95	Water	927			Mixed mopane	268
14							Exposed soils	95
Total		10266		5211		8079		3248

Table 3.1: Data sets used in the experiments

ters by pixel. The calibrated data are available online (along with detailed ground-truth information) from http://cobweb.ecn.purdue.edu/˜biehl/. The whole scene, consisting of the full 145×145 pixels, which contains 16 classes, ranging in size from 20 to 2468 pixels. 13 classes were selected for the experiments, see Fig. 3.8.

KSC data set: was acquired by NASA AVIRIS instrument over the Kennedy Space Center (KSC), Florida in 1996 and consist of 224 bands of 10 nm width with center wavelengths from 0.4-2.5μm. The data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. Several spectral bands were removed from the data due to noise and water absorption phenomena, leaving a total of 176 bands to be used for the analysis. For classification purposes, 13 classes representing the various land cover types that occur in this environment were defined for the site, Fig. 3.9 shows an RGB composition with the labeled classes highlighted. For more information, see [140] and http://www.csr.utexas.edu/hyperspectral/.

DC Mall data set: was collected with an airborne sensor system over the Washington DC Mall, with 1280×307 pixels and 210 spectral bands in the 0.4-2.4μm region. This data set consists of 191 spectral bands after elimination of water absorption and noisy bands and is available at http://cobweb.ecn.purdue.edu /˜biehl/. 7 land cover/use classes are labeled and are highlighted in the Fig. 3.10.

Botswana data set: was acquired over the Okavango Delta, Botswana in May 31, 2001 by the NASA EO-1 satellite, with 30 m pixel resolution over a 7.7 km strip in 242 bands covering the 0.4-2.5μm portion of the spectrum in 10 nm windows. Uncalibrated and noisy bands that cover water absorption features were removed, leaving a total of 145 radiance channels to be used in the experiments. The data consists of observations from 14 identified classes intended to reflect the impact of flooding on vegetation, Fig. 3.11 shows an RGB composition with the labeled classes highlighted. For more information, see [140] and http://www.csr.utexas.edu/hyperspectral/.

3.4.2 Experimental setup

The training set $\mathbf{X}$ is made up of labeled subset $\mathbf{X}_{labeled}$ and unlabeled subset $\mathbf{X}_{unlabeled}$ (such that $\mathbf{X} = \mathbf{X}_{labeled} \cup \mathbf{X}_{unlabeled}$, and $\mathbf{X}_{labeled} \cap \mathbf{X}_{unlabeled} = \emptyset$). A number of unlabeled samples $u = 1500$ was randomly selected from the image parts with no labels to compose $\mathbf{X}_{unlabeled}$. The training of the classifiers (estimation of the SVM parameters) was carried out using the labeled subset $\mathbf{X}_{labeled}$. In our experiments, 70% randomly chosen samples from the labeled data set was initially assigned to the training set and the remaining 30% was used as the test set. In order to investigate the influence of the training set size on the classifier performance, the initial training set (consisting of 70% of the labeled samples) was further subsampled randomly to compose the labeled subset $\mathbf{X}_{labeled}$, with sample size conforming to one of the following two distinct cases:

- *Case* 1 ($n_k = 10$) in ill-posed condition: $n < d$ and $n_k < d$.
- *Case* 2 ($n_k = 40$) in poor-posed condition: $n > d$ and $n_k < d$.

We used three common classifiers: 1-nearest neighbor (1NN) like in [56, 60, 68], quadratic discriminant classifier (QDC) [141], and support vector machines (SVM) [142]. The SVM classifier with radial basis function (RBF) kernels in Matlab SVM Toolbox, LIBSVM [143], is applied in our experiments. SVM with RBF kernels has two parameters: the penalty factor C and the RBF kernel widths γ. We apply a grid-search on C and γ using 5-fold cross-validation to find the best C within the given set $\{10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and the best γ within the given set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$.

All classifiers were evaluated against the test set. We use overall classification accuracy (OCA) to evaluate the feature extraction results. The results were averaged over ten runs, we compare the resulting classification accuracies using the proposed SELD method with those resulting from the following methods: Raw data, where the classification is simply performed on the original data sets without dimensionality reduction; PCA [37]; LDA [57]; LLFE [52–54] (including NPE [53], LPP [54], LLTSA [55]); NWFE [35]; SDA [67], of which the parameter α is optimized with 5-fold cross-validation within the given set $\{0.1, 0.5, 2.5, 12.5, 62.5\}$; and SELF [69], where the parameter β is chosen from $\{0, 0.1, 0.2, ..., 0.9, 1\}$ by 5-fold cross validation.

3.4.3 Results and discussion

Table 3.2 and Table 3.3 display the classification accuracies of testing data in cases 1, 2, respectively. The best accuracy of each data set (in column) is highlighted in bold font.

From these tables, we have the following findings:

1. The results confirm that feature extraction can improve the classification performance on hyperspectral images. Most information can be preserved even with a few extracted features. Especially for the raw data set with QDC classifier, the results can be improved a lot by using feature extraction as a preprocessing. SVM classifier with RBF kernel function did not perform well in the raw data set of Indian Pine, this can be improved by using feature extraction.

2. When the number of labeled samples is very limited such as in *Case* 1, the supervised LDA perform much worse than other methods. By considering the local neighborhood information inferred from both labeled and unlabeled samples, SDA improves over LDA. However, one limitation of both LDA and SDA methods is that the number of extracted features depends on the number of classes.

3. By selecting $\beta = 1$ optimized with 5-fold cross-validation within the given set $\{0, 0.1, 0.2, ..., 0.9, 1\}$, SELF performs like PCA in both cases. For the Botswana data set with QDC classifier in *Case* 1, SELF and PCA give a better performance when small number of bands are used, while for the KSC data set in *Case* 2 (Fig. 3.7), SELF and PCA perform worse than other methods when small number of bands are used. It should be noted though that for a small number of features the OCA's are usually very small and useless in practice.

4. The proposed SELD outperforms the other feature extraction methods in both cases. In the ill-posed classification problems (*Case* 1, $n_k = 10 < n < d$), the highest OCA in Indian Pine, KSC, DC Mall and Botswana data sets are 0.698 ($SELD_{NPE}$ with 1NN classifier), 0.874 ($SELD_{NPE}$ with SVM classifier), 0.976 ($SELD_{LPP}$ with 1NN classifier) and 0.91 ($SELD_{LLTSA}$ with SVM classifier), respectively. In *Case* 2 ($n_k = 40 < d < n$), the highest OCA among for the same four images are 0.792 ($SELD_{NPE}$ with 1NN classifier), 0.936 ($SELD_{NPE}$ with SVM classifier), 0.998 ($SELD_{NPE}$ with SVM classifier) and 0.951 ($SELD_{NPE}$ with 1NN classifier), respectively.

In ill-posed (*Case* 1) and poor-posed (*Case* 2) classification problems, the QDC classifier cannot be developed to the raw data sets since the input dimension is higher than the number of available training samples. In these situations, 1NN and SVM classifier show better performances than QDC. The results in Table 3.2 and Table 3.3 show that the proposed method yields best OCA on all four data sets.

Feature Extraction	Classifier	Data Set			
		Indian Pine	KSC	DC	Botswana
Raw	QDC	0.14	0.146	0.474	0.084
	1NN	0.524	0.728	0.965	0.835
	SVM	0.475	0.846	0.948	0.876
PCA	QDC	0.568(5)	0.703(6)	0.969(3)	0.836(4)
	1NN	0.52(20)	0.726(19)	0.965(12)	0.833(20)
	SVM	0.583(6)	0.808(16)	0.946(2)	0.878(5)
LDA	QDC	0.14(4)	0.146(5)	0.474(4)	0.124(4)
	1NN	0.108(12)	0.30(12)	0.409(4)	0.151(10)
	SVM	0.129(6)	0.393(12)	0.476(5)	0.218(13)
NPE	QDC	0.521(6)	0.71(5)	0.969(3)	0.833(4)
	1NN	0.596(15)	0.84(16)	0963(6)	0.873(7)
	SVM	0.633(12)	0.839(18)	0.966(13)	0.895(9)
LPP	QDC	0.523(6)	0.731(5)	0.97(4)	0.795(4)
	1NN	0.612(10)	0.833(12)	0.966(7)	0.848(5)
	SVM	0.643(10)	0.84(20)	0.957(10)	0.867(11)
LLTSA	QDC	0.56(5)	0.666(3)	0.969(3)	0.815(5)
	1NN	0.563(20)	0.816(14)	0.965(2)	0.864(5)
	SVM	0.604(20)	0.82(19)	0.967(2)	0.898(7)
NWFE	QDC	0.574(5)	0.763(5)	0.967(3)	0.828(4)
	1NN	0.661(10)	0.833(18)	0.97(17)	0.881(17)
	SVM	0.624(7)	0.858(17)	0.957(2)	0.891(8)
SDA	QDC	0.413(5)	0.68(5)	0.889(5)	0.704(5)
	1NN	0.539(10)	0.817(12)	0.857(6)	0.77(13)
	SVM	0.483(7)	0.811(12)	0.817(6)	0.811(6)
SELF	QDC	0.568(5)	0.703(6)	0.969(3)	0.836(4)
	1NN	0.52(20)	0.726(19)	0.965(12)	0.833(20)
	SVM	0.583(6)	0.808(16)	0.946(2)	0.878(5)
$SELD_{NPE}$	QDC	0.551(7)	0.771(4)	0.965(3)	0.826(4)
	1NN	**0.698(18)**	0.863(20)	0.974(20)	0.903(20)
	SVM	0.648(12)	**0.874(19)**	0.959(18)	0.905(9)
$SELD_{LPP}$	QDC	0.541(5)	0.758(5)	0.969(4)	0.793(4)
	1NN	0.656(16)	0.844(20)	**0.976(15)**	0.873(18)
	SVM	0.645(11)	0.857(20)	0.959(3)	0.876(7)
$SELD_{LLTSA}$	QDC	0.531(5)	0.755(4)	0.953(4)	0.829(4)
	1NN	0.667(20)	0.852(20)	0.964(8)	0.899(19)
	SVM	0.642(18)	0.833(19)	0.948(12)	**0.91(9)**

Table 3.2: Highest OCA Using Extracted Features (The Number of Extracted Features is Written in the Back Brackets) Applied to Four Different Data Sets in Case 1

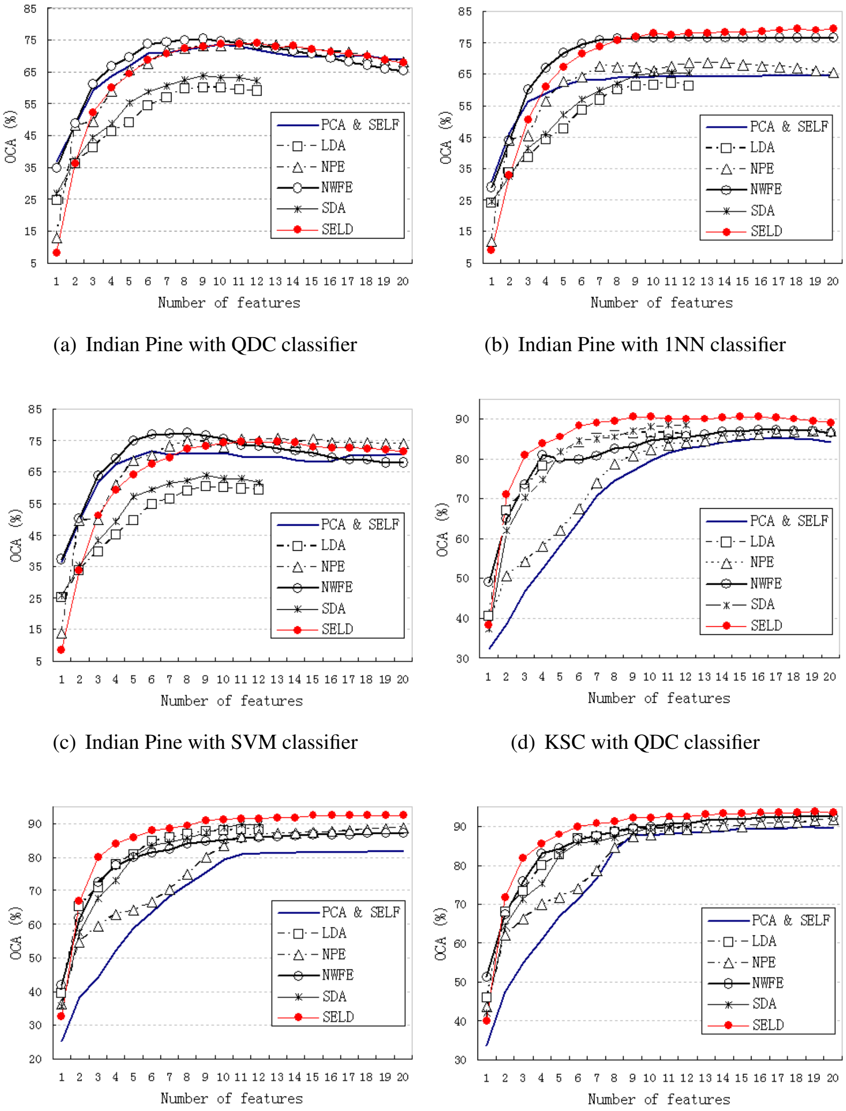
The experimental results in Table 3.2 and Table 3.3 also show that none of the three classifiers achieves the highest accuracy on every data set. This can also be seen in Fig. 3.7. The reason may be that the distributions of data sets are very different as was mentioned in [60, 144, 145]. In the following, we take the Indian Pine and KSC images in *Case* 2 as examples to explore the performances of different methods when the number of extracted features increases, the results were

Feature Extraction	Classifier	Data Set			
		Indian Pine	KSC	DC	Botswana
Raw	QDC	0.14	0.146	0.474	0.084
	1NN	0.65	0.818	0.983	0.902
	SVM	0.622	0.924	0.983	0.931
PCA	QDC	0.736(10)	0.853(17)	0.997(7)	0.937(8)
	1NN	0.646(20)	0.816(20)	0.983(14)	0.90(17)
	SVM	0.717(5)	0.896(19)	0.99(12)	0.94(7)
LDA	QDC	0.601(10)	0.864(10)	0.954(6)	0.909(6)
	1NN	0.621(11)	0.881(11)	0.975(6)	0.932(12)
	SVM	0.604(9)	0.895(12)	0.98(6)	0.901(8)
NPE	QDC	0.738(11)	0.87(20)	0.997(17)	0.941(8)
	1NN	0.687(13)	0.889(20)	0.988(13)	0.941(8)
	SVM	0.757(13)	0.916(20)	0.987(15)	0.945(8)
LPP	QDC	0.727(12)	0.891(13)	0.996(19)	0.927(12)
	1NN	0.71(10)	0.886(13)	0.985(12)	0.93(7)
	SVM	0.751(10)	0.92(20)	0.989(3)	0.925(12)
LLTSA	QDC	0.749(11)	0.872(18)	0.997(15)	0.935(7)
	1NN	0.644(19)	0.884(16)	0.982(7)	0.932(6)
	SVM	0.753(20)	0.908(18)	0.984(2)	0.932(6)
NWFE	QDC	0.752(9)	0.871(16)	0.997(13)	0.943(10)
	1NN	0.767(12)	0.87(20)	0.99(15)	0.921(19)
	SVM	0.775(8)	0.924(18)	0.988(16)	0.938(8)
SDA	QDC	0.636(9)	0.885(12)	0.993(6)	0.915(11)
	1NN	0.655(12)	0.897(11)	0.969(6)	0.939(12)
	SVM	0.637(9)	0.898(12)	0.978(6)	0.905(12)
SELF	QDC	0.736(10)	0.853(17)	0997(7)	0.937(8)
	1NN	0.646(20)	0.816(20)	0.983(14)	0.90(17)
	SVM	0.717(5)	0.896(19)	0.99(12)	0.94(7)
$SELD_{NPE}$	QDC	0.74(12)	0.906(9)	0.997(13)	0.935(8)
	1NN	**0.792(20)**	0.924(20)	0.992(20)	**0.951(18)**
	SVM	0.747(13)	**0.936(19)**	**0.998(12)**	0.948(9)
$SELD_{LPP}$	QDC	0.742(12)	0.904(15)	0.997(13)	0.931(13)
	1NN	0.785(12)	0.918(19)	0.993(19)	938(12)
	SVM	0.76(12)	0.931(18)	0.99(17)	0.933(11)
$SELD_{LLTSA}$	QDC	0.734(9)	0.911(11)	0.997(12)	0.945(11)
	1NN	0.779(19)	0.913(9)	0.992(18)	0.947(20)
	SVM	0.757(20)	0.925(19)	0.98(14)	0.949(13)

Table 3.3: Highest OCA Using Extracted Features (The Number of Extracted Features is Written in the Back Brackets) Applied to Four Different Data Sets in Case 2

shown in Fig. 3.7. The statistical significance of differences was computed using McNemar's test, which is based upon the standardized normal test statistic [103], Table 3.4-Table 3.9 show the results using the best results of each method in the same bands over ten runs.

1. On Indian Pine data set, NWFE outperforms the other methods for QDC and SVM classifiers, the difference is statistically significant, with $|Z| >$



(a) Indian Pine with QDC classifier

(b) Indian Pine with 1NN classifier

(c) Indian Pine with SVM classifier

(d) KSC with QDC classifier

(e) KSC with 1NN classifier

(f) KSC with SVM classifier

Figure 3.7: Performance of each feature extraction method in Case 2 for Indian Pine and KSC data sets. Each experiment was repeated 10 times, the average was acquired. By selecting $\beta = 1$ optimized with 5-fold cross-validation within the given set $\{0, 0.1, 0.2, \ldots, 0.9, 1\}$, SELF has the same performance as PCA. The proposed SELD method is the one which combines LDA and NPE.

Z_{rc}	Indian Pine using 9 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	21.6	-1.1	-7.7	18.1	0	-2.4
LDA	-21.6	0	-22.6	-27.8	-4.8	-21.6	-24
NPE	1.1	22.6	0	-6.1	19.1	1.1	-1.2
NWFE	7.7	27.8	6.1	0	24.8	7.7	5
SDA	-18.1	4.8	-19.1	-24.8	0	-18.1	-20.8
SELF	0	21.6	-1.1	-7.7	18.1	0	-2.4
SELD	2.4	24	1.2	-5	20.8	2.4	0

Table 3.4: Statistical significance of differences in classification (Z) with QDC classifier in Case 2. Each case of the table represents Z_{rc} where r is the row and c is the column. The best results of each method over ten runs are used.

Z_{rc}	Indian Pine using 9 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	0.9	-9.9	-26.7	-5.8	0	-29
LDA	-0.9	0	-10.3	-22.5	-9.3	-0.9	-28.3
NPE	9.9	10.3	0	-13.7	3.5	9.9	-20.2
NWFE	26.7	22.5	13.7	0	16.2	26.7	-6.7
SDA	5.8	9.3	-3.5	-16.2	0	5.8	-22.1
SELF	0	0.9	-9.9	-26.7	-5.8	0	-29
SELD	29	28.3	20.2	6.7	22.1	29	0

Table 3.5: Statistical significance of differences in classification (Z) with 1NN classifier in Case 2. Each case of the table represents Z_{rc} where r is the row and c is the column. The best results of each method over ten runs are used.

1.96. For 1NN classifier, the proposed SELD method yields the highest OCA of 79.2%, which is better than NWFE with SVM classifier 77.5%. The difference between the best results of SELD with 1NN classifier and NWFE with SVM classifier is statistically significant ($Z = 3.86$).

2. On KSC data set, SELD performs better than the other methods with all the three classifiers. The statistical difference of accuracy $|Z| > 1.96$ clearly demonstrates the efficiency of the proposed SELD.

3. Using only $C-1$ features may not be enough in some real situation, which is one limitation of both LDA and SDA. NPE can improve its performance by using more extracted features, as shown in Fig. 3.7(f). When more features are used, the overall classification accuracy can be improved.

The results in Table 3.4-Table 3.9 and in Fig. 3.7 show that SELD with 1NN classifier can have a better performance in Indian Pine image, while in the KSC image, SELD with SVM classifier will be a better choice.

In order to compare the classified maps visually, we generate classification maps with the combination of the highest OCA using different methods and classi-

Z_{rc}	Indian Pine using 9 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	15.5	-13.8	-17.2	7.8	0	-12.2
LDA	-15.5	0	-27.6	-30.5	-10.5	-15.5	-26.6
NPE	13.8	27.6	0	-3.9	20.4	13.8	1.9
NWFE	17.2	30.5	3.9	0	23.6	17.2	5.4
SDA	-7.8	10.5	-20.4	-23.6	0	-7.8	-19.3
SELF	0	15.5	-13.8	-17.2	7.8	0	-12.2
SELD	12.2	26.6	-1.9	-5.4	19.3	12.2	0

Table 3.6: Statistical significance of differences in classification (Z) with SVM classifier in Case 2. Each case of the table represents Z_{rc} where r is the row and c is the column. The best results of each method over ten runs are used.

Z_{rc}	KSC using 12 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	-1.6	-4.8	-5.9	-4.7	0	-9.2
LDA	1.6	0	-2.5	-2.9	-4.3	1.6	-7
NPE	4.8	2.5	0	-0.5	-0.7	4.8	-5.7
NWFE	5.9	2.9	0.5	0	-0.3	5.9	-4.6
SDA	4.7	4.3	0.7	0.3	0	4.7	-3.7
SELF	0	-1.6	-4.8	-5.9	-4.7	0	-9.2
SELD	9.2	7	5.7	4.6	3.7	9.2	0

Table 3.7: Statistical significance of differences in classification (Z) with QDC classifier in Case 2. Each case of the table represents Z_{rc} where r is the row and c is the column. The best results of each method over ten runs are used.

Z_{rc}	KSC using 12 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	-12.2	-13.8	-13.4	-13.7	0	-19
LDA	12.2	0	-1.5	2.8	-2.7	12.2	-7.5
NPE	13.8	1.5	0	4.4	-0.6	13.8	-6
NWFE	13.4	-2.8	-4.4	0	-4.7	13.4	-10.7
SDA	13.7	2.7	0.6	4.7	0	13.7	-5.4
SELF	0	-12.2	-13.8	-13.4	-13.7	0	-19
SELD	19	7.5	6	10.7	5.4	19	0

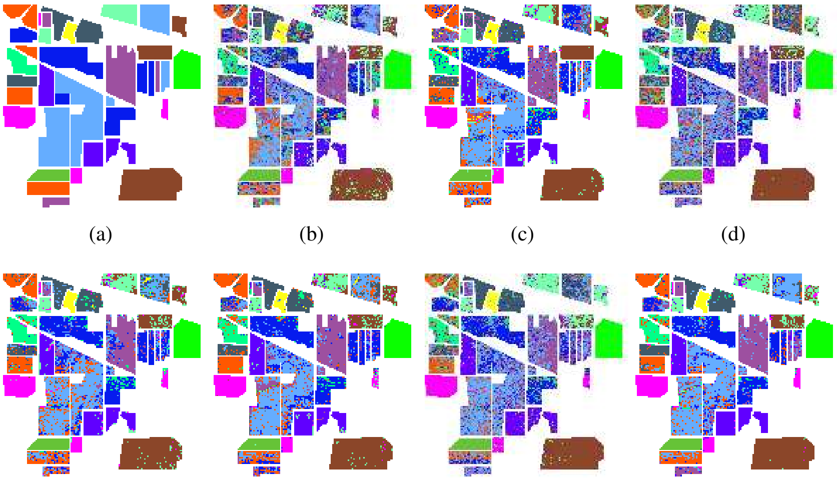
Table 3.8: Statistical significance of differences in classification (Z) with 1NN classifier in Case 2. Each case of the table represents Z_{rc} where r is the row and c is the column. The best results of each method over ten runs are used.

fiers in *Case* 2 ($n_k = 40$), displayed in Fig. 3.8-Fig. 3.11. The results demonstrate that:

1. By incorporating the local neighborhood information of the data, SELD preserves well spatial consistency in the classification maps, for example, the "Grass" in DC Mall image (Fig. 3.10). SELD also produces smoother homogeneous regions in the classification maps, which is particularly significant

Z_{rc}	KSC using 12 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	0.1	-5.7	-7.1	-2.2	0	-11
LDA	-0.1	0	-4.8	-7.5	-3.4	-0.1	-11.9
NPE	5.7	4.8	0	-2.2	2.6	5.7	-6.8
NWFE	7.1	7.5	2.2	0	5.2	7.1	-6.2
SDA	2.2	3.4	-2.6	-5.2	0	2.2	-9.8
SELF	0	0.1	-5.7	-7.1	-2.2	0	-11
SELD	11	11.9	6.8	6.2	9.8	11	0

Table 3.9: Statistical significance of differences in classification (Z) with SVM classifier in Case 2. Each case of the table represents Z_{rc} where r is the row and c is the column. The best results of each method over ten runs are used.


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3.8: Classification maps for Indian Pine with $n_k = 40$ (Case 2) (a) Ground truth of the area with 13 classes, and thematic map using (b) 1NN classifier without feature extraction ($r = 220$), (c) PCA and SELF features and QDC Classifier ($r = 10$), (d) LDA features and 1NN Classifier ($r = 11$), (e) NPE features and SVM Classifier ($r = 13$), (f) NWFE features and SVM Classifier ($r = 8$), (g) SDA features and 1NN Classifier ($r = 12$), and (h) The proposed $SELD_{NPE}$ features and 1NN Classifier ($r = 20$).

when classifying the “Stone-steel towers” and “Grass/Trees” in the Indian Pine image (Fig. 3.9).

2. SELD also yields good class discrimination. For Indian Pine image, it is easy to find that SELD outperforms other feature extraction methods in “Grass/Pasture”, “Grass/Trees”, “Soybeans-notill” and “Soybeans-clean” parts (Fig. 3.9). For DC Mall image, SELD discriminates “Water” better

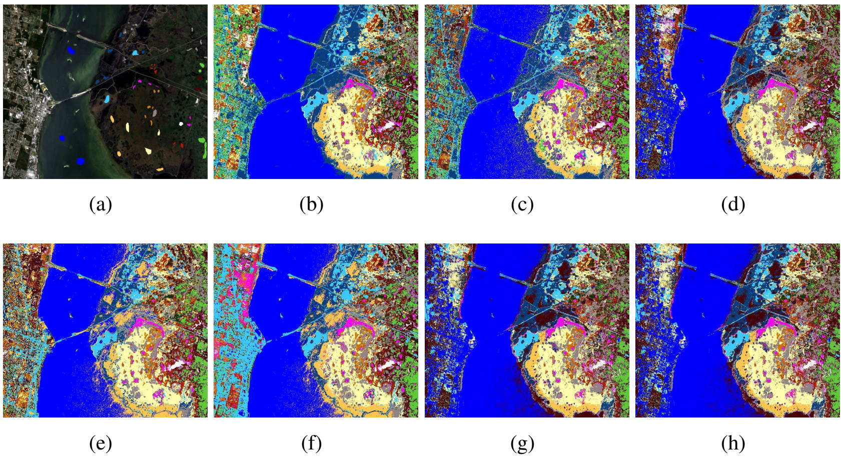


Figure 3.9: Classification maps for KSC with $n_k = 40$ *(Case 2) (a) RGB composition with 13 classes labeled and highlighted in the image, and thematic map using (b) SVM classifier without feature extraction (*$r = 176$*), (c) PCA and SELF features and SVM Classifier (*$r = 19$*), (d) LDA features and SVM Classifier (*$r = 12$*), (e) LPP features and SVM Classifier (*$r = 20$*), (f) NWFE features and SVM Classifier (*$r = 18$*), (g) SDA features and SVM Classifier (*$r = 12$*), and (h) The proposed* $SELD_{NPE}$ *features and SVM Classifier (*$r = 19$*).*

than the other methods (Fig. 3.10).

The plots in Fig. 3.12 and Fig. 3.13 give more insight into class discrimination by different methods. The training and testing samples of three classes of KSC image in *Case* 1 are projected into the feature space formed by the first two eigenvectors of different feature extraction methods. The results in Fig. 3.12 and Fig. 3.13 show that LDA has overfitting problems, because in *Case* 1 ($n < d$, and $n_k < d$), both the *within-classs* scatter matrix $\mathbf{S}_w$ and the *between-class* scatter matrix $\mathbf{S}_b$ are singular, $\mathbf{S}_w$ cannot be inverted, and both $\mathbf{S}_w$ and $\mathbf{S}_b$ are not accurate. By considering the local neighborhood information inferred from both labeled and unlabeled samples, SDA improves over LDA, but the test data are projected with different classes mixed. The distributions of projected data obtained by SELD are more concentrated and more distinct as compared with those of PCA, LDA, NPE, NWFE and SDA. This explains also classification improvement in Table 3.2 and Table 3.3.

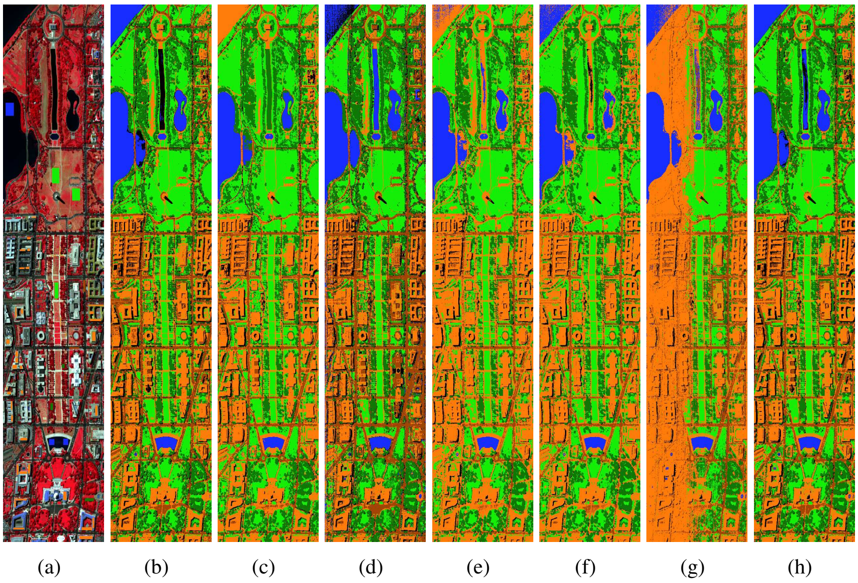


Figure 3.10: Classification maps for DC Mall with $n_k = 40$ (Case 2) (a) RGB composition with 7 classes labeled and highlighted in the image, and thematic map using (b) SVM classifier without feature extraction ($r = 191$), (c) PCA and SELF features and QDC Classifier ($r = 7$), (d) LDA features and SVM Classifier ($r = 6$), (e) NPE features and QDC Classifier ($r = 17$), (f) NWFE features and QDC Classifier ($r = 13$), (g) SDA features and QDC Classifier ($r = 6$), and (h) The proposed $SELD_{NPE}$ features and SVM Classifier ($r = 12$).

3.5 Algorithm analysis

The computational complexity of the proposed SELD is mainly in finding the e nearest neighbors for all the selected unlabeled training samples. To find the e nearest neighbors for u selected unlabeled training samples in the d dimensional Euclidean space, the complexity is O(du^2). However, some methods can be used to reduce the complexity of searching the e nearest neighbors, such as K-D trees [146]. $SELD_{NPE}$ and $SELD_{LLTSA}$ have additional complexities over $SELD_{LPP}$ in calculating the reconstruction weights, which is O(due^3). For storing the matrix $\overline{\mathbf{C}}$ or $\underline{\mathbf{C}}$ in equation (3.8) and (3.9), the complexity is O(N^2), where N is the total training samples including labeled and unlabeled ones. For example, if we use all the samples in the Botswana data set to train, $N = 1476 \times 256$, this will exceed the memory capacity of an ordinary PC even though the matrix is sparse. In order to reduce the computational complexity and memory consumption, some of unlabeled samples were selected in our experiments.

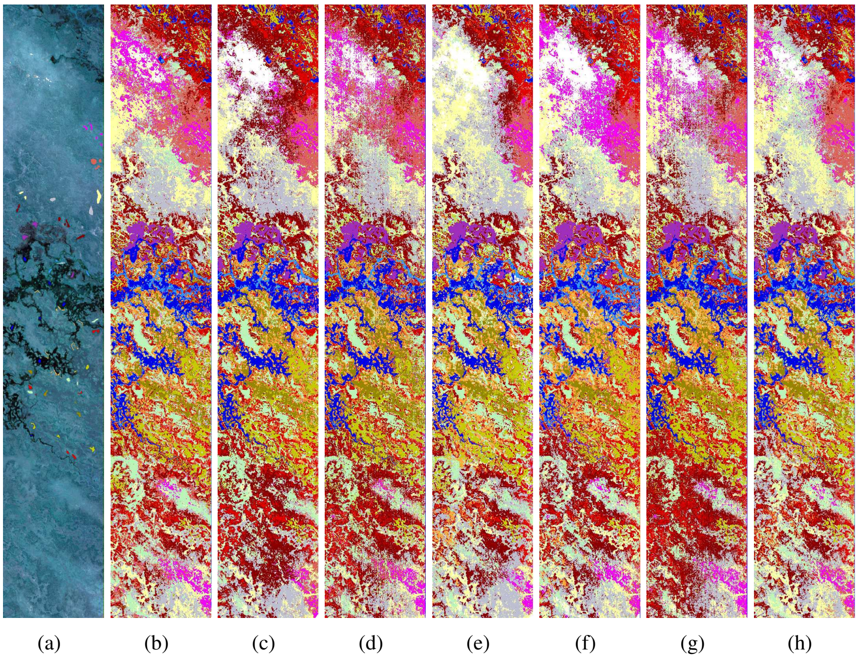
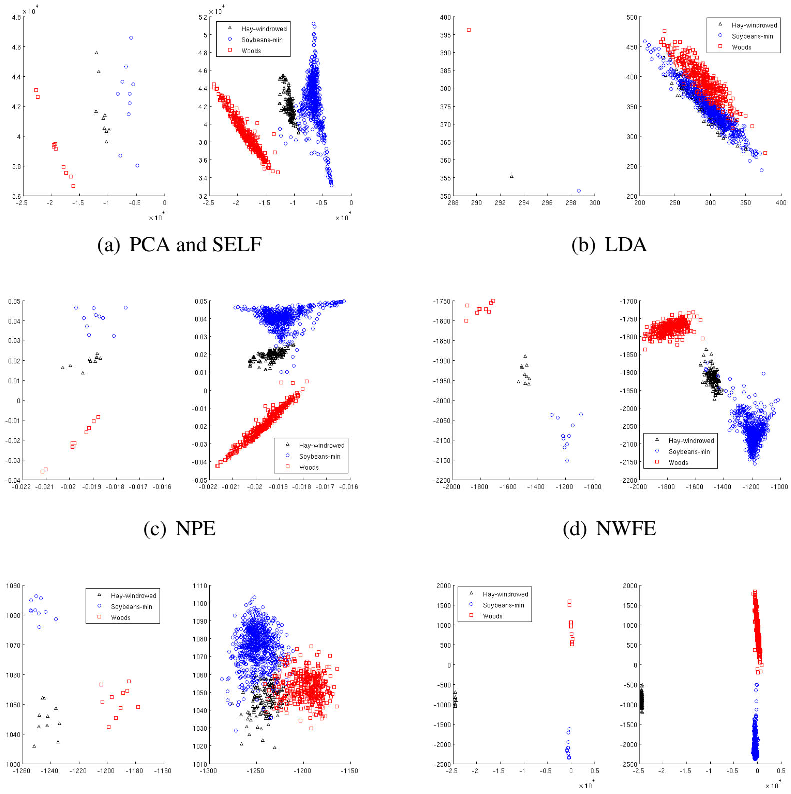


Figure 3.11: Classification maps for Okavango Delta, Botswana with $n_k = 40$ (Case 2) (a) RGB composition with 14 classes labeled and highlighted in the image, and thematic map using (b) SVM classifier without feature extraction ($r = 145$), (c) PCA and SELF features and SVM Classifier ($r = 7$), (d) LDA features and 1NN Classifier ($r = 12$), (e) NPE features and SVM Classifier ($r = 8$), (f) NWFE features and QDC Classifier ($r = 10$), (g) SDA features and SVM Classifier ($r = 6$), and (h) The proposed $SELD_{NPE}$ features and 1NN Classifier ($r = 18$).

3.5.1 Computational cost

We compared the computational cost of different approaches. All the methods were implemented in Matlab. The experiments were carried out on 64-b, 2.67 GHz Intel i7 920 (8 core) CPU computer with 12 GB memory, Fig. 3.14 shows the computational time of different approaches, and the OCA with 1NN classifier. The recorded times were only consumed in the process of feature extraction. This included the time consumed on the parameter determination of some methods (such as α in SDA, and β in SELF). We can see that PCA and LDA are the fastest, and the proposed SELD is more efficient than NWFE, SDA and SELF as the number of training samples increases. The reason is that the parameter determination in SDA and SELF is time consuming.



(a) PCA and SELF

(b) LDA

(c) NPE

(d) NWFE

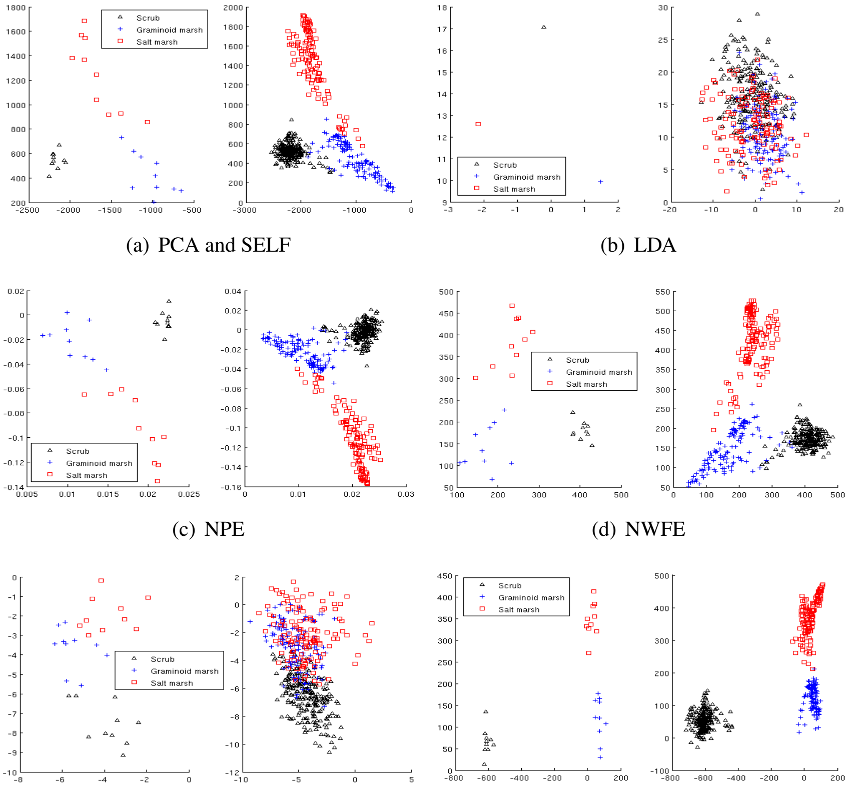
(e) SDA

(f) SELD

Figure 3.12: Distributions of training samples and testing samples for "Hay-windrowed", "Soybeans-Min" and "Woods" of Indian Pine data set using the first two significant features obtained from different methods. In each method, the left scatter plot is for training data and the right one is for testing data ($n_k = 10$, Case 1).

3.5.2 Selection of unlabeled samples

The choice of unlabeled samples is very important step in the semi-supervised methods. Selection of too many unlabeled samples will increase computational complexity, while a small number of unlabeled samples is not sufficient to exploit the local neighborhood information of the data sets. One easy solution is selecting unlabeled samples randomly from the whole image. Fig. 3.15(a) shows an example of the performances with different number of labeled and unlabeled samples. The number of unlabeled samples was evaluated from 200 to 3000 with a step of 200. Fig. 3.15(b) shows the corresponding computation times. The classification accuracy of SELD will be improved as more unlabeled samples are used, particularly



(a) PCA and SELF (b) LDA

(c) NPE (d) NWFE

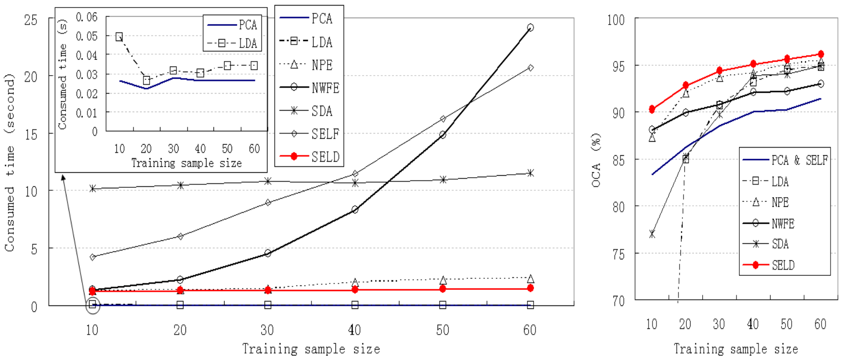
(e) SDA (f) SELD

Figure 3.13: Distributions of training samples and testing samples for "Scrub", "Graminoid marsh" and "Salt marsh" of KSC data set using the first two significant features obtained from different methods. In each method, the left scatter plot is for training data and the right one is for testing data ($n_k = 10$, Case 1).

in ill-posed (Case 1) classification problems. Generally, semi-supervised methods can achieve better classification results by using more unlabeled samples than labeled ones [147, 148]. However, the usage of a large number of unlabeled samples will cause problems in computational complexity and memory consumption. This may be improved by using some spatial selection methods [137].

3.5.3 Selection of nearest neighbors

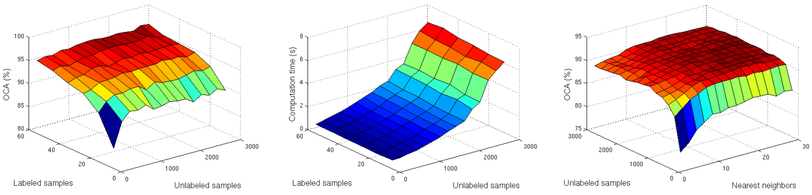
In graph-based feature extraction methods, the number of nearest neighbors (e) is an important parameter. We can employ cross-validation to optimize e. However, we found in our experiments that our approach produces consistently good results



(a) Computational cost

(b) OCA

Figure 3.14: Comparision of computational time (second) and OCA with different sample size, $e = 12$ *and* $u = 1500$*. The experiments was repeated 10 times, the average was acquired. The highest OCA with* r *changing from 1 to 20 is recorded.*



(a)

(b)

(c)

Figure 3.15: Surface of (a) the OCA as a function of labeled and unlabeled samples, $r = 13$ *and* $e = 12$*; (b) the computation time as a function of labeled and unlabeled samples,* $r = 13$ *and* $e = 12$*; (c) the OCA as a function of unlabeled samples and nearest neighbors,* $n_k = 10$ *and* $r = 13$*.*

over a large range of e values, which suggests insensitivity to this parameter in a broad range. Fig. 3.15(c) illustrates the performance with different number of unlabeled samples and nearest neighbors when e is changed from 2 to 30 with a step of 2. Note that the maximal dimensionality of $SELD_{LLTSA}$ was set to $e - 2$ (e should be greater than r [52]).

3.6 Conclusion

In this Chapter, we presented a new semi-supervised feature extraction method and we applied it to classification of hyperspectral images. The main idea of the proposed method is to divide first the samples into the labeled and the unlabeled sets. The labeled samples are employed through the supervised LDA only and the unlabeled ones through the unsupervised method only. We combine the two in a non-linear way, which makes full use of the advantages of both approaches. Experimental results on hyperspectral images demonstrate advantages of our method and improved classification accuracy compared to some related feature extraction methods. Moreover, we do not need to optimize any tuning parameters, which makes our method more efficient. Also the new method removes the limitation of LDA and SDA in terms of the number of extracted features.

4
GSELD

When using morphological features for the classification of high resolution hyperspectral images from urban areas, one should consider two important issues. The first one is that classical morphological openings and closings degrade the object boundaries and deform the object shapes. Morphological openings and closings by reconstruction can avoid this problem, but this process leads to some undesirable effects. Objects expected to disappear at a certain scale remain present when using morphological openings and closings by reconstruction. The second one is that the morphological profiles (MPs) with different structuring elements and a range of increasing sizes of morphological operators produce high-dimensional data. These high-dimensional data may contain redundant information and create a new challenge for conventional classification methods, especially for the classifiers which are not robust to the Hughes phenomenon.

In this Chapter, we first apply morphological profiles with partial reconstruction and directional MPs for the classification of high resolution hyperspectral images from urban areas. Secondly, we develop a semi-supervised feature extraction to reduce the dimensionality of the generated morphological profiles for the classification, see Fig. 4.1. To the best of our knowledge the use of semi-supervised FE methods for the generated morphological profiles has not been investigated yet. Experimental results on real urban hyperspectral images demonstrate the efficiency of the considered techniques.

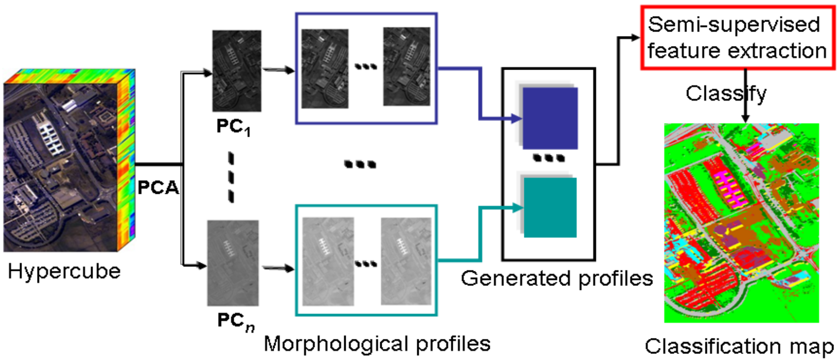


Figure 4.1: Diagram of proposed semi-supervised FE for MPs.

4.1 Introduction

Recent advances in sensors technology have led to an increased availability of hyperspectral data from urban areas at very high both spatial and spectral resolutions. Many techniques are developed to explore the spatial information of the high resolution remote sensing data, in particular, mathematical morphology [70, 71] is one of the most popular methods. Pesaresi and Benediktsson [72] proposed the use of morphological transformations to build a morphological profile (MP). Bellens et al. [73] further explored this approach by using both disk-shaped and linear structuring elements to improve the classification of very high-resolution panchromatic urban imagery. The approach of [17] extended the method in [70] for hyperspectral data with high spatial resolution. The resulting method built the MPs on the first principal components (PCs) extracted from a hyperspectral image, leading to the definition of extended MP (EMP). The appoach of [39] performs spectral-based morphology using the full hyperspectral image without dimensionality reduction. In [28], kernel principal components are used to construct the EMP, with significant improvement in terms of classification accuracies compared with the conventional EMP built on PCs. In [74], the attribute profiles (APs) [75] were applied to the first PCs extracted from a hyperspectral image, generating an extended AP (EAP). The approach of [76] improved the classification results by constructing the EAP with the independent component analysis.

When using MPs, one should consider two important issues. The first one is that classical morphological openings and closings degrade the object boundaries and deform the object shapes, which may result in losing some crucial information and introducing fake objects in the image. To avoid this problem, one often uses morphological openings and closings by reconstruction [17, 18, 72, 77, 78], which can reduce some shape distortions in the image. However, morphological openings and closings by reconstruction lead to some unexpected results in the resulting images, such as over-reconstruction [73]. Objects which are expected to disappear at a certain scale remain present when using morphological openings and closings by reconstruction. The approach of [73] proposed a partial reconstruction for the classification of very high-resolution panchromatic urban imagery. Morphological openings and closings by partial reconstruction can solve the problem of over-reconstruction while preserving the shape of objects as much as possible. They limit the extent of the reconstruction. The edges of simple objects are reconstructed well, but a full retrieval of complex elongated shapes might not be obtained. For simple objects like rectangles for example, the reconstruction is complete. Since, in urban remote sensing scenes, most objects are not very complex and are often simply even rectangular shaped, partial reconstruction is very well suited.

The second problem is that the resulting data sets may contain redundant information, because the construction of the generated profiles is based on different structuring elements (SEs) and a range of increasing sizes of morphological operators. Furthermore, the increase in the dimensionality of the generated profiles may create a new challenge for conventional classification methods, especially for the classifiers which are not robust to the Hughes phenomenon [1] (for a limited number of training samples, the classification accuracy decreases as the dimension increases). Although some advanced classifiers, such as neural networks [17], SVM [18, 19] and random forest classifiers [19], are shown to deal efficiently with these high dimensional data sets, common statistical classifiers are often limited in this context. For this reason, feature extraction (FE), aiming at reducing the dimensionality of data while keeping as much intrinsic information as possible, is a desirable preprocessing tool to reduce the dimensionality of the generated profiles for classification. Relatively few bands can represent most information of the data, making feature extraction very useful for classification of remote sensing data [28, 29]. The effect of different FE methods on reducing the dimensionality of the generated profiles for classification of hyperspectral data from urban areas has been discussed in several studies [17–19, 149].

However, to the best of our knowledge the use of semi-supervised FE methods for the generated morphological profiles has not been investigated yet. In many real world applications, it is usually difficult, expensive and time-consuming to collect sufficient amount of labeled samples. Meanwhile, it is much easier to obtain unlabeled samples. For this reason, semi-supervised methods [62–66, 68], which aim at improved classification by utilizing both unlabeled and limited labeled data gained popularity in the machine learning community.

In this Chapter, we first investigate the effect of the morphological profiles with partial reconstruction and the effect of directional morphological profiles [73] on the classification of hyperspectral images from urban areas. Secondly, we develop a semi-supervised FE method as a preprocessing to reduce the dimensionality of the generated morphological profiles for classification.

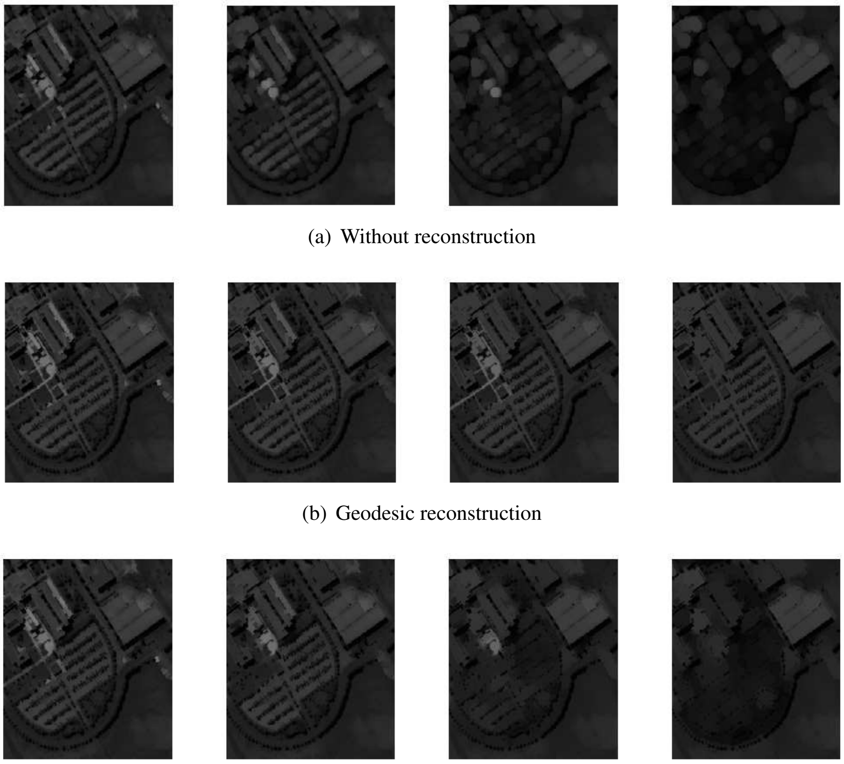
4.2 Morphological features

Morphological operators act on the values of the pixels according to transformations that consider the neighborhood (with a given size and shape) of the pixels. The basic operators are dilation and erosion [70]. These operators are applied to an image with a set of known shapes, called the structuring elements. In the case of erosion, a pixel takes the minimum value of all the pixels in its neighborhood, defined by the SE. By contrast, dilation takes the maximum value of all the pixels in its neighborhood. Dilation and erosion are usually employed in pairs, either dilation of an image followed by erosion of the dilated result, or erosion of an image followed by dilation of the eroded result. These combinations are known as opening and closing. An opening acts on bright objects compared with their surrounding, while closings act on dark objects. For example, an opening deletes (this means the pixels in the object take on the value of their surrounding) bright objects that are smaller than the SE. The term scale of an opening or closing is referred to the size of SE.

4.2.1 Disk-based and linear-based structure elements

Because of its isotropic character morphological openings and closings with disk-shaped SEs are the most popular methods used in current literature [17–19]. Objects where the SE (disk shape with a radius R) does not fit are deleted from the image. Fig. 4.2(a) shows an image and two openings with disk-shaped SEs of different sizes. Objects with a width smaller than $2R$ are deleted from the image.

Aside from the disk-shaped SEs, we can also use linear SEs [73]. A line has a certain orientation θ and length L, i.e., the Euclidean distance between the two endpoints of the line (rounded off). Fig. 4.2(b) shows three closings with linear SEs of length L using the features extracted from Fig. 4.7(a). A pixel is deleted if there exist no line of length L and orientation θ that goes through that pixel. This means that an object that is smaller than L in the orientation θ is removed. Objects which are smaller than L in all directions are removed from all these openings or closings. Therefore, the maximum of all openings with a linear SE of length L and different orientations (analogously the minimum of all closings) removes objects with a maximum dimension smaller than L. The number of orientations used to determine this maximum of openings should be chosen as high as possible, taking into account the computation time. For more details on linear-based SE, the readers should consult [73], which we applied in our experiments.



(a) Without reconstruction

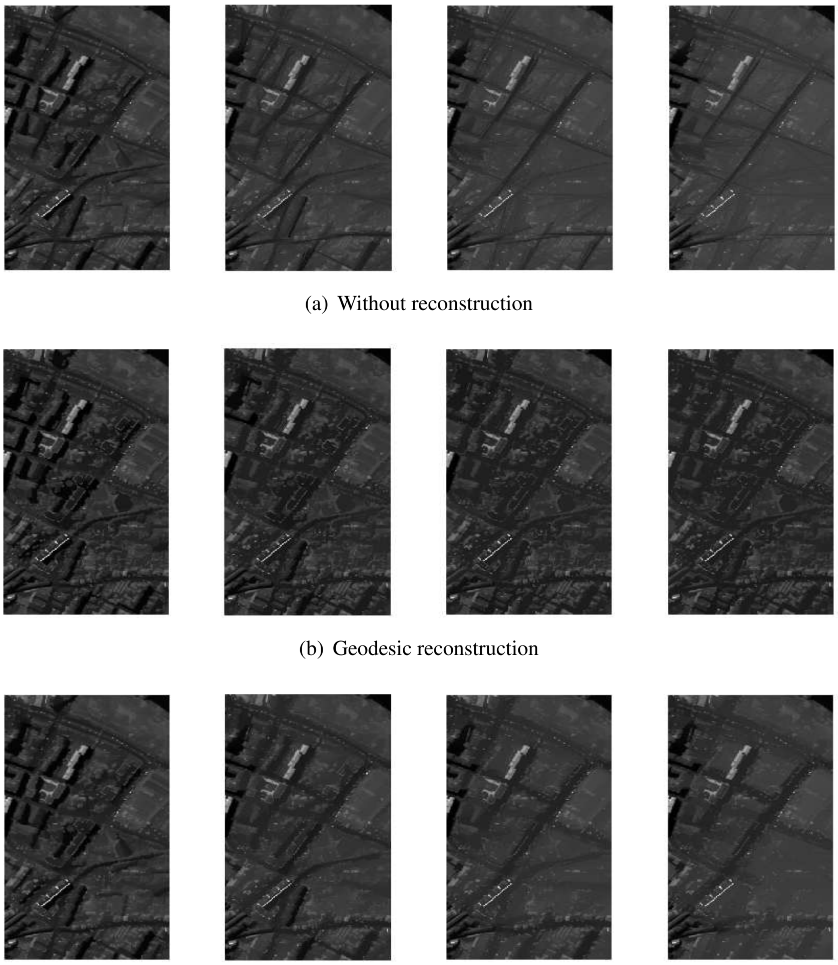
(b) Geodesic reconstruction

(c) Partial reconstruction

Figure 4.2: Openings with disk-shaped SEs of increasing size. The scales of SEs vary from 2 to 8, with step 2. The image processed is part of the first PC extracted from University Area data set in Fig. 4.6(a).

4.2.2 Reconstruction and Partial reconstruction

Aside from deleting objects smaller than the SE, morphological openings and closings also deform the objects which are still present in the image, see Fig. 4.2(a) and Fig. 4.3(a), the corners of rectangular objects in Fig. 4.2(a) (square object on the top right) are rounded. To preserve the shapes of objects, morphological openings and closings by reconstruction are generally the tool of choice [18, 77]. This process reconstructs the whole object if at least one pixel of the object survives the opening or closing. We can see the results in Fig. 4.2(b) and Fig. 4.3(b), the shapes of the objects are well preserved, and some small objects disappear as the scale (here the scale is related to the size of the SE) increases. However, morphological openings and closings with reconstruction will lead to some undesirable effects (such as over-reconstruction), a lot of objects that disappeared in the morphological openings and closings without reconstruction remain present in that



(a) Without reconstruction

(b) Geodesic reconstruction

(c) Partial reconstruction

Figure 4.3: Closings with linear SEs of increasing size. The scales of SEs vary from 20 to 80, with step 20. The image processed is part of the first PC extracted from Pavia Centre data set in Fig. 4.7(a).

with reconstruction. Objects which are expected to disappear in the image at a low scale, are still present at the highest scales, as shown in Fig. 4.2(b) (small bright road on the middle left) and Fig. 4.3(b) (small black road on the middle right).

The approach of [73] proposed a partial reconstruction to solve the problem of over-reconstruction while preserving the shape of objects as much as possible, and made a great improvement in the classification of very high-resolution panchromatic urban imagery. In the partial reconstruction process, a pixel is only reconstructed if it is connected to a pixel that was not erased, and this second pixel lies within a certain geodesic distance $dist$ from the pixel. The geodesic distance between two pixels is the length of the shortest path between the two pixels that lie completely within the object. The parameter $dist$ sets the amount of reconstruction. For disk shaped SE, this amount can be chosen such that rectangular objects are completely reconstructed. For linear SE, the choice of a good value is more difficult. However, 10% of the length of the SE seems a good value [73]. Fig. 4.2(c) and Fig. 4.3(c) show the results of morphological openings and closings with partial reconstruction in different scales. The shapes of objects are better preserved with partial reconstruction compared to the morphological openings and closings without reconstruction. Some of the more complex shapes are not so well preserved as with geodesic reconstruction. On the other hand, a lot of small objects which erroneously remain present in the profiles with reconstruction, disappear correctly at the right scale in the partial reconstruction profiles. Basically this is because in remote sensing (urban) scenes different objects lie closely together and because of noise and other effects, different objects are often connected by a sequence of pixels with similar (or more extreme) pixel values. Therefore, reconstruction considers all those connected objects as a single object and objects will only disappear when the SE does not fit the broadest (for disk shapes) or longest (for directional) part of the connected object, even though this part might be far away from the actual object. Partial reconstruction only reconstructs the immediate surrounding of the surviving part, and avoid thereby most of these errors.

4.3 Extended morphological profiles with partial reconstruction

A morphological profile (MP) consists of the opening profile (OP) and the closing profile (CP), see Fig. 4.4. For disk shaped SE this means objects where the smallest objects size (i.e. the width) is smaller than the diameter of the disk. Closings and openings with disk-shaped SEs thus act on the minimum size of objects. This results in an disk-based MP carrying information about the minimum size of objects. Fig. 4.2(c) shows the result of the opening transform with partial reconstruction for different-sized, disk-shaped SEs. As the size of the SE increases, more and more bright objects disappear in the dark background. The size of the SE that makes objects disappear corresponds to the minimum size of the object.

In [73], directional MP was proposed to obtain an indication of the maximum size of objects. With a linear structuring element of length L and orientation θ, an opening (resp. closing) deletes bright (resp. dark) objects (or object parts) which are smaller than that length in that direction. When performing such openings (or closings) with different orientations, objects which are shorter than L will be completely removed in all of these images. The maximum (resp. minimum)

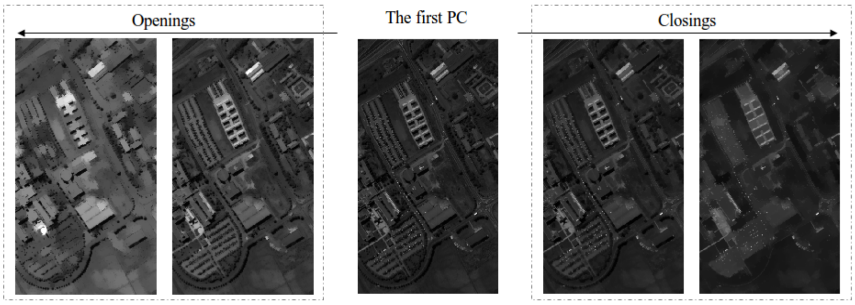


Figure 4.4: Morphological profile with 2 openings and 2 closings by partial reconstruction. Disk SEs are used with radius $R = 2$ and $R = 6$. The image processed is the first PC extracted from University Area dataset in Fig. 4.6(a).

over all of these openings (resp. closings) will therefore remove the short objects (or object parts) and keep the long objects. Creating multiple such maximum or minimum images for different lengths L gives you the directional MP. Thus the directional MP carries information about the maximum size of objects. This information can be used for detecting linear objects (roads), since these objects have large maximum sizes and small minimum sizes. Fig. 4.3(c) shows an example of the directional MP with partial reconstruction. Note that individual houses disappear at lower scales, while roads and apartment buildings with a more elongated shape have almost constant intensities. For more details on MP with partial reconstruction and directional MP, the readers should consult [73]. By increasing the size of the SE, more and more objects are removed. We will use the term scale of an opening or closing to refer to this size. A vector containing the pixel values in openings and closings by reconstruction of different scales is called the morphologic profile. The MPs carries information about the size and the shape of objects in the image.

For the panchromatic image, MP is built on the original single band image directly. The OP with n scales at pixel **x** forms n-dimensional vector, and so as the CP. By incorporating the OP and the CP, the morphological profile of pixel **x** is defined as $(2n+1)$-dimensional vector. When applying MP to the hyperspectral data, feature extraction is used as a pre-processing to reduce the dimensionality of the high-dimensional original data. MP built on different features has been discussed in several studies [17,28,87], Fig. 4.4 shows a MP with partial reconstruction built on the first PC. By applying MP to each extracted feature independently, extended morphological profile (EMP) is formed as a stacked vector which is constructed from all the morphological profiles. Fig. 4.5 shows an EMP with partial reconstruction built on the first two PCs.

The morphologic profiles with a certain SE produce a vector of values, each value corresponding with the feature output for a specific scale. While morpho-

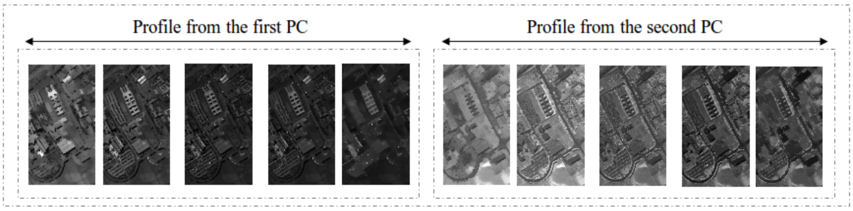


Figure 4.5: Extended morphological profile built on the first two PCs with 2 openings and 2 closings by partial reconstruction. Disk SEs are used with radius $R = 2$ and $R = 6$.

logic profiles with different SEs will then be a high-dimensional stacked vector. The resulting high-dimensional data may contain redundant information. Furthermore, if we use these high-dimensional data as an input feature for classification, this may create a challenge for conventional classification methods. Therefore, we will use feature extraction as a preprocessing to reduce the dimensionality of the generated morphological profiles before classification.

4.4 Generalized SELD for feature extraction of MPs

A number of approaches exist for feature extraction of the generated morphological profiles [17–19, 149], ranging from unsupervised methods to supervised ones. One of the best known unsupervised methods is Principle Component Analysis (PCA) [37], which is widely used [17, 18, 78]. Green et al. [150] introduced the minimum noise fraction (MNF) transformation. Recently, some local methods, which preserve the properties of local neighborhoods were used to reduce the dimensionality of hyperspectral images [48, 49, 56], such as Locally Linear Embedding [48], Neighborhood Preserving Embedding (NPE) [53]. By considering neighborhood information around the data, these local methods can preserve local neighborhood information and detect the manifold embedded in the high-dimensional feature space.

We addressed supervised FE methods in section 2.4 with special attention to LDA and NWFE. Many extensions to these two methods have been proposed in recent years, such as modified Fisher's linear discriminant analysis [58], regularized linear discriminant analysis [36], modified nonparametric weight feature extraction using spatial and spectral information [59], and kernel nonparametric weighted feature extraction [60].

However, in real-world applications, labeled samples are usually very limited, while unlabeled ones are available in large quantities at very low cost. Recently, some semi-supervised feature extraction methods were proposed to reduce the dimension of hyperspectral data sets. The approach of [68] proposed a general semi-supervised dimensionality reduction framework based on pairwise constraints, which employs regularization with sparse representation. In an earlier work [79], we proposed a semi-supervised local discriminant analysis (SELD) method, which combines LDA and NPE, to extract features from the original hyperspectral data. In this paper, we propose a generalized SELD (GSELD) to extract features from the generated morphological profiles.

Let $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in R^d$ denote high-dimensional data, $\{\mathbf{z}_i\}_{i=1}^N$, and $\mathbf{z}_i \in R^r$ the low-dimensional representations of the high-dimensional data $r \leq d$. In our application, *d* is the dimensionality of the generated profiles, and *r* is the dimensionality of the extracted features. The goal of linear feature extraction is to find a $d \times r$ projection matrix $\mathbf{W}$, which can map every high-dimensional data $\mathbf{x}_i$ to $\mathbf{z}_i = \mathbf{W}^T\mathbf{x}_i$ such that most information of the high-dimensional data is kept in a much lower dimensional feature space.

Focusing on class discrimination, LDA is in general well suited to preprocessing for the task of classification, since the transformation improves class separation. However, when only a small number of labeled samples are available, LDA tends to perform poorly due to overfitting (see Fig. 3.3-Fig. 3.6. Moreover, as the rank of the between-class scatter matrix S_b is $C-1$, the LDA can extract at most $C-1$ features, which is not always sufficient to represent essential information of the original data. NPE works directly on the data without any ground truth, and incorporates the local neighborhood information of data points in its feature extraction process. In Chapter 3, we combined LDA and NPE in a new framework, and proposed a semi-supervised local discriminant analysis (SELD) method to extract features from the original hyperspectral data. SELD magnified the advantages of LDA and NPE, and compensated for disadvantages of the two at the same time. In this Chapter, we propose a new semi-supervised method to extract features from the generated morphological profiles. The proposed method extends our SELD method from Chapter 3 with a tunable parameter and we abbreviate this generalized SELD method as GSELD.

Suppose a training data set is made up of the labeled set $\mathbf{X}_{labeled} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{1, 2, \cdots, C\}$, *C* is the number of classes, and unlabeled set $\mathbf{X}_{unlabeled} = \{\mathbf{x}_i\}_{i=n+1}^N$. The *k*th class has n_k samples with $\sum_{k=1}^C n_k = n$. Without loss of generality, we center the data points by subtracting the mean vector from all the sample vectors, and assume that the labeled samples in $\mathbf{X}_{labeled} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ are ordered according to their labels, with data matrix of the *k*th class $\mathbf{X}^{(k)} = [\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \cdots, \mathbf{x}_{n_k}^{(k)}]$ where $\mathbf{x}_i^{(k)}$ is the *i*th sample in the *k*th class. Then the labeled set can be expressed as $\mathbf{X}_{labeled} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(n)}]$, all training set $\mathbf{X} = [\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}]$. The optimization problem of the proposed GSELD is:

$$\mathbf{w}_{GSELD} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}(\alpha\mathbf{P} + \underline{\mathbf{I}})\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}(\alpha(\bar{\mathbf{I}} - \mathbf{P}) + \mathbf{M})\mathbf{X}^T\mathbf{w}} \tag{4.1}$$

where the matrices $\mathbf{P}$ and $(\bar{\mathbf{I}} - \mathbf{P})$ are from the reformulation of LDA part, and the matrices $\underline{\mathbf{I}}$ and $\mathbf{M}$ are from the reformulation of NPE part, for more details, the readers should consult [79]. α is the tunable parameter. When α is set to zero, equation 4.1 reduces to 2.10. When the parameter α is set to 1, the proposed method reduces to SELD [79]. Let $\overline{\mathbf{S}}_{GSELD} = \mathbf{X}(\alpha\mathbf{P} + \underline{\mathbf{I}})\mathbf{X}^T$ and $\underline{\mathbf{S}}_{GSELD} = \mathbf{X}(\alpha(\bar{\mathbf{I}} - \mathbf{P}) + \mathbf{M})\mathbf{X}^T$, we can solve the generalized eigenvalue problem of GSELD as (2.2), and get the projection matrix $\mathbf{W}$.

The algorithmic procedure of the proposed method which uses GSELD to extract features from the generated MPs is formally stated below:

1. Use PCA to extract the most p significant principal components (usually with cumulative variance near to 99%) from the original hyperspectral data sets.

2. Build the MPs on the p extracted PCs. The MPs are defined in the same way as in [17,73]. An MP consists of the original image (one of the PC features) and M openings with SE of increasing size (all applied on the original image) and M closings with the same SE. Then, an Extended Morphological Profile (EMP) is obtained with $d = p \times (2M + 1)$ dimension.

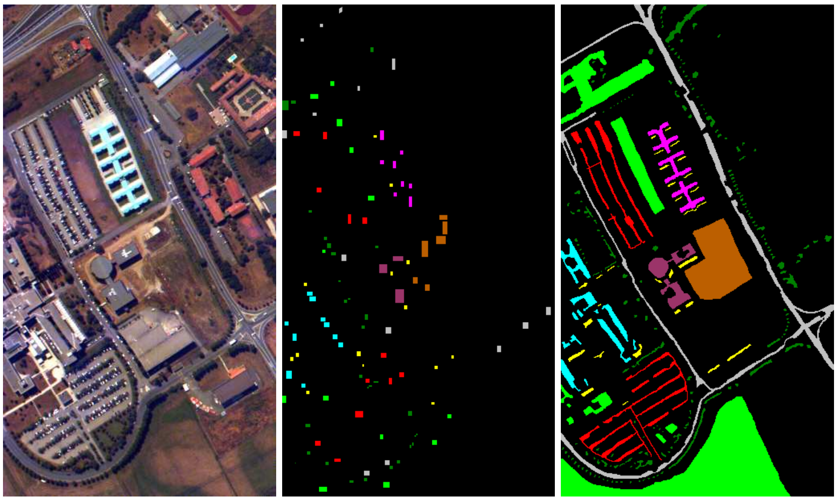
3. Divide the training samples into two subsets. Suppose that the labeled samples in $\mathbf{X}_{labeled} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ are ordered according to their labels, with data matrix of the kth class $\mathbf{X}^{(k)} = [\mathbf{x}_1^{(k)}, \cdots, \mathbf{x}_{n_k}^{(k)}]$ where $\mathbf{x}_i^{(k)}$ is the ith sample in the kth class, then the labeled set can be expressed as $\mathbf{X}_{labeled} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(C)}]$. The unlabeled set is denoted as $\mathbf{X}_{unlabeled} = \{\mathbf{x}_i\}_{i=n+1}^{N}$.

4. Construct the matrices $\mathbf{P}$ and $\bar{\mathbf{I}}$ from the labeled samples, and construct the matrix $\underline{\mathbf{I}}$ and $\mathbf{M}$ from the unlabeled samples in the same way as Chapter 3.

5. Compute the eigenvectors and eigenvalues for the generalized eigenvector problem in (2.2). The projection matrix $\mathbf{W}_{GSELD} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$ is made up by the r eigenvectors of the matrix $\underline{\mathbf{S}}_{GSELD}^{-1}\overline{\mathbf{S}}_{GSELD}$ associated with the largest r eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

6. Project the high dimensional generated morphological profiles ($\mathbf{x}_i \in R^d$) into a lower dimensional subspace ($\mathbf{z}_i \in R^r$) by
$$\mathbf{x} \to \mathbf{z} = \mathbf{W}_{GSELD}^T\mathbf{x}$$

7. Use these extracted features $\mathbf{Z}$ in the lower dimensional subspace as an input to do classification.



(a) False color image (b) Training set (c) Test set

Figure 4.6: University Area data set.

4.5 Experimental results

4.5.1 Hyperspectral data sets

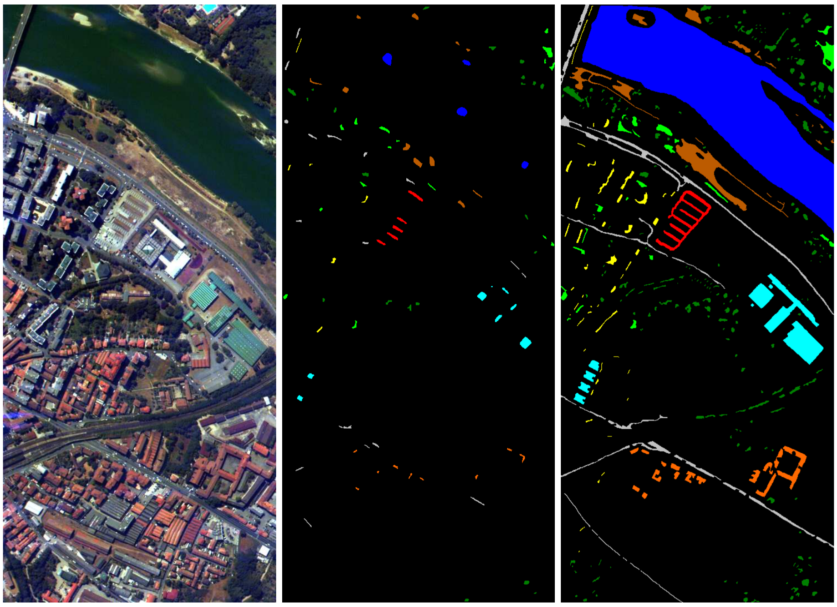
Experiments were run on two data sets, namely the '*Pavia Center*' and '*University Area*'. The data sets are from urban areas in the city of Pavia, Italy. The data were collected by the ROSIS (Reflective Optics System Imaging Spectrometer) sensor, with 115 spectral bands in the wavelength range from 0.43 to 0.86μm and very fine spatial resolution of 1.3 meters by pixel.

Pavia Center: The data with 1096×492 pixels was collected over Pavia city center, Italy. It contains 102 spectral channels after removal of noisy bands (see Fig. 4.7(a) for a color composite). Nine groundtruth classes were considered in experiments, see Table 4.1. Note that the color in the cell denotes different classes in the classification maps (Fig. 4.6 - Fig. 4.9).

University Area: The data with 610×340 pixels was collected over the University of Pavia, Italy. It contains 103 spectral channels after removal of noisy bands (see Fig. 4.6(a) for a color composite). The data also includes 9 land cover/use classes, see Table 4.1.

4.5.2 Experimental setup

To apply the morphological profiles with partial reconstruction and directional morphological profiles of [73] from panchromatic imagery to hyperspectral images, principal component analysis (PCA) was first applied to the original hyper-



(a) False color image (b) Training set (c) Test set

Figure 4.7: Pavia Center data set.

Pavia Center			University Area		
Class Name	# Training set	# Test set	Class Name	# Training set	# Test set
Water	745	65278	Asphalt	548	6641
Trees	785	6508	Meadows	540	18649
Meadows	797	2905	Gravel	392	2099
Bricks	485	2140	Trees	524	3064
Soil	820	6549	Metal Sheets	265	1345
Asphalt	678	7585	Soil	532	5029
Bitumen	808	7287	Bitumen	375	1330
Tiles	223	3122	Bricks	514	3682
Shadows	195	2165	Shadows	231	947

Table 4.1: Training and test samples for data sets used in the experiments

spectral data set, and the first 3 principal components (PCs) were selected (representing 99% of the cumulative variance) to construct the MPs. For disk-shaped structuring elements, morphological profiles with 15 openings and closings (ranging from 1 to 15 with step size increment of 1) were then computed for each PC. For linear structuring elements, morphological profiles with only 15 closings (ranging from 10 to 150 with step size increment of 10) were constructed for each PC, since objects like roads in the extracted PCs proved to be mostly dark compared to the background, we only made use of closing transforms. As a result, each disk-based profile was made up of 31 bands and the final disk-based MPs, constructed using three principal components, consisted of 93 bands. The final MPs based on both disk and linear SEs were 138 bands.

We used three common classifiers: 1-nearest neighbor (1NN), linear discriminant classifier (LDC) [141], and support vector machines (SVM) [142]. The SVM classifier with radial basis function (RBF) kernels in Matlab SVM Toolbox, LIBSVM [143], is applied in our experiments. SVM with RBF kernels has two parameters: the penalty factor C and the RBF kernel width γ. We apply a grid-search on C and γ using 5-fold cross-validation to find the best C within the given set $\{10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and the best γ within the given set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$.

In order to investigate the influences of the training samples size in more detail, the training data sets were then randomly subsampled to create samples whose sizes corresponded to five distinct cases: 10, 20, 40, 80 and 160 samples per class, respectively. All classifiers were evaluated against the testing sets, the results were averaged over five runs. The word ‘Reconstruction’ in the tables is shortened as ’Re.’.

4.5.3 Results using morphological profiles with partial reconstruction and directional MPs

We compared the MPs with reconstruction, without reconstruction, and with partial reconstruction in both two data sets. We also compared the results with the directional MPs. Since Gaussian Classifier LDC is not efficient to deal with high-dimensional data, we use 1NN and SVM classifiers in this experiment. The resulting accuracies are shown in Table 4.2-Table 4.3. The best overall accuracy (OA) of each data set in each training sample size is highlighted (in column) in bold font.

From these tables, we have the following findings:

1. The results confirm that the MPs (without reconstruction, with reconstruction, and with partial reconstruction) can improve the classification performance on hyperspectral images. By building the extended morphological profiles on the first 3 principal components, the results can be improved a lot. Compared to the situation with only spectral bands in each training sample size, the OA of Pavia Center and University Area data sets with MPs have 0.2%-2.6% and 12.4%-20% improvements for the 1NN classifier, respectively. For SVM classifier, these improvements are 2%-3.3% and 1.5%-25.5%, respectively.

Dataset	Methods	Classifier	Training Set Size				
			10	20	40	80	160
Pavia Center	Spectral Only	1NN	0.941	0.948	0.952	0.956	0.961
		SVM	0.935	0.942	0.946	0.947	0.948
	No Re.	1NN	0.943	0.956	0.96	0.964	0.965
		SVM	0.961	0.968	0.974	0.977	0.979
	Re.	1NN	0.96	0.97	**0.976**	**0.981**	**0.983**
		SVM	**0.966**	0.968	0.974	0.979	0.981
	Partial Re.	1NN	0.949	0.963	0.967	0.971	0.973
		SVM	0.963	**0.971**	**0.976**	0.98	0.981
University Area	Spectral Only	1NN	0.626	0.637	0.644	0.678	0.69
		SVM	0.653	0.729	0.725	0.734	0.787
	No Re.	1NN	0.818	0.826	0.833	0.841	0.837
		SVM	0.825	0.884	0.886	0.896	0.894
	Re.	1NN	0.75	0.782	0.786	0.823	0.825
		SVM	0.709	0.766	0.799	0.797	0.802
	Partial Re.	1NN	0.806	0.806	0.809	0.829	0.821
		SVM	**0.835**	**0.894**	**0.909**	**0.916**	**0.917**

Table 4.2: Overall Accuracy in a Classification with Spectral Only Compared to Classifications with Disk-based MPs without Reconstruction, with Reconstruction, and with Partial Reconstruction

Dataset	Methods	Classifier	Training Set Size				
			10	20	40	80	160
Pavia Center	No Re.	1NN	0.953	0.962	0.968	0.971	0.973
		SVM	0.966	0.972	0.976	0.981	0.981
	Re.	1NN	0.961	0.971	0.976	0.981	**0.983**
		SVM	0.967	0.972	0.975	0.979	0.981
	Partial Re.	1NN	0.957	0.968	0.974	0.976	0.978
		SVM	**0.968**	**0.975**	**0.979**	**0.982**	0.982
University Area	No Re.	1NN	0.826	0.834	0.841	0.85	0.85
		SVM	0.854	0.907	0.912	0.918	0.916
	Re.	1NN	0.757	0.776	0.779	0.814	0.816
		SVM	0.733	0.763	0.815	0.799	0.81
	Partial Re.	1NN	0.83	0.829	0.839	0.857	0.848
		SVM	**0.884**	**0.924**	**0.941**	**0.947**	**0.95**

Table 4.3: Overall Accuracy Comparison in a Classification among Disk- & Linear-based MPs without Reconstruction, with Reconstruction, and with Partial Reconstruction

	Spectral Only	Disk-based MP			Disk- and Linear-based MP		
		No Re.	Re.	Partial Re.	No Re.	Re.	Partial Re.
OA	95.9	97.9	98.2	98	**98.3**	98.1	**98.3**
AA	91	94.2	96.8	96.2	96.2	96.6	**97**
κ	93.4	96.4	96.9	96.6	**97.1**	96.8	**97.1**
std	0.92	0.43	0.55	0.36	0.41	0.44	**0.27**
Water	98.9	**100**	99.3	99.7	99.9	99.3	99.5
Trees	87.2	93	93.1	**93.8**	92.7	93.2	93.2
Meadows	**94.6**	85.5	85	91	90	85.6	93.5
Bricks	62.6	84.1	**99.8**	98.1	96.4	99.6	99.3
Soil	94.8	95.8	96	97	96.7	95.7	**97.8**
Asphalt	94.5	96.2	**98.8**	97.7	98.2	97.8	98.4
Bitumen	86.6	95	97.3	90.2	93.7	**97.9**	92.7
Tiles	99.6	**100**	99.9	**100**	**100**	99.9	**100**
Shadows	**100**	98.7	99.9	98.7	98	**100**	98.9

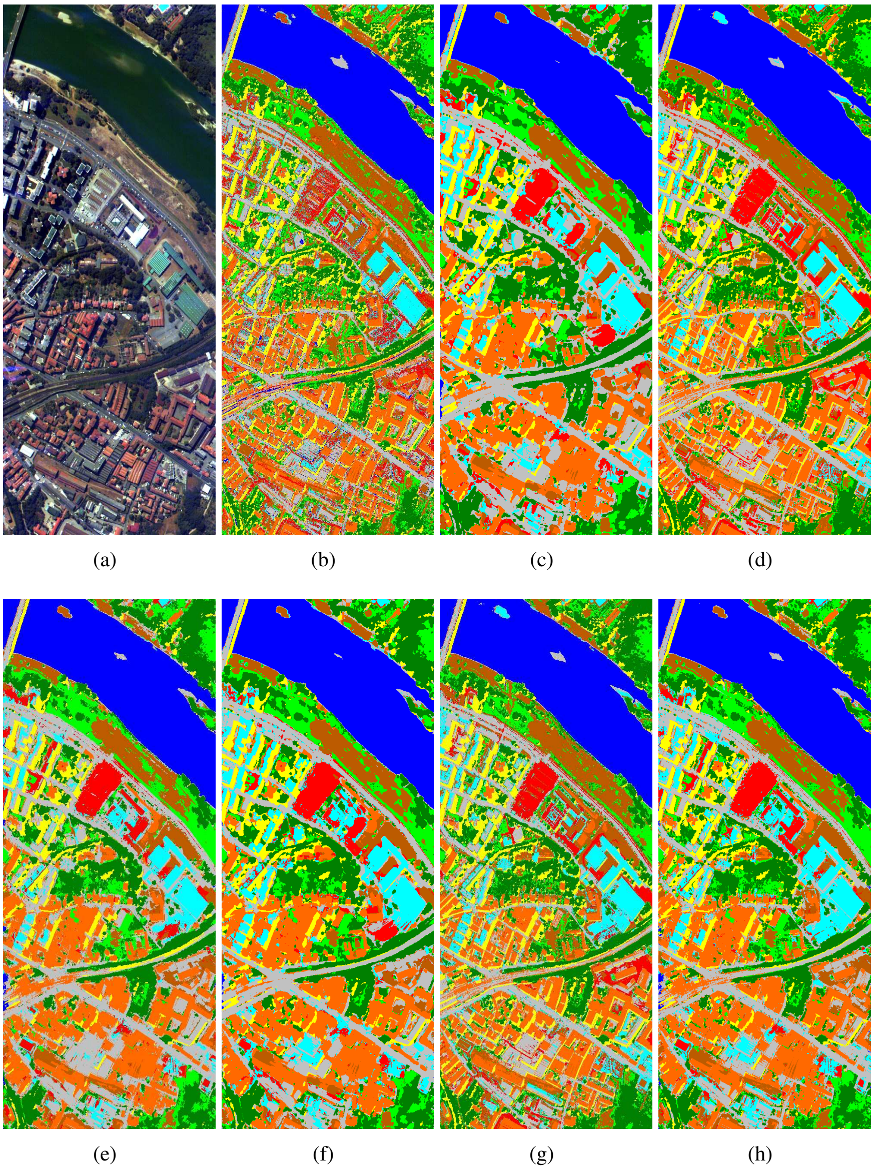
Table 4.4: Pavia Center: Best Classification Accuracy (%) over ten runs for Classification Maps in Fig. 4.8, 20 training samples per class were used.

2. As the number of training samples increases, the OA will increase. Especially for SVM classifier, in the Pavia Center data set, the OA of spectral only has 2% improvements from 10 training samples per class to 160 training samples, this also happens on MPs with nearly 2% improvements; in the University Area data set, the OA of spectral only increases from 65.3% to 78.7% when the number of training samples per class changes from 10 to 160, while MPs with almost 7% improvements.

3. The results can be improved by adding the directional MPs. There is a substantial improvement of the overall accuracy over the classification with only disk-based MPs. However, when using MPs with reconstruction, the classification accuracies by adding the directional MPs improves very little and is comparatively much less than those without reconstruction and with partial reconstruction. This is because the disk-based MPs and linear-based MPs with reconstruction contain much the same information.

4. It is better not to use MPs with reconstruction in some cases. This is in particular the case in University Area data set, where the MPs with reconstruction perform even worse than MPs without reconstruction. The MPs with partial reconstruction and SVM classifier almost gets the best results all the time, this is obvious in University Area data set.

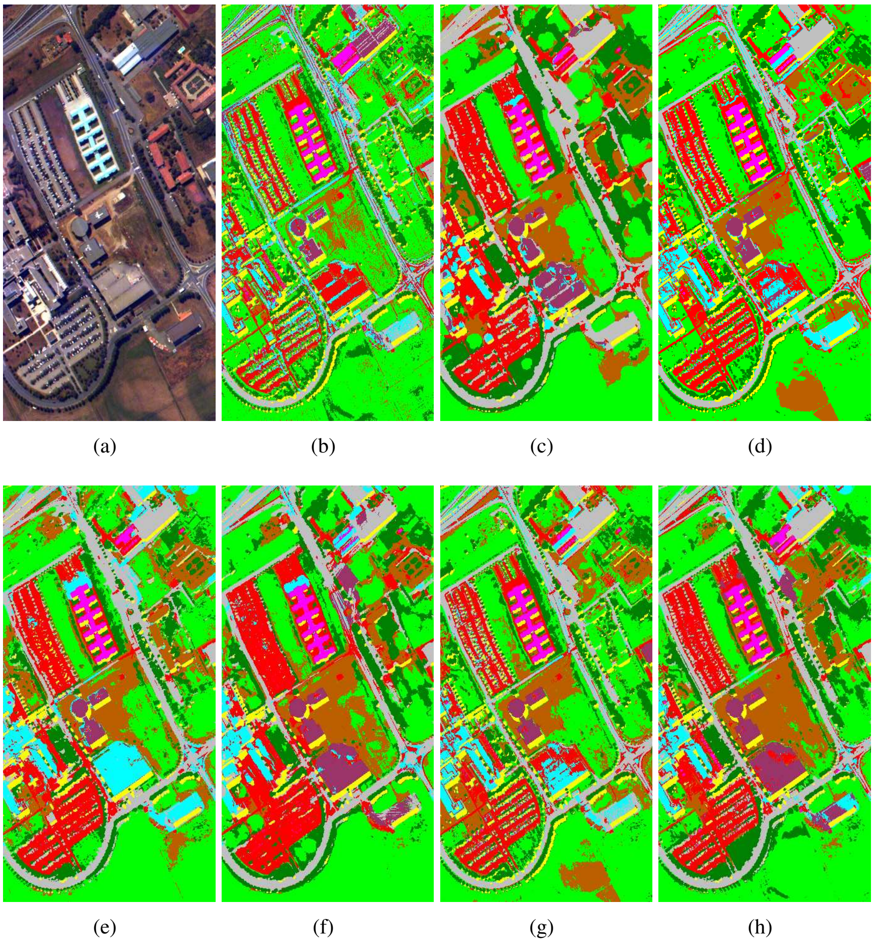
In order to compare the classification results visually, we randomly select 20 training samples per class for training, and use all the samples for testing. The SVM classifier was used. The best results over ten runs are shown in Fig. 4.8-Fig. 4.9 and Table 4.4-Table 4.5. The Z tests [103] were reported Table 4.6-Table 4.7.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4.8: Classification maps for Pavia Center with best classification accuracy over ten runs, 20 training samples per class with SVM classifier were used. (a) False color image, and thematic map using (b) Spectral Only, (c) Disk-based MP without reconstruction, (d) Disk-based MP with reconstruction, (e) Disk-based MP with partial reconstruction, (f) Disk- and linear-based MP without reconstruction, (g) Disk- and linear-based MP with reconstruction, and (h) Disk- and linear-based MP with partial reconstruction.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4.9: Classification maps for University Area with best classification accuracy over ten runs, 20 training samples per class with SVM classifier were used. (a) False color image, and thematic maps using (b) Spectral Only, (c) Disk-based MP without reconstruction, (d) Disk-based MP with reconstruction, (e) Disk-based MP with partial reconstruction, (f) Disk- and linear-based MP without reconstruction, (g) Disk- and linear-based MP with reconstruction, and (h) Disk- and linear-based MP with partial reconstruction.

	Spectral Only	Disk-based MP			Disk- and Linear-based MP		
		No Re.	Re.	Partial Re.	No Re.	Re.	Partial Re.
OA	78.1	89.1	80.3	90.5	92.6	81.7	**93.8**
AA	78.7	85.6	86.7	89.2	90.2	88	**93.7**
κ	71	85.3	74.5	87.2	90.1	76.3	**91.9**
std	4.23	0.77	2.42	**0.73**	1.21	4.15	1.14
Asphalt	59.3	88.6	80.7	86.5	82.9	87.9	**89.6**
Meadows	89.6	98.1	77.8	98.4	**99.5**	77.5	94.6
Gravel	53.2	43.9	74.2	70.7	66.4	**77.1**	69
Trees	91.2	95.7	92.8	86.7	93.6	97.7	**97.8**
Metal Sheets	98.9	**99.9**	99.6	99.6	99.6	99.2	99.5
Soil	48.8	60.9	56.9	65.6	84.5	58.6	**97.5**
Bitumen	86.2	90.4	99.6	97.4	98	99.1	**99.9**
Bricks	81	96.2	**98.6**	98.2	95.7	95.1	98.4
Shadows	**100**	97.2	99.9	99.8	91.9	99.9	97.6

Table 4.5: University Area: Best Classification Accuracy (%) over ten runs for Classification Maps in Fig. 4.9, 20 training samples per class were used.

1. The MPs (without reconstruction, with reconstruction, and with partial reconstruction) can preserve well spatial information on hyperspectral images. The classification maps with MPs produce much smoother homogeneous regions than that of spectral only, which is particularly significant when using MPs with no reconstruction and with partial reconstruction, see Table 4.6-Table 4.7. The statistical difference of accuracy $|Z| > 1.96$ clearly demonstrates the benefit of using the MPs with no reconstruction and with partial reconstruction rather than the spectral only.

2. The classification maps using the MPs with reconstruction look much noisier because of the over reconstruction problems. The MPs with no reconstruction deform the objects, see Fig. 4.9(c) and Fig. 4.9(f), the borders of some objects are deformed. While small objects might be fused together (e.g., the buildings and shadows in the bottom part of the Pavia center image) when using the MPs with partial reconstruction and no reconstruction, in this case, the MPs with full reconstruction perform better.

3. When using both disk-based and directional MPs with partial reconstruction, we get the best OA, AA and $Kappa$ for both data sets, and relative lower standard deviation (std). For University Area data set, the difference is statistically significant. For Pavia Center data set, the difference is not statistically significant with $|Z| < 1.96$.

4.5.4 Results using semi-supervised feature extraction to reduce the dimensionality of the generated MPs

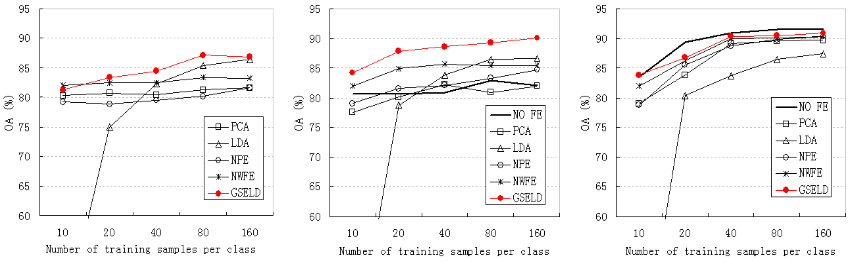
We compare the resulting classification accuracies using the proposed GSELD method to extract features from the generated morphological profiles with those

Table 4.6: University Area: Statistical Significance of Differences in Classification (Z) over ten runs. Each case of the table represents Z_{rc} where r is the row and c is the column, 20 training samples per class with SVM classifier were used.

Z_{rc}		Spectral Only	Disk-based MP			Disk- and Linear-based MP		
			No Re.	Re.	Partial Re.	No Re.	Re.	Partial Re.
Spectral Only		0	-4.0065	-1.1304	-4.5819	-4.7267	-0.8276	-5.3427
Disk-based MP	No Re.	4.0065	0	4.1963	-1.8734	-2.1650	2.6714	-3.7365
	Re.	1.1304	-4.1963	0	-5.0917	-5.1235	0.1011	-6.0577
	Partial Re.	4.5819	1.8734	5.0917	0	-0.8098	3.2139	-2.3441
Disk- and Linear-based MP	No Re.	4.7267	2.1650	5.1235	0.8098	0	3.4092	-1.2275
	Re.	0.8276	-2.6714	-0.1011	-3.2139	-3.4092	0	-3.9860
	Partial Re.	5.3427	3.7365	6.0577	2.3441	1.2275	3.9860	0

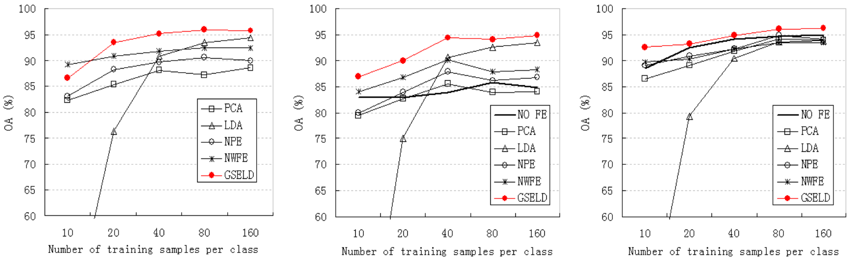
Table 4.7: Pavia Center: Statistical Significance of Differences in Classification (Z) over ten runs. Each case of the table represents Z_{rc} where r is the row and c is the column, 20 training samples per class with SVM classifier were used.

Z_{rc}		Spectral Only	Disk-based MP			Disk- and Linear-based MP		
			No Re.	Re.	Partial Re.	No Re.	Re.	Partial Re.
Spectral Only		0	-2.8358	-2.5593	-2.9667	-3.1620	-2.7805	-3.3847
Disk-based MP	No Re.	2.8358	0	0.1722	-0.0756	-0.4919	0.0670	-0.6849
	Re.	2.5593	-0.1722	0	-0.2473	-0.5993	-0.1123	-0.7600
	Partial Re.	2.9667	0.0756	0.2473	0	-0.4601	0.1472	-0.6827
Disk- and Linear-based MP	No Re.	3.1620	0.4919	0.5993	0.4601	0	0.5541	-0.1174
	Re.	2.7805	-0.0670	0.1123	-0.1472	-0.5541	0	-0.7521
	Partial Re.	3.3847	0.6849	0.7600	0.6827	0.1174	0.7521	0



(a) LDC classifier (b) 1NN classifier (c) SVM classifier

Figure 4.10: Highest OA of University Area in different samples size with partial reconstruction based on only disk-based MPs, the number of extracted features changed from 1 to 20, each experiment was repeated 5 times, the average was acquired.



(a) LDC classifier (b) 1NN classifier (c) SVM classifier

Figure 4.11: Highest OA of University Area in different samples size with partial reconstruction based on both disk-based and linear-based MPs, the number of extracted features changed from 1 to 20, each experiment was repeated 5 times, the average was acquired.

resulting from the following methods: PCA [37]; LDA [57]; NPE [53]; NWFE [35]. The data sets of *University Area* is used. In our experiments, $u = 1500$ unlabeled samples are randomly selected for training the proposed GSELD, the parameter α in (4.1) is set as $\alpha = \frac{u}{n}$ (n is the number of labeled training samples), which can change automatically according to the ratio of the number of unlabeled and labeled samples while increasing the class separability. 20 features (except only for the $C - 1$ features in LDA) are extracted, then, the testing accuracies of each employed number of features are calculated respectively. The highest OA with three classifiers in different samples size are shown in Fig. 4.10-Fig. 4.11, the number of extracted features changed from 1 to 20, each experiment was repeated 5 times, the average was acquired.

1. The results confirm that feature extraction can improve the classification performance. Especially for conventional classifiers (such as LDC classifier), FE makes the classification possible. For 1NN classifier, the results of Uni-

versity Area data set can be improved a lot by using FE as a preprocessing.

2. SVM classifier is more efficient to deal with the high dimensional data, this is obvious in University Area data set, see Fig. 4.10(c) and Fig. 4.11(c). In some cases, it can achieve even better performances than those using FE as a preprocessing, see Fig. 4.10(c). When using only the disk-based MPs, SVM classifier with no FE outperforms those with FE as a preprocessing.

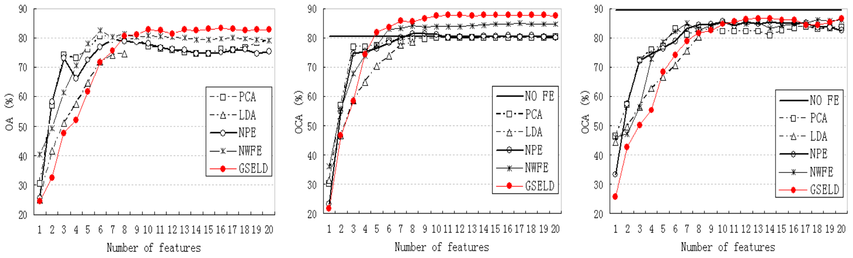
3. For the LDC and the SVM classifiers, as the number of training samples per class increases, the OA of each method will increase. This is particular for the supervised LDA method, when the number of training samples per class is 10, the OA is much lower than 60%. When the number of training samples per class is more than 80, the OA of LDA increases above 80%.

4. For low resolution Indian Pine data set, when the training sample size increases, the proposed GSELD with LDC classifier outperforms the other methods with LDC classifier. While NWFE with KNN classifier performs a little bit better than GSELD with KNN classifier. When using SVM classifier, the OA of GSELD is similar with that of NWFE.

5. For high resolution urban data set (University Area), when using both the disk-based and linear-based morphological features, the proposed GSELD gets the highest OA in different samples size. The highest OA for training samples size with 10, 20, 40, 80 and 160 are 92.5% (GSELD with SVM classifier), 93.4% (GSELD with LDC classifier), 95.1% (GSELD with LDC classifier), 96% (GSELD with SVM classifier) and 96.2% (GSELD with SVM classifier), respectively.

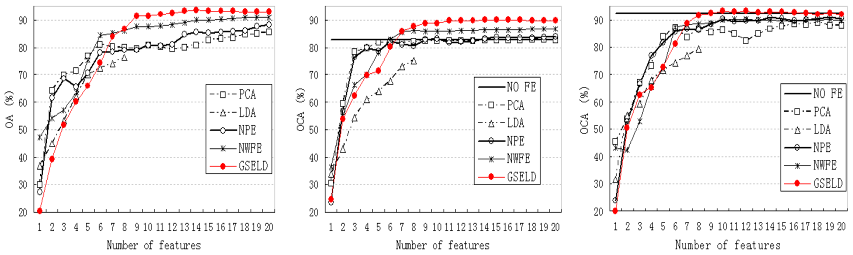
The experiments were carried out on 64-b, 2.67 GHz Intel i7 920 (8 core) CPU computer with 12 GB memory, the time was only consumed in the process of feature extraction for MPs based on both disk and linear SEs with Partial Reconstruction. When the training sample size of University Area data set changes from 80 to 160, the consumed time of NWFE increases from 14.8 seconds to 113.9 seconds, while for the proposed GSELD, the consumed time increases from 2.4 seconds to 5.3 seconds. Fig. 4.12 and Fig. 4.13 show the performances with different number of extracted features when 20 training samples per class are used as training set. The Z tests using MPs based on both disk and linear SEs with Partial Reconstruction were reported in Table 4.8-Table 4.10. The results confirm some findings in Fig. 4.10 and Fig. 4.11, moreover we find the following:

1. Most information of the generated MPs can be preserved even with a few extracted features. For 1NN classifier, when the number of extracted features is more than 7, the results of NWFE and GSELD are better than that without FE. When using both disk-based and linear-based MPs, the difference is



(a) LDC classifier (b) 1NN classifier (c) SVM classifier

Figure 4.12: Performance of each feature extraction method using 20 training samples per class for University Area data set, the MPs are based on only disk SE with Partial Reconstruction. Each experiment was repeated 5 times, the average was acquired.



(a) LDC classifier (b) 1NN classifier (c) SVM classifier

Figure 4.13: Performance of each feature extraction method using 20 training samples per class for University Area data set, the MPs are based on both disk and linear SEs with Partial Reconstruction. Each experiment was repeated 5 times, the average was acquired.

statistically significant with $|Z| > 1.96$. For the SVM classifier using both disk-based and linear-based MPs, the proposed GSELD gets better result even with 9 extracted features.

2. Using only $C - 1$ features may not be enough in some situation, which is one limitation of LDA. PCA and NPE can improve their performances by using more extracted features, as shown in Fig. 4.12 and Fig. 4.13. When more features are used, the overall classification accuracy can be improved with statistical significance ($|Z| > 1.96$).

3. The proposed GSELD outperforms the other feature extraction methods with all these three classifiers, with $Z > 0$. When using both the disk-based and linear-based morphological features, the proposed GSELD gets the highest OA for all these three classifiers. The highest OA for LDC classifier, 1NN classifier and SVM classifier are 93.4% (GSELD with 14

Z_{rc}	PCA	LDA	NPE	NWFE	GSELD
PCA	0	2.6330	-1.2986	-1.4557	-2.7927
LDA	-2.6330	0	-3.8497	-3.9939	-5.4819
NPE	1.2986	3.8497	0	-0.1588	-1.2281
NWFE	1.4557	3.9939	0.1588	0	-1.0275
GSELD	2.7927	5.4819	1.2281	1.0275	0

Table 4.8: LDC classifier: Statistical Significance of Differences in Classification (Z) over five runs. Each case of the table represents Z_{rc} where r is the row and c is the column, 20 training samples per class were used. The MPs are based on both disk and linear SEs with Partial Reconstruction.

Z_{rc}	NO FE	PCA	LDA	NPE	NWFE	GSELD
NO FE	0	0.0368	1.3690	-0.3187	-1.0036	-2.7382
PCA	-0.0368	0	1.3503	-0.3581	-1.0452	-2.8113
LDA	-1.3690	-1.3503	0	-1.5848	-2.0409	-3.2841
NPE	0.3187	0.3581	1.5848	0	-0.7027	-2.3453
NWFE	1.0036	1.0452	2.0409	0.7027	0	-1.3659
GSELD	2.7382	2.8113	3.2841	2.3453	1.3659	0

Table 4.9: 1NN classifier: Statistical Significance of Differences in Classification (Z) over five runs. Each case of the table represents Z_{rc} where r is the row and c is the column, 20 training samples per class were used. The MPs are based on both disk and linear SEs with Partial Reconstruction.

Z_{rc}	NO FE	PCA	LDA	NPE	NWFE	GSELD
NO FE	0	0.6960	3.5715	0.3081	0.6967	-0.1704
PCA	-0.6960	0	3.3101	-0.3829	0.1476	-1.1095
LDA	-3.5715	-3.3101	0	-3.4251	-2.8042	-4.1037
NPE	-0.3081	0.3829	3.4251	0	0.4424	-0.5633
NWFE	-0.6967	-0.1476	2.8042	-0.4424	0	-0.9627
GSELD	0.1704	1.1095	4.1037	0.5633	0.9627	0

Table 4.10: SVM classifier: Statistical Significance of Differences in Classification (Z) over five runs. Each case of the table represents Z_{rc} where r is the row and c is the column, 20 training samples per class were used. The MPs are based on both disk and linear SEs with Partial Reconstruction.

extracted features), 90% (GSELD with 14 extracted features) and 93.2% (GSELD with 10 extracted features), respectively.

4.6 Conclusion

In this Chapter, we first investigated the morphological profiles with partial reconstruction and directional morphological profiles for the classification of high resolution hyperspectral images from urban areas. We showed on two real urban hyperspectral data sets that the MPs with partial reconstruction are more competitive than those with no reconstruction and with reconstruction, and some classes like road are classified better with the directional morphological features. Secondly, we developed a semi-supervised feature extraction as a preprocessing tool to reduce the dimensionality of the generated morphological profiles for classification. The results show that feature extraction can improve significantly the performance for some classifiers, and the proposed semi-supervised method compares favorably with conventional feature extraction methods as preprocessing approaches for the morphological profiles generated on high resolution hyperspectral data from the urban area.

5

Kernel features

A fast iterative Kernel Principal Component Analysis (KPCA) is proposed to extract features from hyperspectral images. The proposed method is a kernel version of the Candid Covariance-Free Incremental Principal Component Analysis, which solves the eigenvectors through iteration. Without performing eigen decomposition on Gram matrix, our method can reduce the space complexity and time complexity greatly. Experimental results were validated in comparison with the standard KPCA and linear version methods.

We investigated the influence of morphological features with different reconstruction (including with no reconstruction, with reconstruction and with partial reconstruction) for the classification of high resolution hyperspectral images from urban areas. To apply morphological profiles on hyperspectral data, we first reduced the dimensionality of hyperspectral data by feature extraction, then built the extended morphological profiles on the extracted features. We showed on two real hyperspectral data sets that KPCA is more efficient to extract features for constructing EMP. In many cases, the most widely used EMP with reconstruction can not get a satisfied result, because of over-reconstruction problems. EMP with partial reconstruction built on KPCs is more competitive than those of EMP with no reconstruction and with reconstruction built on other different features.

5.1 Introduction

As it was already explained in Chapter 1, it is possible nowadays to collect hyperspectral images with hundreds of bands [28], while hyperspectral images contain much more information than regular RGB images, most of their information content can be explained by a small amount of the well extracted chosen features. The complexity of hyperspectral image processing techniques usually depends on the number of spectral bands in the acquired data. Therefore, it is necessary to find methods which can transform these high-dimensional HyperCube data into a lower dimensional space with reduced dimensionality, while at the same time, preserving as much information content as possible. In the previous Chapters we treated linear feature extraction methods, and in this Chapter we turn to nonlinear feature extraction (FE) methods

Conventional dimensionality-reduction techniques include unsupervised approaches such as Principal Component Analysis (PCA) [151], Minimum Noise Fraction [152] and independent component analysis [40], as well as supervised approaches, such as Fisher's linear discriminant analysis (LDA) [90]. Due to its low complexity and the absence of parameters, these linear methods have been widely used for feature extraction in hyperspectral images [17, 40, 153, 154]. However, they all depend on linear projection and can result in a loss of nonlinear properties of the original data after reduction of dimensionality. They are expected to be suboptimal (and even entirely fail) for nonlinear classification tasks (i.e., when the data distributions are such that the resulting decision boundaries are highly nonlinear).

Nonlinear FE methods attempt to address these problems. In the last decade, a large number of nonlinear techniques for dimensionality reduction have been proposed. See for an overview, e.g., [155–159]. In contrast to the traditional linear techniques, the nonlinear techniques have the ability to deal with complex nonlinear data. In particular for real world data, the nonlinear dimensionality reduction techniques may offer an advantage, because real world data is likely to form a highly nonlinear manifold. Previous studies have shown that nonlinear techniques outperform their linear counterparts on complex artificial tasks. For instance, the Swiss roll dataset comprises a set of points that lie on a spiral-like two-dimensional manifold that is embedded within a three-dimensional space. A vast number of nonlinear techniques are perfectly able to find this embedding, whereas linear techniques fail to do so.

5.1.1 Manifold learning and nonlinear dimensionality reduction

Two of the leading non-linear algorithms, in the field of dimensionality reduction are Isomap and Local Linear Embedding. Isomap [97] is a global non linear technique that operates on geodesic distances between data sets. Isomap first constructs the nearest neighbor graph of each point and then calculates the shortest paths. Each data point is connected with its nearest points with an edge that has as weight their Euclidean distance. However, global pairwise distances are calculated based on the shortest paths between all points (geodesic distance). The low dimensional mapping of the dataset is derived by the application of classic metric multidimensional scaling [160] on the geodesic distance matrix. The original Isomap approach exhibits a number of deficiencies when encountering curved manifolds or projecting large datasets. Towards solving these problems, de Silva and Tenenbaum introduced a improvement of their original algorithm, namely C-Isomap [161]. C-Isomap employs a different edge weighting scheme by taking also into account the mean distance of each point to its local neighbors.

Contrary to Isomap, LLE [98, 99] is a local non linear method that produces a number of local mappings based on each point's nearest neighbors. Additionally, LLE does not require all data to exist in a single coordinate system, only the existence of a relation between a point and its neighbors. The mode of operation is similar to that of Isomap. At first the nearest neighbors of each point are identified and based on them the linear reconstruction of the point is calculated. The embedding is derived by an eigen decomposition of the various reconstructions. More recently, Chang and Yeung [101] proposed robust locally linear embedding for nonlinear dimensionality reduction, and they demonstrated that the method is better suited for outlier problem. Chen and Qian [48] improved the existing LLE by introducing a spatial neighborhood window for hyperspectral dimensionality reduction. Isomap and LLE have been also employed for semi-supervised classification in the context [162].

Same as LLE, Laplacian Eigenmaps [51] operate on a geodesic distance matrix defined by the nearest neighbors of each point. However, unlike LLE, the embedding is derived by the eigenvectors of the graphs's Laplacian matrix. Local Tangent Space Alignment (LTSA) [52] is another unsupervised method for nonlinear dimension reduction. It describes local properties of the high-dimensional data using the local tangent space of each datapoint, and performs eigen decomposition on a matrix defined by the orthogonal basis of local data neighborhoods. Contrary to LTSA, its supervised version, S-LTSA [100] makes use of a-priori knowledge (i.e. data class membership) and is suitable for a continuous a changing environment. The approach of [49] proposed a supervised local manifold learning weighted KNN classifier for the classification of hyperspectral images, which combine local manifold learning (LLE, LTSA and LE) and the k-nearest-neighbor (kNN) classifier.

Another early and prominent method that has been extensively used in the area of supervised learning is the self organization technique of T. Kohonen [157]. Self Organizing Map (SOM) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space. Curvilinear Component Analysis [163] (CCA) is an enhancement of SOM and operates on the points pairwise distance matrix. An adaptation of CCA, Curvilinear Distance Analysis [163] (CDA) follows the principles of [164] and operates on the geodesic distance matrix of the points.

5.1.2 Kernel-based methods

Another approach to nonlinear FE is to kernelize classical linear methods. KPCA Kernel Principal Component Analysis (KPCA) [84] is a nonlinear version of PCA, which is more suitable to describe higher-order complex and nonlinear distributions. In [28], KPCA is used to extract features from hyperspectral images and performs well in terms of accuracy, comparing to the PCA. The results in [93,165] demonstrate a superior classification performance of GDA over LDA if the data in the input space possesses nonlinear class separation. GDA has been successfully employed for hyperspectral-data classification in [112]. In [166], Prasad and Bruce incorporated GDA within a multi-classifier and decision-fusion framework for HSI target recognition. The kernel-based LDFA [102] was applied to dimensionality reduction for hyperspectral image classification. Lai and Fyfe [167] described kernel canonical correlation analysis (CCA), and Bach and Jordan [168] described kernel independent component analysis (ICA) based upon kernel CCA. Excellent general references for kernel methods are [84, 113]. Kernel methods among many other subjects are described in [169,170]. In [171], kernel PCA is used for change detection in univariate image data. In [172], the kernel-based maximum autocorrelation factor and kernel MNF transformations were applied to change detection in hyperspectral data. In [60], the kernel method is applied to extend NWFE to kernel-based NWFE, with improved classification accuracies.

The central idea behind kernel-based methods is to map the input data onto an intermediate feature induced space (potentially possessing a much higher dimensionality), such that complex nonlinear decision boundaries in the input space become simpler linear decision boundaries in the kernel-induced space. Ham et al. [173] proposed a kernel interpretation of KPCA, Isomap, LLE, and Laplacian Eigenmap and demonstrated that they share a common KPCA formulation with different kernel definitions. However, the computational complexity of nonlinear methods are very intensive in computation and memory consumption. Taking KPCA as an example, in order to capture these nonlinear kernel principal components, a large number of training samples are required, particularly for the data embedded in a high dimensional space. This leads to problems for KPCA, since it has to store and manipulate the Gram matrix by calculating the kernel matrix. For example, if there are N samples in the training dataset, then the size of the

Gram matrix is N^2. Hence, the space complexity of storing the Gram matrix is $O(N^2)$, while the time complexity (performing eigen decomposition on a $N \times N$ Gram matrix) is $O(N^3)$ [174]. Using KPCA to extract features from hyperspectral images will cause some problems on storage resources and computational load. It was reported in [33, 175] that most nonlinear methods were incapable to handle hyperspectral images with sizes larger than 70 × 70. Some solutions were to divide hyperspectral images into small blocks, and to perform the KPCA feature extraction on each of these small blocks separately. In [28], some samples selected randomly from the original data are used to compute the Gram matrix, but the problems still exist.

In [176], an iterative kernel principal component analysis was proposed by reformulating the generalized Hebbian algorithm (GHA) [177] in a kernel space to obtain a memory efficient approximation of KPCA. However, the convergence speed of GHA is relatively slow which limits its application. In this Chapter, we first develop a fast iterative KPCA (FIKPCA) by using a different approach. We kernelize the Candid Covariance-Free Incremental PCA (CCIPCA) of [178], which was proved to converge fast [179]. Reformulate the CCIPCA in a kernel space to perform efficient and fast feature extraction from hyperspectral images. Instead of performing eigen decomposition on Gram matrix, the proposed FIKPCA solves eigenvectors through iteration, which can reduce the space and time complexities greatly, and it can process the hyperspectral images larger than 70 × 70 efficiently.

When applying morphological features for the classification of high resolution hyperspectral images from urban areas, one should consider another important issue except the two we considered in Chapter 4. The high dimensionality of these hyperspectral data as well as the redundancy within the bands, make the generation of an MP based on each spectral band seem not feasible. To overcome this problem, feature extraction is firstly used to reduce the dimensionality of these hyperspectral data, and then morphological processing is applied on each extracted feature band independently. Principal component analysis (PCA) [37] is the most popular method used to extract features for building MPs [17,18,39]. [17] extended the method in [78] for hyperspectral data with high spatial resolution by using PCA to reduce the high dimensionality of the data. The resulting method built the MPs on the first principal components (PCs) extracted from a hyperspectral image, leading to the definition of extended MPs (EMP). In [74], the morphological attribute profiles [75] were applied to the first PCs extracted from a hyperspectral image, generating an extended morphological attribute profiles (EAP). However, it was found that too much spectral information were lost during the linear principal component analysis (PCA) transformation [28, 87], as PCA relies on second-order statistics only. By taking high order statistics into account, independent component analysis (ICA) [85] has been studied to reduce the dimensionality of hyperspectral data [40, 180]. [87] built MPs on the first features extracted from original hyperspectral data by ICA, with an improvement in the classification results. [76] improved the classification results by building the EAPs on the first independent components (ICs) comparing to the results of those built on PCs [74]. Kernel PCA [181], which is more suitable to describe higher order complex and nonlinear distributions, has been recently investigated in reducing the dimensionality of hyperspectral remote sensing [19, 182]. In [28], kernel principal components are used to construct the EMP, with significantly improvement in terms of classification accuracies compared with the conventional EMP built on PCs.

5.2 Kernel feature extraction methods for hyperspectral data

5.2.1 Kernel principal component analysis

The standard KPCA solves the eigenvectors by performing eigen decomposition on Gram matrix. Suppose $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ is the matrix of original data, where $\mathbf{x}_n \in R^D$, $n = 1, 2, \cdots, N$, N is the total number of samples. There exists a function φ which can map the original data into a higher or infinite dimensional Hilbert space:

$$\varphi : \begin{array}{c} R^D \to H \\ \mathbf{x}_n \mapsto \varphi(\mathbf{x}_n) \end{array}$$

A new data set can be obtained in the feature space $\mathbf{\Phi} = (\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2), \cdots, \varphi(\mathbf{x}_n))$. φ is an implicit function, which cannot be calculated directly, but some kernel functions can be used by performing inner product between the two samples $\mathbf{x}_m$ and $\mathbf{x}_n$ in the original space, $\kappa_{mn} = \kappa(\mathbf{x}_m, \mathbf{x}_n) = \varphi^T(\mathbf{x}_m)\varphi(\mathbf{x}_n)$. The covariance matrix is defined in the feature space as follows:

$$\mathbf{C} = \frac{1}{N}\sum_{n=1}^{N} \varphi(\mathbf{x}_n)\varphi^T(\mathbf{x}_n) \tag{5.1}$$

It satisfies the secular equation:

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v} \tag{5.2}$$

where $\mathbf{v}$ and λ are the eigenvectors and eigenvalues of the covariance matrix $\mathbf{C}$, and $\mathbf{v}$ can be described in the span of the data set $\mathbf{\Phi} = (\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2, \cdots, \varphi(\mathbf{x}_n))$:

$$\mathbf{v} = \sum_{n=1}^{N} \alpha_n \varphi(\mathbf{x}_n) \tag{5.3}$$

From equations (5.1),(5.2),(5.3),

$$\frac{1}{N}\mathbf{K}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \tag{5.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_N)$, $\mathbf{K}$ is $N \times N$ Gram matrix, $\mathbf{K} = \boldsymbol{\Phi}^T(\mathbf{X})\boldsymbol{\Phi}(\mathbf{X})$, with elements $\kappa_{mn} = \kappa(\mathbf{x}_m, \mathbf{x}_n)$. Then for each testing sample $\mathbf{x}$, its kernel principal component can be calculated as follows:

$$\begin{aligned}(\mathbf{v}, \varphi(\mathbf{x})) &= \sum_{n=1}^{N} \alpha_n(\varphi(\mathbf{x}_n) \cdot \varphi(\mathbf{x})) \\ &= \sum_{n=1}^{N} \kappa(\mathbf{x}_n, \mathbf{x})\end{aligned} \tag{5.5}$$

It is assumed that the Gram matrix $\mathbf{K}$ is zero-mean, otherwise, it can be centered as [181]:

$$\overline{\mathbf{K}} = \mathbf{K} - \mathbf{I}_N\mathbf{K} - \mathbf{K}\mathbf{I}_N + \mathbf{I}_N\mathbf{K}\mathbf{I}_N \tag{5.6}$$

where $\mathbf{I}_N = \frac{1}{N}\mathbf{I}_{N\times N}$, and $\mathbf{I}_{N\times N}$ is the identity matrix of size $N \times N$.

5.2.2 The proposed FIKPCA

The complexities of the KPCA were pointed out in [174, 176]. In order to reduce the space and time complexities, we propose a fast iterative KPCA (FIKPCA) which is different and much faster than the one in [176]. While iterative KPCA of [176] kernelizes the generalized Hebbian algorithm [177], our method kernelizes the CCIPCA [178] resulting in a much faster convergence. The CCIPCA first centers the original data, $\mathbf{u}(n) = \mathbf{x}_n - m_n$, where m_n is the mean of $\mathbf{x}_n$, and initializes the first k dominant eigenvectors $\mathbf{v}_1(n), \mathbf{v}_2(n), \cdots, \mathbf{v}_k(n)$, directly from the $\mathbf{u}(n), n = 1, 2, \cdots, N$. Then it solves the first k dominant eigenvectors as follows:

For $n = 1$ to N, do the followings steps

1. $\mathbf{u}_1(n) = \mathbf{u}(n)$.
2. For $i = 1$ to min$\{k$,$n\}$ do

   (a) If $i = n$ initialize the ith eigenvector as $\mathbf{v}_i(n) = \mathbf{u}_i(n)$.

   (b) Otherwise

$$\mathbf{v}_i(n) = \frac{n-1-l}{n}\mathbf{v}_i(n-1) + \frac{1+l}{n}\mathbf{u}_i(n)\mathbf{u}_i^T(n)\frac{\mathbf{v}_i(n-1)}{||\mathbf{v}_i(n-1)||} \tag{5.7}$$

$$\mathbf{u}_{i+1}(n) = \mathbf{u}_i(n) - \mathbf{u}_i^T(n)\frac{\mathbf{v}_i(n)}{||\mathbf{v}_i(n)||}\frac{\mathbf{v}_i(n)}{||\mathbf{v}_i(n)||} \tag{5.8}$$

where l is the amnesic parameter [178].

Generally speaking, the mapping function φ is implicit, therefore, the sample in feature space $\varphi(\mathbf{x}_n)$ is not suitable to use in equations (5.7) and (5.8) for iteration. Hence, kernelizing the CCIPCA is not simply using $\varphi(\mathbf{x}_n)$ to replace $\mathbf{x}_n$. we use a Gram-power matrix **G** which has the same eigenvectors as Gram matrix **K** [174], to reformulate the CCIPCA in a kernel version.

$$\begin{aligned}\mathbf{G} &= \mathbf{K} * \mathbf{K}^T \\ &= \begin{bmatrix} \kappa_{11} & \dots & \kappa_{1N} \\ \vdots & \ddots & \vdots \\ \kappa_{N1} & \dots & \kappa_{NN} \end{bmatrix} \begin{bmatrix} \kappa_{11} & \dots & \kappa_{1N} \\ \vdots & \ddots & \vdots \\ \kappa_{N1} & \dots & \kappa_{NN} \end{bmatrix}^T \\ &= \sum_{n=1}^{N} \mathbf{K}(\mathbf{x}_n)\mathbf{K}^T(\mathbf{x}_n) \end{aligned} \tag{5.9}$$

where $\mathbf{K}(\mathbf{x}_n) = (\kappa_{1n}, \kappa_{2n}, \cdots, \kappa_{Nn})^T$.

A row of Gram matrix **K** can be used as a sample for each iteration instead of using the vector $\varphi(\mathbf{x})$. At the end of iteration, the approximate eigenvectors of the Gram-power **G** is obtained, so as the eigenvectors of the Gram matrix **K**. Suppose $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N)$ is the egienvector of the Gram-power matrix **G**, the proposed FIKPCA algorithm is as follows:

First, center the $\mathbf{K}(\mathbf{x}_n)$, $\mathbf{K}(\mathbf{x}_n) = \mathbf{K}(\mathbf{x}_n) - \frac{1}{N}\sum_{j=1}^{N} \kappa(\mathbf{x}_j, \mathbf{x}_n)$, $n = 1, 2, \cdots, N$, compute the first k dominant eigenvectors $\mathbf{w}_1(n), \mathbf{w}_2(n), \cdots, \mathbf{w}_k(n)$ directly from the $\mathbf{K}(\mathbf{x}_n)$, where $\mathbf{w}_i(n)$ is the nth step estimate of eigenvector $\mathbf{w}_i$, $i = 1, \cdots, k$, then do as the Algorithm 1.

Algorithm 1 the FIKPCA algorithm

```
1. For $iteration = 1$ to $Q$ (the number of iteration)
2. For $n = 1$ to $N$ (the number of samples)
3. For each input data $\mathbf{x}_n$, calculate the corresponding
    column $\mathbf{K}(\mathbf{x}_n), \mathbf{K}_1(\mathbf{x}_n) = \mathbf{K}(x_n)$.
4. For $i = 1$ to $k$ (the number of extracted discriminant
    vectors)
5. Using equation 5.10 and equation 5.11 to updata the
    eigen vectors and kernel principal components
```

$$\mathbf{w}_i(n) = \frac{n-1-l}{n}\mathbf{w}_i(n-1) + \frac{1+l}{n}\mathbf{K}_i(\mathbf{x}_n)... \\ ...\mathbf{K}_i^T(\mathbf{x}_n)\frac{\mathbf{w}_i(n-1)}{||\mathbf{w}_i(n-1)||} \quad (5.10)$$

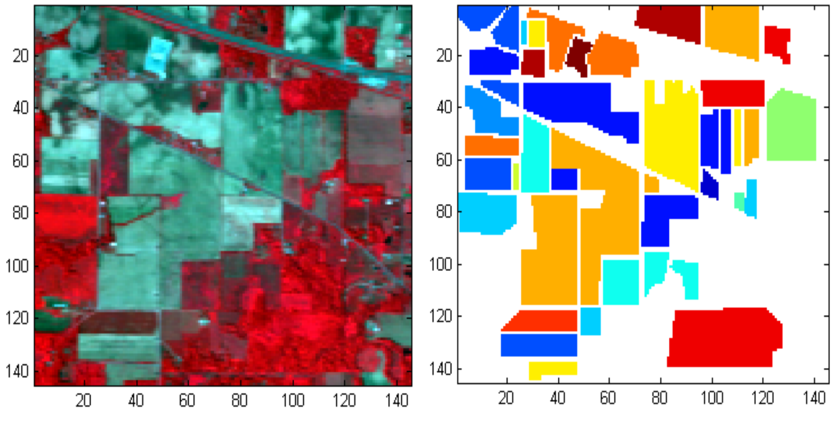
$$\mathbf{K}_{i+1}(\mathbf{x}_n) = \mathbf{K}_i(\mathbf{x}_n) - \mathbf{K}_i^T(\mathbf{x}_n)\frac{\mathbf{w}_i(n)}{||\mathbf{w}_i(n)||}\frac{\mathbf{w}_i(n)}{||\mathbf{w}_i(n)||} \quad (5.11)$$

```
6. Go to step 4
7. Go to step 2
8. Go to step 1
9. Normalize each eigenvector $\mathbf{w}_i = \frac{\mathbf{w}_i}{||\mathbf{w}_i||}$.
```

At each iteration, we need to calculate only a row of Gram matrix, which we need to store the $N \times 1$ vector $\mathbf{K}(\mathbf{x}_n)$. The space complexity is $\mathbf{O}(N)$, which compares favorably with the $\mathbf{O}(N^2)$ complexity of KPCA. In the process of our method, we need time complexity of $\mathbf{O}(QNk)$ to obtain the first k dominant eigenvectors, while the time complexity of KPCA is $\mathbf{O}(N^3)$. This makes the proposed FIKPCA much faster than KPCA, especially for large sample sizes in hyperspectral images. In practice, the proposed method converges after several iterations.

5.2.3 Data sets and experimental setup

We used the AVIRIS Indian Pines image with 220 bands of size 145 lines by 145 samples, originally from Multispec$^{©}$. From this image, 179 bands are selected by removing the noisy channels. The resulting false color composition and the groundtruth are shown in Fig. 5.1. Two subset images were selected to compare the extracted features, the consuming time and the overall classification accuracy (OCA) among PCA, CCIPCA, KPCA and the proposed method. The first subset image consists of pixels [38-87]× [41-90] for a size of 50×50, which contains three labeled classes with number of labeled samples “Corn-notill” (543), “Soybeans-notill” (327), and “Soybeans-min” (1037), the groundtruth is shown in Fig. 5.3. The second subset image consists of pixels [27-86]× [31-90] for a size of 60×60, which contains four labeled classes with number of labeled



(a) (b)

Figure 5.1: (a) False color composition of the AVIRIS Indian Pines Scene; (b) Ground truth containing 16 classes.

samples “Corn-notill” (904), “Grass/Trees” (275), “Soybeans-notill” (425), and “Soybeans-min” (1140), the groundtruth is shown in Fig. 5.5.

In our experiments, the standard KPCA with its matlab codes are available at: http://www.feld.cvut.cz. We don’t compare the iterative KPCA of [176], because of its slow convergence. For the classifier, we used support vector machines (SVMs)optimized by LIBSVM with codes are available in [143], and using 10% of the samples for training SVMs with a linear kernel, the rest of 90% labeled samples are used to test.

5.2.4 Experimental results

Entropy values were used to measure the information content contained in each individual extracted feature band [48]. Extracted feature bands with higher entropy values are selected. Image entropy is defined as:

$$E(A) = -\sum_{A\in\mathbf{G}} P(A) log_2 P(A) \tag{5.12}$$

where **G** is the set which contains the number of all gray levels in image A (all images are linearly stretched between 0 and 255), and $P(A)$ is the probability distribution.

However, extracted feature bands with higher entropy values may be not consistent with the requirement of image classification [183]. It can be easily seen from the Eq. 5.12 that $E(A)$ is calculated only with respect to the single extracted feature band. Therefore, the amount of information content measured by the entropy lacks a point of groundtruth, which cannot guarantee that the extracted feature bands with higher entropy values are useful for classification objective. The Mutual information (MI) is used to measure the similar information that the two images share, and intuitively measures the dependency between the two images, higher MI indicates more dependency between them. The mutual information is defined as:

$$MI(A,B) = -\sum_{A\in\mathbf{G},B\in\mathbf{G}} P(A,B)log_2\frac{P(A,B)}{P(A)\bullet P(B)} \tag{5.13}$$

Or equivalent to:

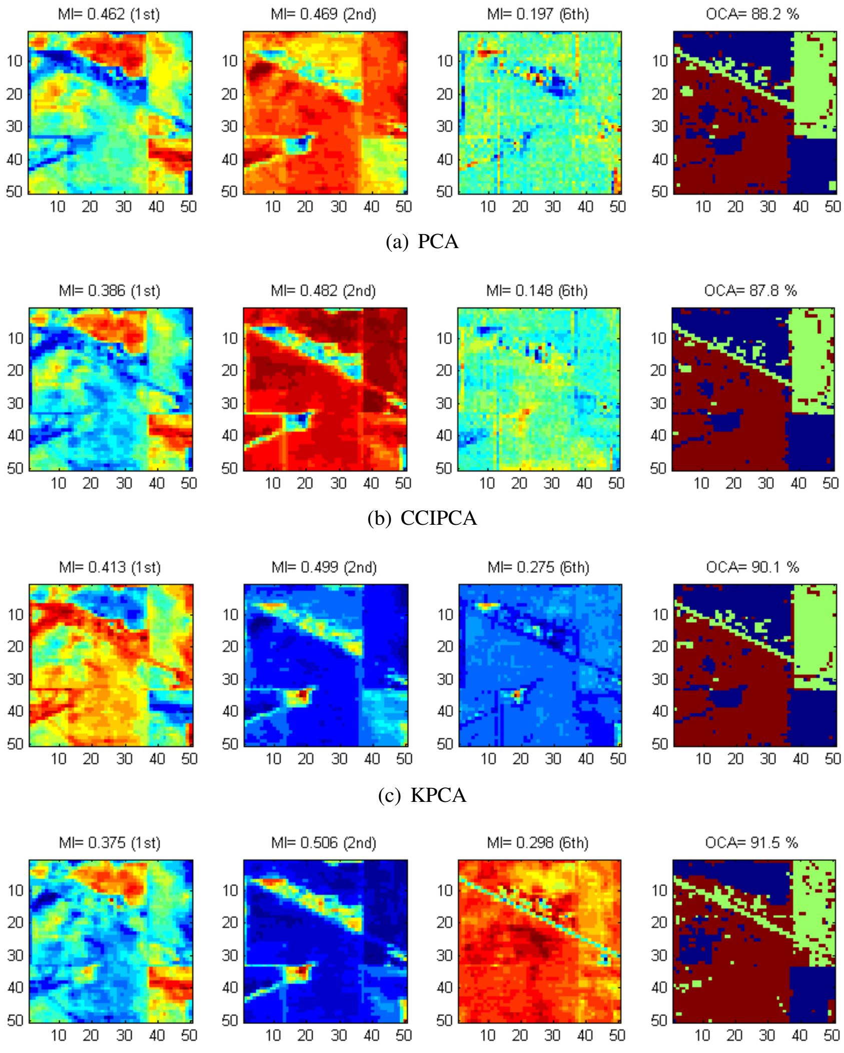
$$MI(A,B) = E(A) + E(B) - E(A,B) \tag{5.14}$$

where $P(A,B)$ is the joint probability distribution of the image A and B, and $E(A,B)$ is their joint entropy.

As the groundtruth implicitly defines the required classification result, we can take the extracted features and the corresponding groundtruth as random images. The MI can be used to estimate the dependency between them, and measures the relative utility of each extracted feature to the classification objective.

From the in Fig. 5.2 and in Fig. 5.4, we can see that nonlinear methods perform better than linear methods in terms of MI as the number of the extracted feature bands increases. This means nonlinear methods can extract more image content than linear methods. For linear methods as PCA and CCIPCA, the sixth extracted feature bands contain much noisy in in Fig. 5.2, while the tenth extracted feature bands mainly contain noisy in in Fig. 5.4. For nonlinear methods as KPCA and FIKPCA, the tenth extracted feature bands still contain some information which is important for classification. The first feature band extracted by PCA method is better than the rest of the methods in terms of MI, this is because the first feature bands extracted by PCA contains most energy of the original image. As the first 6 extracted feature bands in the first subset image and the first 10 extracted feature bands in the second subset image are used as input to do classification, we can see that nonlinear methods perform better than linear methods in term of OCA, and our approach gets the best results.

The results in Fig. 5.3 and in Fig. 5.5 show that the classification accuracies of the kernel methods are much better than those of the linear methods when more than 4 extracted features are used as input to classify. The linear methods are much faster than the kernel ones. Among the kernel methods, the proposed FIKPCA is much faster than the standard KPCA, for 50×50 subset image, KPCA method needs more than 90 seconds to process, while the proposed FIKPCA just uses less than 6 seconds to extract 6 feature bands. For 60×60 subset image, it requires more than 250 seconds for KPCA method, but just 14 seconds for FIKPCA method. As



(a) PCA

(b) CCIPCA

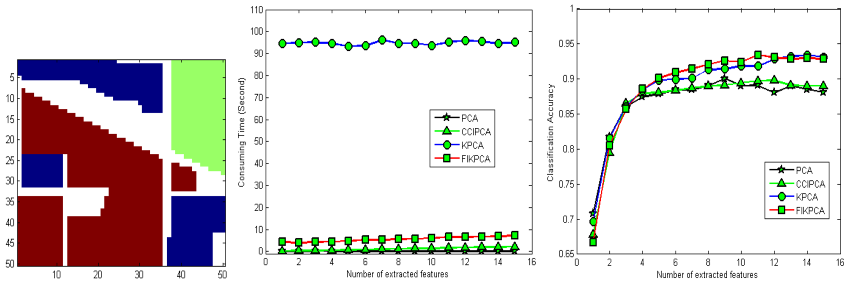
(c) KPCA

(d) Proposed FIKPCA

Figure 5.2: The extracted feature bands and classification maps produced by each method for the 50×50 *subset image, of which the groundtruth is shown in Fig. 5.3. The first, second and sixth extracted feature bands are used to compare. The SVM classifiers were trained with 10% of labeled samples per class randomly selected from the groundtruth, the trained classifier is then applied to the remaining 90% of the known ground pixels in the scene.*

the size of the images is larger than 70×70, KPCA method will be out of memory, while the proposed method is still efficient, see Fig. 5.6.

When the first extracted feature band is used for classification, PCA gives the best results. This is because the first feature band extracted by PCA contains most



(a) (b) (c)

Figure 5.3: (a) Groundtruth of the 50×50 *subset image; (b) comparison of the consuming time and; (c) overall classification accuracy with different number of extracted features bands. Each experiment was repeated 5 times, the average was calculated.*

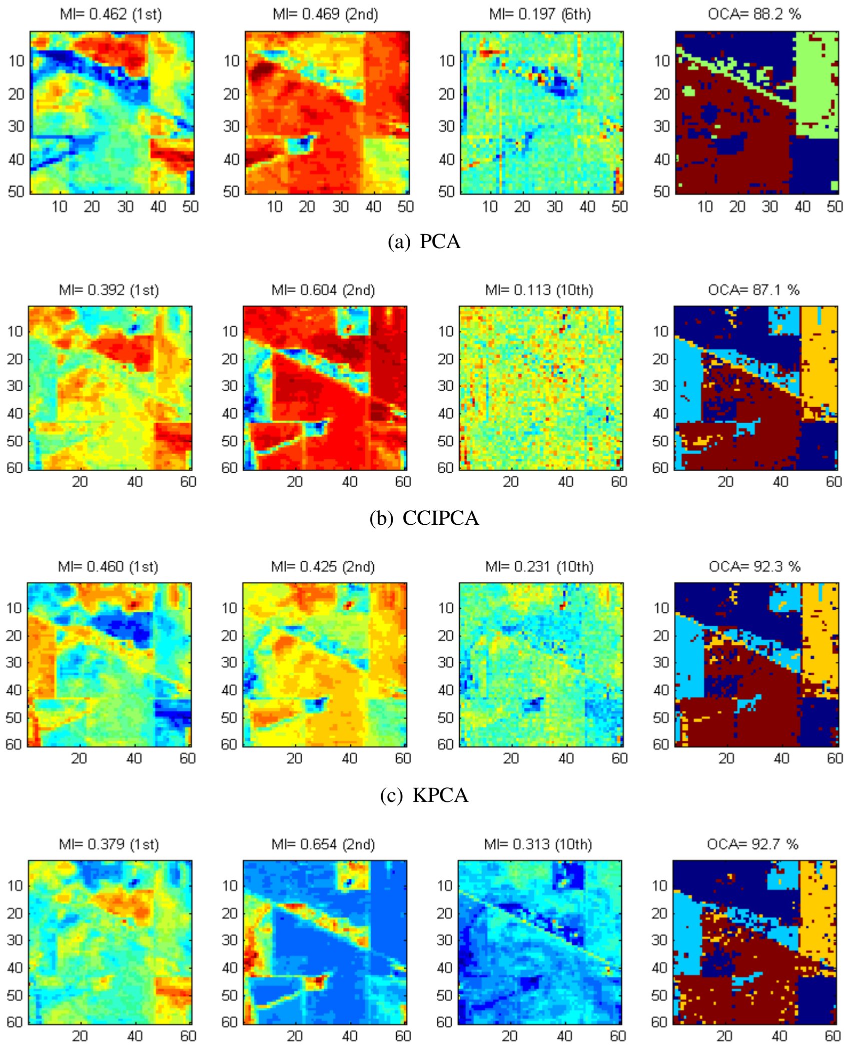
information content of the image, the first 5 extracted feature bands contain more than 99% information content of the image, so adding more extracted features does not change the result much. While using more than 10 features extracted by PCA, the classification rate decreases slightly, as more noisy is added. The classification results of KPCA will be a little better when more extracted feature bands are used. However, in many cases, as more extracted feature bands are used in the following steps, for example, in [28] when performing morphological profiles on extracted feature bands, it will increase the dimensionality for the classification. Moreover, the proposed method is much faster in getting its best result than KPCA, as it is shown in Fig. 5.3 and in Fig. 5.5.

In fact, during the iterations, calculating the corresponding vector $\mathbf{K}(\mathbf{x}_i)$ of the Gram matrix as the input consumes most of the processing time, especially for a large number of samples, see Fig. 5.6. We use a normalized linear kernel function of which the computational cost of kernel is nearly the same as that of a linear kernel and much lower than that of radial basis function and polynomial kernels [184].

As for the whole original hyperspectral image, the KPCA is incapable to handle hyperspectral image with sizes 145145, the proposed FIKPCA is still efficient, and its classification accuracy is a little better than PCA and CCIPCA, as shown in Fig. 5.7.

5.3 Extended morphological profiles generated on KPCs with partial reconstruction

When applying morphological features for the classification of high resolution hyperspectral images from urban areas, one should consider another important issue



(a) PCA

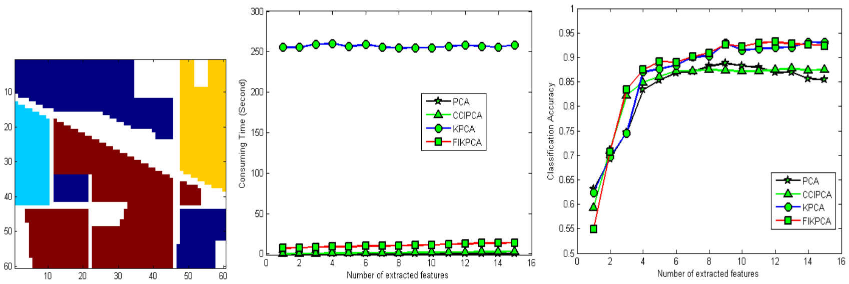
(b) CCIPCA

(c) KPCA

(d) Proposed FIKPCA

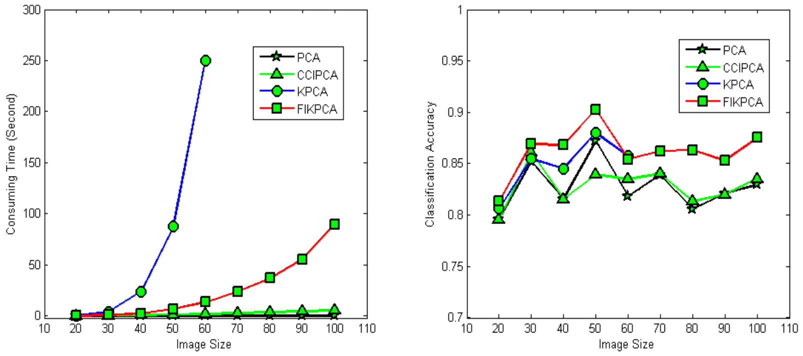
Figure 5.4: The extracted feature bands and classification maps produced by each method for the 60×60 *subset image, of which the groundtruth is shown in Fig. 5.3. The first, second and sixth extracted feature bands are used to compare. The SVM classifiers were trained with 10% of labeled samples per class randomly selected from the groundtruth, the trained classifier is then applied to the remaining 90% of the known ground pixels in the scene.*

except the two we considered in Chapter 4. The high dimensionality of these hyperspectral data as well as the redundancy within the bands, make the generation of an MP based on each spectral band seem not feasible. To overcome this problem, feature extraction is firstly used to reduce the dimensionality of these hyperspec-



(a) (b) (c)

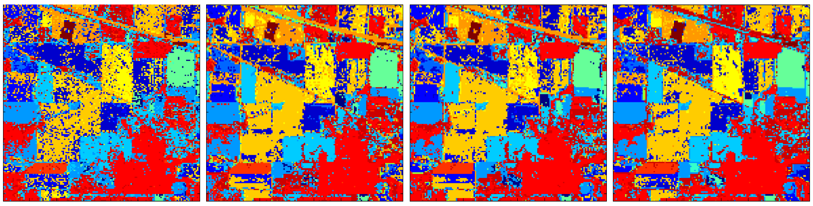
Figure 5.5: (a) Groundtruth of the 60×60 *subset image; (b) comparison of the consuming time and; (c) overall classification accuracy with different number of extracted features bands. Each experiment was repeated 5 times, the average was calculated.*



(a) (b)

Figure 5.6: (a)Consuming time and; (b)overall classification accuracy by each method used to obtain the first 10 dominant eigenvectors, with image sizes increasing. Horizontal axis 20 means square image with size 20×20*. For the classification, the first 10 extracted feature bands were used as input, and the SVM classifiers were trained with 10% of labeled samples per class randomly selected from the groundtruth, the trained classifier is then applied to the remaining 90% of the known ground pixels in the scene.*

tral data, and then morphological processing is applied on each extracted feature band independently. Principal component analysis (PCA) [37] is the most popular methods used to extract features for building MPs [17, 18, 39]. [17] extended the method in [78] for hyperspectral data with high spatial resolution by using PCA to reduce the high dimensionality of the data. The resulting method built the MPs on the first principal components (PCs) extracted from a hyperspectral image, leading to the definition of extended MPs (EMP). In [74], the morphological attribute profiles [75] were applied to the first PCs extracted from a hyperspectral



(a) No FE (72.1%) (b) PCA (74.0%) (c) CCIPCA (73.3%) (d) FIKPCA (77.4%)

Figure 5.7: Classification accuracy of the whole image using 6 features extracted by PCA, CCIPCA and the proposed method. 5% of the samples of its groundtruth are used to train SVMs with linear kernel, the rest are used to test.

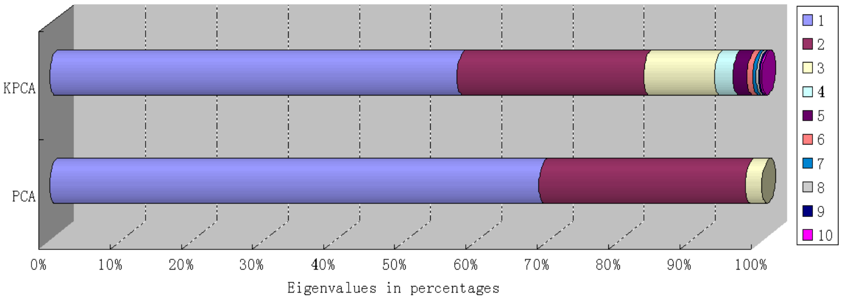


Figure 5.8: Eigenvalues and cumulative variance in percentages for PCA and KPCA.

image, generating an extended morphological attribute profiles (EAP). However, it was found that too much spectral information were lost during the linear principal component analysis (PCA) transformation [28, 87], as PCA relies on second-order statistics only. By taking high order statistics into account, independent component analysis (ICA) [85] has been studied to reduce the dimensionality of hyperspectral data [40, 180]. [87] built MPs on the first features extracted from original hyperspectral data by ICA, with an improvement in the classification results. [76] improved the classification results by building the EAPs on the first independent components (ICs) comparing to the results of those built on PCs [74]. Kernel PCA [181], which is more suitable to describe higher order complex and nonlinear distributions, has been recently investigated in reducing the dimensionality of hyperspectral remote sensing [19, 182]. In [28], kernel principal components are used to construct the EMP, with significantly improvement in terms of classification accuracies compared with the conventional EMP built on PCs.

In this section, we apply morphological profiles with partial reconstruction of [73] to the classification of high resolution hyperspectral images from urban areas. We first extract features from the original hyperspectral data sets by PCA, ICA and KPCA, then build extended morphological profiles on the extracted features with morphological openings and closings by partial reconstruction. Finally, we

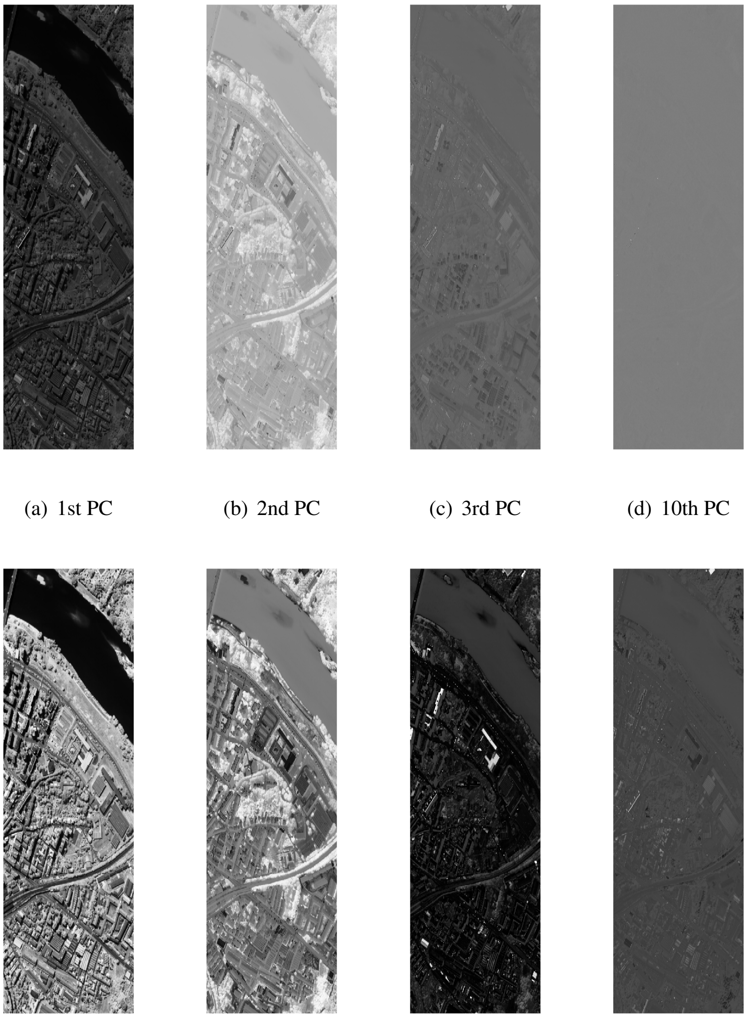


Figure 5.9: Principal components and kernel principal components for Pavia Center data set.

use the extended morphological profiles as the inputs of SVM classifiers to do classification.

5.3.1 Data sets and experimental setup

Hyperspectral Image Data Sets: Experiments were run on two data sets, namely the '*Pavia Center*' and '*University Area*' from urban areas in the city of Pavia, Italy. The data were collected by the ROSIS (Reflective Optics System Imaging Spectrometer) sensor, with 115 spectral bands in the wavelength range from 0.43 to 0.86μm and very fine spatial resolution of 1.3 meters by pixel. Chapter 4 shows the training sets and test sets used in our experiments, which are selected from the data by an expert, corresponding to a predefined species/classes. Note that the color in the cell denotes different classes in the classification maps (Fig. 4.6-Fig. 4.7).

Experimental setup: To apply the morphological profiles with partial reconstruction of [73] from panchromatic imagery to hyperspectral imaging, feature extraction was first applied to to reduce the dimensionality of the original hyperspectral data. For PCA and ICA, the first 3 principal components (PCs) were selected (representing almost 99% of the cumulative variance) to construct the MPs for both data sets. For KPCA, the number of extracted kernel principal components which represent 99% of the cumulative variance depends on the the number of total training samples and the parameters in the selected kernel function, as was also discussed in [19, 28]. In our experiments, 5000 samples were randomly selected to train and construct the training kernel matrix, Gaussian kernel function with $\delta = \frac{4}{n \times n}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{ij}^{2}}$, where n is the total number of the training samples, $\kappa_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ is the element of the kernel matrix. To achieve more than 99% of the cumulative variance, 10 KPCs and 12 KPCs are needed in *University Area* and *Pavia Center* data sets, respectively, see Fig. 5.8. Morphological profiles with 4 openings and closings (ranging from 2 to 8 with step size increment of 2) were then computed for each extracted features. We represent extended morphological profile with no reconstruction, with reconstruction and with partial reconstruction as EMPN, EMPR and EMPP, respectively. EMPP built on features extracted by PCA, ICA and KPCA are denoted as $EMPP_{PCA}$, $EMPP_{ICA}$ and $EMPP_{KPCA}$, respectively.

We used one of the most popular classifiers: support vector machines (SVM) [181], as it performs well even with a limited number of training samples, which can overcome the Huges phenomenon. The SVM classifier with radial basis function (RBF) kernels and linear kernels in Matlab SVM Toolbox, LIBSVM [143], is applied in our experiments. SVM with RBF kernels has two parameters: the penalty factor C and the RBF kernel widths γ. While SVM with linear kernels has only one parameter (the penalty factor C). We apply a grid-search on C and γ using 5-fold cross-validation to find the best C within the given set $\{10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and the best γ within the given set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. The training data sets were randomly subsampled to create samples

FE	Methods	Classifier	Training Set Size				
			10	50	100	150	All
Spectral only	Raw (103)	Linear	66.4	74.1	76.8	77.2	78.2
		RBF	63.5	75.5	78.8	79.9	80.2
	PCA (3)	Linear	58.6	69.3	73.2	71.6	73.6
		RBF	61.3	64.1	66	66.4	66.5
	ICA (3)	Linear	61.8	69.3	73	71.5	73.6
		RBF	63.5	66	66.1	65.7	66.9
	KPCA (12)	Linear	62.7	75.1	78.8	79.4	81.3
		RBF	64.3	73.2	77.7	77.9	80.3
EMPN	PCA (27)	Linear	82.6	86	85.4	85	86.3
		RBF	82.1	85.2	84.5	85	85.8
	ICA (27)	Linear	80.1	85.6	84.7	84.3	84.9
		RBF	80.3	86.3	84.7	85.5	85.2
	KPCA (108)	Linear	84.4	91	91.4	91.5	91.3
		RBF	84.2	91	91.5	91.8	91.5
EMPR	PCA (27)	Linear	76.9	81.1	82.2	80.8	82.6
		RBF	74.1	80.4	81.6	80.4	80.3
	ICA (27)	Linear	75.4	84.3	87.5	86.5	84
		RBF	73.7	82	84	82.8	83.1
	KPCA (108)	Linear	71.2	88.9	92.7	93.2	**95.7**
		RBF	71.7	89	93	92.6	94.8
EMPP	PCA (27)	Linear	81.9	88	87.8	87.4	87.8
		RBF	82.5	89.1	89.2	89	88.7
	ICA (27)	Linear	81.4	87.9	87.8	87.5	88.9
		RBF	80.5	89.3	89.3	89.7	89.2
	KPCA (108)	Linear	86	**93.9**	**94.5**	**94.7**	94.4
		RBF	**86.2**	93	94	94.2	94.3

Table 5.1: University Area. Overall accuracy (%) in a classification with spectral features compared to classifications with EMPN, EMPR and EMPP built on different features (# bands)

FE	Methods	Classifier	Training Set Size				
			10	50	100	150	All
Spectral Only	Raw (102)	Linear	94	95.1	95.5	95.8	96
		RBF	94	94.6	95.6	95.6	96.5
	PCA (3)	Linear	94.1	94.8	94.6	94.7	95.2
		RBF	94.2	94.5	94.7	94.7	95.1
	ICA (3)	Linear	94.2	94.7	94.6	94.8	95.2
		RBF	94.1	94.7	94.8	94.9	95.3
	KPCA (10)	Linear	95.5	95.8	96.3	96.3	96.7
		RBF	95.5	96.2	96.3	96.7	97
EMPN	PCA (27)	Linear	95.7	96.4	96.5	96.5	96.7
		RBF	95.8	95.6	95.1	95.9	96.4
	ICA (27)	Linear	95.8	96.7	96.4	96.5	96.7
		RBF	96	96.4	96	96.4	96.5
	KPCA (90)	Linear	97.7	98.5	98.4	98.5	98.6
		RBF	97.4	98.5	98.2	98.3	98.6
EMPR	PCA (27)	Linear	95.7	97.2	97.8	98	98.3
		RBF	95.7	97.3	97.5	98.3	98
	ICA (27)	Linear	96.1	97.4	97.4	97.7	97.8
		RBF	96.4	97	96.8	97.9	98.2
	KPCA (90)	Linear	96.8	97.7	98.3	98.6	98.8
		RBF	96.8	97.7	98.3	98.6	98.8
EMPP	PCA (27)	Linear	96	97.1	97.3	97.2	97.4
		RBF	96.1	97.3	97.4	97.4	97.6
	ICA (27)	Linear	96.1	97	97.3	97.3	97.2
		RBF	96.1	97.1	96.9	97.4	97.5
	KPCA (90)	Linear	**97.8**	**98.9**	**99**	**99.1**	**99.1**
		RBF	97.5	98.5	**99**	**99.1**	99

Table 5.2: Pavia Center. Overall accuracy (%) in a classification with spectral features compared to classifications with EMPNs, EMPRs and EMPPs built on different features (# bands)

whose sizes corresponded to five distinct cases: 10, 50, 100, 150 samples per class, respectively. The classifiers were evaluated against the testing sets, the results were averaged over five runs.

5.3.2 Experimental results

Table 5.1 and Table 5.2 display the classification accuracies of testing data in different sample size. The best accuracy in each sample size (in column) is highlighted in bold font. From these tables, we have the following findings:

(1) Morphological features can improve the classification results. The results using morphological features are much better than those using the original hyperspectral data and the spectral features only. By building the extended morphological profiles on the first few extracted features, the results can be improved a lot. Compared to the situation with the original hyperspectral data and the spectral features only in each training sample size, the OA of *University Area* with morphological profiles built on features extracted by PCA, ICA and KPCA have 0.5%-20%, 2.9%-20% and 2.9%-20% improvements, respectively. For *Pavia Center*, these improvements are 0%-3.6%, 0%-3.1% and 1.3%-4.3%, respectively.

(2) The classification results with the features (representing almost 99% of the cumulative variance) extracted by KPCA are better than those with features extracted by PCA and ICA. For *University Area* dataset, the OA with KPCs has 1.4%-14.4% and 0%-14.7% improvements, compared to the results with features extracted by PCA and ICA in each training sample size. For *Pavia Center*, the improvements are 1%-2% and 1.1%-1.9%, respectively.

(3) It is better not to use MPs with reconstruction in some cases. This is in particular for the *University Area* data set, where the MPs with reconstruction perform even worse than MPs with no construction. By using EMPP built on PCA and ICA, the results can be improved a lot. Compared to the results using EMPR, the OA of EMPP built on features extracted by PCA and ICA have 5%-8.7%, 0.3%-7.7% improvements, respectively. For $EMPP_{KPCA}$, these improvements are obvious in small sample size, with 14.3% improvements when using 10 training samples per class.

(4) EMP built on nonlinear features (KPCs) perform better than those built on linear features (PCs and ICs), the improvements are 2%-5% and 3%-7% for the data sets of *University Area* and *Pavia Center*. The performances of EMP built on PCs are similar to those built on ICs for *Pavia Center* data set. While for the *University Area* data set, the performance of EMPN and EMPP are similar for PCA and ICA. In the case of using EMPR for the classification of the *University Area* data, ICA performed slightly better than PCA, the enhancement is between 1.4 and 6.5 points when the training sample sizes are more than 50.

(5) As the number of training samples increases, the OA will increase, this is obvious when the training samples size increases from 10 to 50. When the training samples size is larger than 50, the performances using EMPN and EMPP keep stable for both of two hyperspectral data sets. For example in *University Area* data set with 50 training samples per class, we get 93.9%

OA by using $EMPP_{KPCA}$. To achieve similar OA, the $EMPR_{KPCA}$ in this case requires more than three times training samples.

(6) When using $EMPP_{KPCA}$, we get almost the highest OA for both data sets in different training sample size. For *University Area* data set, the highest OA with the training samples size of 10, 50, 100, 150 and all are 86.2% ($EMPP_{KPCA}$ and SVM classifier with RBF kernel), 93.9% ($EMPP_{KPCA}$ and SVM classifier with linear kernel), 94.5% ($EMPP_{KPCA}$ and SVM classifier with linear kernel), 94.7% ($EMPP_{KPCA}$ and SVM classifier with linear kernel) and 95.7% ($EMPR_{KPCA}$ and SVM classifier with linear kernel), respectively. For *Pavia Center* data set, the highest OA in different training samples size are 97.8% ($EMPP_{KPCA}$ and SVM classifier with linear kernel), 98.9% ($EMPP_{KPCA}$ and SVM classifier with linear kernel), 99% ($EMPP_{KPCA}$ and SVM classifier with both RBF and linear kernels), 99.1% ($EMPP_{KPCA}$ and SVM classifier with both RBF and linear kernels) and 99.1% ($EMPP_{KPCA}$ and SVM classifier with linear kernels), respectively.

For the classification purpose, we cannot always get the best classification results with the number of extracted features which represent almost 99% of the cumulative variance, as were also suggested in [18, 19]. Cross-validation may be a good solution to determine the optimum number of features extracted by PCA, ICA and KPCA. Compared to the linear feature extraction such as PCA and ICA, KPCA [181] which takes higher order statistics and nonlinear distributions of the data into account, is more suitable to model and extract features from the original hyperspectral data sets. As KPCA maps the input data into a high-dimensional feature space and performs eigen decomposition on Gram matrix, the number of its extracted features depends on the number of the total training samples. When the number of training samples is larger than the dimensionality of the original hyperspectral data sets, we can even extract more features than the total bands of original data. However, there are some important issues to be considered when using KPCA, such as the choice of kernel function, the optimization of parameters in kernel functions and the computational load both in terms of CPU and memory.

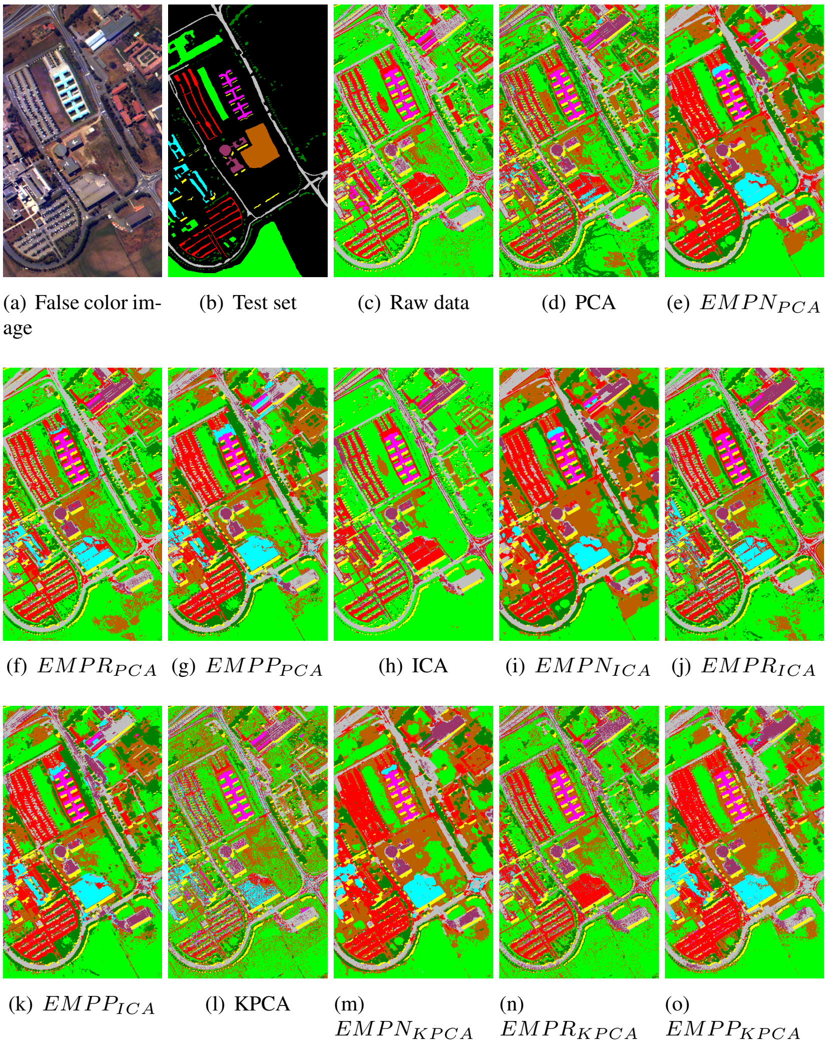
In high-resolution hyperspectral remote sensing imagery from urban areas, spectral characteristics of some surface materials are so similar that they cannot be separated using only spectral information. Morphological profiles which carry information about the size and the shape of objects in the images can explore the spatial information and improve the classification accuracies. In order to compare the classified maps visually, we generate classification maps using the best combinations of SVM classifiers when 10 training samples per class are used, displayed in Fig. 5.10-Fig. 5.11. Their corresponding classification accuracies are shown in Table 5.3 and Table 5.4, averaged over five runs.

Table 5.3: Classification Accuracy (%) for University Area using the best combinations of SVM classifiers when 10 training samples per class are used.

	Spectral Only				EMPN			EMPR			EMPP		
	Raw	PCA	ICA	KPCA	PCA	ICA	KPCA	PCA	ICA	KPCA	PCA	ICA	KPCA
OA	66.4	61.3	63.5	64.3	82.6	80.3	84.4	76.9	75.4	71.7	82.5	81.4	**86.2**
AA	72.7	69.1	68.4	71.4	81.1	79.9	84	83.3	80.4	76.3	82.6	81.5	**84.8**
1	64.4	48.9	50.7	53.1	67.5	63.1	77.5	**82**	65.8	62.2	68.4	63.9	77.8
2	64.2	62.3	70.6	64.1	92.5	88.4	91.9	73.4	79.4	73.6	91.7	88	**94.1**
3	49	46.2	47.1	59.4	45.6	42.4	57.9	66.9	**72.9**	54.9	62.1	59.5	67.3
4	68.3	69.1	59.8	55	84.7	84.7	75.6	85.2	**87.2**	60.3	78.6	79.9	56.7
5	98.5	99.3	99.4	99.6	99.7	99.7	**99.9**	99.4	99.4	99.7	99.3	99.4	99.7
6	62.2	45.6	38.6	63.3	61	64.6	65.9	55	46.9	66.9	57.6	61.1	**77.3**
7	77.2	79	82.1	78	81.3	78.8	94.2	**94.3**	90.2	88.7	88.6	87.2	94.1
8	70.5	71.8	67.6	71.1	**98.1**	97.7	93.5	93.7	83.3	80.9	97.5	97.5	97
9	99.7	99.5	99.5	99.4	**99.8**	99.3	99.7	**99.8**	98.9	99.5	**99.8**	96.7	99.4

Table 5.4: Classification Accuracy (%) for Pavia Center using the best combinations of SVM classifiers when 10 training samples per class are used.

	Spectral Only				EMPN			EMPR			EMPP		
	Raw	PCA	ICA	KPCA	PCA	ICA	KPCA	PCA	ICA	KPCA	PCA	ICA	KPCA
OA	94	94.2	94.2	95.5	95.8	96	97.7	95.7	96.4	96.8	96.1	96.1	**97.8**
AA	89.7	89.2	88.8	90.7	91.6	91.4	95.4	92.6	93.9	93.7	92.3	92.8	**95.7**
1	98.1	98.4	98.4	99.2	99.3	99.3	**99.8**	98.9	98.4	99.7	99.3	99	**99.8**
2	83.1	84.1	84.1	86.6	88.1	**96.1**	93.4	84.2	88.4	91.2	86.6	89.2	93.1
3	**93.9**	90.3	89.6	91.3	87.3	76	93.6	90.9	88.2	91.7	89.3	88.6	91.1
4	76.1	75.1	68.2	76.3	84	86.7	94.3	91.2	96.8	89.9	85.4	89.2	**97.3**
5	86.1	84.5	84.9	89.3	94	92.7	**94.4**	89	89.5	91	93.9	89.6	93.7
6	97.4	95.6	95	96.7	92.6	97.1	94.7	97	96.8	94.4	**98**	97.7	96.8
7	74.8	75.9	79.6	79.6	79.8	76	89.4	82.2	87.2	85.4	78.8	83	**90.7**
8	99.2	99.4	99.5	97	99.9	99.2	**100**	99.9	99.9	99.8	99.9	99.9	**100**
9	99.9	99.8	99.8	99.9	99.1	99.3	99	99.8	99.9	**100**	99.6	99	99.1



(a) False color image (b) Test set (c) Raw data (d) PCA (e) $EMPN_{PCA}$

(f) $EMPR_{PCA}$ (g) $EMPP_{PCA}$ (h) ICA (i) $EMPN_{ICA}$ (j) $EMPR_{ICA}$

(k) $EMPP_{ICA}$ (l) KPCA (m) $EMPN_{KPCA}$ (n) $EMPR_{KPCA}$ (o) $EMPP_{KPCA}$

Figure 5.10: Classification maps for University Area data set using the best combinations of SVM classifiers when 10 training samples per class are used.

From these figures and tables, we can find that the EMP (without reconstruction, with reconstruction, and with partial reconstruction) can preserve well spatial information on hyperspectral images. The classification maps with EMP produce much smoother homogeneous regions than those with the raw data and only spectral features, which is particularly significant when using EMPN and EMPP. The classification maps using the EMPR looks much noisy because of the over reconstruction problems, for *University Area* data set, see Fig. 5.10(g), Fig. 5.10(k) and Fig. 5.10(o); for *Pavia Center* data set, see Fig. 5.11(g), Fig. 5.11(k) and Fig. 5.11(o). The EMPN deform the objects, see Fig. 5.10(f), Fig. 5.10(j) and Fig. 5.10(n), the borders of some objects are deformed. The shapes of objects are better preserved with EMPP. Simultaneously, the spatial information are better preserved with EMPP, which can be seen in Fig. 5.10(h), Fig. 5.10(l) and Fig. 5.10(o).

For some objects with large homogeneous regions, EMP (including EMPN, EMPR and EMPP) perform better than only spectral features, while EMPN and EMPP are more efficient than EMPR, this is obvious in case of *University Area* data set. The class 2 "*Meadows*" in *University Area* data set are better preserved with EMP. There are 2.8%-31.8% improvements compared with only spectral features, see Table 5.3. While EMPN and EMPP have 8.6%-20.7% improvements than EMPR. For class 1 "*water*" and class 5 "*Bitumen*" in *Pavia Center* data set, the results with EMP built on different features are better than those with their corresponding spectral features, see Table 5.4. The EMPN and EMPP get the best result for both of these classes.

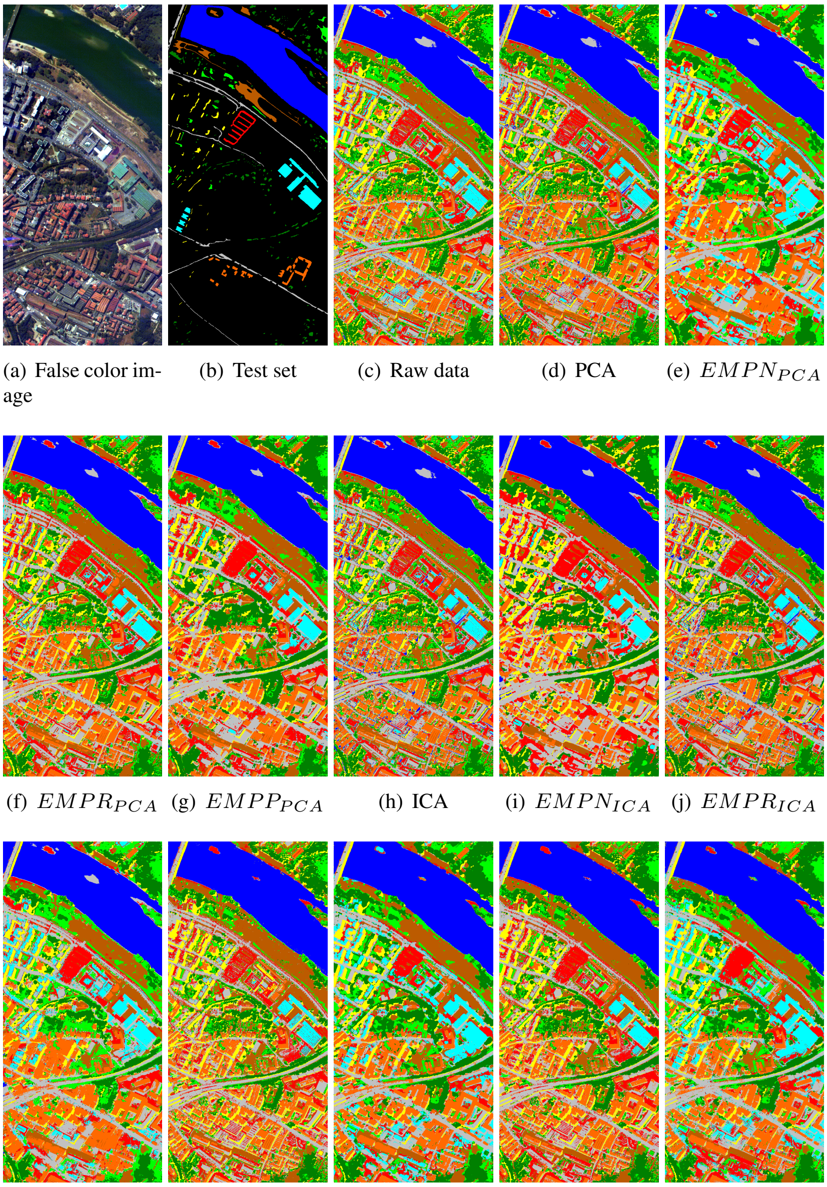
For some objects with rectangular shape, EMPN produces worse results than EMPR and EMPP, see class 3 "*Gravel*" in *University Area* data set in Table 5.3. Because morphological openings and closings degrade the object boundaries and deform the object shapes, which also can be seen in Fig. 5.10(f), Fig. 5.10(j) and Fig. 5.10(n).

$EMPP_{KPCA}$ get the best results, with 86.2% OA and 84.8% AA for *University Area* data set, and with 97.8% OA and 95.7% AA for *Pavia Center* data set.

5.4 Conclusion

Instead of solving eignvectors by eigen decomposition on Gram matrix, the proposed FIKPCA obtains the eigenvectors through iteration, which reduces the space and time complexity greatly. The experimental results show that the proposed FIKPCA is much faster than KPCA, and it can handle hyperspectral images larger than 70 × 70 efficiently, which is of particular interest in remote sensing. The classification results using the features extracted by the proposed FIKPCA are better than PCA and CCIPCA.

We investigated the influence of morphological features with different reconstruction (including with no reconstruction, with reconstruction and with partial reconstruction) for the classification of high resolution hyperspectral images from urban areas. To apply morphological profiles on hyperspectral data, we first reduced the dimensionality of hyperspectral data by feature extraction, then built the extended morphological profiles on the extracted features. We showed on two real hyperspectral data sets that KPCA is more efficient to extract features for constructing EMP. In many cases, the most widely used EMP with reconstruction can



(a) False color image (b) Test set (c) Raw data (d) PCA (e) $EMPN_{PCA}$

(f) $EMPR_{PCA}$ (g) $EMPP_{PCA}$ (h) ICA (i) $EMPN_{ICA}$ (j) $EMPR_{ICA}$

(k) $EMPP_{ICA}$ (l) KPCA (m) $EMPN_{KPCA}$ (n) $EMPR_{KPCA}$ (o) $EMPP_{KPCA}$

Figure 5.11: Classification maps for Pavia Center data set using the best combinations of SVM classifiers when 10 training samples per class are used.

not get a satisfied result, because of over-reconstruction problems. EMP with partial reconstruction built on KPCs is more competitive than those of EMP with no reconstruction and with reconstruction built on other different features.

6 Conclusions and future work

Recent advances in sensors technology have led to an increased availability of hyperspectral remote sensing data at very high both spectral and spatial resolutions. Many techniques are developed to explore the spectral information and the spatial information of these data. In particular, feature extraction is one of methods to preserve the spectral information, while morphological profile is the most popular methods used to to explore the spatial information. In this dissertation, two semi-supervised feature extraction methods and a kernel method were proposed, morphological profiles with partial reconstruction were applied to the hyperspectral data.

1. The proposed semi-supervised local discriminant analysis (SELD) is effective and powerful to deal with both the cases of ill-posed and poorly posed in statistics. SELD provides a new framework to combine supervised and unsupervised methods for semi-supervised feature extraction. The advantages of the proposed SELD are as follows:

   - No tradeoff parameters optimization. The main idea of SELD is to divide first the samples into the labeled and the unlabeled sets. The labeled samples are employed through the supervised LDA only and the unlabeled ones through the unsupervised method only. We combine the two in a non-linear way without any tuning parameters;
   - Discrimination maximized and local neighborhood information well preserved. Experimental results on the synthetic data demonstrate the results of SELD;

- Statistically significance in overall classification accuracy. The McNemar's tests based upon the standardized normal test statistic [103] were computed, and the experimental results on real hyperspectral data show the statistical significance of the proposed SELD, compared to some related methods;
- Less computational cost. Experimental results on real hyperspectral data show the efficiency of the proposed SELD compared with NWFE, SDA and SELF as the number of training samples increases.

2. Morphological profiles with partial reconstruction and directional morphological profiles are applied to explore the spatial information of very high resolution hyperspectral data from the urban area. We first use PCA to reduce the dimensionality of the hyperspectral data. Then, we construct MPs with partial reconstruction on the first few PCs. With partial reconstruction, the MPs can preserve size and shape information better. On the other hand, a lot of small objects which remain present in that with reconstruction, now disappear in the case with partial reconstruction. Compared to MPs by reconstruction, experimental results on real hyperspectral data demonstrate that the overall classification accuracy and *kappa* with partial reconstruction have more than 10% improvements in some cases. Simultaneously, The McNemar's tests show the statistical significance of the MPs with partial reconstruction.

3. The generated morphological profiles (MPs) with different structuring elements and a range of increasing sizes of morphological operators produce high-dimensional data. The proposed generalized semi-supervised local discriminant analysis (GSELD) compares favorably with conventional feature extraction methods as preprocessing approaches for the morphological profiles generated on high resolution hyperspectral data from the urban area. Experimental results on real hyperspectral data with three different classifiers (LDC, kNN and SVM) show the proposed GSELD outperforms the other feature extraction methods, with McNemar's tests larger than zeros. For high resolution urban data set (University Area), when using both the disk-based and linear-based morphological features, the proposed GSELD gets the highest OA in different samples size. The highest OA for training samples size with 10, 20, 40, 80 and 160 are 92.5% (GSELD with SVM classifier), 93.4% (GSELD with LDC classifier), 95.1% (GSELD with LDC classifier), 96% (GSELD with SVM classifier) and 96.2% (GSELD with SVM classifier), respectively.

4. A fast iterative kernel principal component analysis (FIKPCA) was proposed to extract features from hyperspectral images. Instead of solving

eignvectors by eigen decomposition on Gram matrix, the proposed FIKPCA obtains the eigenvectors through iteration, which reduces the space and time complexity greatly. The experimental results show that the proposed FIKPCA is much faster than KPCA, and it is more effective than linear methods. Moreover, it can handle hyperspectral images larger than 70×70 efficiently, which is of particular interest in remote sensing.

5. Extended MPs with partial reconstruction built on kernel principal components is investigated. Traditional linear features, on which the morphological profiles usually are built, lose too much spectral information. Nonlinear features are more suitable to describe higher order complex and nonlinear distributions. In particular, kernel principal components is one of the nonlinear features we used to built MPs with partial reconstruction, with significantly improvement in terms of classification accuracies. Experimental results show that EMP built on nonlinear features (KPCs) perform better than those built on linear features (PCs and ICs), the improvements are 2%-5% and 3%-7% for the data sets of *University Area* and *Pavia Center*.

On the basis of the study, the analysis and the experiments carried out in the framework of this thesis, we identified some interesting directions of research as future developments of this work.

1. A kernel version of SELD aimed at exploiting nonlinear properties of the data can be developed. Moreover, our fast iterative kernel principal component analysis [82], which solves the eigenvectors through iteration, can reduce the space complexity and time complexity greatly.

2. Further investigations on using morphological profiles for classification of hyperspectral data should be carried out. The performances of classification are better by using the morphological profiles with partial built on kernel principal components than those built on linear features. However, to get the same percentages of cumulative variance of eigenvalues, KPCA needs more features, which will produce high-dimensional data when constructing MPs. We believe semi-supervised FE or kernel semi-supervised FE will perform better to reduce the dimensionality of generated MPs.

3. A technique based on denoising the extracted features in low-dimensional subspace can be developed. Nowadays, some hyperspectral data from the agriculture have more than a thousand spectral bands, denoising on the original hyperspectral data may result in high cost on storage resources and computational time. In our opinion, we can first use feature extraction (e.g. PCA and kernel methods) to reduce the dimensionality of the original data, and

then denoising on the extracted features in the much lower dimensional subspace. This can combine spectral and spatial information of the hyperspectral data better to get a higher classification accuracy.

Bibliography

[1] G.F. Hughes. *On the mean accuracy of statistical pattern recognizers*. IEEE Transactions on Information Theory, 14:55–63, 1968.

[2] J.R. Schott. *Remote sensing: the image chain approach (2nd ed.)*. Oxford University Press, 2007.

[3] J.H. Schurmer. *Hyperspectral imaging from space, Air Force Research Laboratories Technology Horizons*, Dec. 2003.

[4] R.B. Smith. *Introduction to hyperspectral imaging with TMIPS*.

[5] F.M. Lacar et al. *Use of hyperspectral imagery for mapping grape varieties in the Barossa Valley*. IEEE International Geoscience and remote sensing symposium (IGARSS'01), 6:2875–2877, 2001.

[6] J.G. Ferwerda. *Charting the quality of forage: measuring and mapping the variation of chemical components in foliage with hyperspectral remote sensing*. Wageningen University, ITC Dissertation, 2005.

[7] A.K. Tilling et al. *Remote sensing to detect nitrogen and water stress in wheat*.

[8] J.A.F. Pierna et al. *Combination of Support Vector Machines (SVM) and Near Infrared (NIR) imaging spectroscopy for the detection of meat and bone meat (MBM) in compound feeds*. Journal of Chemometrics, 18:341–349, 2004.

[9] H. Werff. *Knowledge based remote sensing of complex objects: recognition of spectral and spatial patterns resulting from natural hydrocarbon*. Utrecht University, ITC Dissertationn, 2006.

[10] M.F. Noomen. *Hyperspectral reflectance of vegetation affected by underground hydrocarbon gas seepage*. Enschede, ITC Dissertation, 2007.

[11] P. Tremblay, S. Savary, and M. Rolland et al. *Standoff gas identification and quantification from turbulent stack plumes with an imaging Fourier-transform spectrometer*. In Proceedings of SPIE 2010, volume 7673, 2010.

[12] S.J. Raudys and A.K. Jain. *Small smple size effects in statistical pattern recognition: recommendations for practitioners*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13:252–264, 1991.

[13] J.A. Benediktsson, J.R. Sveinsson, and K. Arnason. *Classification and feature extraction of AVIRIS data*. IEEE Transactions on Geoscience and Remote Sensing, 33:1194–628, 1995.

[14] P.J. Curran and J.L. Dungan. *Estimation of signal-to-noise: A new procedure applied to AVIRIS data*. IEEE Transactions on Geoscience and Remote Sensing, 27:620–628, 1989.

[15] R.O. Green et al. *Imaging spectrosopy and the airborne visible/infrared imaging spectrometer (AVIRIS)*. Remote Sensing of Environment, 65:227–248, 1998.

[16] C. Wallays, B. Missotten, J. De Baerdemaeker, and W. Saeys. *Hyperspectral waveband selection for on-line measurement of grain cleanness*. Biosystems Engineering, 104:1–7, 2009.

[17] J.A. Benediktsson, J.A. Palmason, and J.R. Sveinsson. *Classification of hyperspectral data from urban areas based on extended morphological profiles*. IEEE Trans. Geosci. Remote Sens., 43:480–491, 2005.

[18] M. Fauvel, J.A. Benediktsson, J. Chanussot, and J.R. Sveinsson. *Spectral and Spatial Classification of Hyperspectral Data Using SVMs and Morphological Profile*. IEEE Trans. Geosci. Remote Sens., 46:3804–3814, 2008.

[19] T. Castaings, B. Waske, J.A. Benediktsson, and J. Chanussot. *On the influence of feature reduction for the classification of hyperspectral images based on the extended morphological profile*. Int. J. Remote Sens., 31:5921–5939, 2010.

[20] H.H. Ho, C.H. Li, B.C. Kuo, and Y.Y. Chang. *Novel nearest neighbor classifier based on adaptive nonparametric separability*. Lecture Notes in Artificial Intelligence, 4304:204–213, 2006.

[21] L. Breiman. *Bagging predictors*. Machine Learning, 24:123–140, 1996.

[22] Y. Freund and R.E. Schapire. *A decsion-theoretic generalization of on-line learning and an application to boosting*. In In Proc. Second Eur. Conf. Computat. Learning Th., pages 23–37, 1995.

[23] Y. Freund and R.E. Schapire. *Decisio-theoretic generalization of on-line learning and an application to boosting*. Journal of Computer and System Sciences, 55:119–139, 1997.

[24] T.K. Ho. *The random subspace mehod for constructing decision forests*. IEEE Transaction on Pattern Analysis and Machine Intelligence, 20:832–844, 1998.

[25] T.G. Dietterich, M. Kearns, and Y. Mansour. *Applying the weak learning framework to understand the improve c4.5*. In Proceedings 13t International Conference on Machine Learning, pages 96–104, 1996.

[26] C. Ji and S. Ma. *Combinations of weak classifiers*. IEEE Transactions on Neural Networks, 8:32–42, 1997.

[27] M. Skruichina and R.P.W. Duin. *Bagging, boosting and random subspace method for linear classifiers*. Patter Analysis & Applications, 5:121–135, 2002.

[28] M. Fauvel, J. Chanussot, and J.A. Benediktsson. *Kernel principal component analysis for the classification of hyperspectral remote-sensing data over urban areas*. EURASIP Journal on Advances in Signal Processing, 2009, 2009.

[29] L. Bruzzone and C. Persello. *A Novel Approach to the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images With Improved Generalization Capability*. IEEE Trans. Geosci. Remote Sens., 47:3180–3191, 2009.

[30] F. van der Heiden, R.P.W. Duin, D. de Ridder, and D.M.J. Tax. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Chichester: John Wiley & Sons, 2004.

[31] S.B. Serpico and L.Bruzzone. *A new search algorithm for feature selection in hyperspectral remote sensing images*. IEEE Trans. Geosci. Remote Sens., 39:1360–1367, 1994.

[32] S.B. Serpico and G. Moser. *Extrction of spectral channels from hyperspectral images for classification purposes*. IEEE Trans. Geosci. Remote Sens., 45:484–495, 2007.

[33] M. Fong. *Dimension reduction on hyperspectral images*. University of California, Los Angeles, United States, Report, 2007.

[34] K.-S. Park, S. Hong, P. Park, and W.-D. Cho. *Spectral content characterization for efficient image detection algorithm design*. EURASIP Journal on Advances in Signal Processing, 2007, 2007.

[35] B.C. Kuo and D.A. Landgrebe. *Nonparametric weighted feature extraction for classification*. IEEE Trans. Geosci. Remote Sens., 42:1096–1105, 2004.

[36] T.V. Bandos, L. Bruzzone, and G. Camps-Valls. *Classification of hyperspectral images with regularized linear discriminant analysis*. IEEE Trans. Geosci. Remote Sens., 47:862–873, 2009.

[37] H. Hotelling. *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, 24:417–441, 1933.

[38] J. Schott. *Remote sensing: the image chain approach*. Oxford University Press, 1996.

[39] A. Plaza, P. Martinez, J. Plaza, and R. Perez. *Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations*. IEEE Trans. on Geoscience and Rem. Sensing, 43:466–479, 2005.

[40] J. Wang and C.I. Chang. *Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis*. IEEE Trans. on Geoscience and Rem. Sensing, 44:1586–1600, 2006.

[41] S. Kaewpijit, J.L. Moigne, and T. El-Ghazawi. *Automatic reduction of hyperspectral imagery using wavelet spectral analysis*. IEEE Trans. on Geoscience and Rem. Sensing, 41:863–871, 2003.

[42] L.M. Bruce, C.H. Koger, and J. Li. *Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction*. IEEE Trans. on Geoscience and Rem. Sensing, 40:2331–2338, 2002.

[43] J.C. Harsanyi. *Detection and classification of subpixel spectral signatures in hyperspectral image sequences*. Ph.D. dissertation, University of Maryland, 1993.

[44] R.D. Phillips, W.T. Watson, R.H. Wynne, and C.E. Blinn. *Feature reduction using a singular value decomposition for the iterative guided spectral class rejection hybrid classifier*. ISPRS J. Photogramm. Remote Sens., 64:107–116, 2009.

[45] M. He and S. Mei. *Dimension reduction by random projection for endmember extraction*. In In Proc. IEEE Conf. Ind. Electron. Appl., pages 2323–2327, 2010.

[46] N. Renard, S. Bourennane, and J. Blanc-Talon. *Denoising and dimensionality reduction using multilinear tools for hyperspectral images*. IEEE Geosci. Remote Sens. Lett., 5:138–142, 2008.

[47] R. Dianat and S. Kasaei. *Dimension reduction of optical remote sensing images via minimum change rate deviation method*. IEEE Trans. on Geoscience and Rem. Sensing, 48:198–206, 2010.

[48] G. Chen and S.-E Qian. *Dimensionality reduction of hyperspectral imagery using improved locally linear embedding*. Journal of Applied Remote Sensing, 1:1–10, 2007.

[49] L. Ma, M.M. Crawford, and J. Tian. *Local manifold learning-based k-nearest-neighbor for hyperspectral image classification*. IEEE Trans. Geosci. Remote Sens., 48:4099–4109, 2010.

[50] C.M. Bachmann, T.L. Ainsworth, and R.A. Fusina. *Exploiting manifold geometry in hyperspectral imagery*. IEEE Trans. Geosci. Remote Sens., 43:441–454, 2005.

[51] M. Belkin and P. Niyogi. *Laplacia Eigenmaps and Spectral Techniques for Embedding and Clustering*. Advances in Neural Information Processing Systems 14, MIT Press, British Columbia, 2002.

[52] Z.Y. Zhang and H.Y. Zha. *Principal manifolds and nonlinear dimensionality reduction via tangent space alignment*. SIAM J. Sci. Comput., 26:313–338, 2004.

[53] X.F. He, D. Cai, S.C. Yan, and H.J. Zhang. *Neighborhood preserving embedding*. In Tenth IEEE International Conference on Computer Vision 2005, volume 2, pages 1208–1213, 2005.

[54] X.F. He and P. Niyogi. *Locality preserving projections*. Advances in Neural Information Processing Systems 16, MIT Press, British Columbia, 2004.

[55] T. Zhang, J. Yang, D. Zhao, and X. Gea. *Linear local tangent space alignment and application to face recognition*. Neurocomputing Letters, 70:1547–1553, 2007.

[56] H.Y. Huang and B.C. Kuo. *Double nearest proportion feature extraction for hyperspectral-image classification*. IEEE Trans. Geosci. Remote Sens., 48:4034–4046, 2010.

[57] C.-I Chang and H. Ren. *An experiment-based quantitative and comparative analysis of target detection and image classification algorithms for hyperspectral imagery*. IEEE Trans. Geosci. Remote Sens., 38:1044–1063, 2000.

[58] Q. Du. *Modified Fisher's linear discriminant analysis for hyperspectral imagery*. IEEE Geosci. Remote Sens. Lett., 4:503–507, 2007.

[59] B.C. Kuo, C.W. Chang, C.C. Hung, and H.P. Wang. *A modified nonparametric weight feature extraction using spatial and spectral information*. Proceedings of International Geoscience and Remote Sensing Symposium, 2006.

[60] B.C. Kuo, C.H. Li, and J.M. Yang. *Kernel nonparametric weighted feature extraction for hyperspectral image classification*. IEEE Trans. Geosci. Remote Sens., 47:1139–1155, 2009.

[61] B.C. Kuo, C.H. Chang, T.W. Sheu, and C.C. Hung. *Feature extractions using labeled and unlabeled data*. In In Proc. IGARSS, 2005.

[62] A. Blum and T. Mitchell. *Combining labeled and unlabeled data with co-training*. In Proceedings of the 11th Annual Conference on Computational Learning Theory, pages 92–100, 1998.

[63] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[64] L. Bruzzone, M. Chi, and M. Marconcini. *A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images*. IEEE Trans. Geosci. Remote Sens., 44:3363–3373, 2006.

[65] G. Camps-Valls, T. Bandos, and D. Zhou. *Semisupervised graph-based hyperspectral image classification*. IEEE Transactions on Geoscience and Remote Sensing, 45:3044–3054, 2007.

[66] X. Zhu. *Semi-supervised learning literature survey*. Univ. Wisconsin Madison, Madison, WI, Comput. Sci., 2005.

[67] D. Cai, X.F. He, and J.W. Han. *Semi-supervised discriminant analysis*. In IEEE 11th International Conference on Computer Vision 'ICCV07', pages 1–7, 2008.

[68] S.G. Chen and D.Q. Zhang. *Semisupervised Dimensionality Reduction with Pairwise Constraints for Hyperspectral Image Classification*. IEEE Geoscience and Remote Sensing Letters, 8:369–373, 2011.

[69] M. Sugiyama, T. Ide, S. Nakajima, and J. Sese. *Semi-supervised local Fisher discriminant analysis for dimensionality reduction*. Machine Learning, 78:35–61, 2010.

[70] P. Soille. *Morphological Image Analysis, Principles and Applications,2nd ed.* Berlin, Germany: Springer-Verlag, 2003.

[71] P. Soille and M. Pesaresi. *Advances in Mathematical Morphology Applied to Geoscience and remote Sensing*. IEEE Trans. Geosci. Remote Sens., 40:2042–2055, 2002.

[72] M. Pesaresi and J.A. Benediktsson. *A new approach for the morphological segmentation of high-resolution satellite imagery*. IEEE Trans. Geosci. Remote Sens., 39:309–320, 2001.

[73] R. Bellens, S. Gautama, L. Martinez-Fonte, W. Philips, J.C.-W. Chan, and F. Canters. *Improved classification of VHR images of urban areas using directional morphological profiles*. IEEE Trans. Geosci. Remote Sens., 46:2803–2812, 2008.

[74] M. Dalla Mura, J.A. Benediktsson, B. Waske, and L. Bruzzone. *Extended profiles with morphological attribute filters for the analysis of hyperspectral data*. Int. J. Remote Sens., 31:5975–5991, 2010.

[75] M. Dalla Mura, J.A. Benediktsson, B. Waske, and L. Bruzzone. *Morphological attribute profiles for the analysis of very high resolution images*. IEEE Trans. Geosci. Remote Sens., 48:3747–3762, 2010.

[76] M. Dalla Mura, A. Villa, J.A. Benediktsson, J. Chanussot, and L. Bruzzone. *Classification of Hyperspectral Images by Using Extended Morphological Attribute Profiles and Independent Component Analysis*. IEEE Geosci. Remote Sens. Lett., 8:541–545, 2011.

[77] J. Crespo, J. Serra, and R. Shafer. *Theoretical aspects of morphological filters by reconstruction*. Signal Process., 47:201–225, 1995.

[78] J.A. Benediktsson, M. Pesaresi, and K. Arnason. *Classification and feature extraction for remote sensing images from urban areas based on morphological transformations*. IEEE Trans. Geosci. Remote Sens., 41:3747–3762, 2003.

[79] W.Z. Liao, A. Pizurica, W. Philips, and Y.G. Pi. *Feature extraction for hyperspectral image based on semi-supervised local discriminant analysis*. In in Proc. IEEE Joint Urban Remote Sensing Event (JURSE 2011), pages 401–404, 2011.

[80] Wenzhi Liao, Aleksandra Pižurica, Paul Scheunders, Wilfried Philips, and Youguo Pi. *Semi-Supervised Local Discriminant Analysis for Feature Extraction in Hyperspectral Images*. IEEE Trans. Geosci. Remote Sens., accepted.

[81] W. Liao, R. Bellens, A. Pizurica, W. Philips, and Y. Pi. *Classification of Hyperspectral Data over Urban Areas Using Directional Morphological Profiles and Semi-supervised Feature Extraction*. In IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5:14 pages, 2012.

[82] W. Liao, A. Pizurica, W. Philips, and Y. Pi. *A Fast Iterative PCA Feature Extraction for Hyperspectral Images*. In In Proc. of the 2010 IEEE International Conference on Image Processing (ICIP 2010), pages 1317–1320, 2010.

[83] Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Wilfried Philips, and Youguo Pi. *Classification of hyperspectral data over urban areas based on extended morphological profile with partial reconstruction*. In In Proc. Advanced Concepts for Intelligent Vision Systems (ACIVS) 2012, page submitted, 2012.

[84] B. Scholkopf and A. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press Series, 2002.

[85] A. Hyvarinen and E. Oja. *Independent component analysis: algorithms and applications*. Neural Networks, 13:411–430, 2000.

[86] X. Song, G. Fan, and M. Rao. *Automatic CRP mapping using nonparametric machine learning approaches*. IEEE Trans. Geosci. Remote Sens., 43:888–897, 2005.

[87] J.A. Palmason, J.A. Benediktsson, J.R. Sveinsson, and J. Chanussot. *Classification of hyperspectral data from urban areas using morphological preprocessing and independent component analysis*. In In Proc. IGARSS'05, volume 1, pages 176–179, 2005.

[88] R.A. Fisher. *The use of multiple measurements in taxonomic problems*. Ann. Eugenics, 7:179–188, 1936.

[89] R.A. Fisher. *The statistical utilization of multiple measurements*. Ann. Eugenics, 7:376–386, 1938.

[90] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. CA: Academic, 1990.

[91] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice-Hall, 2007.

[92] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[93] G. Baudat and F. Anouar. *Generalized discriminant analysis using a kernel approach*. Neural Comput., 12:2385–2404, 2000.

[94] C. Lee and D.A. Landgrebe. *Feature extraction based on decision boundaries*. IEEE Trans. Pattern Anal. Mach. Intell., 15:388–400, 1993.

[95] M.M. Dundar and D. Landgrebe. *Toward an optimal supervised classifier for the analysis of hyperspectral data*. IEEE Trans. Geosci. Remote Sens., 42:271–277, 2004.

[96] P.F. Hsieh, D.S. Wang, and C.W. Hsu. *A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information, extraction*. IEEE Trans. Pattern Anal. Mach. Intell., 28:223–235, 2006.

[97] J.B. Tenenbaum, V. de Silva, and J.C. Langford. *A global geometric framework for nonlinear dimensionality reduction*. Science, 290:2319–2323, 2000.

[98] S.T. Roweis and L.K. Saul. *Nonlinear dimensionality reduction by locally linear embedding*. Science, 290:2323–2326, 2000.

[99] L.K. Saul and S.T. Roweis. *Think globally, fit locally: unsupervised learning of low dimensional manifolds*. Journal of Machine Learning Research, 4:119–155, 2003.

[100] Hongyu Li, Wenbin Chen, and I-Fan Shen. *Supervised local tangent space alignment for classification*. In IJCAI, pages 1620–1621, 2005.

[101] H. Chang and D.Y. Yeung. *Robust locally linear embedding*. Pattern Recognition, 39:1053–1065, 2006.

[102] M. Sugiyama and S. Roweis. *Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis*. Journal of Machine Learning Research, 8:1027–1061, 2007.

[103] G.M. Foody. *Thematic map comparison: evaluating the statistical significance of differences in classification accuracy*. Photogrammetric Engineering & Remote Sensing, 70:627–633, 2004.

[104] J.A. Richards and Xiuping Jia. *Remote Sensing Digital Image Analysis, An Introduction*. Springer-Verlag, Berlin, Heidenberg, 3rd edition, 1999.

[105] D.L. Civco. *Artificial neural networks for land-cover classification and mapping*. International Journal of Geophysical Information Systems, 7:173–186, 1993.

[106] H. Bischof and A. Leona. *Finding optimal neural networks for land use classification*. IEEE Transactions on Geoscience and Remote Sensing, 36:337–341, 1998.

[107] H. Yang, F. van der Meer, W. Bakker, and Z.J. Tan. *A back-propagation neural network for mineralogical mapping from AVIRIS data*. International Journal of Remote Sensing, 20:97–110, 1999.

[108] B.E. Bose, I.M. Guyon, and V.N. Vapnik. *A training algorithm for optimal margin classifiers*. In In 5th Annual ACM Workshop on COLT, D. Haussler, Ed., Pittsburgh, pages 144–152, 1992.

[109] R. F. Cromp J. A. Gualtieri, S. R. Chettri and L. F. Johnson. *Support vector machine classifiers as applied to AVIRIS data*. In In Proceedings of the 8th JPL Airborne Geoscience Workshop, 1999.

[110] C. Huang, L.S. Davis, and J.R.G. Townshend. *An assessment of support vector machines for land cover classification*. International Journal of Remote Sensing, 23:725–749, 2002.

[111] G. Camps-Valls, L. Gomez-Chova, J. Calpe, E. Soria, J.D. Martin, L. Alonso, and J. Moreno. *Robust support vector method for hyperspectral data classification and knowledge discovery*. IEEE Transactions on Geoscience and Remote Sensing, 42:1530–1542, 2004.

[112] G. Camps-Valls and L. Bruzzone. *Kernel-based methods for hyperspectral image classification*. IEEE Transactions on Geoscience and Remote Sensing, 43:1351–1362, 2005.

[113] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[114] H. Caillol, A. Hillion, and W. Pieczynski. *Fuzzy random fields and unsupervised image segmentation*. IEEE Transactions on Geoscience and Remote Sensing, 31:801–810, 1993.

[115] P. Masson and W. Pieczynski. *SEM algorithm and unsupervised statistical segmentation of satellite images*. IEEE Transactions on Geoscience and Remote Sensing, 31:618–633, 1993.

[116] S. Le Hegarat-Mascle, I. Bloch, and D. Vidal-Madjar. *Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing*. IEEE Transactions on Geoscience and Remote Sensing, 35:1018–1031, 1997.

[117] P.B.G. Dammert, J.I.H. Askne, and S. Kuhlmann. *Unsupervised segmentation of multitemporal interferometric SAR images*. IEEE Transactions on Geoscience and Remote Sensing, 37:2259–2271, 1999.

[118] T. Yamazaki and D. Gingras. *Unsupervised multispectral image classification using MRF models and VQ method*. IEEE Transactions on Geoscience and Remote Sensing, 37:1173–1176, 1999.

[119] Y. Zhong, L. Zhang, B. Huang, and L. Pingxiang. *An unsupervised artificial immune classifier for multi/hyperspectral remote sensing imagery*. IEEE Transactions on Geoscience and Remote Sensing, 44:420–431, 2006.

[120] Matthias Seeger. *Learning with labeled and unlabeled data*, 2001.

[121] Xiaojin Zhu. *Semi-supervised learning literature survey*. Computer Sciences, University of Wisconsin-Madison, 2005.

[122] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning 1st edition*. MIT Press, Cambridge, Massachusetts and London, England, 2006.

[123] N.M. Dempster, A.P. Laird, and D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, 39:1–38, 1977.

[124] Q. Jackson and D.A. Landgrebe. *An adaptive classifier design for high-dimensional data analysis with a limited training data set*. IEEE Transactions on Geoscience and Remote Sensing, 39:2664–2679, 2001.

[125] L. Bruzzone, M. Chi, and M. Marconcini. *Transductive SVMs for semisupervised classification of hyperspectral data*. In In International Geoscience and Remote Sensing Symposium, IGARSS2005, pages 144–152, 2005.

[126] F. Chung. *Spectral graph theory*. In in CBMS Regional Conference Series in Mathematics, 1997.

[127] M.I. Jordan. *Learning in Graphical Models (1st edition)*. MIT Press, Cambridge, Massachusetts and London, England, 1999.

[128] K. Bennet and A. Demiriz. *Semi-supervised support vector machines*. in Advances in Neural Information Processing Systems (NIPS), Cambridge, 1998.

[129] T. Joachims. *Making large-scale support vector machine learning practical*. MIT Press, 1999.

[130] O. Chapelle and A. Zien. *Semi-supervised classification by low density separation*. In In International Conference on Artificial Intelligence and Statistics (AISTATS), pages 57–64, 2005.

[131] I.W. Tsang and J.T. Kwok. *Large-scale sparsified manifold regularization*. In in Advances in Neural Information Processing Systems (NIPS), pages 1401–1408, 2006.

[132] G. Gasso, K. Zapien, and S. Canu. *L1-norm regularization path for sparse semi-supervised Laplacian SVM*. In in International Conference on Machine Learning and Applications (ICMLA), 2007.

[133] A. Erkan and Y. Altun. *Semi-supervised learning via generalized maximum entropy*. In in International Conference on Artificial Intelligence and Statistics (AISTATS), volume 9, pages 209–216, 2010.

[134] T. Bandos, D. Zhou, and G. Camps-Valls. *Semi-supervised hyperspectral image classification with graphs*. In in International Geoscience and Remote Sensing Symposium, IGARSS2006, 2006.

[135] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Scholkopf. *Learning with local and global consistency*. In Advances in Neural Information Processing Systems, NIPS2004, 2004.

[136] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla. *Composite kernels for hyperspectral image classification*. IEEE Geoscience and Remote Sensing Letters, 3:93–97, 2006.

[137] L. Capobianco, A. Garzelli, and G. Camps-Valls. *Target detection with semisupervised kernel orthogonal subspace projection*. IEEE Trans. Geosci. Remote Sens., 47:3822–3833, 2009.

[138] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. *Dimensionality Reduction: A Comparative Review*. In Tilburg University Technical Report, 2009.

[139] D.A. Landgrebe. *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.

[140] J. Ham, Y. Chen, M. Crawford, and J. Ghosh. *Investigation of the random forest framework for classification of hyperspectral data*. IEEE Trans. Geosci. Remote Sens., 43:492–501, 2005.

[141] R.P.W. Duin, P. Juszczak, D. de Ridder, P. Paclik, E. Pekalska, and D.M.J. Tax. *PRTools, A Matlab Toolbox for Pattern Recognition*, 2004.

[142] Christopher J.C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery, 2:121–167, 1998.

[143] C.C. Chang and C.J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001.

[144] J. Munoz-Marf, L. Bruzzone, and G. Camps-Valls. *A support vector domain description approach to supervised classification of remote sensing images*. IEEE Trans. Geosci. Remote Sens., 45:2683–2692, 2008.

[145] B. Mojaradi, H. Abrishami-Moghaddam, M.J.V. Zoej, and R.P.W. Duin. *Dimensionality Reduction of Hyperspectral Data via Spectral Feature Extraction*. IEEE Trans. Geosci. Remote Sens., 47:2091–2105, 2009.

[146] De Berg and Mark et al. *Computational Geometry: Algorithms and Applications (3rd Edition)*. Springer, 2008.

[147] V. Castelli and T. Cover. *The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter*. IEEE Trans. Inf. Theory, 42:2102–2117, 1996.

[148] K. Sinha and M. Belkin. *The value of labeled and unlabeled examples when the model is imperfect*. Cambridge, MA: MIT Press, 2008.

[149] M. Dalla Mura, J. Benediktsson, and L. Bruzzone. *Classification of hyperspectral images with extended attribute profiles and feature extraction techniques*. In in Proc. IGARSS, pages 76–79, 2010.

[150] A.A. Green, M. Berman, P. Switzer, and M.D. Craig. *A transformation for ordering multispectral data in terms of image quality with implications for noise removal*. IEEE Trans. Geosci. Remote Sens., 26:65–74, 1988.

[151] I.T. Jolliffe. *Principal Component Analysis*. Springer, New York, 2002.

[152] J.W. Boardman and F.A. Kruse. *Automated spectral analysis: A geologic example using AVIRIS data, north Grapevine Mountains, Nevada*. In Proceedings of the Tenth Thematic Conference on Geologic Remote Sensing, pages I–407, 1994.

[153] S. Prasad and L.M. Bruce. *Decision fusion with confidence based weight assignment for hyperspectral target recognition*. IEEE Trans. Geosci. Remote Sens., 46:1448–1456, 2008.

[154] M.D. Farrel and R.M. Mersereau. *On the impact of PCA dimension reduction for hyperspectral detection of difficult targets*. IEEE Geosci. Remote Sens. Lett., 2:192–195, 2005.

[155] M. Niskanen and O. Silven. *Comparison of dimensionality reduction methods for wood surface inspection*. In In Proceedings of the 6th International Conference on Quality Control by Artificial Vision, pages 178–188, 2003.

[156] C. J. C. Burges. *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, chapter Geometric Methods for Feature Selection and Dimensional Reduction: A Guided Tour*. Kluwer Academic Publishers, 2005.

[157] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, New York, 2007.

[158] L.K. Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee. *Spectral methods for dimensionality reduction*. In Semisupervised Learning, MIT Press, 2006.

[159] J. Venna. *Dimensionality reduction for visual exploration of similarity structures*. PhD thesis, Helsinki University of Technology, 2007.

[160] W.S. Togerson. *Theory and methods of scaling*. Wiley, 1958.

[161] Vin de Silva and Joshua B. Tenenbaum. *Global versus local methods in nonlinear dimensionality reduction*. In In Advances in Neural Information Processing Systems (NIPS), pages 705–712, 2002.

[162] Xin Yang, Haoying Fu, Hongyuan Zha, and Jesse Barlow. *Semi-supervised nonlinear dimensionality reduction*. In ACM ICML, pages 1065–1072, 2006.

[163] Marc Lennon, Gregoire Mercier, Marie-Catherine Mouchot, and Laurence Hubert-Moy. *Curvilinear component analysis for nonlinear dimensionality reduction of hyperspectral images*. Image and Signal Processing for Remote Sensing VII, 4541:157–168, 2002.

[164] D.R. Olsen and K. Fukunaga. *Representation of nonlinear data surfaces*. IEEE Transactions on Computers, 22:915–922, 1973.

[165] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller. *Fisher discriminant analysis with kernels*. In in Proc. IEEE Neural Netw. Signal Process, pages 41–48, 1999.

[166] S. Prasad and L.M. Bruce. *Information fusion in kernel-induced spaces for robust subpixel hyperspectral ATR*. IEEE Geosci. Remote Sens. Lett., 6:572–576, 2009.

[167] P.L. Lai and C. Fyfe. *Kernel and nonlinear canonical correlation analysis*. Int. J. Neural Syst., 10:365–377, 2000.

[168] F.R. Bach and M.I. Jordan. *Kernel independent component analysis*. J. Mach. Learn. Res., 3:1–48, 2002.

[169] C.M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[170] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes: The Art of Scientific Computing, 3rd ed.* Cambridge, U.K.: Cambridge Univ. Press, 2007.

[171] A.A. Nielsen and M.J. Canty. *Kernel principal component analysis for change detection*. In in Proc. SPIE Eur. Remote Sens. Conf., volume 7109A, 2008.

[172] A.A. Nielsen. *Kernel maximum autocorrelation factor and minimum noise fraction transformations*. IEEE Tran. on image processing, 20:612–624, 2011.

[173] J. Ham, D. Lee, S. Mika, and B. Scholkopf. *A Kernel View of the Dimensionality Reduction of Manifolds*. In Proc. Int'l Conf. Machine Learning, pages 47–54, 2004.

[174] W.Y. Shi, Y.F. Guo, C. Jin, and X.Y. Xue. *An improved generalized discriminant analysis for large-scale data set*. In Seventh International Conference on Machine Learning and Applications, pages 769–772, 2008.

[175] K. Burgers, Y. Fessehatsion, S. Rahmani, and J.Y. Seo. *A comparative analysis of dimension reduction algorithms on hyperspectral data*, 2009.

[176] K.I. Kim, M.O. Franz, and B. Scholkopf. *Iterative kernel principal component analysis for image modeling*. IEEE Trans Pattern Analysis and Machine Intelligence, 27:1351–1366, 2005.

[177] T. Sander. *Optimal unsupervised learning in a single-layer linear feedforward neural network*. Neural Network, 12:459–473, 1989.

[178] J. Weng, Y. Zhang, and W.S. Huang. *Candid covariance-free incremental principal component analysis*. IEEE Trans Pattern Analysis and Machine Intelligence, 25:1034–1040, 2003.

[179] Y. Zhang and J. Weng. *Covergence analysis of complementary candid incremental principal component analysis*, 2001.

[180] P.K. Varshney and M.K. Arora. *Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data*. Berlin, Germany: Springer-Verlag, 2003.

[181] B. Scholkopf, A.J. Smola, and K.R. Muller. *Nonlinear component analysis as a kernel eigenvalue problem*. Neural Computation, 10:1299–1319, 1998.

[182] M. Fauvel, J. Chanussot, and J.A. Benediktsson. *Kernel principal component analysis for feature reduction in hyperspectrale images analysis*. In

Proceedings of the 7th Nordic Signal Processing Symposium, pages 238–241, 2006.

[183] Baofeng Guo, Steve Gunn, Bob Damper, and James Nelson. *Band selection for hyperspectral image classification using mutual information*. IEEE Geoscience and Remote Sensing Letters, 3:522–526, 2006.

[184] K. Hotta. *Local normalized linear summation kernel for fast and robust recognition*. Pattern Recognition, 43:906–913, 2010.