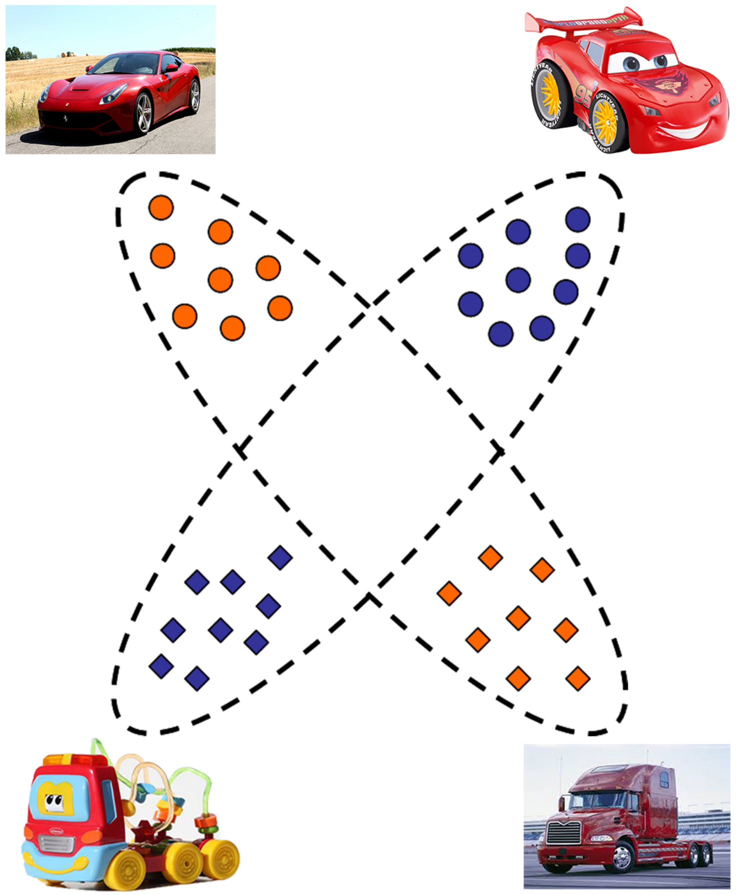
The front cover picture illustrates an example of learning from weakly representative data. Assume that we want to classify real-world images of cars and trucks using a large training (labeled) set of car and truck images collected from animation movies. Although the objects of interest (cars and trucks) look similar in real-world and animated images, there are key differences between the two worlds, e.g., animated objects are personified to make them look like human. Thus, one needs to take into account those differences (i.e. domain shift) in order to apply a classifier trained with animated images to real-world images.

The front cover picture is dedicated to my six-month old son, who is very fascinated by toys and cartoons.

Learning from Weakly Representative Data and Applications in Spectral Image Analysis

Cuong Viet Dinh

Learning from Weakly Representative Data
and
Applications in Spectral Image Analysis



Cuong Viet Dinh

LEARNING FROM WEAKLY REPRESENTATIVE DATA AND APPLICATIONS IN SPECTRAL IMAGE ANALYSIS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 10 oktober 2013 om 15:00 uur
door

CUONG VIET DINH
(In Vietnamese: ĐINH VIỆT CƯỜNG)

Master of Engineering
geboren te Phuc Yen, Vinh Phuc province, Vietnam

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. M.J.T. Reinders

Copromotor: Dr. M. Loog – Professor at the University of Copenhagen

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr.ir. M.J.T. Reinders,	Technische Universiteit Delft, promotor
Dr. M. Loog,	Technische Universiteit Delft, copromotor
Prof.dr. T. Heskes,	Radboud University Nijmegen
Prof.dr. F. Pla,	Universitat Jaume I, Castelló, Spain
Prof.dr. E.O. Postma,	Tilburg University
Prof.dr.ir. L.J. van Vliet,	Technische Universiteit Delft
Prof.dr.ir. H. Russchenberg,	Technische Universiteit Delft
Prof.dr.ir. B.P.F. Lelieveldt,	Technische Universiteit Delft, reservelid

This research is supported by Carinthian Tech Research AG, Austria through the Austrian COMET funding program.

Advanced School for Computing and Imaging

This work was carried out in graduate school ASCI.
ASCI dissertation series number 288.

ISBN 978-94-6186-221-1
© 2013, Cuong V. Dinh, all rights reserved.

To my wife, Minh Anh and my son, Viet Tung.

In the memory of my father in law.

CONTENTS

1. Introduction 1
- 1.1 Context of the research work 1
  - 1.1.1 Introduction to spectral imaging systems 1
  - 1.1.2 Spectral signature 3
  - 1.1.3 Applications of spectral imaging 4
- 1.2 Thesis's objectives 5
- 1.3 Background 6
  - 1.3.1 Transfer learning 6
  - 1.3.2 Semi-supervised learning 8
- 1.4 Thesis's contributions and outline 9

2. SEDMI: Saliency-based Edge Detection in Multispectral Images 15
- 2.1 Introduction 16
- 2.2 Motivation and related work 17
  - 2.2.1 Edge detection as detecting salient features 17
  - 2.2.2 Towards clustering-based edge detection 18
  - 2.2.3 Related work on ensemble clustering 18
- 2.3 SEDMI method 19
  - 2.3.1 Constructing the feature space 19
  - 2.3.2 Performing ensemble clustering 19
  - 2.3.3 Calculating edge strength map 20
  - 2.3.4 Connection with the evidence accumulation clustering 20
- 2.4 Experimental results 21
  - 2.4.1 Artificial data 22
  - 2.4.2 Real-world data sets 28
- 2.5 Discussion 32
- 2.6 Conclusions 35

3. Training Data Selection for Cancer Detection in Spectral Endoscopy Images 37
- 3.1 Introduction 38
- 3.2 System Overview 39
- 3.3 Materials 41
  - 3.3.1 Data acquisition 41

3.3.2 Spectral visualization . . . 42
3.4 Methods . . . 42
3.4.1 Calibration and pre-processing . . . 42
3.4.2 Data selection for training . . . 44
3.5 Experimental Results . . . 44
3.5.1 All available data sets are used for training . . . 45
3.5.2 Training data selection by the Gaussian domain description . . . 46
3.6 Conclusions . . . 48

4. FIDOS: A Generalized Fisher-based Feature Extraction Method for Domain Shift . . . 49
4.1 Introduction . . . 50
4.2 Fisher-based Feature Extraction for Domain Shift . . . 52
4.2.1 Notations . . . 52
4.2.2 Feature Extraction using Classical Fisher Criterion . . . 52
4.2.3 FIDOS . . . 53
4.3 Experimental Setup . . . 55
4.3.1 Data Description . . . 55
4.3.2 Method Evaluation . . . 57
4.4 Experimental Results and Discussion . . . 59
4.4.1 Artificial Data Set . . . 59
4.4.2 Real Data . . . 61
4.5 Conclusions . . . 64
4.6 Acknowledgements . . . 65

5. Semi-supervised Learning . . . 67
5.1 A study on semi-supervised dissimilarity representation . . . 68
5.1.1 Introduction . . . 68
5.1.2 Semi-supervised dissimilarity representation . . . 69
5.1.3 Experiments . . . 70
5.1.4 Discussion and Conclusions . . . 73
5.2 Semi-supervised hyperspectral pixel classification using clustering-based mode selection . . . 74
5.2.1 Introduction . . . 74
5.2.2 Mode seek clustering . . . 75
5.2.3 Feature extraction . . . 75
5.2.4 Semi-supervised classification . . . 76
5.2.5 Data set . . . 76
5.2.6 Experiments, Results and evaluation . . . 77
5.2.7 Conclusions . . . 78

6. Discussion . . . 81
6.1 Overall Conclusion . . . 81

6.2 Outlook 82
6.2.1 Data visualization 82
6.2.2 Data representation for object classification 82
6.2.3 Soft training set selection for transfer learning 83
6.2.4 Extended dissimilarity representation 84

Appendices 85

Bibliography 93

Summary 101

Samenvatting 103

Acknowledgements 107

Curriculum Vitae 109

Publications 111

1

INTRODUCTION

1.1 Context of the research work

1.1.1 Introduction to spectral imaging systems

Spectral imaging, also known as imaging spectroscopy, is concerned with the measurement, analysis, and interpretation of spectra acquired from a given scene or specific object at a distance by an airborne or a satellite sensor [93, 111, 118]. Figure 1.1 illustrates the principle of a spectral imaging system in the case of satellite remote sensing. First, the incident radiation of the source of illumination, the sun in this case, propagates through the atmosphere that modifies its intensity and spectral distributions. The materials interact with this radiation and then reflect, transmit, and/or absorb it. The reflected radiation then passes back through the atmosphere and finally reaches the sensor [75].

At each scanning time, the sensor simultaneously collects the reflected radiation over a line of spatial resolution elements (pixels) in the image. By scanning the image line by line, the sensor collects the reflected radiation over the entire image. As a result, the resulting spectral data is a 3D cube in which the first two dimensions correspond to the spatial location of the scene and the third one shows the spectrum corresponding to each pixel.

Spectral images have been classified into two main categories: multispectral images and hyperspectral images. Traditional multispectral imaging systems, such as the Landsat Thematic Mapper and SPOT XS, capture image data from a few number of carefully chosen spectral bands spread across the visible and infrared regions of the electromagnetic spectrum [42, 113]. This crude spectral detail limits the number and the details of classes that can be discriminated. With advances in sensor technology, a new class of sensors, i.e., hyperspectral imagers, has emerged. These new systems are able to collect image data simultaneously in dozens or hundreds of narrow, adjacent spectral

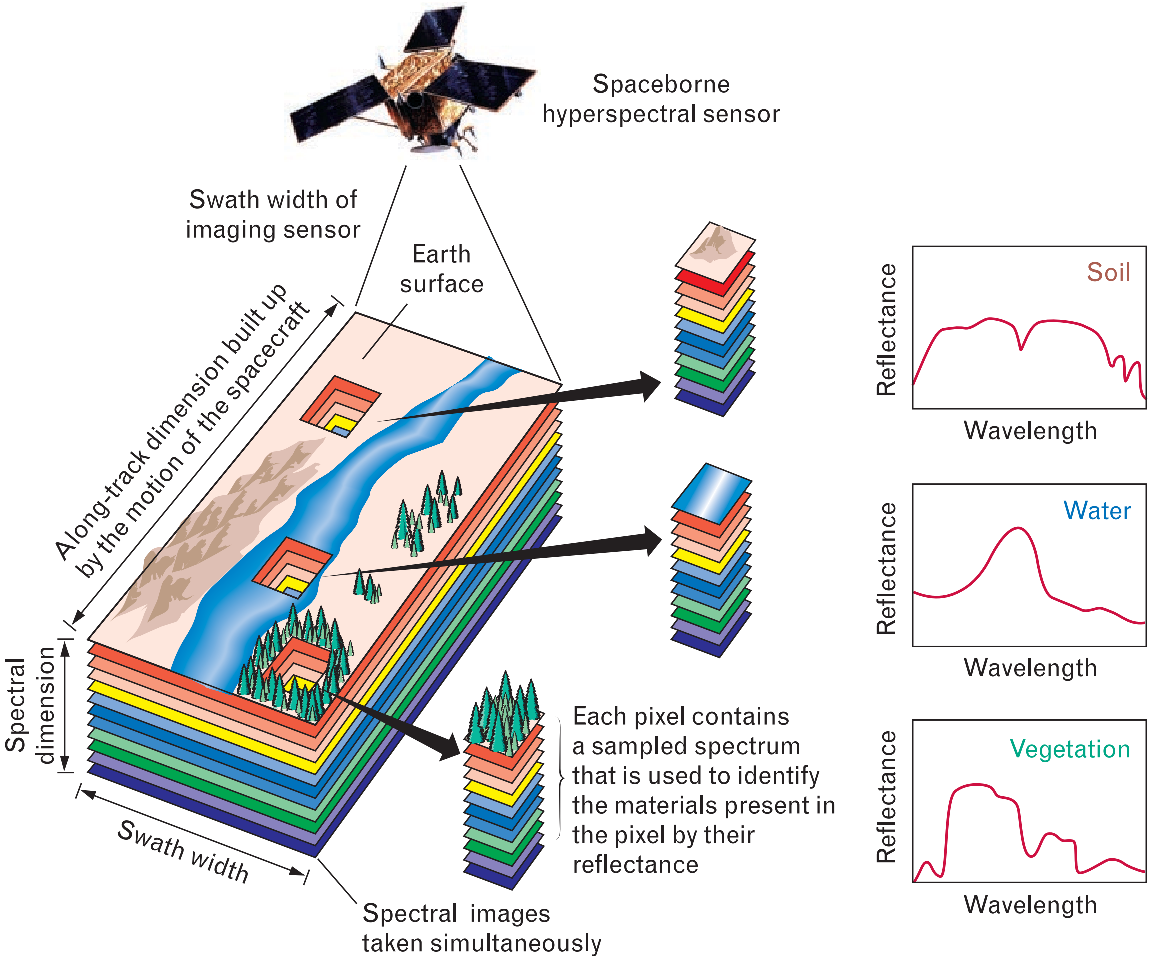

Spaceborne hyperspectral sensor
Swath width of imaging sensor
Earth surface
Along-track dimension built up by the motion of the spacecraft
Spectral dimension
Swath width
Spectral images taken simultaneously
Each pixel contains a sampled spectrum that is used to identify the materials present in the pixel by their reflectance
Reflectance
Soil
Wavelength
Reflectance
Water
Wavelength
Reflectance
Vegetation
Wavelength

Figure 1.1: Principle of a satellite remote sensing system (picture taken from [111]). Left panel: The sensor collects reflected radiation from the scene. Each pixel contains a spectrum that is used to identify the materials present in the pixel. Right panel: Reflectance spectrum is plotted against wavelength for three materials: soil (top), water (middle), and vegetation (bottom).

bands. Thus, compared to multispectral images, hyperspectral images provide significantly more detailed spectral information and can be used to detect and to identify a variety of natural and man-made materials [61, 111].

In this thesis, we use the common term “spectral images" when the distinction between the two categories is irrelevant. When such a distinction is needed, we use the more specific terms, i.e., multispectral images and hyperspectral images.

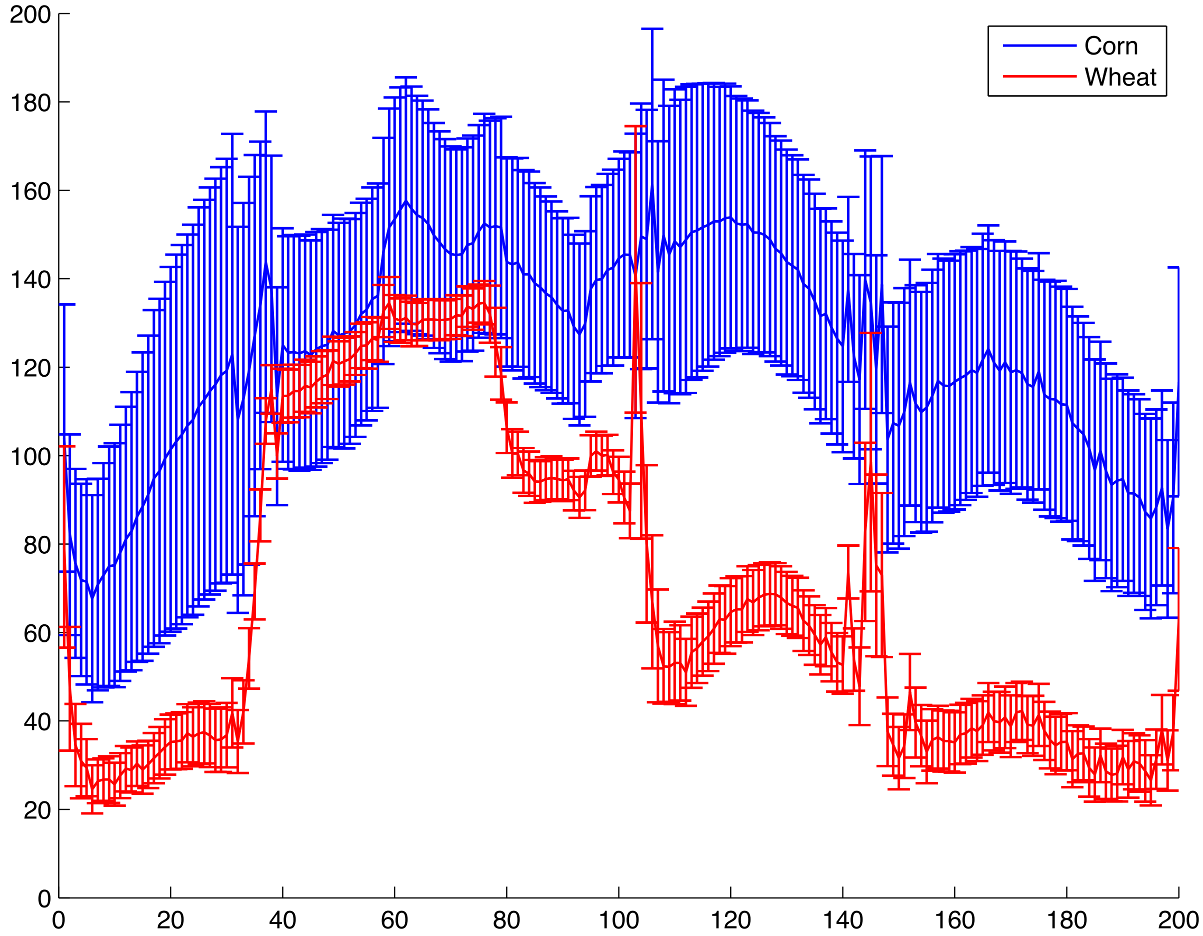


Figure 1.2: Spectral mean and variance of corn (blue) and wheat (red) in an AVIRIS remote sensing data set [45].

1.1.2 Spectral signature

Spectral signature, or reflectance spectrum, is a signal's property of interest in spectral imaging. Reflectance spectrum is defined by the ratio of the reflected radiation to the incident radiation as a function of wavelength. For most materials, their reflectances vary with respect to wavelengths since energy at different wavelengths is scattered or absorbed to different levels [113]. Studies in spectral imaging commonly assume that the reflectance spectrum of every material is unique and, thus, represent a means for uniquely identifying materials [75, 111]. Figure 1.1 (right panel) demonstrates that the reflectance spectral curves of different materials, in this case soil, water, and vegetation exhibit different characteristics and are significantly different from one another.

The term “spectral signature" suggests a unique correspondence between a material and its reflectance spectrum. However, in field data as well as laboratory data, variability in the reflectance spectrum is observed within each material. Many factors may be responsible for such variability, such as the variations in atmospheric conditions, sensor noise, material composition, and the surrounding materials [54, 100, 111]. In addition, the spectra of materials themselves might change over time. For instances, it is easy to notice the spectral difference of a forest between seasons.

Figure 1.2 shows the spectral mean and variance of corn and wheat classes in an Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) remote sensing data set [45]. The data were acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. The figure shows that corn and wheat (two different classes of materials) have different spectral variances. In addition, the within-class variances of corn at several spectral bands are even larger than that between corn and wheat. Thus, it is often a challenging task to distinguish a class of material from other classes especially when they exhibit similar spectral responses and the within-class variation is comparable to the between-class differences at most spectral bands.

1.1.3 Applications of spectral imaging

Although originally developed for mining and geology, spectral imaging has been recently applied to many fields, such as agriculture, environmental monitoring, and biomedical diagnostics.

In agriculture, spectral remote sensing systems are widely used to check soil conditions for potential problems such as moisture deficiencies, to identify potential land yield, and to monitor the development and health of crops. This facilitates the prevention of the spread of disease to ensure the crop's quality. In addition, spectral imaging is also an important tool for food quality and safety inspection of poultry [12], fruits [80], and vegetables [8, 43]. Spectral imaging enables the determination of the composition as well as the distribution of chemical components in food products. Thus, food products can be scanned for disease conditions, ripeness, tenderness, grading, and contamination.

In environmental monitoring, spectral remote sensing systems can be used to identify objects over an area. This allows for analyses of the growth of urban areas and measuring the sensitivity of different areas to natural risks. In addition, spectral remote sensing systems can also be used to investigate changes in the land and coastal-ocean ecosystems [118].

In biomedical diagnostics, spectral imaging has been applied to analyze many different types of samples, ranging from in vivo biochemical species to organs of living people [124]. These studies have given rise to new methods and instrumentations to facilitate early, noninvasive diagnosis of various medical conditions, such as cancer, arteriosclerosis [102], and retinal disease [14]. For early cancer diagnosis, previous studies [48, 88, 127] have shown that there is a significant difference in the fluorescent properties, such as their spectral shape and intensity, between malignant and normal tissues.

Therefore, autofluorescence spectroscopy techniques have been used to identify early instances of diseases in organs such as colon [127], larynx [48], and lung [60]. The advantage of these techniques over the current gold standard technique, i.e., histopathological analysis of biopsies, lies in their potential to perform in vivo detection without the need for tissue removal [124]. Thus, they facilitate the determination of the dysplastic and malignant regions for the biopsy to be performed afterward. These spectroscopic diagnosis techniques are often referred to as point-measurement methods as they attempt to obtain the spectra of a single tissue. Multi/hyper-spectral endoscopy techniques developed recently provide three-dimensional images of the area of interest in both spatial and spectral domains [63, 78, 124]. For instance, [63] demonstrates a real time hyperspectral imaging system for cancer video diagnosis. As a result, spectral images provide richer information than point-measurement techniques as they can acquire the spectra of thousands to millions of cancer and normal tissues at the same time.

1.2 Thesis's objectives

On the one hand, advances in sensor technology enable spectral imaging systems to provide fine spectral resolution needed to characterize the spectral properties of materials. On the other hand, the volume of data in a single scene can seem overwhelming. The spectral difference between two adjacent wavelength bands is typically very small. Therefore, "much of the data in a scene would seem to be redundant, but embedded in it is critical information that can be used to identify materials" [113]. Finding appropriate approaches for data visualization and object classification from this rich source of information remain key challenging topics for research in spectral imaging [95].

This thesis aims at facilitating the analysis in spectral imaging by making use of pattern recognition techniques, on the one hand, to improve visualization, and on the other hand, to directly solve classification problems.

Spectral image visualization by mean of edge detection. Edge detection in spectral images is of interest since it helps to roughly localize target objects present in the image. This might contribute significantly to the success of applications such as target detection in spectral images, which allows users to visualize the scene immediately without having to go through all hundreds of spectral channels to identify the target objects.

It often happens with spectral images that the target objects only appear in a few bands. Consequently, their edges are visible in just a few bands, too. Hence, detecting edges in spectral images are more difficult than detecting edges in grayscale and natural color images in which color intensities are expected to change simultaneously. Therefore, edge detection in spectral images is challenging.

Object classification in the case of small training set size. Classification tasks in spectral imaging applications often have to deal with small training size (also known as small sample size) situations due to the fact that collecting *labeled* data samples is time consuming and expensive. For example, in medical applications, the labeling process is often done by experts who have prior knowledge about the problem. The experts need to assign a label to each sample (pixel in the image) carefully to avoid wrong assignments. In remote sensing applications, collecting reliable labeled samples requires a terrestrial campaign, which is often costly in terms of both time and human resources [31].

Small training set size together with the high dimensionality of spectral data pose the curse of dimensionality problem, which is known to hamper robust statistical estimates of classifiers [93, 95]. In this thesis, we investigate two approaches recently used in the machine learning community to address the small training set size problem namely *transfer learning* and *semi-supervised learning* approaches. To improve the classification performance on a given task, transfer learning approach focuses on how to transfer knowledge learnt from related tasks. Whereas, semi-supervised learning approach concentrates on how to leverage knowledge learnt from a large amount of unlabeled samples belonging to the same task.

In medical imaging applications, for example, in order to reduce the labeling cost, one might develop a diagnostic program that determines the disease area either (i) by reusing the labeled samples assigned by the experts from other patients with similar profiles (transfer learning); or (ii) by utilizing a large amount of unlabeled samples in addition to a few labeled samples done by the experts from the same patient (semi-supervised learning).

1.3 Background

This thesis is concerned with three issues: edge detection, transfer learning, and semi-supervised learning. Background and related works to edge detection in spectral imaging is presented in Chapter 2. Below we provide a brief introduction to transfer learning and semi-supervised learning.

1.3.1 Transfer learning

“Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learnt" [114, Chapter 11]. Here, tasks mean the classification tasks need to be solved, e.g. classifying the ten digits using a certain type of data. Traditional machine learning algorithms address isolated tasks in which training and testing data come from the same domain. (A domain refers to the data feature space and the data marginal probability distribution). This limits the adaptability of the methods when encountering data from a related domain or a related

Table 1.1: Relationship between traditional machine learning and different transfer learning settings (Table taken from [87]).

Learning Settings		Source and Target Domains	Source and Target Tasks
Traditional Machine Learning		the same	the same
Transfer Learning	Inductive Transfer Learning / Unsupervised Transfer Learning	the same	different but related
		different but related	different but related
	Transductive Transfer Learning	different but related	the same

task for which the underlying distribution typically deviate from the one encountered in the original setting. Transfer learning attempts to improve the performance within a new task by transferring knowledge learnt from one or more related source tasks (or domains) to the new target task (or domain) [114, Chapter 11],[87]. Transfer learning is categorized into three settings [87]: inductive transfer learning, unsupervised transfer learning, and transductive transfer learning. The three settings and their relation with the traditional machine learning are summarized in Table 1.1.

Traditional machine learning requires both (i) source and target tasks are the same and (ii) source and target domains are the same. The first two transfer learning settings, inductive transfer and unsupervised transfer, relax the second requirement by allowing that source and target tasks may be not the same but related to each other. These settings are useful for cases in which we want to transfer *knowledge*, e.g. data representation learnt from a digit classification task (source task) to an alphabetical character classification task (target task). The main difference between these two settings is the availability of the labeled samples in the target domains. Inductive transfer learning assumes that there are a few labeled samples in the target domain; whereas, unsupervised transfer learning assumes that the target domain does not contain any labeled sample.

Transductive transfer learning requires the same learning task in the source and target domains. The source and target domain distributions might be different yet related to each other. In addition, this setting assumes that there is a rich source of labeled samples in the source domain but no labeled sample available in the target domain. Transductive transfer learning is the main focus in this thesis and is, thereafter, referred to as transfer learning.

A typical situation leading to the difference in distribution between source and target domains is domain shift. Domain shift is characterized by the fact that the measurement system, or the method of description, can change. Such a situation arises in a variety of applications, such as in computer vision [26, 41], remote sensing [54, 101], and multivariate time series [117, 125]. For example, in computer vision, the image of an object captured by a digital camera might differ significantly from the image captured by a webcam. In remote sensing, spectra of objects from the same class collected at different times and/or locations can also be different due to environmental changes or changes in object spectra themselves with respect to both spatial and temporal domains.

A related term for domain shift, which is also widely used in literature, is domain adaptation. Both of them rely on the existence of a "good domain embedding" [97, Chapter 5], i.e., there exists a new feature space transformed from the original feature space, under which the source and target distributions are unchanged. Denote by $P_S(X)$ and $P_T(X)$ the distributions of data X in the source and target domains, respectively, and by Y the corresponding labeling of X. For a classification problem, the assumption of the existence of a good domain embedding means there exists a mapping W under which $P_S(Y|W(X)) = P_T(Y|(W(X)))$, although $P_S(X)$ might be different from $P_T(X)$. Many methods have been proposed to learn such a transformed feature space, such as by metric learning [108] or feature extraction [86]. For more details about these methods, readers are referred to Chapter 4 of this thesis.

There are three main issues in transfer learning [87]: (i) What to transfer?; (ii) How to transfer?; and (iii) When to transfer? The first two questions consider two issues: what the common knowledge among domains is and how to transfer that knowledge from one domain to another domain, respectively. The third question investigates in which situations the transfer process should (or should not) be done. The effectiveness of any transfer method depends on the relatedness between to the source and target domains. If the source domain is not sufficiently related to the target domain, the system might fail to improve its performance [87, 114]. In the worst case, this can even lead to negative transfer [106]. For a more comprehensive review readers are referred to [87, 97, 114].

1.3.2 Semi-supervised learning

Semi-supervised learning is concerned with the improvement of classification performance through the use of a large amount of unlabeled data in addition to the available labeled samples. Unlabeled samples are often much cheaper and easier to obtain than labeled samples. Semi-supervised learning, which yields high accuracy even in cases of small sample size is, therefore, of interest in both theory and practice [135].

Although transfer learning and semi-supervised learning are both proposed to deal with the small sample size problem, they do differ in the type of resource they use to improve the classification performance. While semi-supervised learning utilizes unlabeled samples from the same classification task and having the same distribution, transfer learning relies on related domains and samples do not necessarily follow the same distribution within same tasks.

The unlabeled samples provide semi-supervised learning with extra information on the marginal density of the classification task and many techniques have been proposed to leverage such information. These techniques can be classified into four categories: generative models, graph-based methods, low density separation, and change of representation [10, Chapter 1]. The four categories differ in the additional assumption they made on the data. For example, low density separation techniques assume that the decision boundary should lie in a low-density region. A typical example of these techniques is the Transductive SVM (TSVM) [53]. The goal of TSVM is to find a labeling of the unlabeled data such that a linear boundary has the maximum margin overall data. Intuitively, it makes use of unlabeled data to guide the linear boundary away from dense regions [135].

Differently, change of representation techniques are based on a smoothness assumption, i.e., if two points in a high density region are close, their corresponding outputs should be close, too. These techniques follow two learning steps: (1) perform an unsupervised learning algorithm on the whole data, i.e., including labeled and unlabeled samples, to construct a new data representation; and (2) perform a supervised learning algorithm on the newly constructed data presentation. The change in data representation in step (1) is made in a way that small distances in high-density regions are preserved [10, Chapter 1].

The additional assumptions made in semi-supervised learning are essential because a bad matching between the problem structure and model assumption can lead to degradation in classification performance [135]. Consequently, like in transfer learning, negative effects with respect to the classification performance also happen in semi-supervised learning, which has been observed in several studies, e.g. in [68, 69], [10, Chapter 4]. For a more comprehensive review of semi-supervised learning techniques and the assumptions they make, readers are referred to, e.g. [10, 135].

1.4 Thesis's contributions and outline

This thesis contributes to the field of spectral imaging by studying the visualization (by using edge detection) and classification (by studying transfer and semi-supervised learning) of spectral imaging data. Here we provide a more detailed overview of the contributions in these two directions.

Edge detection in spectral images. Detecting edges in spectral images is difficult as spectra may differ in just a few bands. Existing approaches calculate the edge strength of a pixel locally, based on the variation in intensity between this pixel and its neighbors. They often fail to detect the edges of objects embedded in background clutter, or objects which appear in only some of the bands. We propose a method that aims to overcome this problem by considering the salient properties of edges in an image.

Based on the observation that edges are rare events in the image, we recast the problem of edge detection into the problem of detecting events that have a small probability in a newly defined feature space constructed by the spatial gradient magnitude in all spectral channels. As edges are often confined to small, isolated clusters in this feature space, the edge strength of a pixel, or the confidence value that this pixel is an event with a small probability, can be calculated based on the size of the cluster to which it belongs.

Based on the edge strength map, the final binary edge map can be then generated by applying a thresholding algorithm. Experimental results on a number of multispectral data sets and a comparison with other methods demonstrate the robustness of the proposed method in detecting objects embedded in background clutter or appearing only in a few bands.

This work is presented in Chapter 2 and has been published as [22] and [21]:

Cuong V. Dinh, Raimund Leitner, Pavel Paclik, Marco Loog, and Robert P. W. Duin. SEDMI: Saliency based edge detection in multispectral images, *Image and Vision Computing*, 29(8): 546-556, 2011.

Cuong V. Dinh, Raimund Leitner, Pavel Paclik, and Robert. P. W. Duin. A Clustering Based Method for Edge Detection in Hyperspectral Images, *the 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, 580-587, 2009.

Training set selection for learning from multiple source domains. As mentioned in Section 1.3, the effectiveness of any transfer method depends on the relatedness between the source and target domains. Thus, it is not always wise to use all labeled samples that come from a set of source domains in the training process as irrelevant source domains included in the training data might deteriorate the classification performance.

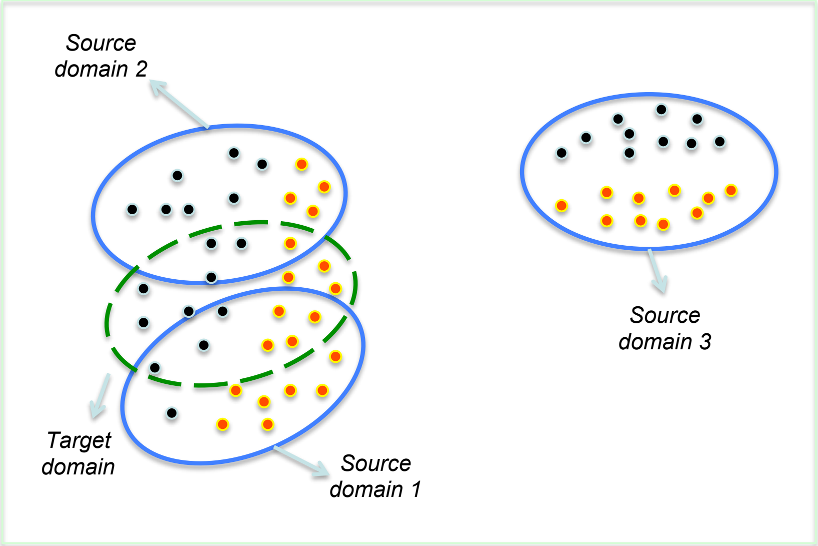


Figure 1.3: Domain selection for transfer learning

Figure 1.3 illustrates an example in which we want to classify objects in a target domain marked by the dashed, green ellipse using knowledge learnt from three source domains marked by the solid blue ellipses. Black and red points represent samples from the two classes of the same classification task. The figure shows that the first two source domains are relevant and provide similar information in terms of the discriminant between classes with respect to the target domain. The third source domain is not related to the target domain. Including this domain in the training set would hamper the classification performance.

We propose a method to select suitable source domains given a target domain based on a similarity measurement between data domains. We evaluated our method on spectral endoscopy images, a relatively novel imaging technique that could be potentially used for early stage cancer detection. The data under consideration include different types of cancer, which poses a challenge for the detection as different cancer types often exhibit different spectral signatures. Our results on this data set demonstrate that the classification is significantly improved when a few source domains that are presumably similar to a given target domain are selected for training instead of using all available source domains.

This work is presented in Chapter 3 and has been published as [23] and [64]:

Cuong V. Dinh, Marco Loog, Raimund Leitner, Olga Rajadell, Robert P.W. Duin. Training Data Selection for Cancer Detection in Multispectral Endoscopy Images, *the 21st International Conference on Pattern Recognition (ICPR)*, Tokyo, Japan, 2012.

Raimund Leitner, Martin De Biasio, Thomas Arnold, Cuong V. Dinh, Marco Loog, Robert P. W. Duin. Multi-spectral video endoscopy system for the detection of cancerous tissue, *Pattern Recognition Letter*, 2012.

Feature extraction method for domain shift problem. How to handle domain shift, which often happens in spectral images, is a major concern in transfer learning. We propose FIDOS, a generalization of the well-known Fisher feature extraction method, that aims at finding a transformation of the original feature space such that source and target domains are matched. The proposed method maximizes the between-class scatter and at the same time minimizes a convex combination of the within-class and between-domain scatters. To this end, FIDOS constructs a subspace that reduces the drift in the distributions across different domains whilst preserving the discriminants among classes. Our results on both artificial and real world data confirmed that learning invariant features with respect to the domains is essential to deal with domain shift problems.

This work is presented in Chapter 4 and has been published as [20]:

Cuong V. Dinh, Robert P. W. Duin, Ignacio Piqueras-Salazar, and Marco Loog. FIDOS: A generalized Fisher based feature extraction method for domain shift, *Pattern Recognition*, 46(9): 2510–2518, 2013.

It should be noted that the two methods proposed in Chapter 3 and 4 both relate to transfer learning. However, they address different questions in transfer learning:

“When to transfer?" and “What and how to transfer?", respectively. They can also be used as two separated steps of a transfer learning system. As an example, consider the remote sensing application where we want to do a classification task on a target domain using data from the source domains, which were collected at different times in the past. Among the source domains, some of them have been collected recently while others were collected longer ago. As mentioned earlier, two types of shift might happen in this scenario: (i) shift due to environmental change, e.g. difference in lighting condition, even when domains are collected at a similar time in a year; and (ii) the spectral signature of objects themselves change after a long period. For the latter shift, it is not possible to transfer knowledge between domains. Thus, the first step of a possible classification system would be selecting relevant domains by removing non-relevant ones. Then, in the second step, feature extraction is applied to construct a subspace in which all the relevant domains are aligned to remove the effect of environmental change.

Semi-supervised dissimilarity representation. In dissimilarity representation [30], objects are represented by their dissimilarities with respect to a representation set, rather than by features. It is based on the idea that a class is constituted by objects having similar characteristics. The dissimilarity is small between objects of the same class and large between objects from different classes. Therefore, dissimilarities can be used as discriminant features for classification. The key advantage of the dissimilarity representation approach is that it bridges the gap between structural and statistical approaches [30]. For example, in spectral object classification problems, this makes it possible to embed knowledge of structural information about the data, such as the spectral shape information [85], into powerful feature-based statistical approaches.

Up to now, in dissimilarity representation, the representation or prototype set has been usually selected from the training data. For small training set situations, the representation set selected from such limited labeled data might miss important prototypes. This limits the different aspects that can be captured in the data and might result in poor performance. Based on the fact that it is not necessary to know the labels of samples used in the representation set, we investigate the performance change if the representation set is extended by also including test data in a semi-supervised manner. Our semi-supervised method for the dissimilarity representation can be classified into the “Change of Representation" category in semi-supervised learning as it aims at enhancing the data representation by making use of unlabeled data.

This work is presented in Section 5.1 and has been published as [19]:

Cuong V. Dinh, Robert P.W. Duin, Marco Loog. A study on semi-supervised dissimilarity representation, *the 21st International Conference on Pattern Recognition (ICPR)*, Tokyo, Japan, 2012.

Training sample selection for semi-supervised learning. Until now, most studies in semi-supervised learning for remote sensing image classification focus on optimizing the classification performance given a training set generated by randomly selecting samples from the sample distribution. This random selection strategy may be inefficient for problems containing unbalanced classes as it tends to select samples belonging to classes that are dominant in the sample distribution. We propose a new strategy to select training samples that are representative for the problem needed to be solved. We select training samples as the cluster center points resulted from a clustering of all available samples in a feature space constructed by both spectral and spatial features. Experiments on a remote sensing data set show that we can achieve state-of-the-art results using much less number of training samples.

This work is presented in Section 5.2 and has been published as [98]:

Olga Rajadell-Rojas, P. Garcia-Sevilla, C. V. Dinh, and Robert P. W. Duin. Semi-supervised hyperspectral pixel classification using interactive labeling, *3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers)*, Lisbon, Portugal, 2011.

2

SEDMI: SALIENCY-BASED EDGE DETECTION IN MULTISPECTRAL IMAGES

Detecting edges in multispectral images is difficult because different spectral bands may contain different edges. Existing approaches calculate the edge strength of a pixel locally, based on the variation in intensity between this pixel and its neighbors. Thus, they often fail to detect the edges of objects embedded in background clutter or objects which appear in only some of the bands.

We propose SEDMI, a method that aims to overcome this problem by considering the salient properties of edges in an image. Based on the observation that edges are rare events in the image, we recast the problem of edge detection into the problem of detecting events that have a small probability in a newly defined feature space. The feature space is constructed by the spatial gradient magnitude in all spectral channels. As edges are often confined to small, isolated clusters in this feature space, the edge strength of a pixel, or the confidence value that this pixel is an event with a small probability, can be calculated based on the size of the cluster to which it belongs.

Experimental results on a number of multispectral data sets and a comparison with other methods demonstrate the robustness of the proposed method in detecting objects embedded in background clutter or appearing only in a few bands.

Published as Cuong V. Dinh, Raimund Leitner, Pavel Paclik, Marco Loog, and Robert P. W. Duin. SEDMI: Saliency based edge detection in multispectral images, *Image and Vision Computing*, 29(8): 546-556, 2011.

2.1 Introduction

Edge detection for gray-scale images has been thoroughly studied and is well established. However, for color images and especially for multispectral images, this topic is still in its infancy and even defining edges for these images is a challenge [58]. There are two main approaches to detect edges in multi-channel images based on either monochromatic [50, 103] or vector techniques [35, 121, 133]. The monochromatic approaches apply a gray-scale edge detection to each band and then combine the results over all the bands. Several combination rules have been used, e.g. the summation rule [50], the maximum rule [55], and the OR operation [36]. A more sophisticated combination technique is to fuse the individual responses using different weights [7].

Vector-based approaches consider each pixel in a multispectral image as a vector in the spectral domain, then perform edge detection in this domain. These approaches can be further divided into two categories: multidimensional gradient [11, 15, 133] and vector order statistic [35, 120, 121]. The multidimensional gradient approach extends the gray-scale definition of gradient magnitude and direction to multispectral images. Di Zenzo [133] defines the gradient direction at a pixel as the direction in which its vector in the spectral domain has the maximum rate of change. Hence, an eigenvalue decomposition is applied to the set of partial derivatives at a pixel to determine the largest eigenvalue and its corresponding eigenvector. The largest eigenvalue is then considered as the edge magnitude and the eigenvector as the edge direction of this pixel. The disadvantage of this method is its sensitivity to texture because the gradient-based operators are sensitive to small change in intensity.

The vector order statistic approach follows the use of morphological operators for edge detection in gray-scale images [47], which calculates gradients as the difference between a dilation and an erosion. Trahanias et al. [121] order the pixels within a small window by the aggregate distances of each pixel to the others. Then, the edge strength of the pixel located at the center of the window is calculated as the deviation between the vector with the highest rank and the median vector. Evans and Liu [35] improve this method by defining the edge strength of a pixel as the maximum distance between any two pixels in its surrounding window. This helps to localize edge locations more precisely.

In the approaches discussed above, the edge strength of each pixel is computed locally based on the variations in the intensities of the pixels within a small, surrounding window. Consequently, besides extracting meaningful and useful edges, these approaches also extract many other spurious edges that arise from noise and background clutter [89, 107]. For gray scale images, a common method to overcome this problem is based on the salient characteristic of edges in images [105, 107]. This stems from visual attention theory that structurally salient features such as edges, blobs, and circles are pre-attentively distinctive. They attract our attention without the need to scan the entire image in a systematic manner [110]. The saliency of an edge can be defined as its stability of occurrence over scales [77] or the maximum over scales of normalized derivatives [66]. Saliency, according to information theory, is also related to the frequency or the probability of occurrence, i.e. events that occur rarely are more informative [73, 123].

Motivated from these approaches, we recast the edge detection problem in multispectral images into detecting events that occur with a small probability in a newly defined feature space. The feature space is constructed by spatial gradient magnitudes of all pixels over all spectral bands (thereafter referred to as gradient magnitude feature space). We then introduce a saliency-based edge detection in multispectral images (SEDMI) to detect such events.

The prominent characteristic of the gradient magnitude feature space is that edge pixels often fall in small, isolated clusters. The saliency (or the edge strength) of a pixel is then defined as the confidence value that this pixel belongs to a small cluster and subsequently, can be calculated based on the size of the cluster containing the pixel. As the constructed gradient magnitude feature space utilizes the global, structural image information, SEDMI recovers edges of objects surrounded by background clutter or objects appearing in a few bands of a multispectral image.

The rest of this chapter is organized as follows. Section 2.2 provides additional motivation for SEDMI and discusses related work. Section 2.3 presents the SEDMI method. To demonstrate the effectiveness of our method, experimental results and a comparison with other methods are presented in Section 2.4. Sections 2.5 and 2.6 discuss related issues and draw conclusions.

2.2 Motivation and related work

2.2.1 Edge detection as detecting salient features

Salient features are image features assumed to be able to capture the most prominent structures in an image [73]. They may provide crucial clues for image analyses such as image matching and object detection. Salient features are often defined as the local extrema of some functions in the image. Thus, corners, junctions, blobs, and edges (local maxima of gradient magnitudes) can be considered as salient features [92].

According to information theory, saliency is related to the frequency of appearance: events that occur rarely are more informative [73, 123]. Thus, salient features correspond to the events with small probabilities in a feature space defined by, for example, differential invariant features of the pixels over a range of scales [126]. Salient features can then be detected by applying a novelty detection technique to the constructed feature space [67]. Inspired by this approach, we recast the edge detection problem in multispectral images into detecting events with small probability (thereafter referred to as small probability events) in the feature space composed of the gradient magnitudes of the pixels in all channels.

The main assumption made in our method is that edges in a multispectral image are rare events. This assumption is reasonable because the frequency of occurrence of edges in an image is typically small ($O(m)$ in an $m \times m$ image). In addition, spectral differences on edges between objects are often systematic. This yields a similarity in the gradient magnitudes between these edge pixels. Therefore, they form a small, isolated cluster in the gradient magnitude feature space.

2.2.2 Towards clustering-based edge detection

As discussed earlier, the prominent characteristic of the gradient magnitude feature space is that edge pixels often fall in small, isolated clusters. Therefore, the cluster-based novelty detection approach, which is based on the size of the cluster, is suitable for detecting edge pixels in the gradient magnitude feature space [33, 49]. The smaller the cluster size corresponding to a pixel, the more likely this pixel is a small probability event. The cluster size of a pixel p can be defined as either the number of pixels in the cluster containing it [49] or the number of pixels within a hyper-sphere centered at p with radius w. w is determined by learning from a training set [33]. In our method, we use the former definition.

It should be noted that clustering methods often require prior knowledge about the data, such as the number of clusters and cluster shapes. For edge detection, however, such *a prior* knowledge is typically unavailable. To overcome this obstacle, we use ensemble clustering that is well known for its stability and robustness without any prior knowledge [39, 115].

2.2.3 Related work on ensemble clustering

The main aim of data clustering is to partition an unlabeled data set into homogenous regions. However, it is an ill-posed problem due to the lack of prior information about the underlying data distribution [39, 115]. By utilizing the fact that different clusterings (difference in algorithms or in the setting of each algorithm) applied to the same data set are able to capture different structures in the data, ensemble clustering has been shown to be a powerful method for improving the cluster result in terms of both robustness and stability.

In [115], a set of clustering results is transformed into a hyper-graph representation. In the hyper-graph, each vertex corresponds to a point in the data set and each hyper-edge, which can connect any set of vertices, represents a cluster in a clustering. Based on this representation, different consensus functions, e.g. Cluster-based Similarity Partitioning

Algorithm (CSPA), HyperGraph Partitioning Algorithm (HGPA), and Meta-CLustering Algorithm (MCLA), can be used to produce the final clustering result.

In [39], an evidence accumulation clustering algorithm is proposed. In the algorithm, the results of multiple clusterings are summarized into a Co-Association (CA) matrix, in which each element is the number of times a pair of points is assigned to the same cluster. Subsequently, the final clustering can be computed by applying a hierarchical clustering to the CA matrix. In fact, the CA matrix can be considered as a similarity measurement between points. The more frequently two points are in the same cluster, the more similar they are.

It should be noted that we use ensemble clustering in our method to estimate the cluster size corresponding to a pixel but not to generate the final clustering as in the above methods. As demonstrated in Section 2.3.4, the estimated cluster size of a pixel is equal to the sum of the co-association values of this pixel with respect to all pixels in the multispectral images. This provides a strong connection between our method and the evidence accumulation clustering method.

2.3 SEDMI method

2.3.1 Constructing the feature space

For each channel of an $n-$channel multispectral image, we compute its gradient magnitude using a Gaussian derivative [6]. Each pixel is then represented by an $n-$component vector composed of the gradient magnitudes of this pixel over all channels. Thus, the gradient magnitude feature space contains M such vectors, where M is the number of pixels in the image.

2.3.2 Performing ensemble clustering

We perform ensemble clustering in the gradient magnitude feature space to estimate the cluster size for the pixels in the image. One important requirement in ensemble clustering is the diversity in the clustering results. This requirement is needed to ensure that different clusterings preserve different structures in the image and do not yield identical data partitions. Therefore, we use a simple k-means as the base clustering algorithm. At each clustering, we randomly choose the number of clusters and the initial cluster centers.

After each clustering, we calculate for each pixel the size of the cluster containing it. The estimated (expected) cluster size of a pixel p_i, denoted as $EC(p_i)$, is then calculated as the mean (average) of the size of the clusters containing p_i generated by N clusterings:

$$EC(p_i) = \frac{\sum_{t=1}^{N} C_{i,t}}{N} \tag{2.1}$$

where $C_{i,t}$ is the size of the cluster containing pixel p_i at clustering t.

2.3.3 Calculating edge strength map

We calculate the edge strength of a pixel based on its cluster size estimated by the ensemble clustering. A pixel is an edge pixel or an event with small probability if it belongs to small clusters. Therefore, the smaller the expected cluster size of a pixel, the more probable this pixel is a small probability event. Thus, the confidence value that a pixel p_i is a small probability event, or the edge strength of p_i, denoted as $ES(p_i)$, can be calculated as follows:

$$ES(p_i) = 1 - \frac{EC(p_i)}{M} \tag{2.2}$$

It should be noted that an image with high spatial resolution may cause a high computational cost because of the ensemble clustering procedure. In this case, we may reduce the computational cost by (i) randomly selecting a subset of pixels, (ii) performing the ensemble clustering on this subset to compute the edge strength for the pixels in this subset, and (iii) using a regression algorithm, e.g. knn-regression [28], to estimate the edge strength for the remaining pixels in the image.

2.3.4 Connection with the evidence accumulation clustering

Our algorithm to compute the cluster size for a pixel is strongly connected with the evidence accumulation clustering. We will show that the estimated cluster size of a pixel in our algorithm is equal to the sum of the co-association values between this pixel and all the pixels including itself. The following deduction demonstrates this claim. Denote $a_{ij,t}$ the association value between pixels p_i and pixel p_j at clustering t. $a_{ij,t}$ equals 1 if p_i and p_j are in the same cluster and 0 otherwise. Note that $a_{ii,t} = 1$. From (1), the estimated cluster size of p_i is:

$$EC(p_i) = \frac{\sum_{t=1}^{N} C_{i,t}}{N} = \frac{\sum_{t=1}^{N} \sum_{j=1}^{M} a_{ij,t}}{N} = \sum_{j=1}^{M} \frac{\sum_{t=1}^{N} a_{ij,t}}{N} \tag{2.3}$$

Denote $CA(i,j)$ the co-association value between pixels p_i and p_j after N clusterings. $CA(i,j)$ is the number of times the two pixels being assigned to the same cluster normalized by N. Then (3) becomes:

$$EC(p_i) = \sum_{j=1}^{M} CA(i,j) \tag{2.4}$$

Table 2.1: Co-association between a pixel p_i and all the pixels in the feature space.

	p_1	p_2	$\cdots$	p_M	Sum (Cluster size)
Clustering 1	$a_{i1,1}$	$a_{i2,1}$	$\cdots$	$a_{iM,1}$	$C_{i,1}$
Clustering 2	$a_{i1,2}$	$a_{i2,2}$	$\cdots$	$a_{iM,2}$	$C_{i,2}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Clustering N	$a_{i1,N}$	$a_{i2,N}$	$\cdots$	$a_{iM,N}$	$C_{i,N}$
Sum	$N \times CA(i,1)$	$N \times CA(i,2)$	$\cdots$	$N \times CA(i,M)$	$\boldsymbol{N \times EC(p_i)}$

A graphical illustration of our claim is shown in Table 2.1. The sum across a row t ($t = 1 \cdots N$) corresponds to the size of the cluster containing the pixel under consideration (p_i) at clustering t; while the sum across a column j ($j = 1 \cdots M$) is equal to the co-association value between the pixels p_i and p_j times N. It is obvious that the sum across all rows equals to the sum across all columns in a matrix. Thus, (4) is deduced.

On the other hand, the co-association of two pixels represents the similarity, or the inverse distance, between them. From this point of view, the way a pixel is considered as a small probability event in our method is confirmed by the R-ordering in statistics [3]. The greater the distance between a point of interest and all other points in the feature space, the more likely this point is an event that has small probability.

It should be noted that although the estimated cluster size of a pixel can be calculated from the co-association matrix, we do not need to generate the co-association matrix explicitly. Thus, it avoids the problem of quadratic memory required to store the $M \times M$ matrix for large M in the evidence accumulation clustering algorithm.

2.4 Experimental results

We compare the edge detection results between the SEDMI and two other methods: the Di Zenzo method [133] and the Robust Color Morphological Gradient (RCMG) method proposed by Evans and Liu [35]. We select these two methods for comparison as they represent two main approaches for edge detection in multispectral images: multidimensional gradient and vector order statistics, respectively.

For the RCMG method [35], the mask size is set to 5×5 and the number of rejected vector pairs is set to 8 as recommended by the authors. For the SEDMI method, the gradient magnitude for each pixel is computed by a Gaussian derivative with $\sigma = 1$. In the ensemble clustering, the number of clusterings is set to 200. At each clustering, the cluster centers are randomly selected and the number of clusters varies from 3 to 15. We use this configuration for all of the studied multispectral data sets.

Table 2.2: Properties of the four data sets used in experiments

Data sets	No. channels	Spatial Resolution
AI I	20	100 × 100
AI II	20	100 × 100
SEM/EDX	8	128 × 128
Scene	31	820 × 820

Four multispectral data sets are used for the evaluation: two artificial images (AI I and AI II) and two from real-world applications (SEM/EDX and Scene). The properties of these data sets are shown in Table 2.2.

We evaluate edge detection results in term of both quantitative and subjective measurements. For the quantitative measurement, we use the area under the ROC curves (AUC) criteria following [5, 57]. The receiver operating characteristic (ROC) curve [44] is a plot of the true positive edge rate against the false positive edge rate with regards to different thresholds.

For each multispectral data set, we first apply the three methods to generate the corresponding edge strength maps. We then put these edge strength maps into the same thinning process introduced in [65]. In this process, a pixel is only considered as an edge if its edge strength is a local maximum in the horizontal or vertical direction. Finally, we generate the binary edge maps by thresholding the corresponding edge strength maps.

Using the ROC curve, the best threshold is typically determined at the point which yields the minimum sum of false positive and false negative rates [59]. For edge detection problems, however, this threshold often results in many false positive edge pixels because the number of background pixels is normally substantially larger than that of edge pixels (e.g. 9800 background pixels versus 200 edge pixels in the AI I data set). Therefore, we select the threshold that yields the minimum total number of false positive and false negative edge pixels for the artificial data. For the real data, we select the threshold at which a best subjective result is obtained.

2.4.1 Artificial data

2.4.1.1 Objects surrounded by background clutter

Using the AI I data set, we investigate the behavior of the three edge detection methods when the objects in an image are embedded by severe noise or background clutter. We generated a multispectral image composed of 20 channels. We used the same binary image of size 100 × 100 with intensity of 0.7 in the object region and 0.3 in the background region for each channel. The content of the synthetic image without noise is

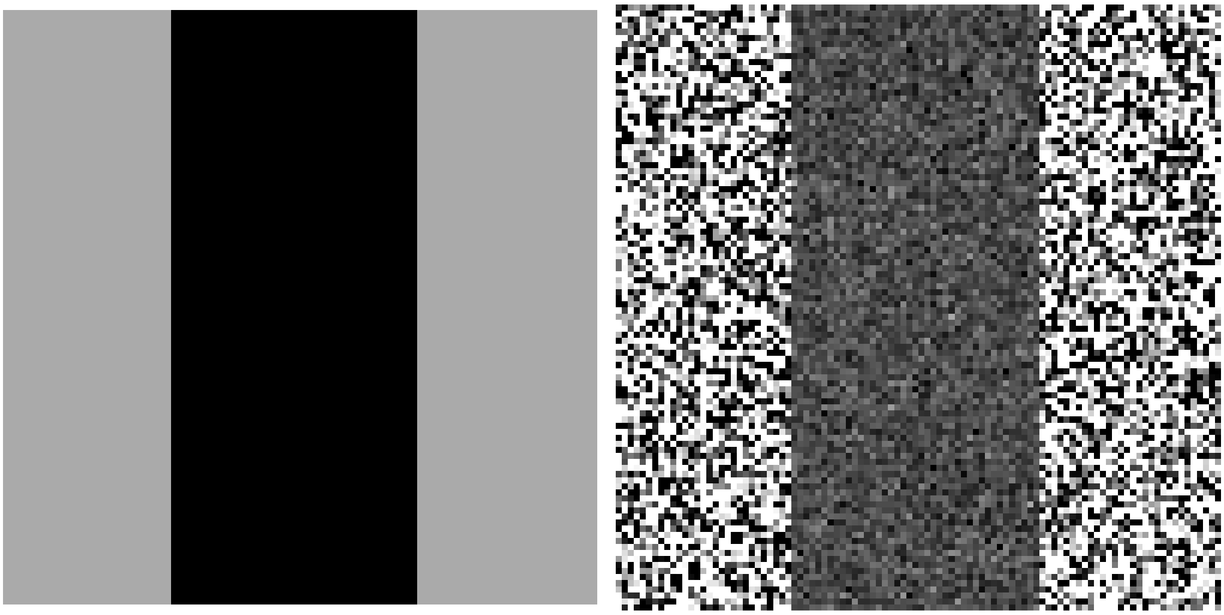

(a) (b)

Figure 2.1: A channel in the AI I data set. (a) The content of the synthetic image without noise (object is located in the middle) and (b) a corrupted image with the SNRs of 16 dB in the object region and 0.2 dB in the background region. Dark color means high intensity.

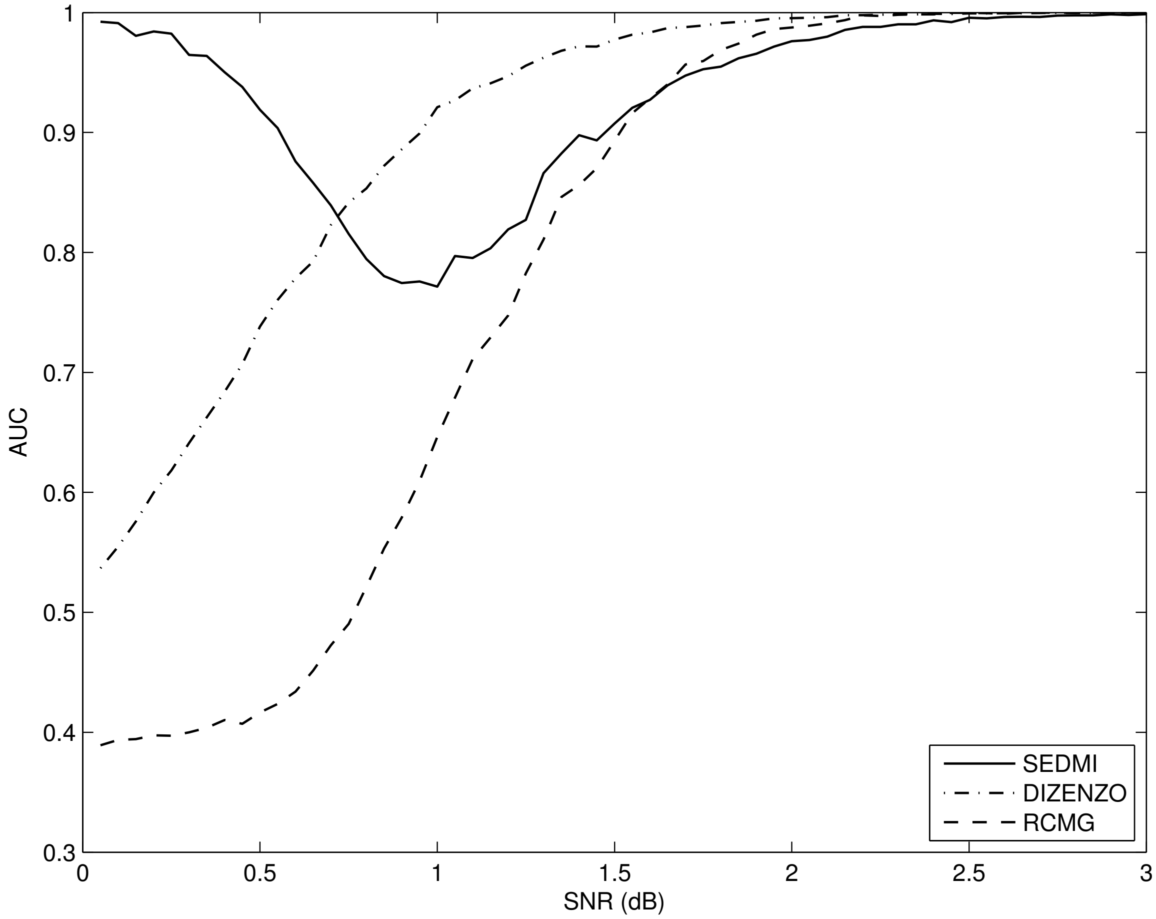


Figure 2.2: AUC curves produced by SEDMI (solid line), Di Zenzo's method (dot dashed line), and the RCMG method (dashed line) for the AI I data set. The horizontal axis shows the SNR with respect to the background noise level.

shown in Figure 2.1a. The object is located in the middle from column 30 to column 70. Thus, edge pixels are located at columns 30 and 70. A fixed, low Gaussian noise level corresponding to a signal to noise ratio (SNR) of 16 dB is added to the object region. The noise level in the background region varies with the corresponding SNRs from 0 to 3 dB. Figure 2.1b shows an example of a channel for a SNR of 0.2 dB with respect to the background noise level.

Figure 2.2 depicts the AUC curves produced by (a) SEDMI (solid line), (b) the Di Zenzo method (dot dashed line), and (c) the RCMG method (dashed line). The horizontal axis shows the SNR with respect to the noise level in the background region. The vertical axis displays the AUC value. SEDMI outperforms Di Zenzo's method and the RCMG method for low SNRs (from 0 to 0.75 dB) or high noise levels. As SNR exceeds 0.75 dB, the Di Zenzo method produces the largest AUC value. SEDMI continues performing better than the RCMG method as SNR grows to 1.65 dB. For SNRs between 1.65 dB and

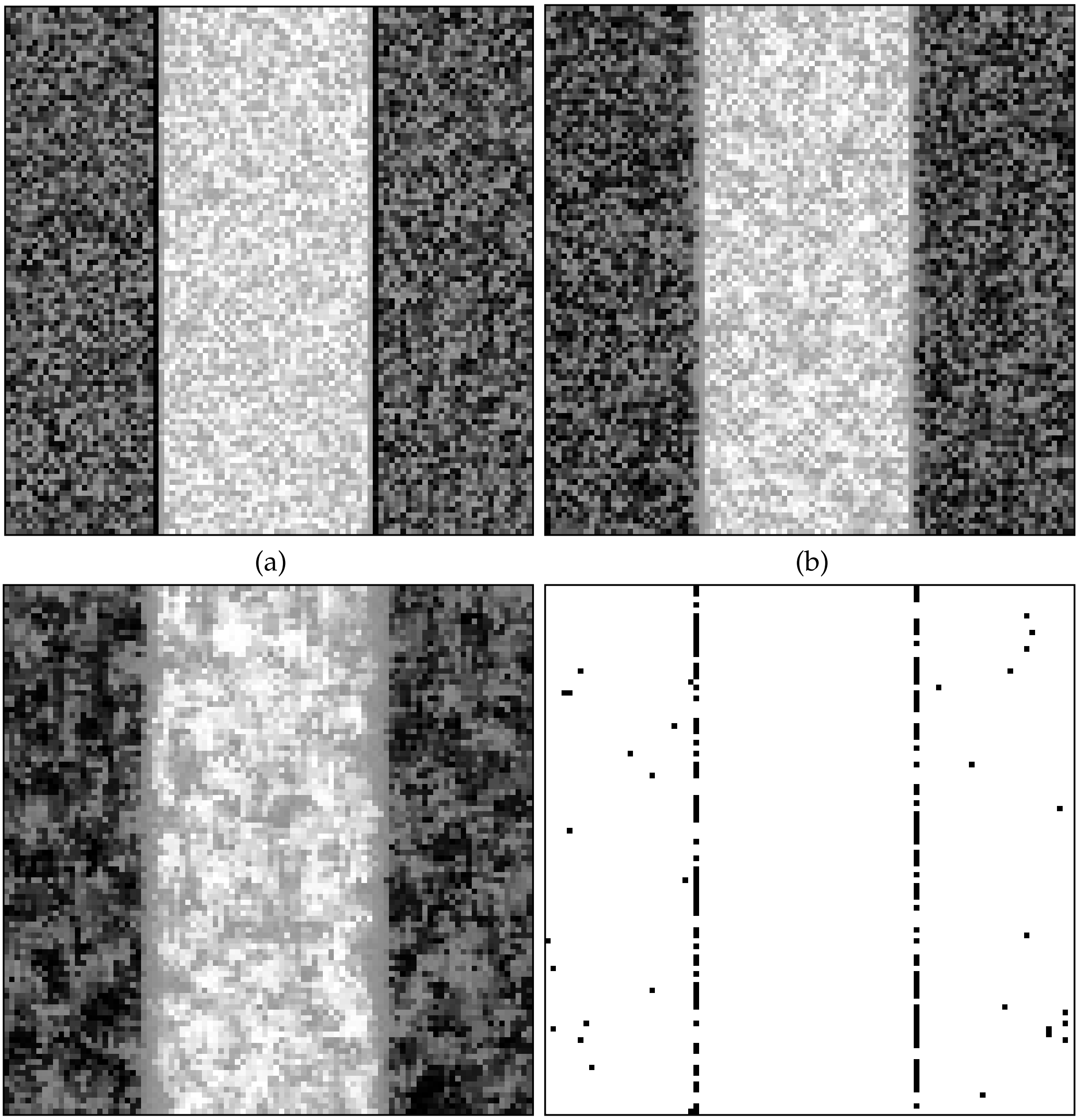
2.5 dB, the other two methods work slightly better than SEDMI.

SEDMI is markedly more robust to severe noise in the background region (background clutter) than the other methods. For high background noise levels (the corresponding SNRs around 0.01 dB), SEDMI yields an AUC value of approximately 1 while the AUC values produced by the Di Zenzo and the RCMG methods are both smaller than 0.6. In this case, the difference in gradient magnitude between the edge pixels is substantially smaller than that between an edge pixel and a pixel in the background region. This leads to the formation of edge pixels as a small, isolated cluster in the global gradient magnitude feature space. Thus, SEDMI detects these edge pixels. Di Zenzo's and the RCMG methods are greatly inferior to SEDMI in dealing with such severe background noise because they do not use the global statistical information in the spatial domain of the image. The edge strength of a pixel is calculated based on a local window. In the background region, a combination between a noisy pixel and its neighbors whose differences in intensities are large leads to a large gradient magnitude for this noisy pixel, even larger than the gradient magnitudes of the true edge pixels. Subsequently, these methods incorrectly determine this noisy pixel as an edge pixel.

The AUC produced by SEDMI decreases to a minimum value of 0.76 as the SNR increases to around 1.0 dB, and then increases again to 1. For small SNRs, only a small number of pixels in the background exhibit similar intensities with those of pixels in the object region. When the SNRs increase, the number of background and object pixels having similar intensities increases too, and thus more background and edge pixels exhibit similar gradient magnitudes. They are then grouped into the same clusters. This makes it difficult for SEDMI to estimate the cluster size for these edge pixels correctly.

The robustness to background clutter of the SEDMI method is further illustrated by the edge strength map and the binary edge map in Figure 2.3. Figures 2.3a-c show the gradient maps generated by (a) SEDMI, (b) the Di Zenzo method, and (c) the RCMG method for a SNR of 0.2 dB. The darker a pixel in the edge strength map, the higher the edge strength in that location. Most of the true edge pixels dominate the highest edge strength values in the edge strength map produced by SEDMI (the corresponding AUC is 0.98). As a result, these edge pixels are correctly selected when thresholding the gradient map. The best binary edge map created by SEDMI using the minimum total number of false positive and false negative edge pixels criterion is depicted in Figure 2.3d.

In contrast, Di Zenzo's and the RCMG methods calculate substantially smaller edge strengths for the true edge pixels than for the noisy pixels in the background region (the corresponding AUCs are 0.59 and 0.39, respectively). These noisy pixels then dominate the binary edge map. Consequently, the best binary edge maps generated according to the above criterion assign all pixels to background. We note that if the threshold is determined by the point in the ROC curve that gives the minimum sum of false positive and false negative rates, then most of the noisy pixels are classified as edge pixels by these two methods.

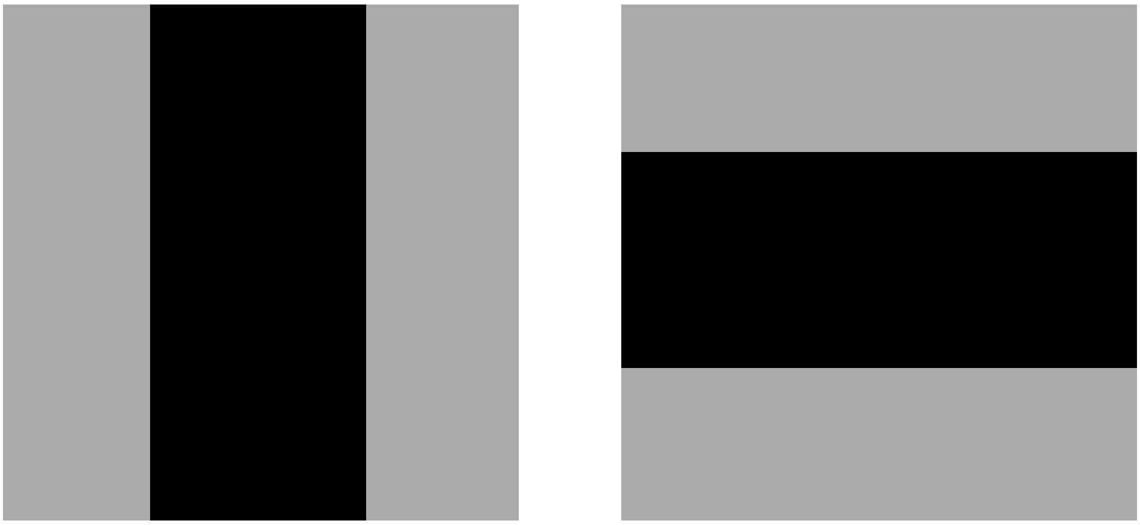


(a) (b) (c) (d)

Figure 2.3: Edge strength maps generated by (a) SEDMI (0.98), (b) Di Zenzo's method (0.59), and (c) the RCMG method (0.39) for the AI I data set with the background noise level corresponding to a SNR of 0.2 dB. Dark color means high edge strength. The corresponding AUC values are shown in brackets. Figure (d) shows the best binary edge map generated by the SEDMI method.

2.4.1.2 Objects occurring in a few bands

The AI II data set contains objects appearing in a few spectral bands. There are two objects of interest a vertical bar and a horizontal bar. The objects have the same intensity values in the images. The vertical bar appears in the first two bands whilst the horizontal bar appears in the remaining eighteen bands. The contents of the synthetic images without noise containing the vertical and the horizontal objects are shown in Figure 2.4a-b. All of the bands in the data set are then corrupted by the independent Gaussian noise. It should be noted that applying a thinning process to this data set will generate offset edges because in the case of binary image corrupted by noise, edge strengths at two sides of the edges differ from each other only due to noise. Therefore, for all methods, we exclude the thinning process from this experiment.



(a) (b)

Figure 2.4: Two representative channels in the AI II data set. (a) a channel with the vertical bar object and (b) a channel with the horizontal bar object.

Figure 2.5 shows the AUC curves produced by the three methods with respect to various levels of Gaussian noise. For SNR lower than 6.0 dB, the RCMG method outperforms both the Di Zenzo method and SEDMI. As SNR exceeds 6.0 dB, SEDMI performs better than the other two methods.

Figure 2.6 shows the best binary edge maps generated by (a) SEDMI, (b) the Di Zenzo method, and (c) the RCMG for the SNR of 16 dB. The corresponding AUCs are 0.998, 0.972 and 0.993, respectively. All three methods detect the horizontal bar well as it appears in most of the bands (18/20). The Di Zenzo method detects many noisy pixels close to the horizontal bar while the vertical bar exhibits discontinuous edges. Compared with SEDMI, the RCMG method misses more edge pixels for the vertical bar as reflected by a slightly lower AUC value.

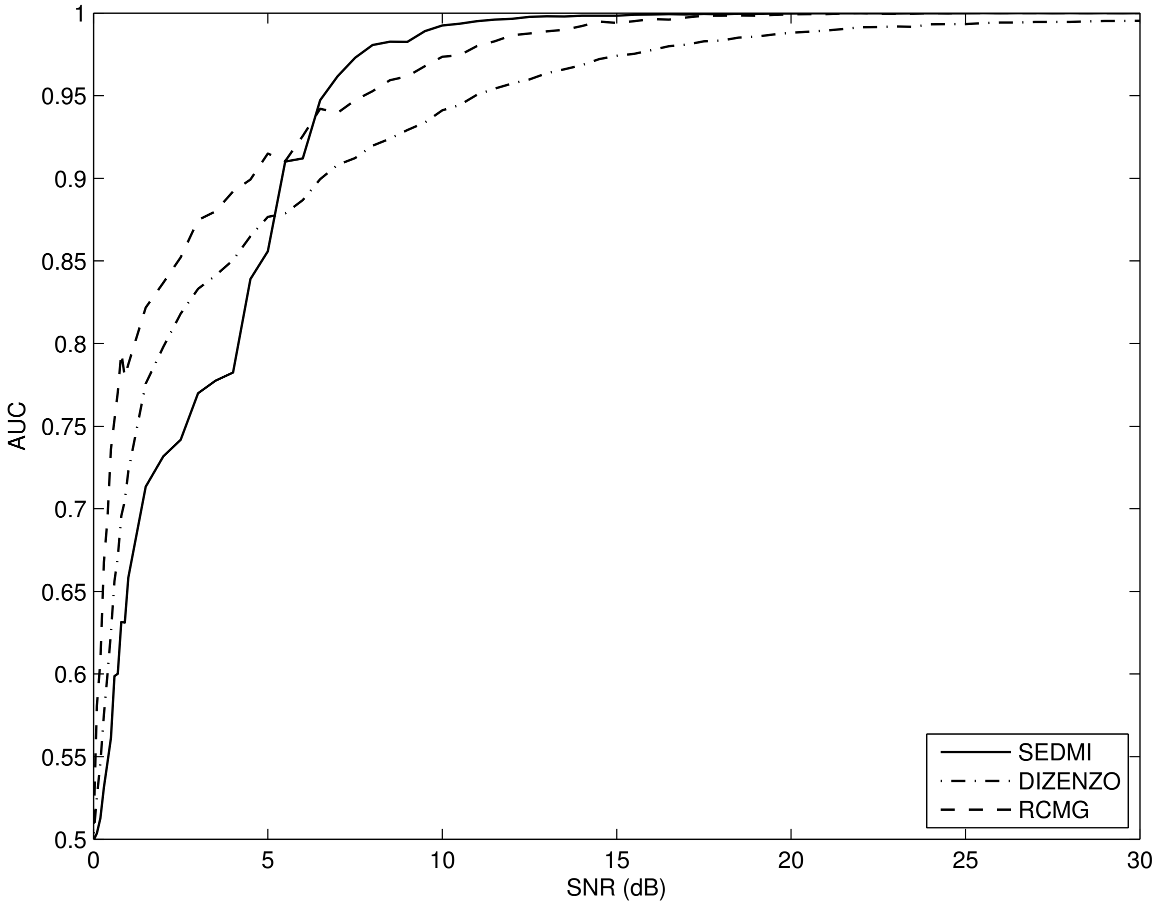


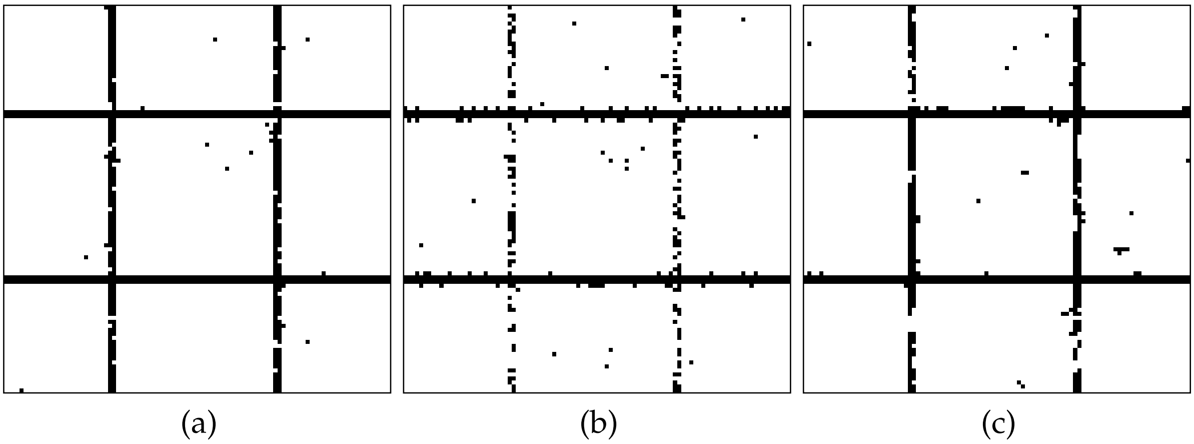
Figure 2.5: AUC curves for the AI II data set generated by SEDMI (solid line), Di Zenzo's method (dot dashed line), and the RCMG method (dashed line). The horizontal axis shows the SNR with respect to the noise level added to the data set.

2.4.2 Real-world data sets

2.4.2.1 SEM/EDX data set

This data set is a collection of scans of detergent powder obtained from a scanning electron microscopy using energy-dispersive X-ray microanalysis (SEM/EDX). The data consists of eight 128×128 images that correspond to particular chemical substances [84]. The data set is noisy in both spatial and spectral domains. Four representative channels are shown in Figures 2.7a-d. The crucial task is to reveal the spatial arrangement of three clusters: the solid, the active, and the porous regions of the detergent powder.

Figures 2.8a-c show the edge strength maps generated for this data set by the evaluated methods. SEDMI exhibits a high contrast between the edge and the background/noisy



(a) (b) (c)

Figure 2.6: The best binary edge maps generated for the AI II data set with the SNR of 16 dB by (a) the SEDMI method (0.998), (b) Di Zenzo's method (0.972), (c) the RCMG method (0.993). The corresponding AUC values are shown in brackets.

pixels. Thus, the method distinguishes edge pixels from noise pixels in the image.

Figures 2.9a-c show the best subjective binary edge results generated by (a) the SEDMI method, (b) the Di Zenzo method, and (c) the RCMG method (binary edge maps based on various thresholds are provided in the Appendix A, Figure S1). The figures demonstrate that the SEDMI method is less affected by noise than the other two methods. SEDMI detects edges along the boundaries between the active and the porosity (particularly in the lower part of the image) whilst the other methods suffer heavily from the noise and fail.

In terms of continuity, edges generated by the RCMG method are more continuous than those generated by the SEDMI and the Di Zenzo methods, e.g. the vertical line on the left side of the image. It is because in the RCMG method, neighbor pixels tend to have similar gradient magnitude values. On the other hand, however, this similarity may result in spurious edges in the noisy region, e.g. the region under the upper curve in the image.

2.4.2.2 Scene data set

Foster's group created a database containing 30 hyperspectral images of natural scenes [83]. Eight representative scenes are available from [38]. We select the fifth scene for our experiment because of two reasons. Firstly, it contains many man-made objects. Therefore, we know exactly their boundary, i.e. we know where edges should be. Secondly, these objects are surrounded by a heavily textured wall. Figure 2.10 shows (a) a gray scale image (channel 28) and (b) the reconstructed color image of the data. Channels 28, 14, and 4 are used, respectively, as the red, green, and blue channels for the reconstruction.

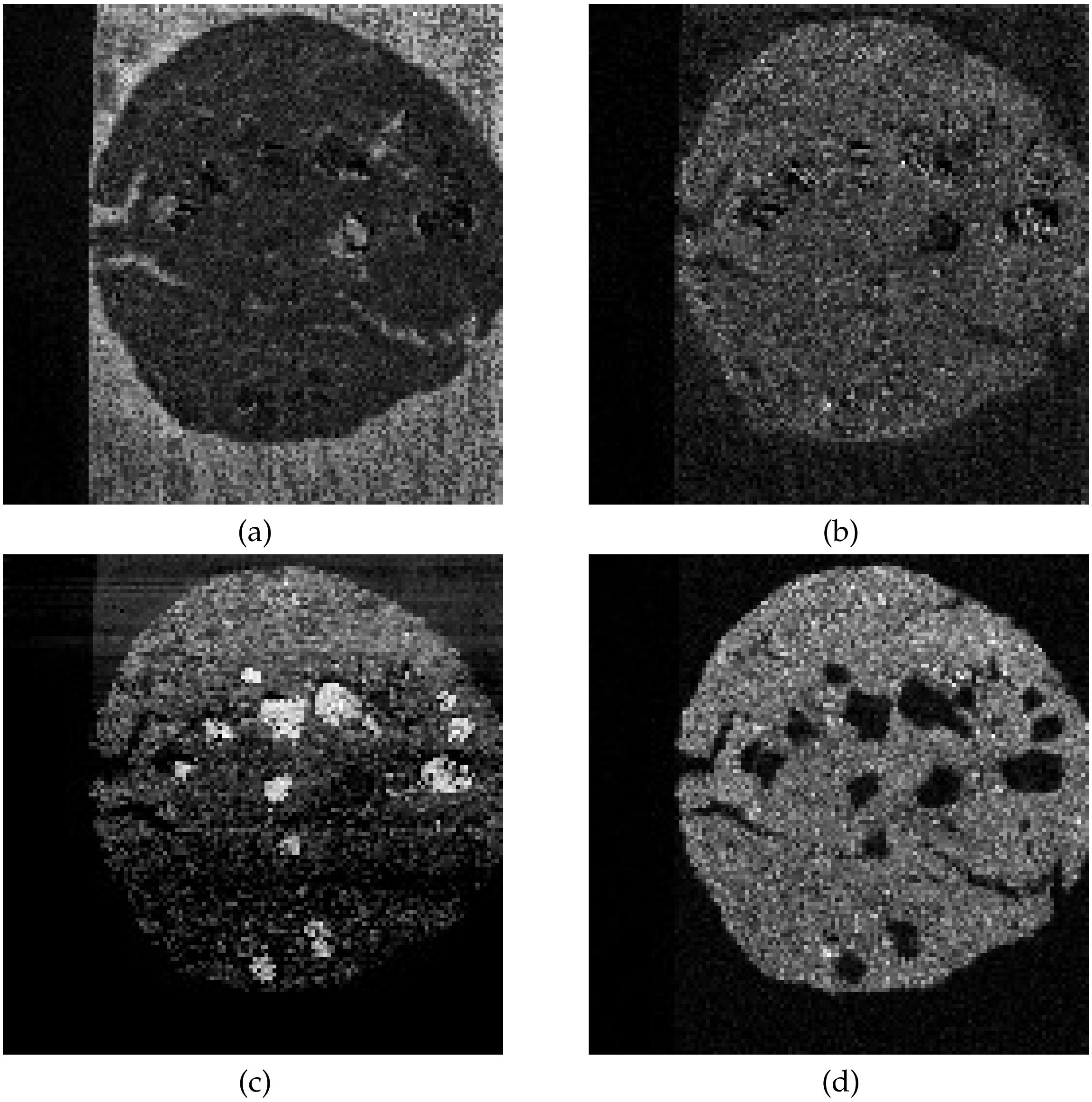
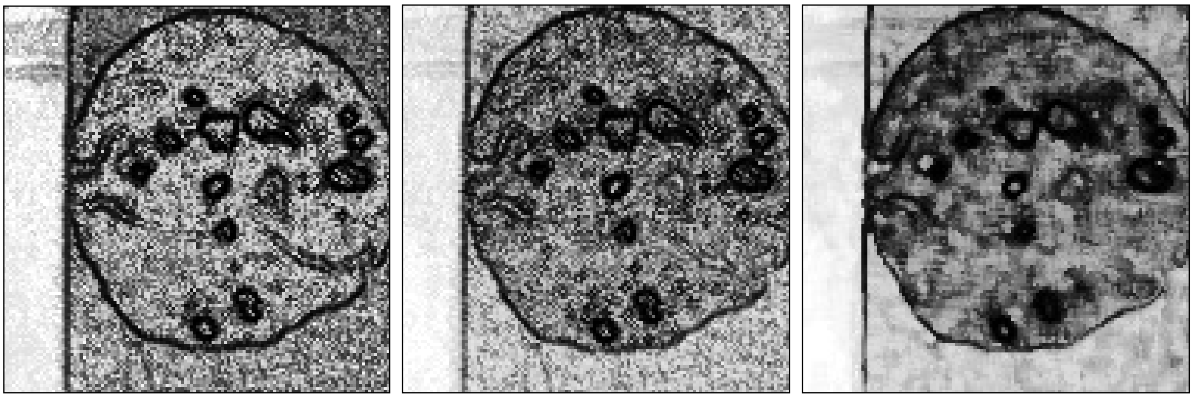

(a) (b) (c) (d)

Figure 2.7: Four channels in the SEM/EDX data set. (a) The second channel, (b) the fourth channel, (c) the sixth channel, and (d) the eighth channel.

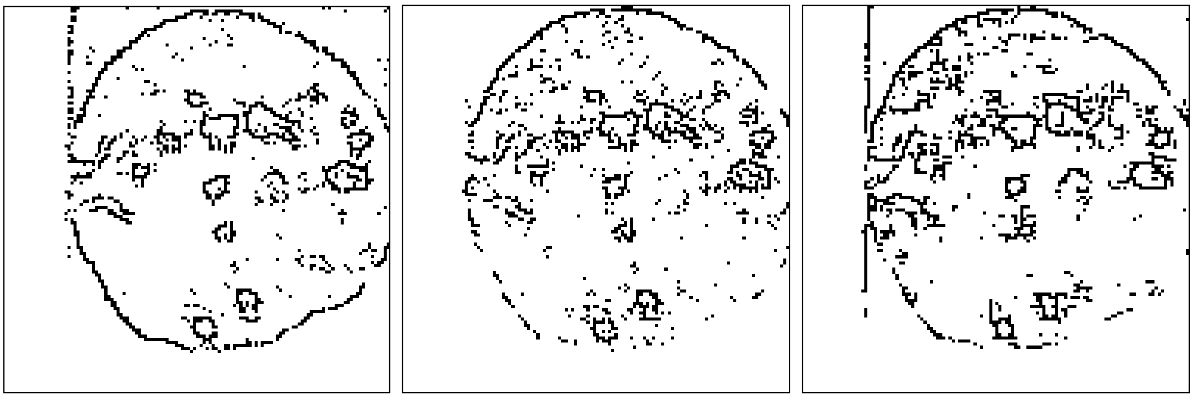
The data set under consideration contains 31 channels with a large spatial resolution of 820 × 820 pixels. As discussed in Section 2.3.3, we reduce the computational cost by computing the edge strength values for 5,000 randomly selected pixels and then estimating the edge strength values for the remaining pixels using a k-NN regression. The number of nearest neighbors used in the k-NN regression is set to 50.

Edge strength maps generated by the three methods are shown in Figures 2.11a-c. Figures 2.12a-c show the best subjective binary edge results by thresholding the three edge strength maps (binary edge maps generated using various thresholds can be found in



(a) (b) (c)

Figure 2.8: Edge strength maps generated on the SEM/EDX data set by (a) SEDMI, (b) Di Zenzo's method, and (c) the RCMG method. Dark color means high edge strength.



(a) (b) (c)

Figure 2.9: The best subjective binary edge maps generated for the SEM/EDX data set by (a) SEDMI, (b) Di Zenzo's method, and (c) the RCMG method.

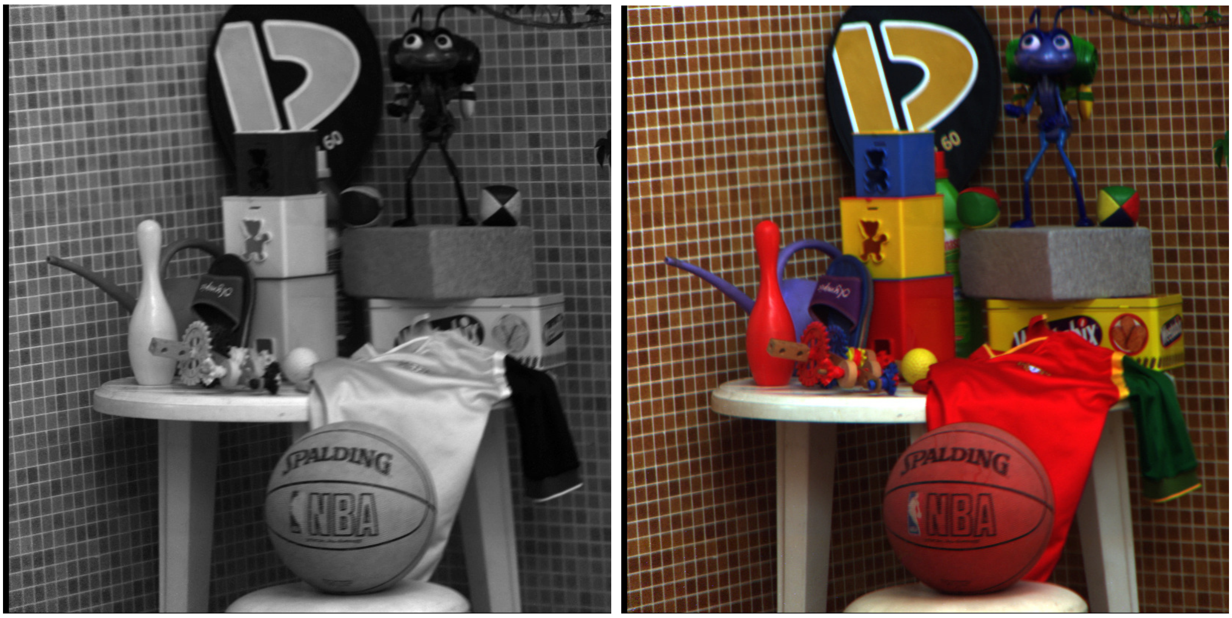
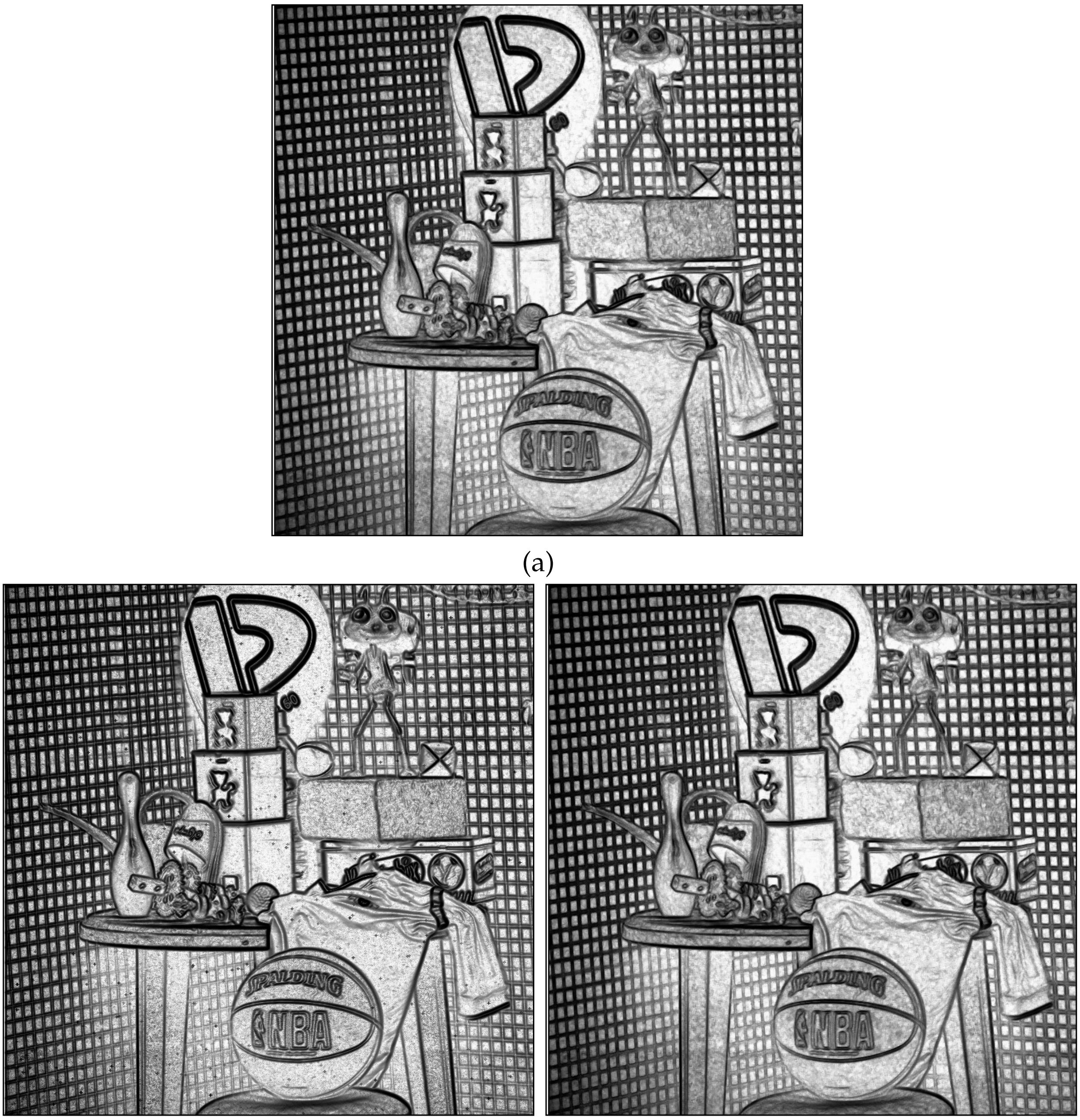

(a) (b)

Figure 2.10: The scene data set. (a) A gray scale image (channel 28) and (b) the reconstructed color image of the data set (channels 28, 14, and 4 are used, respectively, as the red, green, and blue channels for the reconstruction).

the Appendix A, Figure S2). The SEDMI method is able to locate edges of most of the objects such as the text on the ball and the toys on the left side of the table. However, the method does not detect as many edges of the textured wall on the bottom right as the RCMG method does. On the other hand, the Di Zenzo and the RCMG methods generate many spurious edges on the chair and under the ball due to the variance in intensity of the chair's surface. As demonstrated by the AI I data set, SEDMI is better in dealing with such a variance by using the assumption that edges are rare events in the image. Pixels appearing with higher frequency manifest smaller edge strength values; hence, these pixels are not classified as edges.

2.5 Discussion

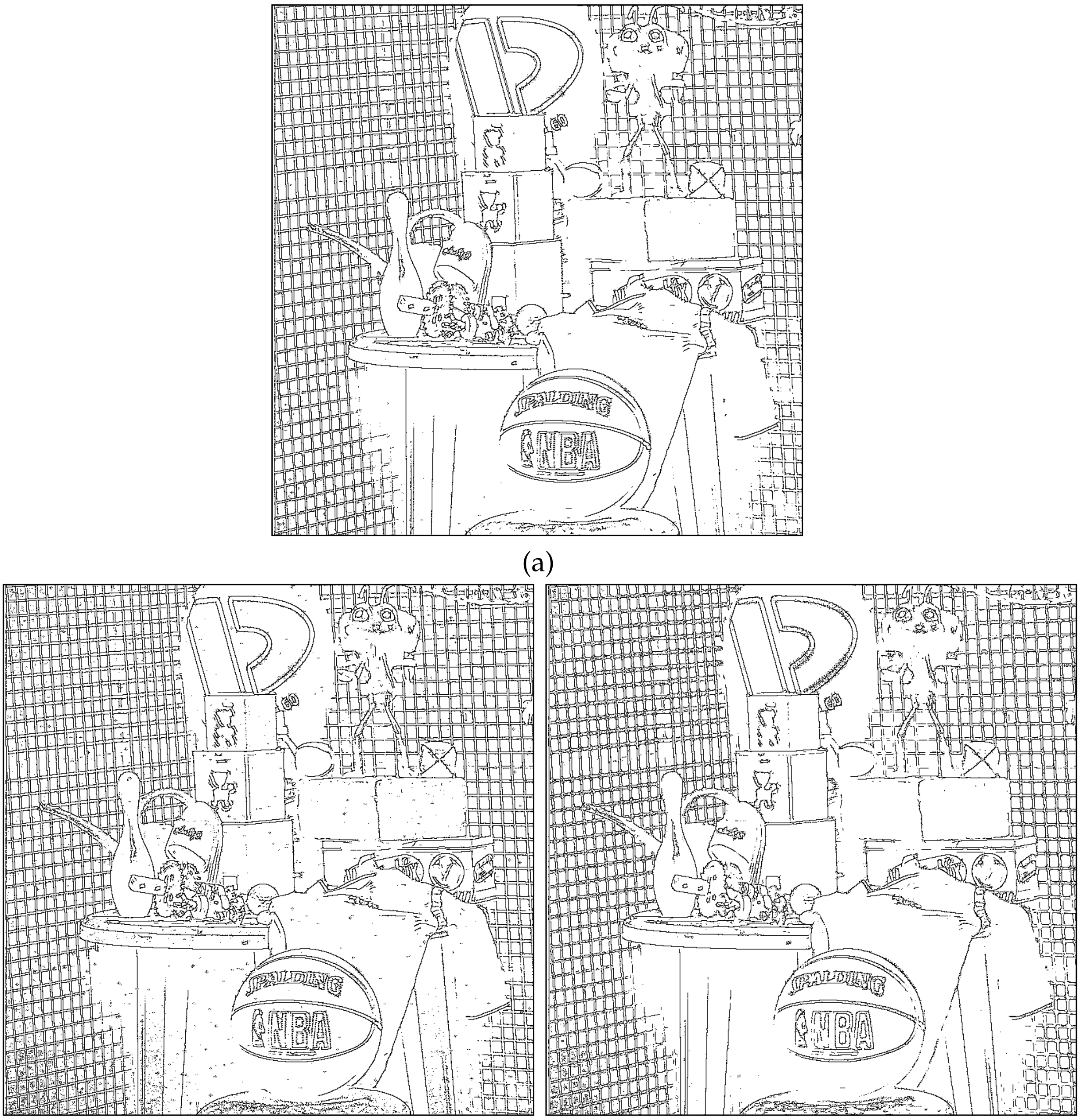
The main advantage of the SEDMI method is the ability to deal with images in which objects are surrounded by severe noise and background clutter. Typical edge detection techniques such as the Di Zenzo method [133] and the RCMG method [35] compute edge strength of a pixel by considering its small, surrounding window. This results in misclassifying the noisy pixels as edge pixels in such a circumstance because the noisy pixels in noisy images may have significantly different intensities compared with their neighbors. Our approach overcomes this problem (c.f. Sections 2.4.1 and 2.4.2) by



(a)

(b) (c)

Figure 2.11: Edge strength maps generated on the scene data set by (a) SEDMI, (b) Di Zenzo's method, and (c) the RCMG method. Dark color means high edge strength.



(a)

(b)

(c)

Figure 2.12: The best subjective binary edge maps generated for the scene data set by (a) SEDMI, (b) Di Zenzo's method, and (c) the RCMG method.

calculating the edge strength of a pixel based on its sparseness in the global gradient magnitude feature space. As a result, edge strengths of noisy pixels, which appear in the image with high frequency, are smaller than those of the edge pixels.

In hyperspectral images, objects appearing in one band may be absent in other bands. Moreover, a broad spectral band might dominate other bands which are more narrow in the spectral domain. Therefore, detecting objects which appear in such narrow spectral bands becomes difficult. As defined by our approach, edge pixels appearing in a few bands exhibit high similarity to one another. These pixels then form a small cluster in the feature space and hence, are detected by SEDMI.

We note that if severe noise (corresponding to low SNRs) is distributed independently in an image where objects appear in all bands, e.g. the AI I data set, Di Zenzo's method and SEDMI perform similarly and both are inferior to the RCMG method. The performance of the RCMG method results from its novel use of the pairwise pixel rejection scheme in calculating the variance in intensities of pixels within a small window. In such a case, applying a smoothing process before using SEDMI will substantially improve the edge detection result.

In terms of complexity, all the three methods are linearly dependent on the number of pixels in the image. SEDMI requires more computation than the other two methods due to the use of the ensemble clustering process. Ensemble clustering requires $O(M \times C \times N)$, where M is the number of pixels in the image; C is the number of clusters in each clustering and N is the number of times doing the clustering. The clustering process can be speeded up by i) performing ensemble clustering on a subset of pixels, which then requires $O(K \times C \times N)$ where K is the number of pixels in the subset; and then ii) generating edge strength map of the whole image using knn-regression algorithm, which requires $O(M \times log(K))$ in average [82]. It should also be noted that the use of the ensemble clustering in our method is to detect events with small probability in the feature space. We are, however, not restricted to using this type of clustering. Other techniques such as the density-based technique can also be employed.

In this chapter we focus on estimating edge strength for every pixel in the image. We note, however, that more sophisticated thinning approaches than the technique used [65] can be applied to multispectral edge detectors that provide accurate edge direction. Therefore, estimating the gradient direction of a pixel is an interesting continuation of the current research.

2.6 Conclusions

We have presented a saliency-based approach for edge detection in multispectral images. First, we constructed the gradient magnitude feature space which contains spatial gradient magnitudes in all spectral channels. This feature space is composed of global information in both spatial and spectral domains. The key characteristic of this feature space w.r.t. the edge detection problem is that edges often stay in small, isolated clusters. Second, based on the assumption that edges are rare events in an image, we recast the edge detection problem into detecting events with small probability in the feature space. This assumption is reasonable as the proportion of edge pixels in an image is generally small.

Using the key characteristic of the feature space, we then estimated the confidence value that a pixel is a small probability event based on the size of the cluster containing it. The estimation is reliably produced by ensemble clustering. The confidence value is then interpreted as the edge strength of the pixel. Thus, the smaller the cluster size corresponding to a pixel, the more probable it is this pixel does belong to an edge in the image.

Experimental results on a number of multispectral data sets show that the proposed method gives promising results, especially in detecting objects embedded in background clutter or appearing in a few bands. The results also confirm that the rarity is an important property of edges in images and this property should be studied further. It also holds for other salient features in an image such as corners, junctions, and blobs. To construct suitable feature space to detect these features in a hyperspectral images may be interesting topics for future research.

3

TRAINING DATA SELECTION FOR CANCER DETECTION IN SPECTRAL ENDOSCOPY IMAGES

Spectral endoscopy images provide considerable potential for early stage cancer detection. This chapter considers this relatively novel imaging technique and presents a supervised method for cancer detection using such spectral data. The data under consideration include different types of cancer. This poses a challenge for the detection as different cancer types may exhibit different spectral signatures. Consequently, it is not always feasible to transfer the knowledge learnt from one data set to another data set. In our approach, we select suitable training data for a given test set based on a similarity measurement between data sets. Our experiments demonstrate that the classification results can be significantly improved if a few data sets that are presumably similar to a given test set are selected for training instead of using all available data sets.

This chapter is based on the works published as Cuong V. Dinh, Marco Loog, Raimund Leitner, Olga Rajadell, Robert P.W. Duin. Training Data Selection for Cancer Detection in Multispectral Endoscopy Images, *the 21st International Conference on Pattern Recognition (ICPR)*, Tokyo, Japan, 2012; and

Raimund Leitner, Martin De Biasio, Thomas Arnold, Cuong V. Dinh, Marco Loog, Robert P. W. Duin. Multispectral video endoscopy system for the detection of cancerous tissue, *Pattern Recognition Letter*, 2012.

3.1 Introduction

Early cancer detection plays an important role in increasing the chance for successful cancer treatment. Video endoscopy combined with histopathological examination of biopsies is currently the gold standard for preventive examinations and diagnosis of cancer and its precursors in otolarygoscopy, gastroenteroscopy and colonoscopy [109, 129, 136]. Taking biopsies requires physical removal of specimens followed by a histopathological analysis [124]. It is difficult to determine the dysplastic and malignant regions for biopsies and therefore the procedure may have to be repeated many times, which delays the necessary treatment [16].

In addition, it may happen that experienced domain experts are not available and the investigation is conducted by less-experienced physicians. A video endoscope that semi-autonomously identifies cancerous and pre-cancerous tissue regions by augmenting the video stream with an overlay highlighting regions that require specific examination would help in these situations. This chapter introduces such a system and discusses the results obtained for the detection of cancerous tissue regions.

Optical techniques, such as the autofluorescence spectroscopy, have been investigated for early cancer diagnosis. Autofluorescence is the light emission of specific substances of biological tissues, e.g. porphyries and proteins if the tissues are excited by a light source. Those substances then emit light of specific wavelengths. The spectra of the tissues then correspond to different wavelengths measured by the spectroscopy. Previous studies, e.g. [48] have shown that there is a significant difference in the fluorescent properties, such as their spectral shape and intensity, between malignant and normal tissues. Therefore, they have been used to identify early instances of diseases in the colon, larynx, lung, and other organs.

The advantage of optical techniques lies in their potential to perform *in vivo* detection without the need for tissue removal. Therefore, they facilitate the determination of the dysplastic and malignant regions for the biopsy. These spectroscopic diagnosis techniques are often referred to as point-measurement methods as they attempt to obtain the spectra of a single tissue.

Multi/hyper-spectral endoscopy techniques developed recently provide three-dimensional images of the area of interest in both spatial and spectral domains [63, 78, 124]. Spectral images provide richer information than point-measurement techniques as they are able to acquire the spectra of thousands to millions of malignant and normal pixels at the same time. The new imaging technique techniques also raise the question on how to effectively exploit this rich source of spectral information. Current spectral endoscopy based approaches for cancer detection are mainly unsupervised methods. For example, [78, 127] use a thresholding algorithm to assign pixels to normal/malignant spectra based on the observation that the intensity of a malignant area is brighter than that of a normal area. In this chapter, we present a supervised method, in particular, we focus on the issue of transferring knowledge among data sets.

The data under consideration consist of eight spectral endoscopy images belonging to different types of cancer. As different cancer types may exhibit different spectral signatures [18], the discriminant information between normal and malignant tissues learnt from a data set may not be applicable to another data set. We address this problem by selecting suitable training data sets for a given test set. Data sets are only selected for training if they are similar to the test set, i.e., they stay close to it in the feature space. Experimental results show that the classifications can be significantly improved if a few data sets which are similar to a test set are selected for training instead of using all data sets.

The rest of the chapter is organized as follows. Section 3.2 introduces an overview of a spectral imaging system for cancer detection. Section 3.3 provides the data acquisition and visualization. Section 3.4 presents our method for training set selection for a given test set. To demonstrate the effectiveness of the proposed method, experimental results are presented in Section 3.5. Section 3.6 discusses related issues and draw conclusions.

3.2 System Overview

In the past spectral image acquisition systems were of limited use for endoscopy due to (i) the necessary spatial scanning of push-broom approaches or (ii) the impractical long switching times of liquid crystal tunable filters. Recent technological advances in the field of tunable filters, in particular the development of fast acousto-optical tunable filters (AOTF), made switching times below 1 ms feasible. Thus, AOTFs represent a suitable technology for the acquisition of spectral image and video data with excellent spatial, spectral and temporal resolution.

This chapter introduces a spectral endoscope system using a fast AOTF synchronized with a highly sensitive electron multiplication charge coupled device camera (EMCCD) that allows the acquisition of both hyperspectral and multispectral video data (see also [64]).

The demonstrator for spectral video endoscopy was designed to fulfill two requirements: (i) acquisition of multispectral videos with up to 8 bands and 40 frames per second, and (ii) the acquisition of hyperspectral images with 5 nm spectral resolution and 1M pixel spatial resolution. Liquid crystal tunable filters (LCTF) have a generally better blocking efficiency, but are with 50 to 150 ms switching time too slow for our purposes. AOTFs achieve switching times around 50 μs and are thus much faster than LCTFs.

To minimize exposure times the demonstrator comprised a high sensitivity EMCCD camera (Andor, Ireland) attached to an AOTF (Brimrose, US) and a 10 mm rigid endo-

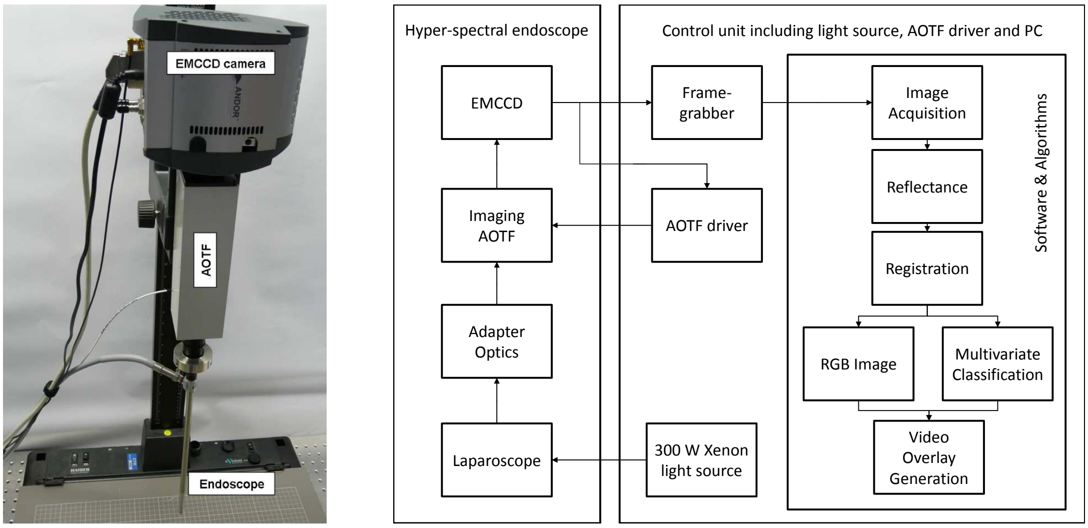


Figure 3.1: The spectral endoscope, its components and data flow. The spectral endoscope comprises a rigid laparoscope attached to an AOTF, adapter optics and an EMCCD camera (left). Light source, AOTF driver and PC are incorporated in a separate housing not shown. A schematic block diagram depicts the connections between the components and the data flow between hardware and software parts (right).

scope including a 300 W Xenon light source (Richard Wolf, Germany) (Figure 3.1). The AOTF was designed for the wavelength range from 400 to 650 nm.

The demonstrator is capable of acquiring images at a frame rate of 40 fps. Either hyperspectral images with up to 51 bands (400 to 650 nm with 5 nm FWHM) can be acquired in 1.25 s. Or, in multispectral video mode, if 8 bands (specific wavelengths) are used, five multispectral images per second for a spatially resolved real-time tissue classification can be acquired. Combined with a live image of the endoscopy, such a system provides an additional, image based, information for the examiner to support the diagnostic decisions. The current limitation is the EMCCD camera: this cannot transmit more than 40 images per second. With a faster camera higher frame rates would be possible. Despite the transfer limit of the EMCCD camera, multispectral videos are feasible due to the fast switching times of AOTFs and, as will be shown in the results sections, the spectral information enables the detection of cancerous tissue regions.

The schematic diagram on the right side of Figure 3.1 shows the connections of the components and the data flow between them. The raw image data is acquired by a frame grabber and transferred into the main memory of the PC. There the reflectance is calculated from the raw image data and two calibration images, a dark noise image and a white reference image cube. As the raw image data and the calibration images have to correspond exactly on a pixel to pixel level, the reflectance calculation is executed before the image registration. Subsequently, the image registration corrects for any non-rigid transformation occurring during the acquisition. The registered images of the image

Table 3.1: Data set overview. Ground truth was available only for a subset of the acquired multispectral images.

Otolaryngology		
Tissue	Datasets	Ground truth
Lymph node	11	
Lingua	6	
Larynx	6	4
Parotid gland	4	1
Pharynx	3	1
Diaphragm oris	2	1
Oesophagus	1	1
Unspecified/mix	3	
Overall	36	8

cube are then used to reconstruct a RGB image for display, while the registered image cube itself serves as the input for the multivariate classification. The classification result is superimposed as an overlay on the RGB image that is finally displayed on a monitor.

3.3 Materials

3.3.1 Data acquisition

The demonstrator has been evaluated during 36 measurements in clinical environment for otolaryngoscopic investigations at the Katharinenhospital (Stuttgart, Germany). Table 3.1 lists the measurements acquired with the demonstrator. The demonstrator was used to acquire spectral images (1004x1002 pixels, 51 bands from 400 to 650 nm with 5 nm FWHM) of biopsies taken after the investigations. The spectral images of the biopsies were acquired within 10-15 minutes after the excision. Within this period it is commonly believed that in-vivo conditions for the tissue are conserved. Once the histopathological findings were available, the measurements were discussed in several workshops with otolaryngoscopic experts. The aim of these workshops was to identify regions that are clearly cancerous or non-cancerous. During the workshops it turned out that only regions that are clearly cancerous can be identified. The margins and if the tumor may have spread can not be answered from the RGB reconstructed images and the available histopathological findings. Only another set of histopathological analyses at the margins could have clarified this question. Unfortunately this was not possible for these measurements. As a consequence, the ground truth of the measurements we report in this chapter comprises regions that are *clearly cancerous* and *maybe non-cancerous*.

Due to the unforeseeable histopathological findings and the difficulty in identifying unambiguous regions, a reliable ground truth was available for only 8 of the acquired 36 measurements (cf. Table 3.1). These 8 data sets were used for the subsequent analysis described in the Section 3.5.

The eight data sets under consideration (called M1, M2, ..., M8) belong to different types of cancer: Larynx (data sets M3, M4, M5, and M8), Pharynx (M1), Esophagus (M2), Diaphragm (M6), and Parotid (M7). For the M4 data set, the exact boundary of the cancer area is unclear since the cancer tissue is under the surface. Therefore, it is not easily detectable by a non-penetrating optical method. The number of spectral bands is 51; however, the first eight bands are removed as they are too noisy and provide no information. Thus, each data set contains 43 spectral bands.

3.3.2 Spectral visualization

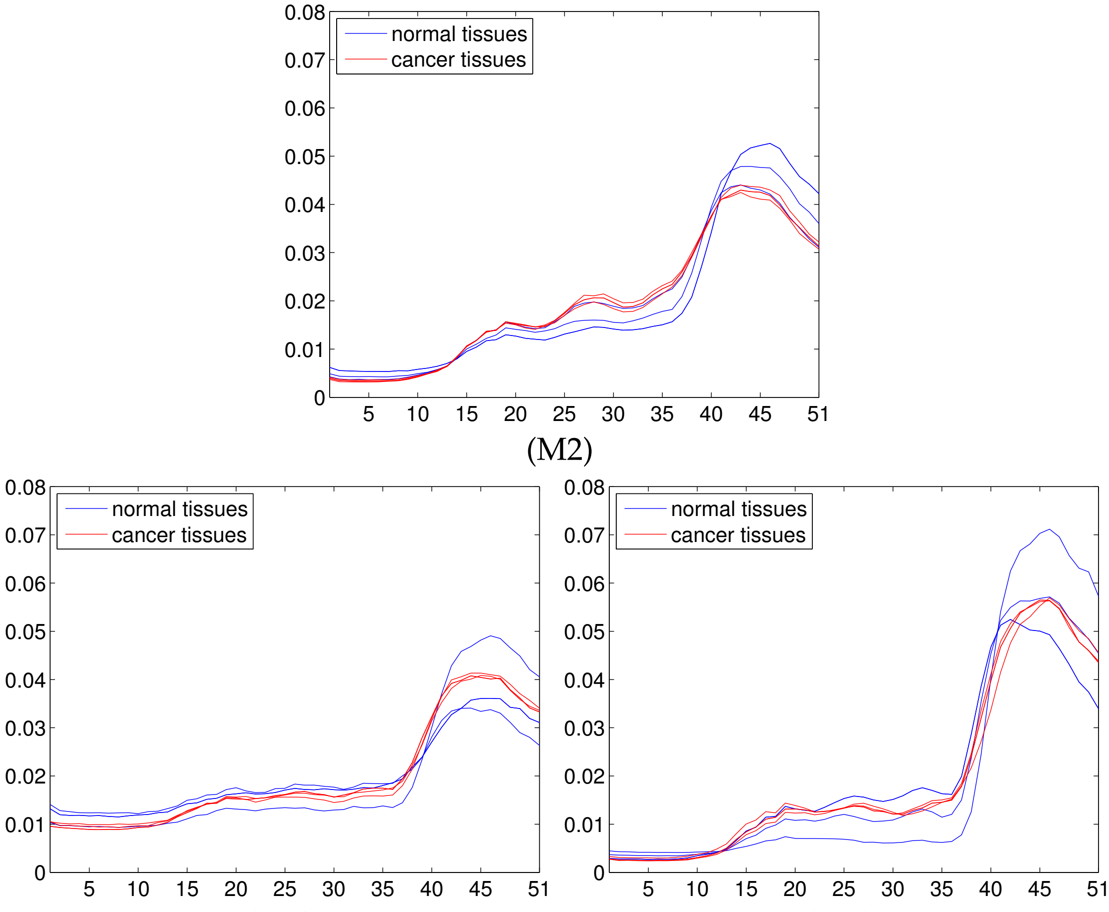
Figures 3.2a-c show examples of the spectra of normal and cancer tissue pixels of three data sets: M2, M5, and M8. The figures display six spectra which characterize different regions (including the normal and abnormal regions) in the data. Spectra of normal and cancer tissues are depicted in blue and red, respectively. The figures show that the first spectral bands (1-8) do not provide any information as normal and cancer tissues have almost the same spectral values. Spectra from both normal and cancer tissues have high responses in the last few bands as the tissues are measured under the white light condition and they are sensitive to red wavelengths.

Within each data set, spectra of cancer tissues are similar to each other. However, they are different from different data sets, especially in the shape of the spectral curves. For example, the spectral curves of the cancer tissues for the M2 data set have two visible peaks in the band ranging from 15 to 35. Whereas, the cancer tissues of M5 have almost flat spectral shape. In addition, normal tissues have a high variance in spectra within each data set as the number of normal tissues within each data set is normally large. Furthermore, the spectra of normal tissues can be heterogenous as they can be effected by operational conditions, e.g. the specularity, and blood in the tissues.

3.4 Methods

3.4.1 Calibration and pre-processing

A spectroscopic measurement system must be precisely calibrated. Our system's characteristics depend on the dark noise and gain of each camera pixel, the transmission characteristics of the AOTF and the emission of the light source. To achieve a homogeneous system performance within the field of view and the selected wavelength range (400-650 nm) the system is calibrated by two reference images, a dark noise image and



(M2)

(M5)

(M8)

Figure 3.2: Examples of spectra of the normal and cancer tissues in three data sets: M2, M5, and M8. The horizontal axis shows the spectral bands, while the vertical axis shows their intensities after the pre-processing.

a white reference image cube. The reflectance in each pixel is determined using the following equation

$$R(x,y,\lambda) = \frac{I_{RAW}(x,y,\lambda) - I_{DN}(x,y)}{I_{WR}(x,y,\lambda) - I_{DN}(x,y)} \cdot R_{WR}(\lambda) \tag{3.1}$$

where R denotes the reflectance, I_{RAW} the uncalibrated (raw) image intensity, I_{DN} the dark noise image, I_{WR} the white reference image cube and R_{WR} the reflectivity of the used white reference.

First, each reflectance spectrum is normalized using the area under the curve normalization in the spectral domain. Second, spectra corresponding to the specular reflection are removed by a simple thresholding algorithm. Third, the principal component analysis (PCA) is used to reduce the number of features from the original space. The reconstruction of all data sets are then based on their first eight eigenvectors which preserve 99%

of the total variance. Finally, a unit variance normalization is applied to each data set so that each spectral band has a zero mean and a unit variance. The main aim of this normalization is to align all the data sets, i.e., to force them to stay close to each other in the feature space.

3.4.2 Data selection for training

As the data sets are different from one to another with respect to their class distributions, the discriminant information between normal and malignant tissues learnt from a data set might not be suitable for another data set. Therefore, it is essential to select suitable training sets for a given test set. We first use the Gaussian data domain description [119] to model the distribution of each data set.

Denote by q the percentage of outliers in each data set, a pixel is considered as an outlier if its probability density $p(x_i)$ is smaller than a threshold θ determined by:

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta - p(x_i)) = q, \tag{3.2}$$

where N is the total number of pixels in the data, $h(.)$ the unit step function, and $p(x_i)$ the probability density of pixel x_i. We then measure the similarity between two data sets by the fraction of pixels they share in their data domain. For two data sets M_i and M_j, we calculate M_{ij} the set of all pixels in M_i that belong to the domain of M_j and M_{ji} the set of all pixels in M_j that belong to the domain of M_i. The similarity between M_i and M_j denoted by S_{ij} is defined as:

$$S_{ij} = |M_{ij}|/|M_i| + |M_{ji}|/|M_j|. \tag{3.3}$$

The similarity among data sets are then used as the criterion to select the training set for a given test set. Note that we model a data set using all the pixels contained. It is therefore possible to measure the similarity between any two data sets, e.g. between a training set and a test set, even when we do not have label information of the test set.

3.5 Experimental Results

Since we do not have prior knowledge about the prior probabilities of normal and malignant classes, we set the prior probabilities equally in all experiments. We intentionally select the simple quadratic discriminant classifier (QDC) to avoid overtraining and evaluate the potential of a multi-variate classification based on the available information in the spectra. In addition, the same number of normal spectra (9300) and malignant spectra (2700) are extracted for training from all data sets.

3.5.1 All available data sets are used for training

We first evaluate the classification results for two training scenarios: i) training and test data are from the same data set, i.e., a part of a data set is used for training and the remainder is for testing; ii) training and test data are from different data sets. For the latter, we follow the leave-one-dataset-out cross validation configuration, i.e., seven data sets are used for training and the remaining data set is used for testing. Moreover, for the second scenario we also investigate the influence of the unit variance normalization in the data preprocessing step. Since the QDC is invariant to affine transformations, the classification results for the first scenario remain unchanged whether this normalization is applied or not. Table 3.2 shows the classification results with respect to different training and normalization options.

In the first scenario, a good classification result is achieved with the average error rate is only 10.8 %. The classifier used just the spectral information contained in each pixel to decide whether the tissue imaged at this pixel is cancerous or healthy. This can be considered as an important result to support the hypothesis that a spectral endoscope can provide additional information about the tissue condition. This information could be displayed as an overlay or a color augmentation during any live endoscopy video. Nevertheless, we do note that these results in this scenario should not be extrapolated as the training and test sets are significantly correlated to each other. The results only indicate that the spectral information within the measurements is consistent. Figure 3.3b shows the classification result in the first scenario on the data set M08 (Figure 3.3a). The overlay in Figure 3.3a shows the regions marked by the medical expert: the blue curve indicates the normal tissue area and the red curve shows the cancer area. The classification results in Figure 3.3b are shown as red and green borders. Regions surrounded by red borders are classified as cancerous tissue and regions surrounded by green borders are classified as healthy tissue. The figure shows that the classifier is able to identify the similar regions as cancerous as the medical expert did.

In the second scenario, the error rates (second and third rows of Table 3.2) increase substantially comparing to the first scenario. In addition, the results demonstrate that the unit variance normalization significantly improves the classification when the training and test data are from different data sets. The average error rate in this case is 29.0 %. As explained in Subsection 3.4.1, this normalization helps to align all the data sets, i.e., to force them to stay close to each other in the feature space. Therefore, we apply the normalization step in all of the following experiments.

The classification result in the second scenario on data set M8 is shown in Figure 3.3c. The predictions tends to identify larger regions expanding into non-cancerous regions. Potentially, the fact that several different tissue types were combined, namely larynx, pharynx, diaphragma oris and parotis, is the reason for the increase of the error rates.

Preliminary experiments with more sophisticated classifiers as SVM and MOGC did

Table 3.2: Error rate (%) for different training and normalization options

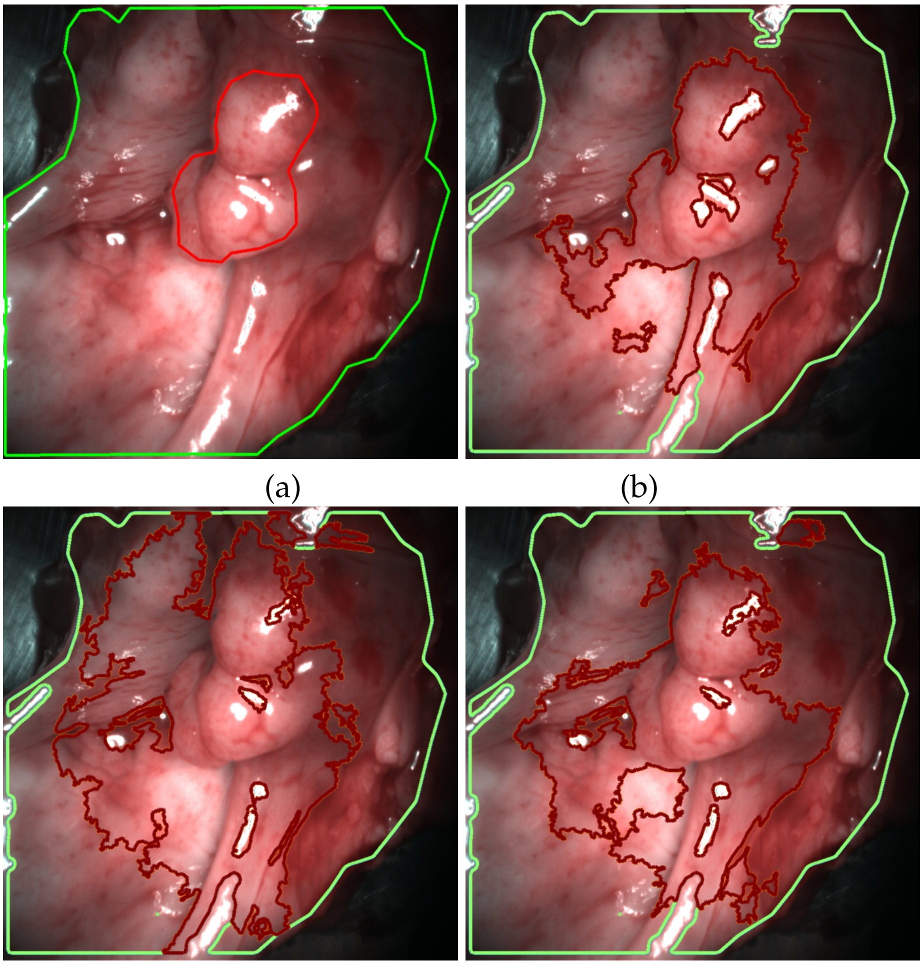
Training scenario	**Norm.**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**Mean**
Same set	No	09.9	11.0	16.0	10.1	05.8	15.8	10.0	07.6	10.8
Different sets	No	39.8	48.9	34.0	28.8	51.6	46.0	30.2	22.8	37.7
Different sets	Yes	**30.1**	**26.5**	**36.0**	**29.6**	**17.6**	**38.1**	**30.5**	**23.3**	**29.0**

not improve the results (cf. Table 3.3), hence the limited model complexity of QDC is not the cause. Generally, it is known that the combination of different tissue types in a spectroscopic model is difficult and this result can be seen as another result supporting this fact. However, these results are important for a comparison to the approach using a training data selection.

3.5.2 Training data selection by the Gaussian domain description

We evaluate the classification results when the training data contain similar data sets for a given test data set. We model the data sets by using the Gaussian domain description in which the percentage of outlier q is set to 0.1. For each data set, we first selected the training data as the most one, two $\cdots$ seven similar data sets (denoted by Case 1, 2 $\cdots$ 7) according to the similarity measurement defined in the Section 3.2. The QDC is then trained on the selected training data and subsequently used for the classification of normal/malignant tissues for the data set under consideration. Table 3.4 shows the error rates for all seven cases. Numbers in bold emphasize the best results achieved for each data set in all cases. Note that Case 7 corresponds to the results shown in the third row of Table 3.2 as all seven data sets are included in the training data. On average, the best classification results are obtained if the two most similar data sets are used for training (Case 2). Increasing the number of training data sets then, in most of the time, worsens the classification as irrelevant data are included in the training process. Case 2 yields the best results for five over eight data sets. Case 2 does not perform well on the data set M6 as the data set itself is challenging: the cancer type (diaphragm) is far different from the other cancer types. Therefore, it is needed to have a sufficient large amount of annotated measurements (which are currently not available in the evaluated data sets) so that the training selection method has more opportunities to select the right suitable training set. Nevertheless, it can be concluded from our experiments that the spectral information to distinguish between dysplastic and healthy tissue is available and consistent among data sets.

The classification result in Case 2 on the M8 is displayed in Figure 3.3d. The figure shows that the training selection scheme does help to improve the result significantly comparing to the case that all other data sets are used for training (Figure 3.3c).



(a) (b)

(c) (d)

Figure 3.3: Classification results. Ground truth labeled by experts (a), classification results corresponding to the first training scenario (b), the second training scenario (c), and training selection scheme (d).

Table 3.3: Error rates (%) using different classifiers

	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**Mean**
QDC	**30.1**	**26.5**	**36.0**	**29.6**	**17.6**	**38.1**	**30.5**	**23.3**	**29.0**
PARZENC	35.1	37.6	42.4	37.3	30.0	43.5	39.2	32.9	37.3
MOGC	34.9	24.5	34.7	28.1	18.1	44.4	27.6	26.7	29.9
LSVM	27.1	48.3	38.9	30.7	20.7	41.0	32.6	21.7	32.6

Table 3.4: Error rate (%) when training data selection is used

	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**Mean**
Case 1	42.0	27.0	45.4	51.2	27.2	42.7	31.9	24.0	36.4
Case 2	**26.8**	25.4	**30.1**	**24.1**	26.5	50.9	**26.5**	**16.0**	**28.3**
Case 3	30.8	27.6	32.6	28.4	20.1	45.6	28.9	18.4	29.1
Case 4	31.8	26.2	39.0	25.2	18.7	44.0	29.8	24.3	29.9
Case 5	**26.8**	25.3	36.9	27.7	19.0	39.6	34.3	25.1	29.3
Case 6	29.2	**24.7**	36.0	28.7	**17.2**	**36.1**	30.9	24.7	28.4
Case 7	30.1	26.5	36.0	29.6	17.6	38.1	30.5	23.3	29.0

3.6 Conclusions

This chapter introduces a spectral imaging system for cancer detection and presents a study of normal/malignant tissue classification for eight spectral endoscopy data sets in a supervised manner. The data are heterogeneous as they are collected from different patients and with different types of cancer. We showed that the classification result is improved if a subset of the data that are similar to the test set is used for training (cf. Table 3.4). In other words, it is not always good to combine all available data for training as the difference between the data sets may result in poor classification.

We introduce an approach to select training data based on the similarity between data sets using the Gaussian data domain description. Experimental results show that the method substantially improves the classification results for our heterogeneous data. Note that we measure the similarity between data sets based on all the pixels, i.e., from both normal and malignant classes. For data from a patient who does not have cancer, all the pixels should fall into the normal region of the selected training data; therefore, our method correctly classifies the data set as normal.

In the present chapter we use PCA to reduce the dimensionality of the feature space. To find subspaces that provide discriminant information between normal and malignant tissues in the data may also improve the performance of the classifiers. Finally, more data sets are essential to fully evaluate the applicability of our method.

4

FIDOS: A GENERALIZED FISHER-BASED FEATURE EXTRACTION METHOD FOR DOMAIN SHIFT

Traditional pattern recognition techniques often assume that the data sets used for training and testing follow the same distribution. However, this assumption is usually not true for many real world problems as data from the same classes but different domains, e.g. data are collected under different conditions, may show different characteristics. We introduce FIDOS, a generalized FIsher-based method for DOmain Shift problem, that aims at learning invariant features across domains in a supervised manner.

Different from classical Fisher feature extraction, FIDOS not only aims to minimize the within-class scatters, but also the difference in distributions between domains. Therefore, the subspace constructed by FIDOS reduces the drift in distributions among different domains and at the same time preserves the discriminants across classes. Another advantage of FIDOS over classical Fisher is that FIDOS extracts more features when multiple source domains are available in the training set; this is essential for a good classification especially when the number of classes is small. Experimental results on both artificial and real world data and comparisons with other methods demonstrate the efficiency of the our method in classifying objects under domain shift situations.

Published as Cuong V. Dinh, Robert P. W. Duin, Ignacio Piqueras-Salazar, and Marco Loog. FIDOS: A Generalized Fisher based Feature Extraction Method for Domain Shift, *Pattern Recognition*, 46(9): 2510-2518, 2013.

4.1 Introduction

Traditional pattern recognition techniques often assume that the data sets used for training and testing follow the same distribution. However, this assumption is usually not true for many real world problems as data from the same classes but different domains may show different characteristics. Consequently, a model learnt from training data that come from one or several source domains might not perform well when applied to test data coming from a target domain.

A typical situation in which the assumption is violated is domain shift. This problem is characterized by the fact that the measurement system, or the method of description, can change [97, Section 1.8]. This problem arises in a variety of applications, such as in computer vision [26, 41], remote sensing [54, 101], and multivariate time series [117, 125]. For example, in computer vision, the image of an object captured by a digital camera may significantly differ from the one captured by a webcam. In remote sensing, spectra of objects from the same class collected at different times and locations can also be different due to environmental changes or changes in the objectǓ’s spectra themselves w.r.t. both spatial and temporal domains. In time series applications, e.g. in measuring the cortical activity by electroencephalography (EEG), the corresponding signals usually appear non-stationary in time, partly due to the inherent non-stationary dynamics in the brain [125].

A related term for domain shift, which is also widely used in the literature, is domain adaptation. Both of them rely on the “existence of good domain embedding" assumption [97, Section 5.5], i.e., there exists a new feature representation transformed from the original feature space under which source and target distributions are unchanged. Denote by $P_S(X)$ and $P_T(X)$ the distributions of data X in the source and target domains, respectively, and by Y the corresponding labeling of X. For a classification problem, the assumption of the “existence of good domain embedding" means there exists a mapping W under which $P_S(Y|W(X)) = P_T(Y|(W(X)))$, although $P_S(X)$ might be different from $P_T(X)$.

In the special case where the transformed space is identical to the original space, the assumption becomes $P_S(Y|X) = P_T(Y|X)$, which is often known as covariate shift situation. To address this problem, importance reweighting [4, 51, 71, 116] is a commonly used technique. It assigns different weights to the training samples so that the training data distribution more closely matches that of the test data.

Many methods have been proposed to learn the transformed feature space. Saenko et al. [108] introduced a metric learning method for the domain shift problem in image classification. The metric is learnt in such a way that samples from the same class but belonging to different domains are similar to one another, whilst samples from different classes are different from each other no matter to which domains they belong. Pan et al. [86] proposed an unsupervised kernel based feature extraction method called transfer component analysis (TCA). TCA determines a kernel space in which the difference in distributions between source and target domains is minimized. In addition, TCA also tries to preserve data variance inferred from both domains. TCA assumes that the combination of the two criteria results in a mapping that approximately satisfies $P_S(Y|W(X)) = P_T(Y|(W(X)))$. In addition, the authors also introduced a partially supervised version of TCA, called semi-supervised transfer component analysis (SSTCA), to make use of the labeling information from the source domain. Compared to TCA, SSTCA requires two additional parameters to account for the label dependence and locality preservation of the data. This increases the complexity of the method and makes it difficult to determine the parameter values, especially for small sample size situations. Tu et al. [122] proposed a supervised method, called transferable discriminative dimensionality reduction (TDDR), for transferring discriminative information from a source domain to a target domain. In their method, an objective function is built that satisfies two criteria: i) it encourages the class separation based on labeled samples from both domains; and ii) it penalizes the distance between source and target distributions.

One aspect which has recently received attention in learning the domain shift is how the algorithm gains extra information when there are a few source domains available instead of one. In principle, any domain shift method designed for single source domain situation can also be used for multiple source domain situations by disregarding the information that training data are from different sources. However, such information might be crucial to reveal the features leading to the shift in the data.

Von Bunau [125] introduced a stationary subspace analysis (SSA) method for multivariate time series applications. In SSA, the signal is split into N consecutive epochs. The stationary subspace corresponds to the space in which the distributions (in terms of mean and covariance) of all epochs remain unchanged. SSA can be applied to the multiple source domain situations when considering each domain as an epoch. Thus, N epochs correspond to N domains. However, the main limitation of SSA is that it often requires a large number of source domains (N is typical hundreds or thousands). Thus, a small number of available source domains may lead to spurious stationary problem. Furthermore, the optimization problem in SSA is non convex.

Another approach is to combine the classifiers learnt from multiple source domains. Luo et al. [74] proposed a method that maximizes the consensus of classifier predictions from multiple sources. Based on the assumption that the target distribution is a mixture of the source distributions, Mansour et al. [76] showed that a combination weighted by the source distributions is favored over the standard convex combinations of the source classifiers.

We introduce FIDOS, a FIsher based feature extraction method for the DOmain Shift problem, that handles both single and multiple source scenarios. FIDOS aims at learning invariant features with respect to different domains in a supervised manner. The subspace constructed from these features minimizes the differences in distributions between all source and target domains while preserving the discriminants across classes.

Like classical Fisher mapping, FIDOS is invariant to linear transformations of the feature space and is computationally efficient as the solution is obtained by solving a generalized eigenvalue problem. In addition, FIDOS in general extracts more features than classical Fisher does as the number of features extracted by classical Fisher is bounded by the number of classes.

The rest of the chapter is organized as follows. Section 4.2 presents our feature extraction method to handle domain shift problem. Section 4.3 introduces the data sets and other methods used for evaluation. Section 4.4 summaries experimental results in comparisons with other methods which demonstrate the effectiveness of FIDOS. Section 4.5 discusses related issues and draws conclusions.

4.2 Fisher-based Feature Extraction for Domain Shift

4.2.1 Notations

We consider the domain shift problem with N source domains and a single target domain. Thus, the input is a training set X^S, the source data, composed of N source domains $X^1, X^2, \cdots, X^N$ and a target domain X^T. Denote by K the number of classes in the classification task, X_i^u the subset of samples in X^u that belong to class i ($u = 1 \cdots N$ and $i = 1 \cdots K$). For each domain u, we have $X^u = \{X_1^u, X_2^u, \cdots, X_K^u\}$. Let p_i^u and μ_i^u be the class prior probability and the mean corresponding to the subset X_i^u, respectively. For each domain u, we have $\sum_{i=1}^{K} p_i^u = 1$. We assume that the class prior is unchanged across domains; thus the prior probability of the subset X_i^u with respect to the whole training set X^S is $\frac{1}{N} \times p_i^u$. The output of the algorithm is a linear mapping W, a $d \times k$-matrix, which transforms the original feature space composed of k features into a reduced feature space of d features ($d < k$).

4.2.2 Feature Extraction using Classical Fisher Criterion

Classical Fisher feature extraction [37] is a well-known method to establish a linear transformation that maximizes the ratio of between-class scatter to average within-class scatter in the lower-dimensional space. Specifically, when ignoring the fact that training data may come from different source domains, classical Fisher maximizing the so-called Fisher criterion J_F

$$J_F(A) = \mathrm{tr}((AS_WA^T)^{-1}(AS_BA^T)), \tag{4.1}$$

where A is a $d \times k$-matrix, $S_B := \sum_{i=1}^{K} p_i(\mu_i - \mu^S)(\mu_i - \mu^S)^T$ and $S_W := \sum_{i=1}^{K} p_i S_i$ are the between-class scatter matrix and the pooled within-class scatter matrix, respectively, and S_i is the within-class covariance matrix of class i. μ_i and p_i are the mean vector and the prior of class i in X^S, i.e., $\mu_i = \frac{1}{\sum_{u=1}^{N} p_i^u} \sum_{u=1}^{N} p_i^u \mu_i^u$ and $p_i = \frac{1}{N} \sum_{u=1}^{N} p_i^u$. The overall mean μ^S in X^S equals $\frac{1}{N} \sum_{u=1}^{N} \sum_{i=1}^{K} p_i^u \mu_i^u$. The solution to this optimization problem is obtained by an eigenvalue decomposition of $S_W^{-1} S_B$ and then taking the rows of the mapping W to equal the d eigenvectors corresponding to the d largest eigenvalues.

It is noteworthy that classical Fisher criterion results in suboptimal solutions in the case of domain shift as it does not take into account the differences in distributions among domains. The best subspace in terms of separation among classes in the training set might not be the desired one for the target set as the target distribution may be different from the source domain distributions. In the following, we present FIDOS that generalizes the Fisher criterion for the domain shift problem.

4.2.3 FIDOS

We define a new between-class scatter matrix S'_B as the weighted average of the between-scatter matrices among the sub-groups that belong to different classes:

$$\begin{aligned} S'_B &= \sum_{i=1}^{i=K-1} \sum_{j=i+1}^{K} \sum_{u=1}^{N} \sum_{v=1}^{N} \frac{p_i^u}{N} \frac{p_j^v}{N} (\mu_i^u - \mu_j^v)(\mu_i^u - \mu_j^v)^T \\ &= \frac{1}{N^2} \sum_{i=1}^{i=K-1} \sum_{j=i+1}^{K} \sum_{u=1}^{N} \sum_{v=1}^{N} p_i^u p_j^v (\mu_i^u - \mu_j^v)(\mu_i^u - \mu_j^v)^T. \end{aligned} \tag{4.2}$$

Note that, sub-groups that belong to the same class but different domains are treated differently as they might not be from the same distribution.

We also introduce a new term called S'_W that takes into account the difference in distributions among all the source domains and the target domain:

$$S'_W = \frac{1}{(N+1)^2} \sum_{u=1}^{N+1} (\mu^u - \mu)(\mu^u - \mu)^T, \tag{4.3}$$

where $\mu^u = \sum_{i=1}^{K} p_i^u \mu_i^u$ is the mean over all classes of a source domain u ($u = 1 \cdots N$). $\mu^{(N+1)}$ and μ are the mean of the target domain X^T and the mean of all data $X = \{X^S, X^T\}$, respectively. S'_W can be considered as the between-domain scatter matrix in which each domain is treated as a separated class.

FIDOS determines a linear mapping W that maximizes the following criterion J'_F

$$J'_F(A) = \mathrm{tr}((A(cS_W + (1-c)S'_W)A^T)^{-1}(AS'_B A^T)), \tag{4.4}$$

where c is a parameter in the range [0,1]. Eq. (4.4) shows that FIDOS tries to achieve two criteria: i) maximizing the discriminant information among classes, and ii) minimizing a convex combination of the within-class scatter matrix and the between-domain scatter matrix. Similar to classical Fisher, the solution to the generalized Fisher criterion is efficiently obtained by an eigenvalue decomposition of $(cS_W + (1-c)S'_W)^{-1}S'_B$ and taking the rows of W to equal the d eigenvectors corresponding to the d largest eigenvalues.

The free parameter c in FIDOS, on the one hand, balances the influence between S_W and S'_W. If c is large, the criterion $J'_F(A)$ concentrates more on reducing the variance within each class. If c is small, $J'_F(A)$ mainly reduces the differences in distributions across domains. On the other hand, the term cS_W can be considered as a regularized term of $(1-c)S'_W$. This is to avoid the deficiency in the rank of the S'_W in the generalized eigenvalue decomposition as the rank of S_W is often higher than that of S'_W.

We now investigate the properties of FIDOS and its relation with classical Fisher in more detail. FIDOS is equivalent to classical Fisher if there is no domain shift because S'_W in Eq. (4.3) equals to zero in this case. In addition, S'_B is equal to S_B. Thus, the criterion $J'_F(A)$ is equivalent to the classical $J_F(A)$. Moreover, similar to classical Fisher, FIDOS is invariant to linear transformation of the feature space.

Importantly, FIDOS extracts more features than classical Fisher does. The maximum number of features extracted by FIDOS and by classical Fisher depends on the ranks of S'_B and S_B, respectively. Eq. (4.2) indicates that the maximum rank of S'_B is $\min\{k, N \times K - 1\}$. Whereas, the rank of S_B is not larger than $\min\{k, K-1\}$. As the result, the number of features extracted by FIDOS is bounded by $\min\{k, N \times K - 1\}$, while for Fisher it is $\min\{k, K-1\}$. Obviously, $\min\{k, N \times K - 1\}$ is always larger than or equal to $\min\{k, K-1\}$. Finally, by using S_B, classical Fisher assumes that all samples from different domains but the same class are sampled from the same distribution; while by using S'_B, FIDOS explicitly models samples from different domains by different distributions even when the samples belong to the same class.

It should be noted, that similar to our method, Tu et al.'s method (TDDR) [122] also minimizes a combination of the within-class scatter matrix and the between-domain scatter matrix. However, TDDR focuses on transferring discriminant information from one source domain to a target domain. Therefore, when applied to multiple source domain scenarios, TDDR is only able to extract the same number of features as that of classical Fisher and less than that of FIDOS. In other words, FIDOS can be considered as a generalized version of TDDR for multiple source domains.

Table 4.1: Artificial example: Mean and covariance matrices of the two classes with respect to shifted and non-shifted features. The mean of the second class for shifted features is distributed randomly in a circle with the center located at the origin and a radius of 4.0

Features	Class	Mean	Covariance
Shifted features	First class	$\begin{pmatrix} 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$
	Second class	$C(\begin{pmatrix} 0 & 0 \end{pmatrix}, 4)$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$
Non-shifted features	First class	$\begin{pmatrix} 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 10 \end{pmatrix}$
	Second class	$\begin{pmatrix} 2 & -4 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 10 \end{pmatrix}$

4.3 Experimental Setup

4.3.1 Data Description

4.3.1.1 Artificial Data

The data set contains two classes with $2k+2$ independent features. The first $2k$ features (k varies) are influenced by domain shift but the last two feature are not. The number of samples in each class of each domain is 100.

Figure 4.1a shows the data scatter plot corresponding to the last two features from both source and target domains. The first and second classes are marked by blue and red, respectively. Both classes follow Gaussian distributions. Their mean and covariance matrices corresponding to the these features are shown in the last two rows of Table 4.1.

For the sake of simplification, we assume that only the second class is shifted due to an “unknown" factor over time and this affects the first $2k$ features. Figure 4.1b displays the data scatter of two consecutive features. The first class is again modeled by a single Gaussian. For the first $2k$ features, the first class stays at the same location; whereas, the second class is influenced by the shift and its mean is randomly distributed on a circle with the center located at the origin and a radius of 4.0. The mean and covariance matrices of the two classes corresponding to the first $2k$ features are shown in the first two rows of Table 4.1.

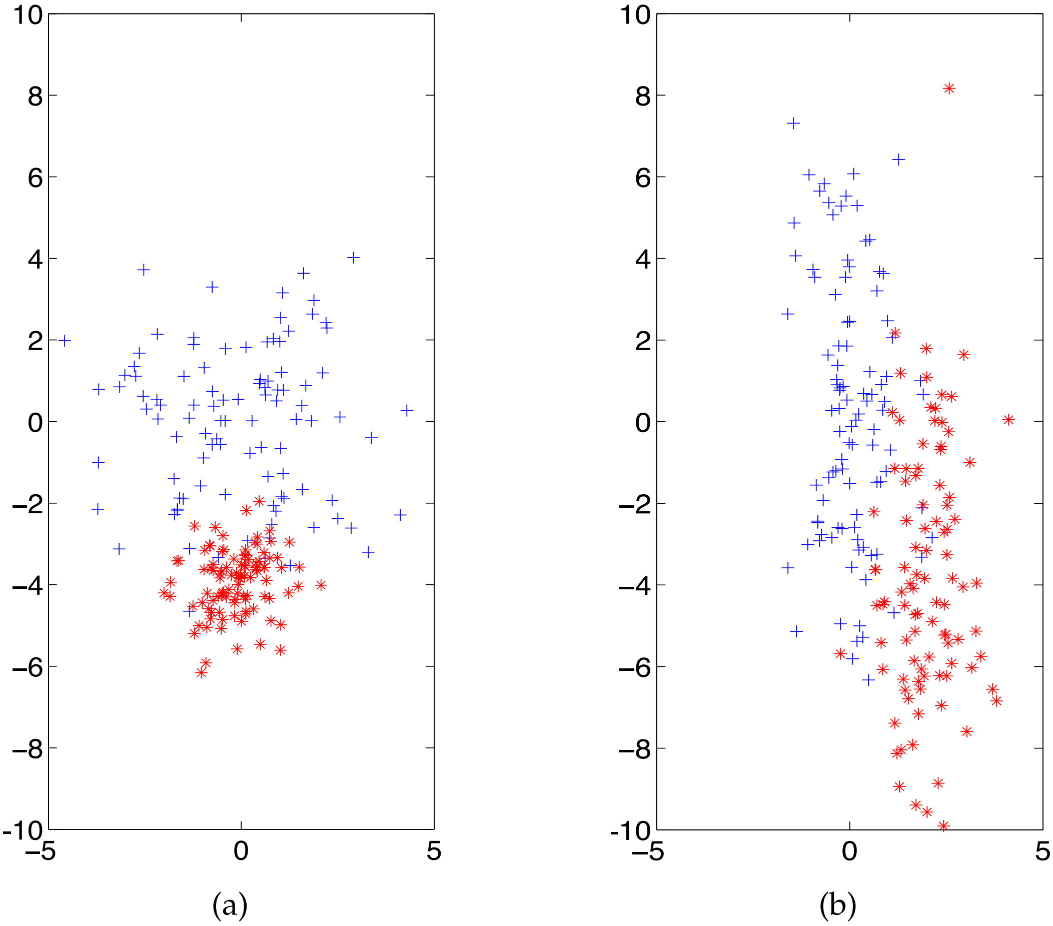


Figure 4.1: Scatter plots of the artificial data set. (a) Scatter plot of any two consecutive features from the first $2k$ features which are influenced by the shift problem and (b) Scatter plot of the last two features which are not influenced by the shift problem.

4.3.1.2 Remote Sensing Data Set

The second data set constitutes 15 multispectral remote sensing images provided by European Space Agency (ESA) in the scope of the Sen3Exp (Sentinel-3 Experiment) campaign [32]. The image dataset, as provided by ESA, was almost ready to be used. A final step of processing was carried out in order to place each pixel of the images in the dataset at its exact Earth location. This way, the images could be used together within a common grid map. In addition, information from the landuse of approximately 70% of the area under study in Barrax, Spain was delivered in the form of a GIS shapefile, from which a groundtruth image file associated with each of the corrected images was obtained. Thus, each pixel is associated with a landuse class.

The data were collected five times a day in 2009 on June 20, 22, and 24. In addition, any two consecutive images collected on the same day do not cover the same area: they overlap by 20-30%. Figure 4.2 shows the reconstructed color images of two consecutive images. As the data set contains images collected at different time points and different locations, it is helpful to study the shift of object's spectra in both temporal and spatial domains. The data set has 63 channels. Among them, channels number 22, 23, 58-63 are removed as they contain only noise. Thus, the number of channels used in our experiments is 55. The entire data are very large and contain many classes. However, not all classes appears in all images. We extract six classes Alfalfa, Bare soil, Barley, Corn, Fallow, and Harvested that appear in all images for our experiments.

4.3.1.3 Lung Data Set

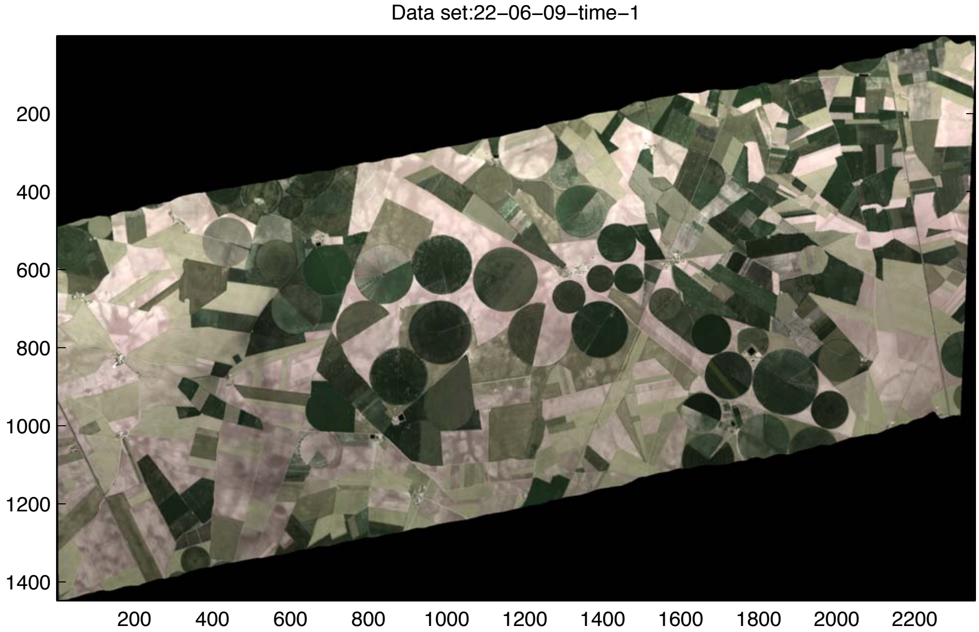
The third data set used to evaluate our approach is the lung data. The data set consists of 30 digitized, standard PA chest radiographs [70] obtained from a publicly available chest radiograph database JSRT [112]. Each image contains three classes: lung, rib, and background. As the radiographs are not calibrated, there might be a shift in intensity between images which is not related to the problem. In this experiment, different images are considered as different source domains. We used the N-jets feature [56, 104] of order six calculated with a scale $\alpha = 3$. Thus, the number of features for each pixel is ten.

4.3.2 Method Evaluation

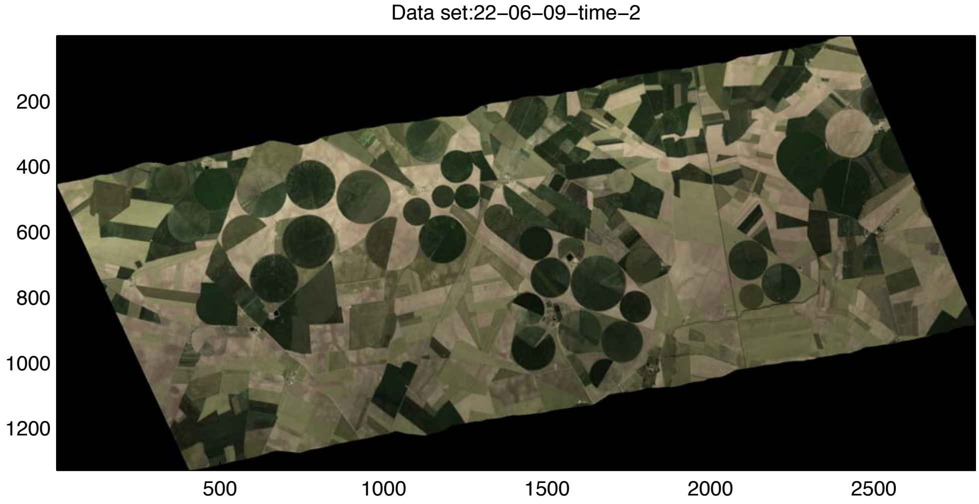
We compare the classification results between FIDOS and five other dimensionality reduction methods: TCA proposed in [86], SSA proposed in [125], and three commonly used methods PCA, classical Fisher, and Chernoff based feature extraction [72]. The first two methods are designed explicitly for domain shift problems. Whereas, the last three methods do not take into account the domain shift information, i.e., they assume that samples from training and test sets follow the same distribution.

Since the optimization problem induced by SSA is non convex, we use its approximated version called Analytic SSA [46] whose solution can be obtained in closed form. For comparison with the other methods, which all provide linear subspaces, we use a linear kernel for TCA, which corresponds to a linear transformation of the original space. **??** provides the explicit form we have used in our experiments.

FIDOS and TCA both require a tuning parameter. In our experiments, this parameter is determined by cross-validation as follows: If the training set contains multiple source domains ($N > 1$), each domain is considered as a fold (there are N folds in total). In each round, classifiers learnt from a training set containing data from $N - 1$ domains are evaluated on the remaining domain. If the training set contains only one source domain ($N = 1$), we divide the training set into five-folds and a standard five-fold cross validation is performed. At the end, the parameter value is chosen as the one that



(a) Reconstructed color image of the remote sensing data set collected at the first time on June 22, 2009



(b) Reconstructed color image of the remote sensing data set collected at the second time on June 22, 2009

Figure 4.2: Reconstructed color images of the two remote sensing data sets collected at two different times on June 22, 2009

yields the smallest average error rate over all folds. Cross-validation ensures that the parameter is selected at which the method is least influenced by a potential shift among folds, especially when each fold corresponds to a domain in the multiple source domain case.

Table 4.2: Error rates for the artificial example with respect to different number of shifted features

Methods	$k = 1$	$k = 5$	$k = 10$	$k = 15$
PCA	25.5	43.5	48.1	48.3
TCA	22.8	17.9	20.1	22.5
SSA	48.9	49.6	49.8	49.8
CHERNOFF	16.7	32.9	42.3	45.8
FISHER	15.3	32	42	45.7
FIDOS	5.5	2.0	0.8	0.5

We use two classifiers, 1-Nearest Neighbor (1NN) and Linear Discriminant Analysis (LDA) for the classification task. The two classifiers are selected as they are representative for parametric and non-parametric approaches in pattern classification. In the following, we show experimental results with respect to the six evaluated methods.

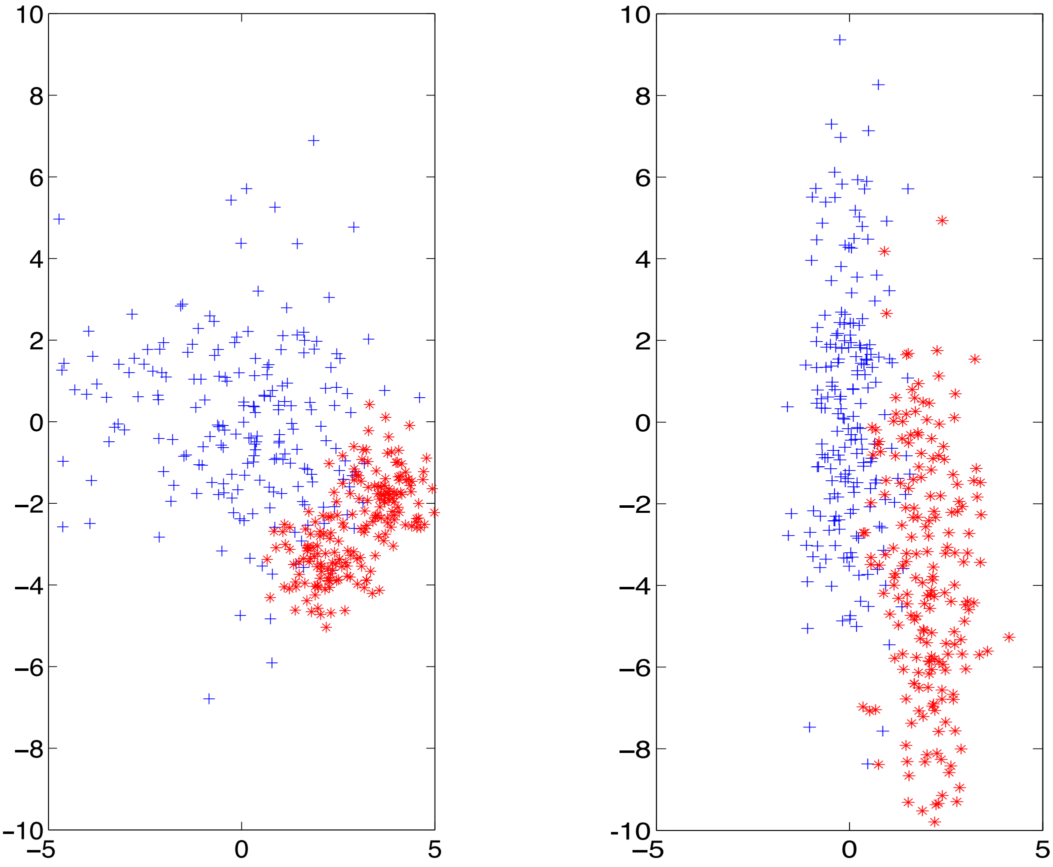
4.4 Experimental Results and Discussion

4.4.1 Artificial Data Set

For a target domain, a feature extraction method learnt from the two source domains is applied to extract one feature and then LDC classifier is used for the classification task. Performance of a feature extraction method is measured as the average classification error over 100 repetitions. To simplify the computation in this example, we use a fixed value for the tuning parameter in TCA and FIDOS that yields the best results in most cases. Specifically, the parameter is set to 1 in TCA and 0.1 in FIDOS.

Table 4.2 shows the results for the six feature extraction methods with respect to different numbers of shifted features (k = 1, 5, 10, and 15). FIDOS handles the domain shift in this artificial example well and constantly performs the best among all the methods tested. By taking into account the difference in distributions between training and target domains, FIDOS suppresses features by which the domains locate at different locations in the feature space. In addition, if by chance the training and target sets stay close to each other in the feature space, FIDOS still retains those features and uses them as additional discriminant information. As a result, increasing the number of shifted features does not deteriorate but instead slightly improve the performance of FIDOS.

In contrast, classical Fisher does not work well on this artificial example. As this method only tries to maximize the ratio between the between-class and within-class scatter matrices, the method is hampered by the shifted features, which coincidentally make classes far from each other such as those in Figure 4.3a. In addition, increasing the



(a)

(b)

Figure 4.3: Data scatter of a training set composed of two data sources. (a) Scatter plot of any two consecutive features from the first $2k$ features which are influenced by the shift problem and (b) Scatter plot of the last two features which are not influenced by the shift problem.

number of shifted features makes the problem more severe. When $k = 15$, the classifier based on classical Fisher feature extraction behaves like a random classifier with the error rate of 45.7%.

SSA does not perform well on this data set. The errors in all cases are larger than 48 %. As mentioned earlier, SSA is developed explicitly for multivariate time series applications in which the number of source domains is significantly large. For a small number of source domain situations, e.g. two source domains in our artificial example, SSA leads to spurious stationary solutions. TCA produces reasonable results in all cases. Similar to FIDOS, TCA also accounts for the differences between distributions across all source and target domains. Therefore, this method performs better than PCA and is not deteriorated by the increase in the number of shifted features. Nevertheless, TCA performs worse than FIDOS because TCA does not exploit the labeling information in the source domains.

4.4.2 Real Data

For the two real data sets, we distinguish two scenarios: i) when there is a single source domain; and ii) when there are multiple (five) source domains available in the training set. The second scenario is investigated to evaluate the performance of the methods when there might be a shift among the source domains themselves.

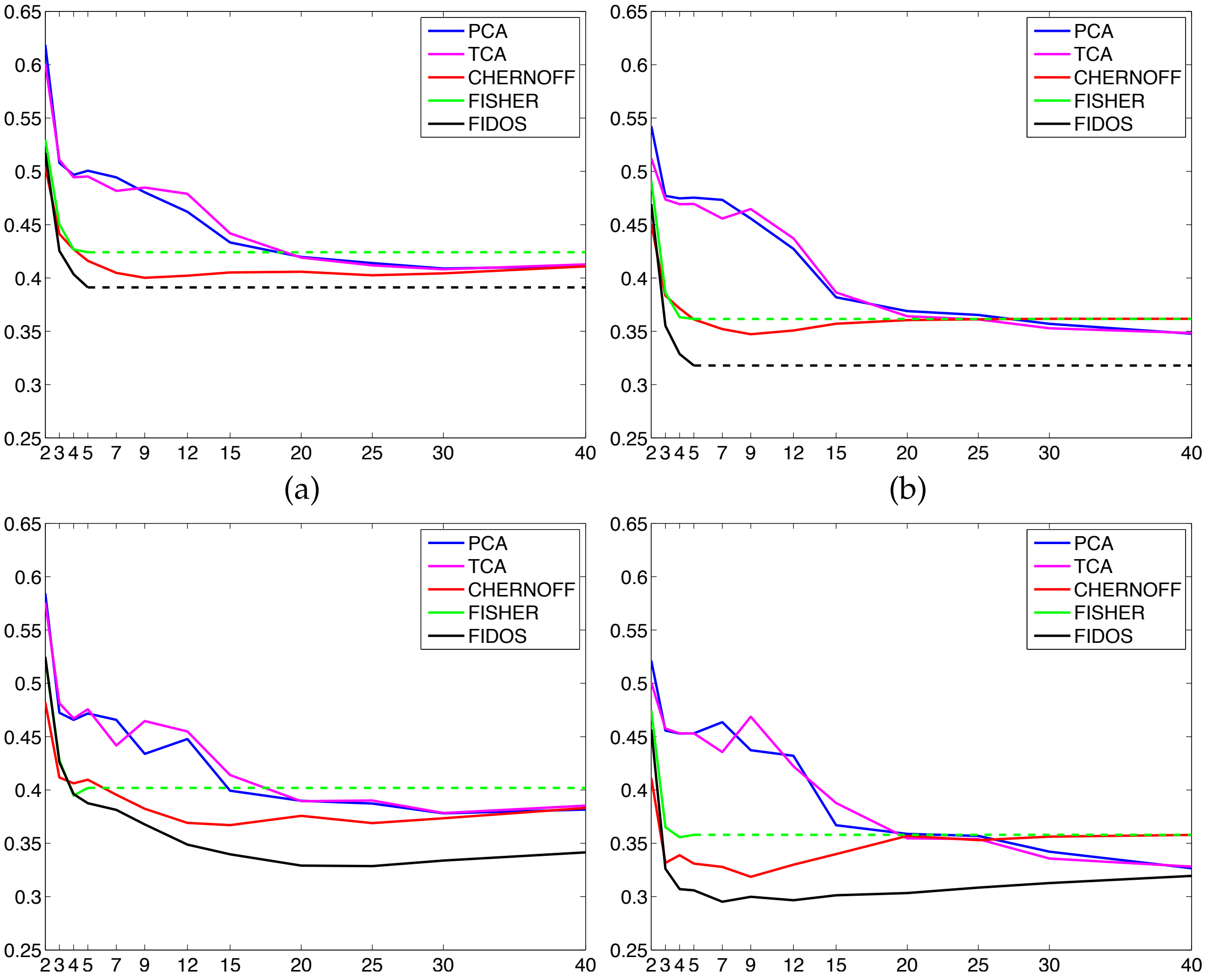
In the first scenario, each image is used for training and the other images are used for testing. The process is repeated for all images. In the second scenario, the data are split into several partitions, each partition contains five images. The number of partitions in the remote sensing and lung data sets is three and five, respectively. At each round, one partition is selected for training and the remaining partitions are used for evaluation. For both cases, the average classification error over all partitionings is then reported.

Figures 4.4 and 4.5 show the results for the remote sensing data and the lung data, respectively. The classification error (vertical axis) is plotted against the number of extracted features (horizontal axis). The first row presents the first scenario (single source domain) and the second row the second scenario (multiple source domains). Left and right panels display the 1-NN and LDC classifiers, respectively.

As stated in Section 4.3.1, we have six classes in the remote sensing data and three in the lung data. Thus, the maximum number of features classical Fisher can extract for the two data sets is five and two, respectively. In the case of single source domain, the maximum number of features extracted by FIDOS is equal to that extracted by classical Fisher. For the plots in Figures 4.4 and 4.5, if the number of features (horizontal axis) is larger than the maximum number of features FIDOS and classical Fisher can extract, we depict the corresponding error rates by dashed lines.

Figures 4.4 and 4.5 clearly show the advantage of FIDOS over classical Fisher. For the remote sensing data set, FIDOS outperforms classical Fisher with respect to both 1NN and LDC classifiers. For the lung data set, FIDOS produces similar results to classical Fisher when 1NN classifier is used and slightly better results when LDC classifier is used. The results on the real data sets confirm our conclusion based on the artificial data set that by taking into account the difference in distributions among domains, FIDOS is able to find features which are invariant with respect to different domains while still preserving the discriminant information.

FIDOS performs the best in all experiments using LDC classifier. When using 1NN classifier, FIDOS and TCA perform the best on the remote sensing and lung data sets, respectively. FIDOS performs substantially better than unsupervised methods, i.e., TCA and PCA, when a small number of features are extracted. This is because FIDOS aims at

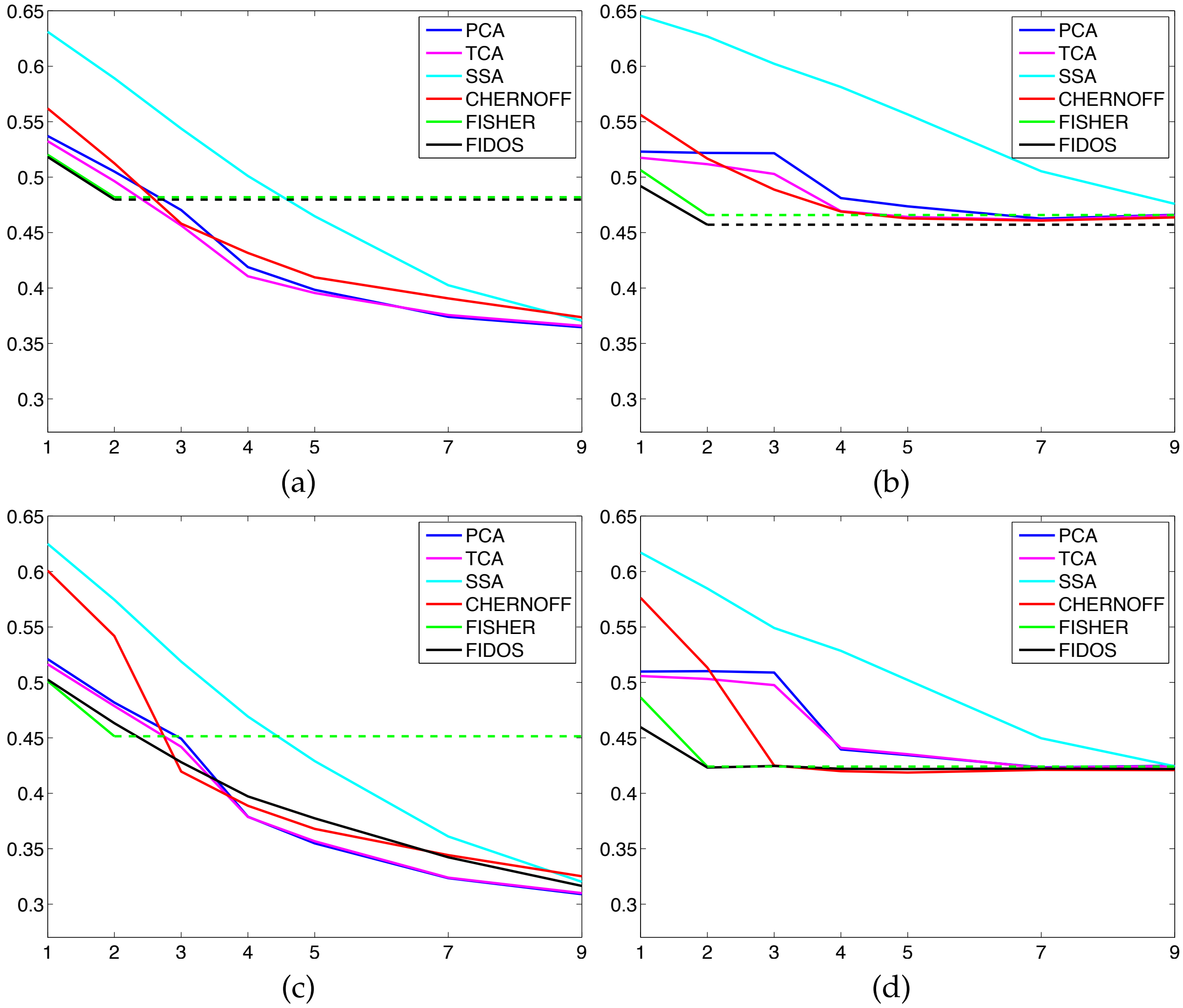


(a) (b) (c) (d)

Figure 4.4: Learning curves of different feature extraction methods for the remote sensing data set with respect to different configurations: a) single source domain and 1NN classifier; b) single source domain and LDC classifier; c) multiple source domains and 1NN classifier; d) multiple source domains and LDC classifier. For FIDOS and classical Fisher, if the number of features (horizontal axis) is larger than the maximum number of features that the two methods can extract, we depict the corresponding error rates by dashed lines.

maximizing the discriminant between classes; whereas, TCA and PCA aim at extracting the features containing high variance in the data. However, capturing high variance in the data does not necessarily lead to a better separability between classes. Similar to the artificial example, SSA performs the worst in all cases. The result for SSA on the remote sensing data set is not shown in Figure 4.4 since the errors for most numbers of extracted features are over 75%.

If multiple domains are included in the training set, FIDOS extracts more features than



(a) (b)

(c) (d)

Figure 4.5: Learning curves of different feature extraction methods for the lung data set with respect to different configurations: a) single source domain and 1NN classifier; b) single source domain and LDC classifier; c) multiple source domains and 1NN classifier; d) multiple source domains and LDC classifier. For FIDOS and classical Fisher, if the number of features (horizontal axis) is larger than the maximum number of features that the two methods can extract, we depict the corresponding error rates by dashed lines.

classical Fisher does. This helps FIDOS gain extra information and improve the classification performance. For example, using FIDOS, the minimum error rate on the remote sensing data set using 1NN and LDC classifiers is achieved when the number of extracted features is twenty and seven, respectively. Note that, the maximum number of features extracted by classical Fisher is five.

If there is a single domain in the training set, FIDOS only extracts the same maximum number of features as classical Fisher does. This might hamper FIDOS on data sets in which the number of classes is small, such as on the lung data set (Figure 4.5a). Therefore, how to extract more features in such a case encourages further investigation. One possibility is to increase the rank of the between-class scatter matrix, e.g. by using a nonparametric method such as the Nonparametric Discriminant Analysis (NDA) [40], or by dividing each class into subclasses and then measuring the difference between the subclasses of different classes [134].

In term of computational complexity, the subspace derived by the approximated version of SSA and the other five methods considered (including FIDOS) is obtained by solving a generalized eigenvalue problem. This requires $O(k^3)$ operations [96, Section 5.9], where k is the number of features in the original feature space. In addition, both FIDOS and TCA require additional computational cost to determine the tuning parameter, e.g. by cross-validation.

4.5 Conclusions

We have presented FIDOS, a generalization of the well known Fisher feature extraction method, to cope with the domain shift problem. FIDOS maximizes the between-class scatters and at the same time minimizes a convex combination of the within-class and between-domain scatters. To this end, FIDOS constructs a subspace that reduces the drift in the distributions across different domains whilst preserving the discriminants among classes.

The experiments on both artificial and real world data demonstrated that learning invariant feature with respect to the domains is essential to deal with domain shifts. FIDOS works better than classical Fisher feature extraction in all experiments. FIDOS also performs the best among the six methods evaluated on the artificial and remote sensing data, and produces comparable results on the lung data set.

FIDOS preserves advantage properties of classical Fisher, such as invariant to linear transformations, and computationally efficient. Furthermore, FIDOS in general extracts more features than classical Fisher. This helps the method gain extra information and thus improving the classification performance.

In this paper, we focused on linear transformations from the original feature space to tackle domain shift. Here, the drift in distributions across different domains, i.e., the between-domain scatter, is measured by the difference between the means of the distributions (Eq. (4.3)). We acknowledge that this might not fully describe the distribution drift across different domains as the mean is not always a good representation of the distribution of a domain in the original feature space. Nevertheless, similar to classical Fisher the final subspace in FIDOS is obtained by also including the data covariance structure (Eq. (4.4)), which also plays a role in representing the data distribution.

The measurement of the drift in domain distributions can be improved by, e.g. kernelizing the original feature space. In such a kernel feature space, the distribution drift can be defined as, e.g. the difference between the means in the reproducing kernel Hilbert space (RKHS), known as maximum mean discrepancy (MMD) [25, 51]. This may provide a better measurement of the distribution drift as in a certain implicit high-dimensional kernel space, one can simply use the mean to represent the distribution of a domain.

4.6 Acknowledgements

The authors would like to thank Marcel J.T. Reinders for useful discussions. We thank ESA for their kind support on providing the remote sensing data set acquired in the framework of the Sen3Exp campaign (Sentinel-3 Experiment). This research is supported by Carinthian Tech Research AG, Austria through the Austrian COMET funding program.

5

SEMI-SUPERVISED LEARNING

This chapter concerns with semi-supervised learning methods for object classification. Semi-supervised learning methods assume that there are only a few labeled samples but a lot of unlabeled samples available for the classification task. The main target is therefore to improve classification performance by using the unlabeled samples. We propose two strategies for semi-supervised learning.

Semi-supervised dissimilarity representation. In the dissimilarity representation approach, objects are represented by their dissimilarities with respect to a representation set instead of features. Up to now, the representation or prototype set has been usually selected from the training data. This limits the different aspects that can be captured, especially when the training data set is small. Based on the fact that it is not necessary to know the labels of samples used in the representation set, we investigate the dissimilarity representation in a semi-supervised setting where the objectŮs representation set is enriched by including also test data. This strategy is presented in Section 5.1.

Training sample selection for semi-supervised learning. Until now, most studies in spectral image classification focus on optimizing the classification performance given a training set generated by randomly selecting samples from the sample distribution. This selection strategy may be inefficient for problems containing unbalanced classes as it tends to select samples belonging to classes that are dominant in the sample distribution. We propose a new strategy to select training samples that are representative for the problem needed to solve. This strategy is presented in Section 5.2.

5.1 A study on semi-supervised dissimilarity representation

In the dissimilarity representation approach, objects are represented by their dissimilarities with respect to a representation set, rather than by features. Up to now, the representation or prototype set has usually been selected from the training data, limiting the different aspects that can be captured, especially when the training data set is small. This paper studies the performance change if the object's representation is extended by including also test data into the representation set in a semi-supervised setting. Experiments on a set of standard data show that the semi-supervised setting can substantially improve the performance of the dissimilarity-based representation especially for the small sample size problem.

5.1.1 Introduction

The dissimilarity representation is an approach in which objects are represented by their dissimilarities with respect to others in a data set. It is based on the idea that a class is constituted by objects having similar characteristics. The dissimilarity is small between objects of the same class and large between objects from different classes. Therefore, dissimilarities can be used as discriminant features for classification [29, 90]. The key advantage of the dissimilarity representation is that it provides a way to embed knowledge about the data structural information into powerful feature-based statistical approaches which are intensively available in machine learning and pattern recognition [29].

In the dissimilarity representation approach, the representation set is a set of objects, often called prototypes, to which other objects in the data set are compared. Based on the representation set, a dissimilarity space is constructed in which each dimension corresponds to the distances of all objects to a prototype. The representation set is traditionally selected as the whole training data set or a part of it. When the training set is sufficient large, selecting a small representation set is of interest since it can reduce the computational cost to compute the dissimilarity matrix. In addition, it is shown in [90] that classifiers, such as the quadratic discriminant classifier, when used with prototype selection usually perform better than the 1-NN classifier using the whole training data set.

The situation, however, is different if the training data set is small. The representation set selected from a small training set might miss important prototypes. Consequently,

---

Section 5.1 was published as Cuong V. Dinh, Robert P.W. Duin, Marco Loog. A study on semi-supervised dissimilarity representation, *the 21st International Conference on Pattern Recognition (ICPR)*, Tokyo, Japan, 2012.

it limits the different aspects that can be captured in the data and might result in an inadequate performance.

We aim at finding an improved data representation for the small sample size problem. A nice property of the dissimilarity representation is that it does not necessitate the availability of the labels of the objects in the representation set. In our approach, we enrich the representation set by including also unlabeled samples from the test set in a semi-supervised setting. The assumption we make is that the test set is available during the training process. Our experiments on several standard data sets demonstrate that including unlabeled samples into the representation set often improves the classification results especially for small training sample size. The semi-supervised dissimilarity representation is therefore useful for the situations in which to obtain labeled data is difficult or expensive.

5.1.2 Semi-supervised dissimilarity representation

Let T and S be the training and test sets. Let R be the representation set composed of k prototypes, $R = \{r_1, r_2, \ldots, r_k\}$. An object x belonging to T or S is represented by: $d_x = [d(x, r_1), d(x, r_2), \ldots, d(x, r_k)]$ in which $d(x, r_i), i = 1 \ldots k$, is the distance between x and the prototype r_i. $d(x, r_i)$ can also be considered as features of x in the constructed dissimilarity space.

In the supervised setting, the representation set is traditionally selected from the training set (i.e., $R \subset T$). In our semi-supervised setting, the representation set also includes objects from the test set (i.e., $R \subset \{T \cup S\}$). The training and test sets in both configurations are the same, only the object representation changes.

Our semi-supervised method for the dissimilarity representation can be categorized as the “Change of Representation" approach [10] which aims at enhancing the data representation using unlabeled data. We enhance the data representation by enlarging the representation set to capture different aspects of the data and thus providing more discriminative information for the classification task under consideration.

We investigate the performance of the supervised and semi-supervised settings in two scenarios:

- All of the available objects are used to build the representation set

$$R := \begin{cases} T & \text{in the supervised setting} \\ T \cup S & \text{in the semi-supervised setting} \end{cases}$$

- A set of informative prototypes is selected, e.g. by using a feature selection technique, from the available prototypes/objects. In this scenario, we examine whether feature selection benefits from the enlargement of its search space to unlabeled data.

Table 5.1: Data sets used in experiments

Data	Distance Measure	#Samples
Polygon	Modified Hausdorff	2 × 2000
38-haus	Hausdorff	2 × 1000
38-eucl	Euclidean	2 × 1000
Zongker	Template-Matching	10 × 200

5.1.3 Experiments

We use two base classifiers: the linear Support Vector Machine (LSVM) with a default tradeoff parameter value (C = 1) and the k-NN classifier with $k = 1$ (1-NN). We divide each data set into training and test sets of various sizes. At each time, the training set is selected randomly based on the data distribution. We repeat the experiments 150 times and average the classification results.

5.1.3.1 Data sets

We have selected standard data sets from two and ten class classification problems. The distance between object is measured in various ways. In this paper, we present four data sets: Polygon, 38-haus, 38-eucl, and Zongker as summarized in Table 5.1. The polygon data set [90] consists of two classes of randomly generated polygons: 2000 convex quadrilaterals and 2000 irregular heptagons. The polygons are first scaled and then their similarity is computed as the modified Hausdorff distances between their vertices. The Zongker data set comes from the NIST digits, originally given as 128 × 128 binary images [128]. The data set is composed of 10 classes, each class consists of 200 samples. The deformable template matching defined by [52] is used as the similarity measure. The 38-haus and 38-eucl data sets also come from the NIST digits but only consider the images of digits ′3′ and ′8′. Each digit class consists of 1000 samples. The similarity measures used in the two data sets are the Hausdorff and Euclidean distances, respectively.

5.1.3.2 Using all available prototypes

The results with respect to different training sizes using the supervised **(SU)** and semi-supervised **(SE)** settings on the four data sets are shown in Figure 5.1. Averaged error rate over 150 repetitions (vertical axis) is plotted against the training set size (horizontal axis) by dashed line for the supervised setting and by solid line for the semi-supervised setting. Red and black display the results for LSVM and 1-NN, respectively.

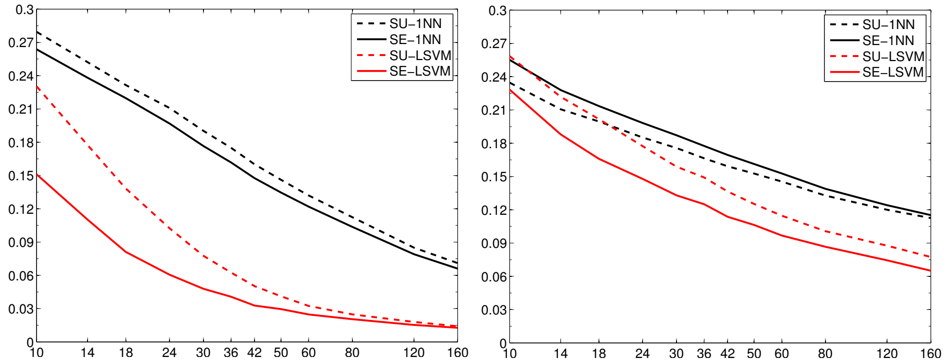
The plots show that when the LSVM is used, the semi-supervised setting often outperforms the supervised setting. The semi-supervised setting performs better than the supervised setting for three data sets Polygon, 38-NIST using Hausdorff distance (38-haus), and Zongker but worse for the 38-NIST data set using the Euclidean distance (38-eulc). The difference in performance between the two settings manifests clearly if the training set size is small. For example, the semi-supervised setting yields a decrease of 14% error rate compared with the supervised setting if the training set size is 30 for the Zongker data set, and a decrease of 8% error rate if the training set size is 10 for the Polygon data set. This verifies our statement that in the case of limited training data, objects are better described by including more prototypes from the test set into the representation set. When the training set is sufficient large, the two settings yield similar performance since the representation set in the supervised setting is large enough to describe the data.

When the 1-NN classifier is used, the semi-supervised setting works slightly worse than the supervised setting for the 38-NIST data set using the Hausdorff or the Euclidean distance. This behavior may be due to the rather high-dimensionality of the dissimilarity space with which the 1-NN has difficulty in dealing [1]. In such a situation, employing a feature selection step might improve the performance of the classifier. This is indeed demonstrated in Section 5.1.3.3.

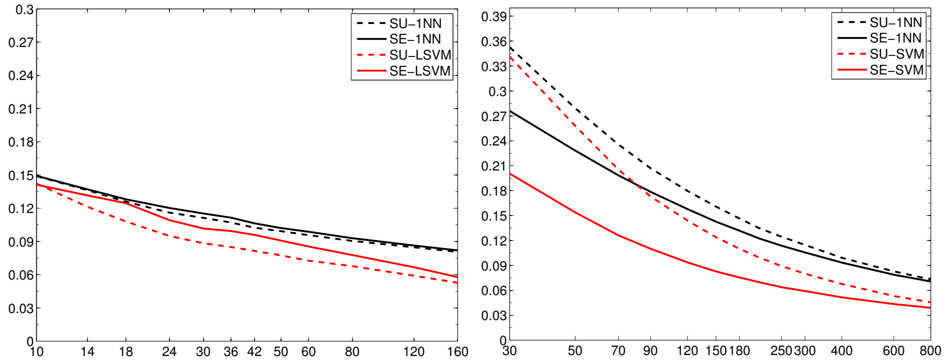
The advantage of the semi-supervised setting is further demonstrated by varying the size of the training set and the representation set. Figure 5.2 presents the results for the Polygon data using the LSVM. The horizontal axis shows the size of the representation set; colors represent the size of the training set. Note that if the size of the representation set is larger than or equal to that of the training set, then the representation set first includes all samples from the training set and the rest is randomly selected from the test set. If the size of the representation set is smaller than that of the training set, then samples of the representation set are randomly selected from the training set. Thus, the supervised setting is equivalent to the case in which the size of the representation set is equal to the size of the training set. As shown in Figure 5.2, for a fixed training set size, increasing the representation set size leads to improvement in the classification result. The results become stable when the representation set size is large, i.e., larger than 400 or 10% of the whole data set in this case. It is worth noting that it is unnecessary to select all samples for representation, for this Polygon data set, just 10% of the data is enough to achieve good classification result.

5.1.3.3 Selecting prototypes from the available set

We use the prototype selection, which employs linear programming, for dissimilarity-based classifiers as presented in [90]. This method recasts the prototype selection as a classification problem that aims at determining a two-class sparse linear classifier. The prototypes associated with the non-zero weights of the classifier are then selected for the “final" representation set.



(a) Polygon (b) 38-haus



(c) 38-eucl (d) Zongker

Figure 5.1: Classification results for the four data sets using the LSVM and 1-NN classifiers. SU and SE mean supervised and semi-supervised settings respectively.

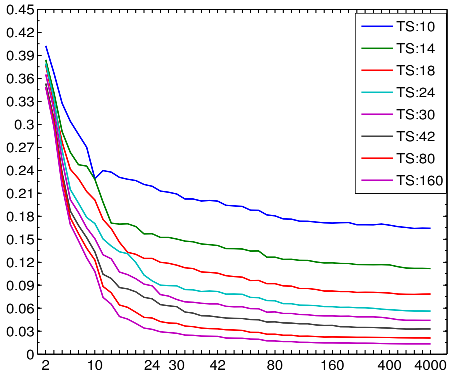


Figure 5.2: Classification results on the Polygon data with varying representation set size and training set size.

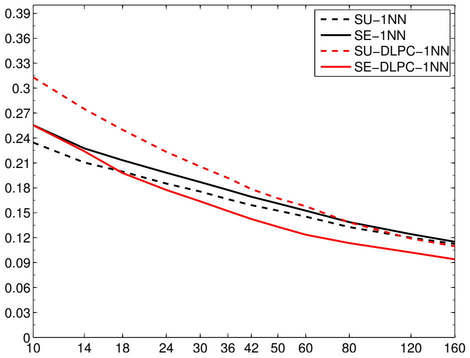


Figure 5.3: Classification results for the 38-haus data using 1-NN classifier with and without prototype selection.

Figure 5.3 shows the classification results for the 38-haus data set where we first employ prototype selection and then 1-NN classifier. The average error rate (vertical axis) is plotted against the size of the training set (horizontal axis) as in Figure 5.1. Results with and without prototype selection are displayed in red and black, respectively.

The semi-supervised setting without prototype selection implies that all available prototypes (objects) are used for the representation. As shown in the figure, prototype selection leads to better classification result. In the supervised setting without prototype selection, just the training set is used for the representation. On the contrary to the semi-supervised setting, the classification result with prototype selection is far worse than without prototype selection if the training set size is less than 100. It is because the small representation set makes it difficult for the prototype selection method to select “good" prototypes for the “final" representation set. It should be noted that if the training set size is larger than 17, the 1-NN classification with prototype selection under the semi-supervised setting performs best.

5.1.4 Discussion and Conclusions

This paper shows a study of dissimilarity-based classification where the representation set is enlarged by including samples from the test set. As a result, the representation set provides a richer description for the objects of interest and helps to significantly improve the classification performance, especially for small sample size problem. Therefore, the semi-supervised setting is helpful for situations in which the availability of labelled data is limited.

If all available objects are included into the representation set, the computation of the distance matrix might, however, become expensive, or in other ways prohibit the direct use of the dissimilarity method. Nevertheless, it is shown in the experiments that in many cases using a subset of the data for the representation provides as good classification result as using all data (cf. Figure 5.2). Prototype selection methods might be used to select such a representation set (Figure 5.3) and, in many cases, still improve upon a representation set merely derived from the training data.

We, however, note that the semi-supervised setting does not always perform better than the supervised setting, e.g. for the 38-eucl data set. It is of future interest to further investigate in which situations the semi-supervised setting does help dissimilarity-based classifiers improve upon the standard supervised setting.

5.2 Semi-supervised hyperspectral pixel classification using clustering-based mode selection

A semi-supervised pixel classification scheme for hyperspectral satellite images is presented. The scheme includes a previous band selection step followed by a clustering process to select modes of interest that will be labeled by the expert. Then pixel classification is performed resulting in a segmentation and classification of the fields appearing in the image. Thanks to the previous clustering step the most suitable pixels are automatically selected to build the classifier. This reduces the expert effort required since less pixels need to be labeled. Pixel classification accuracy obtained outperforms the results of a random selection scheme where more pixels were labeled.

5.2.1 Introduction

Segmentation is a noted non-supervised issue in image analysis research. Lately, this task has also been faced as a semi-supervised task in which experts provide labeled samples that the system can used to classify the pixels as well as to segment the image. To this end pixel classification is widely used but results still need of additional information or process. In this direction, authors have tried to describe the neighborhood of the pixel using spectral/spatial features [99]. Others methods use MRF [2] suffering from the problem of setting a fixed shape. In [81] an adaptive neighborhood was defined to face this problem. Another popular strategy is to define a classification scheme that introduces a previous segmentation task [132] or a post-process improvement [130]. But in all cases training sets are picked randomly over the data set. It is always a drawback to reduce the size of the training set since randomly distributed pixels can lie in non interesting areas and consequently, classes can be missed. On the contrary, the expert action is expected to be minimized in the labeling of the training samples. In this scenario the most interesting samples from the system point of view should be provided to the user instead of the randomly selected ones or the user criteria selection. Tarabalka et al. introduced this idea in [131] focusing in the phase after the pixel classification. This paper introduces a semi-supervised classification scheme aimed at decreasing the training samples before the classification task is performed.

Clustering algorithms analyze the feature space in order to group samples around a representant called mode. Thanks to nonparametric clustering techniques a feature space

---

Section 5.2 was published as O. Rajadell-Rojas, P. Garcia-Sevilla, C. V. Dinh, and R. P. W. Duin. Semi-supervised hyperspectral pixel classification using interactive labeling, *3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers)*, Lisbon, Portugal, 2011. This work is mainly done by Olga Rajadell Rojas; the Ph.D. candidate shares a co-authorship credited for his contribution to discussions during the development of the method and analyzing the results, and reading/correcting the manuscript.

can be analyzed finding their modes in a non-supervised way. In this paper the random selection of samples meant to train the classifier is suggested to be changed for the modes resulting from a clustering process of the samples. This non-supervised selection makes training samples suitable for posterior non-linear classification using a k-nearest neighbor rule.

The chosen clustering method is presented in Section 5.2.2. The database and the features to characterize the database are presented in Sections 5.2.5 and 5.2.3, respectively. Section 5.2.6 presents and discusses the experiments.

5.2.2 Mode seek clustering

Given a hyperspectral image, all pixels can be considered as samples which are characterized by their corresponding feature vectors. The set of features defined is called the feature space and samples (pixels) are represented as points in that multi-dimensional space. A clustering method groups similar objects (samples) in sets that are called clusters. The similarity measure is defined by the cluster algorithm used. A crucial problem lies in finding a good distance measure between the objects represented by these feature vectors. Many clustering algorithms are well known. Among them, K-means is a widely used technique due to its ease of programming and good performance. However, k-means suffers from several drawbacks; it is sensitive to initial conditions, it does not remove undesirable features for clustering, and it is optimal only for hyper-spherical clusters. Furthermore, its complexity can be impractical for large data sets [27]. For such reasons a k-NN modeseeking method is used in this paper. It selects a number of modes, but this number cannot be set. Instead, a neighborhood parameter (s) should be provided, this number controls the amount of modes. For each class object x_j , the method seeks the dissimilarity to its s^{th} neighbors. Then, for the s neighbors of x_j , the dissimilarities to their s^{th} neighbors are also computed. If the dissimilarity of x_j to its s^{th} neighbor is minimum compared to those of its s neighbors, it is selected as prototype [13]. Note that the s parameter only influences the scheme in a way that the bigger it is the less clusters the method will get since more samples will be grouped in the same cluster, that is, less modes will be selected as a result.

5.2.3 Feature extraction

Pixel characterization aims at obtaining one feature vector for each pixel to be used in a pixel classification task in a multidimensional space. When only spectral data is used the feature vector for every pixel is defined as the spectral curve provided by the sensor.

In order to describe the context of a pixel several features have been suggested in the literature [91]. In this paper Gabor filtering will be used as suggested in [99]. In this case, features are obtained by filtering the input image with a set of filters. The set of outputs obtained for each pixel in the image forms its feature vector. Here, the filter bank is defined to be a set of two-dimensional Gabor filters. Each Gabor filter is characterized by a preferred orientation and a preferred spatial frequency (scale) and consists of sine and cosine functions modulated by a Gaussian envelope.

5.2.4 Semi-supervised classification

Here the proposed semi-supervised pixel classification scheme is presented. The scheme proceeds as follows:

1. In order to reduce the number of spectral bands to be used, a set of spectral bands, given a desired number, is selected by using the band selection method proposed in [79].

2. Clustering procedure is applied over the selected spectral bands. An improvement in the clustering process is included by adding as features the spatial coordinates of each pixel in the image. This provides a spatial component suitable for clustering since it is based on distance between samples.

3. The modes resulting from the previous step define the training set for the next step. The expert is involved in this point by providing the corresponding labels of the selected samples. Here the expert is simulated by checking the labels in the ground truth provided for only those samples.

4. A k-NN classifier with $k = 1$ is build with the training set defined above. Note that at this point the spatial coordinates are dismissed as features. The clustering is always performed over the spectral domain but once the modes are obtained the features to be used for the classification step can be the same or changed. In this paper classification step changing the space to spectral/spatial features is also tested.

The parameter s of the clustering algorithm can be tuned to obtained a higher or lower number of interesting points to be labeled. The increase of this parameter is inverse to the number of modes found. As demonstrated later, the number of modes has a direct impact on the performance of the classification but still the results are better than the ones obtained using a random selection.

5.2.5 Data set

A widely used hyper-spectral database has been used. Hyper-spectral image data 92AV3C was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. From the 220 bands that composed the image, 20 are usually ignored because of the noise (the ones that cover the region of water absorption or with low SNR) [62]. The image has a spatial dimension of 145 × 145 pixels. Spatial resolution is 20m per pixel. In it, three different growing states of soya can be found, together with other three different growing states of corn. Woods, pasture and trees are the bigger classes in terms of number of samples (pixels). Smaller classes can be also found such as steel towers, hay-windrowed, alfafa, drives, oats, grass and wheat.

5.2.6 Experiments, Results and evaluation

In Figures 5.4 and 5.5 the performance of the semi-supervised classification scheme is compared with the traditional random selection and classification process. Results are shown as learning curves where error rate is represented as a function of the number of samples used for training. In Figure 5.4 learning curves for different number of spectral features are presented together with the corresponding learning curve when the same amount of pixels are selected at random. It is noticeable that in all cases, when selecting the training set, the classification rate outperforms the result when it is picked at random. The gain reaches 0.3 when a smaller training set is used and decrease to 0.15 when the training set grows, obviously because when the size of the training set grows, random selection has more chances to select samples from all different areas. Also, note that no advantage is obtained in involving a higher number of spectral bands in the process. If the number of spectral bands used to performed the clustering step is fixed to 10, similar conclusions can be obtained from Figure 5.5 where the number of spectral/spatial features is increased in this case. Using more than 24 features leads to higher computational complexity with no performance increase. As a summary also the difference in the error rate between using spectral and spectral/spatial features for classification can be observed in Figure 5.6. Again, in both cases random selection yields higher error rate than the mode selection method. It is remarkable that both kind of features start around the same rate but the difference is quickly introduced when more samples are included and the error rate when using spectral/spatial features decrease considerably.

Showing the results over the image ground-truth, Figures 5.7 and 5.8 show the classification results projected against the image ground truth using 24 spectral/spatial features for the classifying step, when 23 and 104 training samples are selected respectively. In (a) misclassified pixels are represented in white color whereas the rest of the image represent well classified ones and (b) training pixels are presented in white over the ground-truth of the image. In both images background is the black area surrounded the classes and it is considered a non interesting heterogeneous area. It is noticeable that small classes are missed in the mode selection, that means that clustering method cannot detect those areas as independent ones. As a consequence of having no training sample available for that class, classification dismisses it all. As it can be expected, the smaller the number of clusters is, the higher number of small classes are missed. Nevertheless, where a sample is selected, a big area is well classified due to the usage of spectral/spatial features. Figure 5.7 stands for an error rate of 0.41 using only 23 samples as training set. Note that only samples from 10 different classes are selected leading to miss 7 classes missed. However in Figure 5.8, using 104 training samples, the number of modes increases, 15 classes are included in the training set and the error rate decreases to 0.147.

The results may not seem significant in terms of figures. In [34] classification rates reached 95% when the training set size was fixed to 5% of the labeled pixels. There, all spectral bands were used and small classes were dismissed, that is, a 9-class problem was faced. In [99] the 16-class problem was tackled and a smaller number of bands was used but still 5% of the labeled data set was needed to obtain an accuracy of 92%. Note that when random pick is performed a priori probabilities of classes are used so all classes are represented in the training set. Here the 16-class problem is faced with a reduced training set. 104 samples stands for the 1% of the data set. With the selection mode suggested in this paper, an accuracy of 96% can be obtained with 337 modes with only 3.2% of the labeled pixels and using only 3 spectral bands.

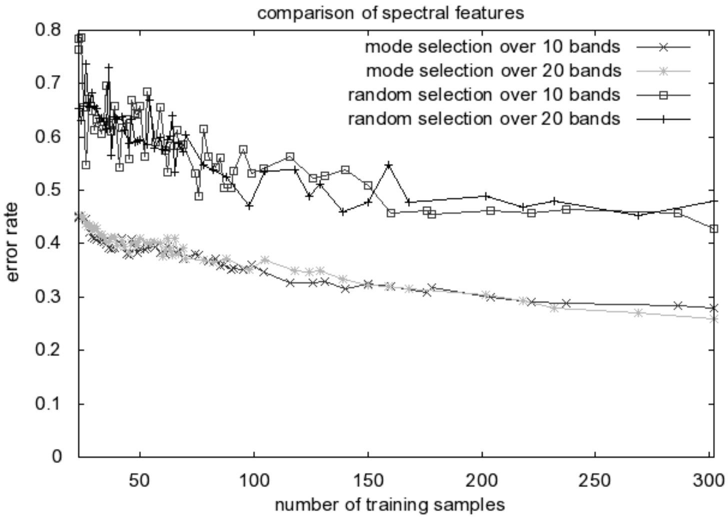


Figure 5.4: Learning curves for different number of spectral features comparing the classification results between two scenarios: using mode selection to select training samples and random selecting training samples.

5.2.7 Conclusions

A semi-supervised segmentation and classification scheme have been suggested. Thanks to the mode selection performed by the clustering process, training samples are selected and only interesting samples are labeled by the expert. In this sense their collaboration is reduced while performance increases in comparison with random selection and classification. Using a clustering method makes the result suitable for classifying with a simple nearest neighbor rule obtaining fairly goods results when fewer initial information is provided. The process is computationally feasible since it has been shown

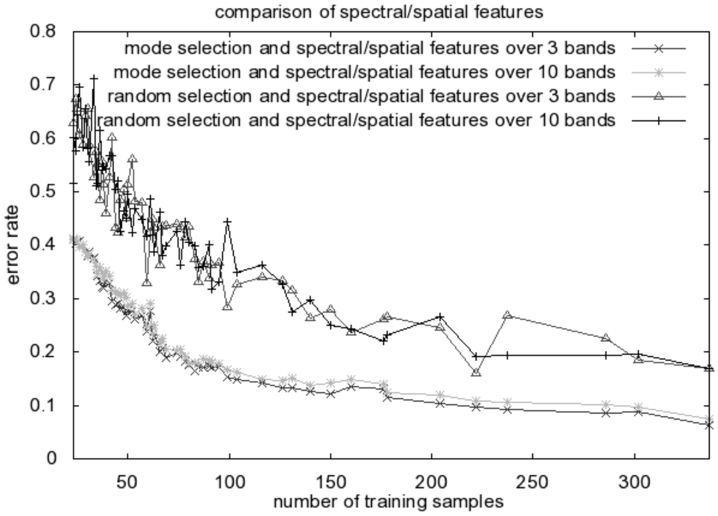


Figure 5.5: Learning curves for different number of spectral bands using spectral/spatial features comparing the classification results between two scenarios: using mode selection to select training samples and random selecting training samples.

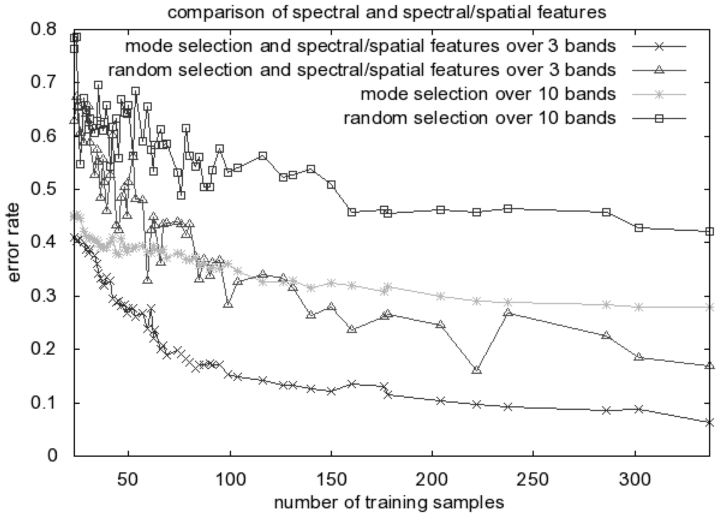
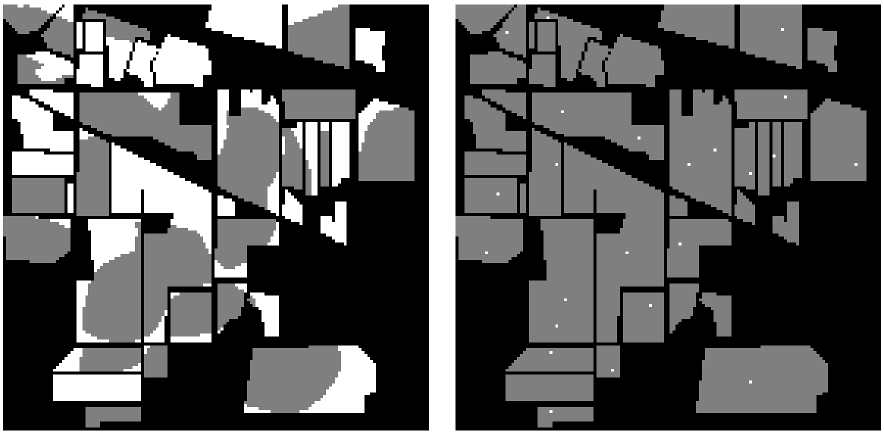


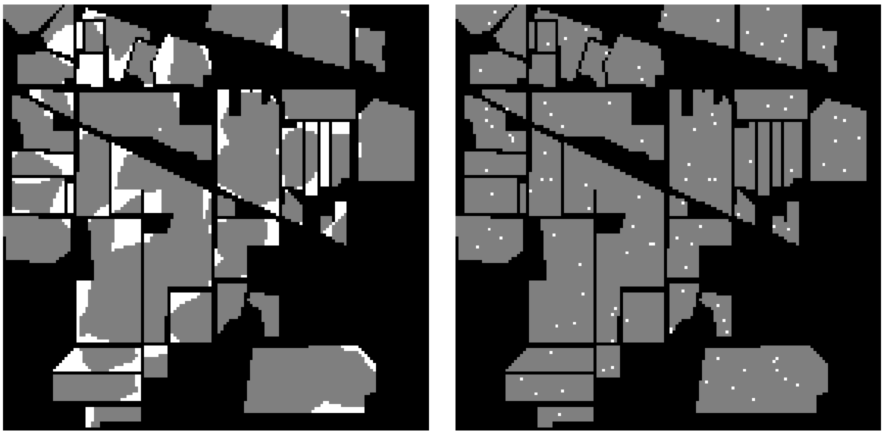
Figure 5.6: Learning curves resulting from selecting the training set and the corresponding number of training samples picked at random for spectral and spectral/spatial features.

that neither all spectral bands nor a high number of features was needed in our experiments. However, small classes may be missed by the clustering procedure and then dismissed in the classification step. To tackle this problem the clustering step should be improved and probably a post-processing technique could also be of interest.



(a) (b)

Figure 5.7: Classification results using 24 spectral/spatial features and 23 selected training samples. (a) representation of misclassified pixels in white and (b) training samples shown in white. Error rate was 0.41.



(a) (b)

Figure 5.8: Classification results using 24 spectral/spatial features and 104 selected training samples. (a) representation of misclassified pixels in white and (b) training samples shown in white. Error rate was 0.147.

6

DISCUSSION

6.1 Overall Conclusion

This thesis investigated two main challenges in spectral imaging analysis: data visualization and spectral classification in the case of a small sample size setting.

For the first challenge, we proposed in Chapter 2 a method to locate objects of interest putting emphasis on edge detection. We focused on the detection of boundaries of objects that are embedded in background clutter or just appearing in a few bands. These situations often happen with spectral images and the detection results might contribute significantly to the success of subsequent applications such as target detection in spectral images.

Spectral classification was exploited using two recently used approaches in machine learning, namely, transfer learning and semi-supervised learning. To this end, four methods have been proposed to (i) select an appropriate training set when learning from multiple source domains (Chapter 3), (ii) extract features in case of a domain shift problem (Chapter 4), (iii) make use of unlabeled samples when classifying using a dissimilarity representation (Chapter 5.1), and (iv) select training samples in a semi-supervised manner (Chapter 5.2).

It is worth mentioning that Chapters 3 and 4 focus on transferring knowledge when considering one or several source domains that possibly match a new target domain. To the best of our knowledge, we are among the first in spectral imaging community to handle multiple source domains. So far most studies in spectral imaging have been devoted to the situation in which training and testing sets are from the same domain. We, however, do think that domain shift is an important issue within spectral imaging and that classification where training and testing sets are from different domains should receive further attention as this manifests the generalization ability of a classification method. In addition, from a practical point of view, this also saves labeling cost needed to perform classification task on a new domain.

It should also be noted that the method proposed in Chapter 2 was developed exclusively for spectral imaging data. Whereas, all the methods presented in Chapters 3 - 5 provide general schemes for transfer learning and semi-supervised learning. They can be applied not only to spectral imaging data but also to other kinds of data. In Chapter 4, we have indeed shown the results on MRI-based medical images in addition to remote sensing data.

In the following we discuss possible extensions of the presented work and indicate directions for future work.

6.2 Outlook

6.2.1 Data visualization

In Chapter 2, we have presented a method for edge detection in multispectral images. In our method, we calculate the edge strength of a pixel by using an ensemble clustering algorithm in the gradient magnitude feature space. This enables the use of globally statistical information from the image to reveal the saliency corresponding to each pixel. However, the spatial connectivity between neighboring pixels is currently not taken into account. Therefore, applying a thresholding algorithm on the resulting edge strength map may lead to discontinuing edges. How to incorporate spatial connectivity information (into the clustering process) needs further investigation.

In addition, we have shown that saliency is an important property of edges in images. Other features in an image such as corners, junctions, and blobs also have this property. Constructing suitable feature spaces to detect these features in a hyperspectral images is also an interesting topic for future research.

6.2.2 Data representation for object classification

In most chapters of this thesis, objects are represented as a vector composed of spectral responses at different wavelengths. We used this simple representation in our studies just to better demonstrate our strategies/methods for analyzing spectral data, such as how to utilize labeled samples from related data domains, or, unlabeled data samples to improve the classification results. However, we believe that more sophisticated data representations might also benefit the classification. In the following, we discuss potential features in both spectral and spatial domains that can be used to enrich the data representation.

Spectral shape information. As mentioned in Chapter 1, hyper-spectral images provides a significantly smoother object's spectrum than multi-spectral images and reflects better the “spectral signature" corresponding to each object. This makes it possible to represent and classify objects in hyperspectral images with respect to their spectral shape signatures [111]. It is shown in [85, 94] that good classification performances are achieved by using the spectral shape information. In their studies, each spectrum is represented by its derivates between adjacent bands. The shape dissimilarity between two spectra is then measured as the sum of the absolute differences between their derivatives. Therefore, it might be of interest to examine if applying spectral shape information to the spectral data used in this thesis would yield better classification results.

Contextual information. Recent advances in sensor technologies allow for high resolutions not only in the spectral domain but also in the spatial domain. For example, it is possible to distinguish small spectral classes like trees in a park or cars on a street in some remote sensing data sets [9, 17, 93]. High spatial resolution also increases the correlation between spectral responses of neighboring pixels. Thus, there is a need to incorporate contextual information, such as shape, texture, and spatial arrangement of pixels, into the analysis of hyperspectral imaging.

In computer vision, contextual features have been studied intensively for grayscale and color images. In a simple form, each pixel can be represented by a set of its neighbor grayscale values. In a more complicated form, each pixel can be represented by its Gaussian derivatives at different scales. An example is the n-jets feature extraction method [56, 104], which is widely used in medical imaging. It is therefore interesting to see how these contextual representations can be applied to spectral imaging.

One direction is to represent each pixel in a spectral image by a set of contextual features concatenated over all bands. It might happen that the number of features becomes too large when the number of bands grows. In such a case, one might need a band selection method [24, 79], as a first step, to select a subset of informative bands.

Another direction is to consider each band as an independent source. A base classifier is applied to the image at each band in which each pixel is represented by its contextual features. Then classification combination techniques can be used to fuse the results over all bands. A simple example is to apply combining rules, such as max, average, and product rules, to the estimations of the base classifiers. This might be potentially helpful to reveal important information, such as which bands are useful to discriminate a particular class from others.

6.2.3 Soft training set selection for transfer learning

Chapters 3 and 4 study the issue of transferring knowledge from a set of related source domains to a new unseen target domain. We have yet not investigated the situation in which there are also a few labeled samples available in the target domain. How to combine labeled samples from both target and source domains is then a question of interest.

A potential approach is to assign different weights to samples from different domains, e.g. samples from a domain will obtain a weight proportional to the similarity between this domain and the target domain. Thus, samples from the target domain have the largest weight. Such an approach can be considered as an extension of the training set selection method proposed in Chapter 3. There, a labeled sample from a source domain has a “hard" assignment, i.e., it is either included into or excluded from the final training set. Instead one would allow each sample to have a “soft" assignment, i.e., a sample is included into the final training set with a weight ranging from 0 to 1 corresponding to its relatedness with the target data.

6.2.4 Extended dissimilarity representation

Chapter 5 shows a study on semi-supervised dissimilarity representation where unlabeled samples from the current classification task are included in the representation set. Unlabeled samples can be used as prototypes because a dissimilarity presentation does not necessitate the availability of the labels of the objects in the representation set. Therefore, one might further extend the representation set by including also background samples, i.e., samples from other classes that are not considered in the current classification task.

For instance, to classify the images of two digits 2 and 8, the representation set can be extended to also include images of other digits and even some letters (from ’A’ to ’Z’). Background samples, e.g. digit 3, may reflect different levels of similarity with respect to the two digits 2 and 8. These differences might provide important discrimination information to classify the two digits. Thus, it is fascinating to investigate whether (and to what extent) “foreign" or background samples help to strengthen the object description in the dissimilarity representation.

To conclude, the work presented in this thesis was developed to facilitate the analysis of spectral images by making use of pattern recognition techniques, on the one hand to improve visualization, and on the other hand, to directly solve classification problems. There is ample room for future extensions with respect to both the data representation and the classification methodology. From the data representation point of view, one might exploit other information, such as the spectral shape, or contextual information, to represent spectral objects. Objects can also be represented in the extended dissimilarity space in which background samples are included into the representation set. From the methodological point of view, transfer learning might benefit from leveraging a soft training set selection strategy. It would be interesting to implement and investigate these extensions in future research.

APPENDICES

A. Statistics-based Edge Detection in Multispectral Images

This Supplementary contains the binary edge maps generated by the three edge detection methods on the two real data sets using different threshold values. The threshold T represents the percentage of edge pixels in the binary edge map.

Figure S1 shows the binary edge results on the SEM/EDX data when T ranges from 6% to 9%. Similarly, figure S2 shows the binary edge results on the Scene data when T ranges from 6% to 15%.

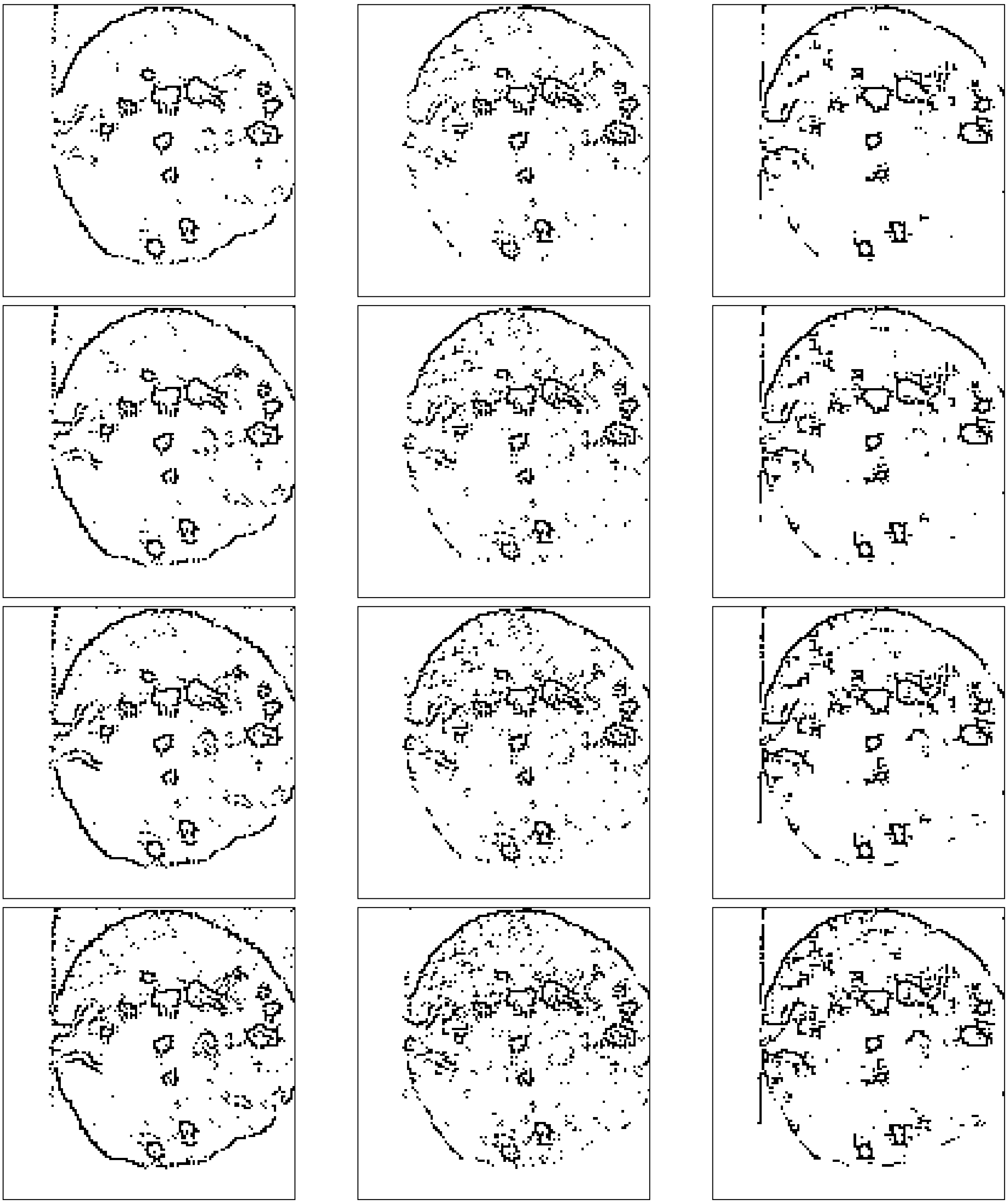


Figure S1: Binary edge maps generated by SEDMI (first column), Di Zenzo's method (second column) and the RCMG method (last column) for the SEM/EDX data set. Rows from 1 to 4 correspond to threshold *T* from 6% to 9%, respectively.

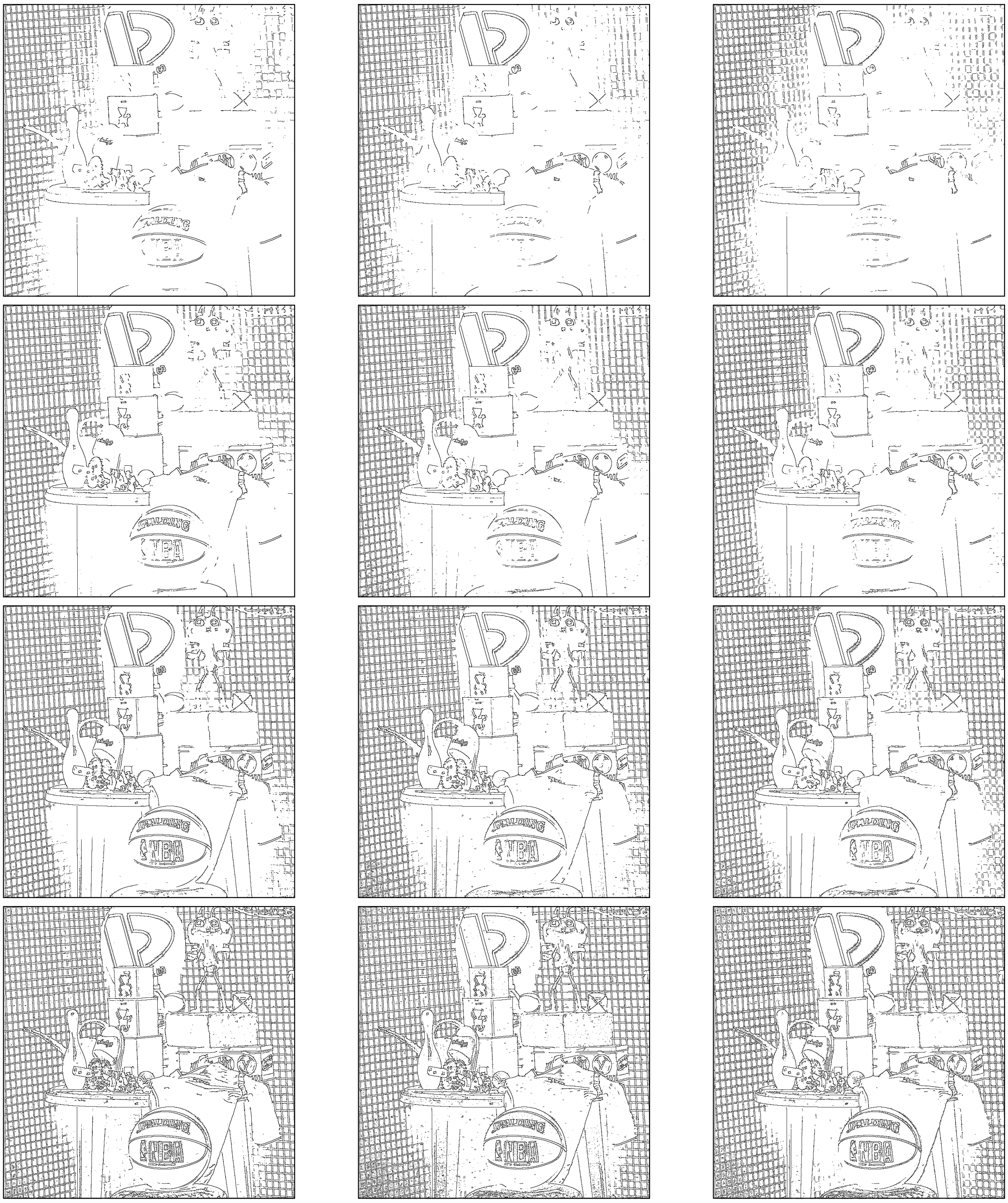


Figure S2: Binary edge maps generated by SEDMI (first column), Di Zenzo's method (second column) and the RCMG method (last column) for the Scene data set. Rows from 1 to 4 correspond to threshold *T* values: 6%, 9%, 12%, and 15%.

B. Approximated version of the linear kernel TCA

Denote by X_S and X_T the data from source and target domains, respectively. X is the set of data from both domains $X = [X_S, X_T]$. The number of samples in X_S, X_T, and X are n_s, n_t, and n, respectively. TCA tries to solve the following optimization $\min_W \left\{\mathrm{tr}(W^T KLKW) + \mu * \mathrm{tr}(W^T W)\right\}$ s.t. $W^T KHKW = I$ in which $H = I - \frac{1}{n_1+n_2} 11^T$ is the centering matrix. L is a matrix in which

$$L_{ij} := \begin{cases} \frac{1}{n_1^2} & \text{if } x_i, x_j \in X_S \\ -\frac{1}{n_1 \times n_2} & \text{if } x_i \in X_S \text{ and } x_j \in X_T \\ \frac{1}{n_2^2} & \text{if } x_i, x_j \in X_T. \end{cases}$$

K is the kernel matrix of X. When a linear kernel is used, $K = X^T X$. The optimization can be rewritten as:

$$\min_W \left\{\mathrm{tr}(W^T X^T XLX^T XW) + \mu * \mathrm{tr}(W^T W)\right\} \tag{.1}$$

subject to

$$W^T X^T XHX^T XW = I. \tag{.2}$$

Note that XHX^T is equal to the total scatter (S_T) of X because

$$\begin{aligned} S_T &= \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = (X - \frac{1}{n} X11^T)(X - \frac{1}{n} X11^T)^T \\ &= X(I - \frac{1}{n} 11^T)(I - \frac{1}{n} 11^T)^T X^T = XHH^T X^T = XHX^T, \end{aligned} \tag{.3}$$

where x_i and μ are a sample and the mean of X, respectively; n is the number of samples in X.

It should be also noted that XLX^T is the between-class scatter matrix (S_B) when considering X_S and X_T as two different classes. Let L_S be the labeling with respect to the first class, $L_{Si} = 1$ if x_i belongs to the source domain, otherwise $L_{Si} = 0$. Let L_T be the labeling with respect to the target class, $L_{Ti} = 1$ if x_i belongs to the target class, otherwise $L_{Ti} = 0$. Denote by μ_S and μ_T the mean vectors of the source and target domains, i.e., $\mu_S = X \frac{1}{n_S} L_S$ and $\mu_T = X \frac{1}{n_T} L_T$, respectively. We have

$$\begin{aligned} S_B &= (\mu_S - \mu_T)(\mu_S - \mu_T)^T = X(\frac{1}{n_S} L_S - \frac{1}{n_T} L_T)(\frac{1}{n_S} L_S - \frac{1}{n_T} L_T)^T X^T \\ &= X(\frac{1}{n_S^2} L_S L_S^T - \frac{1}{n_S n_T} L_S L_T^T - \frac{1}{n_S n_T} L_T L_S^T + \frac{1}{n_T^2} L_T L_T^T) X^T \\ &= XLX^T. \end{aligned} \tag{.4}$$

When ignoring the regularized term ($W^T W$), linear kernel TCA is equivalent to finding a linear mapping W_1 from the original feature space, $W_1 = W^T X$, that satisfies the following optimization $\min_{W_1} \left\{ \mathrm{tr}(W_1^T S_B W_1) \right\}$ subject to $W_1^T S_T W_1 = I$. Thus, TCA tries to maximize the total data scatter matrix and at the same time minimize the distance between the mean of the source domain and the mean of target domain. The regularized term in TCA is important to avoid the rank deficiency of the XLX^T term in the generalized eigenvalue decomposition. However, it is done in the kernel space. In the original feature space, it can be done by adding the identity matrix (I) into the term S_B. Therefore, the linear TCA used in our paper, which can be applied directly to the original feature space, solves

$$\min_{W_1} \left\{ \mathrm{tr}(W_1^T (S_B + \mu I) W_1) \right\} \tag{.5}$$

subject to

$$W_1^T S_T W_1 = I \tag{.6}$$

BIBLIOGRAPHY

[1] C. Aggarwal, A. Hinneburg, and D. Keim. On the surprising behavior of distance metrics in high dimensional space. *Database Theory*, pages 420–434, 2001. Pages: 71

[2] A.Plaza, P.Martínez, J.Plaza, and R.Pérez. Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. *IEEE Trans. on Geoscience & Remote Sensing*, 43:466–479, 2005. Pages: 74

[3] V. Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society*, 139: 318–355, 1976. Pages: 21

[4] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *The Journal of Machine Learning Research*, 10:2137–2155, 2009. Pages: 50

[5] K. Bowyer, C. Kranenburg, and S. Dougherty. Edge detector evaluation using empirical roc curves. *Computer Vision and Image Understanding*, 84(1):77–103, 2001. Pages: 22

[6] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 679–698, 1986. Pages: 19

[7] T. Carron and P. Lambert. Color edge detector using jointly hue, saturation and intensity. In *IEEE International Conference Image Processing*, volume 3, pages 977–981, 1994. Pages: 16

[8] D. Casasent and X.W. Chen. Aflatoxin detection in whole corn kernels using hyperspectral methods. In *Proceedings of SPIE*, volume 5271, pages 275–284, 2003. Pages: 4

[9] J. Chanussot, J.A. Benediktsson, and M. Fauvel. Classification of remote sensing images from urban areas using a fuzzy possibilistic model. *Geoscience and Remote Sensing Letters, IEEE*, 3(1):40–44, 2006. Pages: 83

[10] O. Chapelle, B. Scholkopf, and A. Zien, editors. *Semi-supervised Learning*. MIT Press, 2006. Pages: 8, 9, 69

[11] M. Chapron and C. ENSEA-ETIS. A color edge detector based on statistical rupture tests. In *IEEE International Conference Image Processing*, volume 2, 2000. Pages: 16

[12] Y.R. Chen, K. Chao, and M.S. Kim. Machine vision technology for agricultural applications. *Computers and Electronics in Agriculture*, 36(2):173–191, 2002. Pages: 4

[13] Y Cheng. Mean shift, mode seek, and clustering. *IEEE Transaction on Pattern Analysis and Machine*, 1995. Pages: 75

[14] D. Cohen, M. Arnoldussen, G. Bearman, and WS Grundfest. The use of spectral imaging for the diagnosis of retinal disease. In *LEOS'99. IEEE Lasers and Electro-Optics Society 1999 12th Annual Meeting*, volume 1, pages 220–221. IEEE, 1999. Pages: 4

[15] A. Cumani. Edge detection in multispectral images. *Graphical Models and Image Processing*, 53(1):40–51, 1991. Pages: 16

[16] D.C.G. de Veld, M. Skurichina, M.J.H. Witjes, R.P.W. Duin, H.J.C.M. Sterenborg, and J.L.N. Roodenburg. Clinical study for classification of benign, dysplastic, and malignant oral

lesions using autofluorescence spectroscopy. *Journal of biomedical optics*, 9:940, 2004. Pages: 38

[17] F. Dell'Acqua, P. Gamba, A. Ferrari, JA Palmason, JA Benediktsson, and K. Arnason. Exploiting spectral and spatial information in hyperspectral urban data with high resolution. *Geoscience and Remote Sensing Letters, IEEE*, 1(4):322–326, 2004. Pages: 83

[18] S.G. Demos, R. Gandour-Edwards, R. Ramsamooj, and R. deVere White. Near-infrared autofluorescence imaging for detection of cancer. *Journal of biomedical optics*, 9:587, 2004. Pages: 39

[19] C.V. Dinh, R.P.W. Duin, and M. Loog. A study on semi- supervised dissimilarity representation. In *21th International Conference on Pattern Recognition (ICPR 2012)*. IEEE, 2012. Pages: 12

[20] C.V. Dinh, R.P.W. Duin, I.P. Salazar, and M. Loog. Fidos: A generalized fisher based feature extraction method for domain shift. *Pattern Recognition*, 46(9):2510–2518, 2013. Pages: 11

[21] C.V. Dinh, R. Leitner, P. Paclik, and R.P.W. Duin. A clustering based method for edge detection in hyperspectral images. In *the 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, pages 580–587. Springer, 2009. Pages: 10

[22] C.V. Dinh, R. Leitner, P. Paclik, M. Loog, and R.P.W. Duin. Sedmi: Saliency based edge detection in multispectral images. *Image and Vision Computing*, 29(8):546–556, 2011. Pages: 10

[23] C.V. Dinh, M. Loog, R. Leitner, O. Rajadell, and R.P.W. Duin. Training data selection for cancer detection in multispectral endoscopy images. In *21th International Conference on Pattern Recognition (ICPR 2012)*. IEEE, 2012. Pages: 11

[24] Q. Du and H. Yang. Similarity-based unsupervised band selection for hyperspectral image analysis. *Geoscience and Remote Sensing Letters, IEEE*, 5(4):564–568, 2008. Pages: 83

[25] L. Duan, I.W. Tsang, and D. Xu. Domain transfer multiple kernel learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):465–479, 2012. Pages: 65

[26] L. Duan, D. Xu, I.W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1959–1966. IEEE, 2010. Pages: 7, 50

[27] RO Duda and PE Hart. *Pattern classification*. John-Wiley and Sons, 2001. Pages: 75

[28] 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. `http://www.prtools.org/`. Pages: 20

[29] Robert Duin and E. Pekalska. The dissimilarity representation for structural pattern recognition. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 7042 of *Lecture Notes in Computer Science*, pages 1–24, 2011. Pages: 68

[30] R.P.W. Duin and E. Pekalska. The dissimilarity space: between structural and statistical pattern recognition. *Pattern Recognition Letter*, 33:826–832, 2011. Pages: 12

[31] A.N. Erkan, G. Camps-Valls, and Y. Altun. Semi-supervised remote sensing image classification via maximum entropy. In *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, pages 313–318. IEEE, 2010. Pages: 6

[32] ESA. The sen3exp campaign. `https://earth.esa.int/web/guest/missions/esa-future-missions/sentinel-3`. Pages: 56

[33] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of Data Mining in Computer Security*. Kluwer Academic Pub, 2002. Pages: 18

[34] A.Plaza et al. Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, 113:110–122, 2009. Pages: 78

[35] A. N. Evans and X. U. Liu. A morphological gradient approach to color edge detection. *IEEE Transactions on Image Processing*, 15(6):1454–1463, 2006. Pages: 16, 21, 32

[36] J. Fan, D. K. Y. Yau, A. K. Elmagarmid, and W. G Aref. Automatic image segmentation by integrating color-edge extraction and seeded region growing. *IEEE Transactions on Image Processing*, 10:1454–1466, 2001. Pages: 16

[37] R.A. Fisher. The statistical utilization of multiple measurements. *Annals of Human Genetics*, 8(4):376–386, 1938. Pages: 52

[38] David H. Foster. The hyperspectral images of natural scenes dataset, 2002. http://personalpages.manchester.ac.uk/staff/david.foster/. Pages: 29

[39] Ana L. N. Fred and Anil K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005. Pages: 18, 19

[40] K. Fukunaga. Introduction to statistical pattern recognition. *Academic Press, New York, 2nd edtion*, 1990. Pages: 64

[41] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 999–1006. IEEE, 2011. Pages: 7, 50

[42] M. Govender, K. Chetty, and H. Bulcock. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water SA*, 33(2), 2009. Pages: 1

[43] H. Grahn and P. Geladi. *Techniques and applications of hyperspectral image analysis*. Wiley, 2007. Pages: 4

[44] D.M. Green, J.A. Swets, et al. *Signal detection theory and psychophysics*. Wiley New York, 1966. Pages: 22

[45] R.O. Green, M.L. Eastwood, C.M. Sarture, T.G. Chrien, M. Aronsson, B.J. Chippendale, J.A. Faust, B.E. Pavri, C.J. Chovit, M. Solis, et al. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris). *Remote Sensing of Environment*, 65(3):227–248, 1998. Pages: 3, 4

[46] S. Hara, Y. Kawahara, T. Washio, P. Von BüNau, T. Tokunaga, and K. Yumoto. Separation of stationary and non-stationary sources with a generalized eigenvalue problem. *Neural Networks*, 2012. Pages: 57

[47] R. M. Haralick, S. R. Sternberg, and X. Zhuang. Image analysis using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):532–550, 1987. Pages: 16

[48] M.L. Harries, S. Lam, C. MacAulay, J. Qu, and B. Palcic. Diagnostic imaging of the larynx: autofluorescence of laryngeal tumours using the helium-cadmium laser. *The Journal of Laryngology & Otology*, 109(02):108–110, 1995. Pages: 4, 5, 38

[49] Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003. Pages: 18

[50] M. Hedley and H. Yan. Segmentation of color images using spatial and color space information. *Journal of Electronic Imaging*, 1:374–380, 1992. Pages: 16

[51] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, pages 131–160, 2007. Pages: 50, 65

[52] A.K. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *Pattern Analysis and Machine Intelligence*, 1997. Pages: 70

[53] T. Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning*, pages 200–209, 1999. Pages: 9

[54] G. Jun and J. Ghosh. Spatially adaptive classification of land cover with remote sensing data. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(7):2662–2673, 2011. Pages: 4, 7, 50

[55] T. Kanade and S. Shafer. Image understanding research at cmu. In *Proceedings of Image Understanding Workshop*, pages 32–40, 1987. Pages: 16

[56] J.J. Koenderink and AJ Van Doorn. Representation of local geometry in the visual system. *Biological cybernetics*, 55(6):367–375, 1987. Pages: 57, 83

[57] Scott Konishi, Alan L. Yuille, James M. Coughlan, and Song Chun Zhu. Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):57–74, 2003. ISSN 0162-8828. Pages: 22

[58] A. Koschan and M. Abidi. Detection and classification of edges in color images. *Signal Processing Magazine, Special Issue on Color Image Processing*, 22:64–73, 2005. Pages: 16

[59] W.J. Krzanowski and D.J. Hand. *ROC curves for continuous data*, volume 111. Chapman & Hall/CRC, 2009. Pages: 22

[60] S. Lam, J.Y.C. Hung, S.M. Kennedy, J.C. Leriche, S. Vedal, B. Nelems, C.E. MacAulay, and B. Palcic. Detection of dysplasia and carcinoma in situ by ratio fluorometry. *American Journal of Respiratory and Critical Care Medicine*, 146(6):1458–1461, 1992. Pages: 5

[61] D. Landgrebe. Hyperspectral image data analysis. *Signal Processing Magazine, IEEE*, 19(1): 17–28, 2002. Pages: 2

[62] D. A. Landgrebe. *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 1 edition, 2003. Pages: 76

[63] R. Leitner, T. Arnold, and M. De Biasio. High-sensitivity hyperspectral imager for biomedical video diagnostic applications. In *Proceedings of SPIE*, volume 7674, 2010. Pages: 5, 38

[64] R. Leitner, M.D. Biasio, T. Arnold, C.V. Dinh, M. Loog, and R.P.W. Duin. Multi-spectral video endoscopy system for the detection of cancerous tissue. *Pattern Recognition Letters*, 2012. Pages: 11, 39

[65] Jae S. Lim. *Two dimensional signal and image processing*. Addison-Wesley, 1990. Pages: 22, 35

[66] Tony Lindeberg. Edge detection and ridge detection with automatic scale selection. *Journal of Computer Vision*, 30:117–156, November 1998. ISSN 0920-5691. Pages: 17

[67] Dimitri Lisin, Edward Riseman, and Allen Hanson. Extracting salient image features for reliable matching using outlier detection techniques. In *Proceedings of the third international conference on Computer vision systems*, volume 2626, pages 481–491, 2003. ISBN 3-540-00921-3. Pages: 17

[68] M. Loog. Constrained parameter estimation for semi-supervised learning: The case of the nearest mean classifier. In *Proceedings of ECML PKDD*, LNAI, pages 291–304. Springer, 2010. Pages: 9

[69] M. Loog. Semi-supervised linear discriminant analysis using moment constraints. In *Partially Supervised Learning*, volume 7081 of *LNAI*, pages 32–41. Springer, 2012. Pages: 9

[70] M. Loog and B. Ginneken. Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. *Medical Imaging, IEEE Transactions on*, 25(5):602–611, 2006. Pages: 57

[71] Marco Loog. Nearest neighbor-based importance weighting. In *IEEE International Workshop on MLSP*. IEEE, 2012. Pages: 50

[72] Marco Loog and Robert P.W. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(6):732–739, 2004. Pages: 57

[73] Marco Loog and Francois Lauze. The improbability of harris interest points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1141–1147, 2010. ISSN 0162-8828. Pages: 17

[74] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, and Q. He. Transfer learning from multiple source domains via consensus regularization. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 103–112. ACM, 2008. Pages: 51

[75] D. Manolakis, D. Marden, and G.A. Shaw. Hyperspectral image processing for automatic target detection applications. *Lincoln Laboratory Journal*, 14(1):79–116, 2003. Pages: 1, 3

[76] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21:1041–1048, 2009. Pages: 51

[77] D. Marr and E. Hildreth. Theory of edge detection. In *Proceedings of Royal Society of London*, pages 187–217, 1980. Pages: 17

[78] M.E. Martin et al. Development of an advanced hyperspectral imaging (hsi) system with applications for cancer detection. *Annals of biomedical engineering*, 34(6):1061–1068, 2006. Pages: 5, 38

[79] Adolfo Martínez-Usó, Filiberto Pla, and Pedro García-Sevilla. Clustering-based hyperspectral band selection using information measures. *IEEE Trans. on Geoscience & Remote Sensing*, 45:4158–4171, 2007. Pages: 76, 83

[80] P.M. Mehl, Y.R. Chen, M.S. Kim, and D.E. Chan. Development of hyperspectral imaging technique for the detection of apple surface defects and contaminations. *Journal of Food Engineering*, 61(1):67–81, 2004. Pages: 4

[81] M.Fauvel, J.A.Benediktsson, J.Chanussot, and J.R.Sveinsson. Spectral and spatial classification of hyperspectral data using svms and morphological profiles. *IEEE Trans. on Geoscience & Remote Sensing*, 46(10):3804–3814, 2008. Pages: 74

[82] Andrew William Moore. Efficient memory-based learning for robot control. Technical Report UCAM-CL-TR-209, University of Cambridge, Computer Laboratory, 1990. Pages: 35

[83] S .M .C. Nascimento, F. P. Ferreira, and D. H. Foster. Statistics of spatial cone-excitation ratios in natural scenes. *Journal of the Optical Society of America A*, 19(8):1484–1490, 2002. Pages: 29

[84] P. Paclik, R. P. W. Duin, G .M. P. van Kempen, and R. Kohlus. Segmentation of multispectral images using the combined classifier approach. *Journal of Image and Vision Computing*, 21(6):473–482, 2005. Pages: 28

[85] P. Paclik and R.P.W. Duin. Dissimilarity-based classification of spectra: computational issues. *Real-Time Imaging*, 9(4):237–244, 2003. Pages: 12, 83

[86] S.J. Pan, I.W. Tsang, J.T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, (99):1–12, 2009. Pages: 8, 50, 57

[87] S.J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010. Pages: 7, 8

[88] M. Panjehpour, BF Overholt, T. Vo-Dinh, RC Haggitt, DH Edwards, FP Buckley, et al.

Endoscopic fluorescence detection of high-grade dysplasia in barrett's esophagus. *Gastroenterology*, 111(1):93–101, 1996. Pages: 4

[89] G. Papari and N. Petkov. Adaptive pseudo dilation for gestalt edge grouping and contour detection. *Image Processing, IEEE Transactions on*, 17(10):1950 –1962, 2008. ISSN 1057-7149. Pages: 16

[90] E. Pekalska, R.P.W. Duin, and P. Paclik. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, 2006. Pages: 68, 70, 71

[91] M. Petrou and P. García-Sevilla. *Image Processing: Dealing with Texture*. John-Wiley and Sons, 1 edition, 2006. Pages: 75

[92] J. H. Piater. *Visual feature learning*. PhD thesis, Department of Computer Science, UMASS Amherst, 2001. Pages: 17

[93] A. Plaza, J.A. Benediktsson, J.W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, et al. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009. Pages: 1, 6, 83

[94] Diana Porro-Muñoz, Robert P. W. Duin, Isneri Talavera, and Mauricio Orozco-Alzate. Classification of three-way data by the dissimilarity representation. *Signal Process.*, 91(11): 2520–2529, 2011. Pages: 83

[95] S. Prasad, L.M. Bruce, and J. Chanussot. *Optical Remote Sensing: Advances in Signal Processing and Exploitation Techniques*, volume 3. Springer, 2011. Pages: 5, 6

[96] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*, volume 37. Springer, 2nd edition, 2006. Pages: 64

[97] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009. Pages: 8, 50

[98] O. Rajadell, P. Garcia-Sevilla, VC Dinh, and RPW Duin. Semi-supervised hyperspectral pixel classification using interactive labeling. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2011 3rd Workshop on*, pages 1–4. IEEE, 2011. Pages: 13

[99] Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla. Filter banks for hyperspectral pixel classification of satellite images. In *CIARP 2009, Lecture Notes in Computer Science*, volume 5856, pages 1039–1046. Springer, 2009. Pages: 74, 75, 78

[100] S. Rajan, J. Ghosh, and M.M. Crawford. An active learning approach to hyperspectral data classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(4):1231–1242, 2008. Pages: 4

[101] S. Rajan, J. Ghosh, and M.M. Crawford. An active learning approach to hyperspectral data classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(4):1231–1242, 2008. Pages: 7, 50

[102] L.L. Randeberg, E.L.P. Larsen, A. Aksnes, O.A. Haugen, and L.O. Svaasand. Hyperspectral characterization of atherosclerotic plaques. In *European Conference on Biomedical Optics*. Optical Society of America, 2009. Pages: 4

[103] G.S. Robinson. Color edge detection. *Optical Engineering*, 16:479–484, 1977. Pages: 16

[104] Bart M. Haar Romeny. *Geometry-driven diffusion in computer vision*. Kluwer academic, 1994. Pages: 57, 83

[105] Azriel Rosenfeld, Robert A. Hummel, and Steven W. Zucker. Scene labeling by relaxation operations. *Systems, Man and Cybernetics, IEEE Transactions on*, 6(6):420 –433, 1976. ISSN

0018-9472. Pages: 16

[106] M.T. Rosenstein, Z. Marx, L.P. Kaelbling, and T.G. Dietterich. To transfer or not to transfer. In *NIPS Workshop on Inductive Transfer*, volume 10, 2005. Pages: 8

[107] P.L. Rosin. Edges: saliency measures and automatic thresholding. In *Geoscience and Remote Sensing Symposium, 1995. IGARSS '95. 'Quantitative Remote Sensing for Science and Applications', International*, volume 1, pages 93 –95 vol.1, 1995. Pages: 16

[108] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. *Computer Vision–ECCV 2010*, pages 213–226, 2010. Pages: 8, 50

[109] H. Schimao, Y. Hiki, M. Morise, S. Kikuchi, N. Kobayashi, Y. Sakakibara, A. Kakita, S. Tanabe K. Saigenji, and O. Tsutsumi. Endoscopic treatment for subtypes of early gastric cancer. *Chirur Gastroenterologie*, 16:81–84, 2000. Pages: 38

[110] A. Sha'asua and S. Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *Computer Vision., Second International Conference on*, pages 321 –327, December 1988. Pages: 17

[111] G.A. Shaw and H.K. Burke. Spectral imaging for remote sensing. *Lincoln Laboratory Journal*, 14(1):3–28, 2003. Pages: 1, 2, 3, 4, 83

[112] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi. Development of a digital image database for chest radiographs with and without a lung nodule receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1): 71–74, 2000. Pages: 57

[113] R.B. Smith. Introduction to hyperspectral imaging. *Microimages*, 30:2008, 2006. Pages: 1, 3, 5

[114] E. Soria, J. Martin, R. Magdalena, M. Martinez, and A. Serrano, editors. *Handbook of Research on Machine Learning Applications*. IGI Global, 2009. Chapter 11:Transfer Learning, L. Torrey and J. Shavlik. Pages: 6, 7, 8

[115] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2003. Pages: 18

[116] M. Sugiyama, M. Krauledat, and K.R. Müller. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research*, 8:985–1005, 2007. Pages: 50

[117] M. Tangermann, K.R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K.J. Miller, G.R. Müller-Putz, et al. Review of the bci competition iv. *Frontiers in Neuroscience*, 6, 2012. Pages: 7, 50

[118] Yuliya Tarabalka. *Classification of Hyperspectral Data Using Spectral-Spatial Approaches*. PhD thesis, University of Iceland and Grenoble Institute of Technology, 2010. Pages: 1, 4

[119] D.M.J. Tax. *One-class classification; Concept-learning in the absence of counter-examples (Chapter 3)*. PhD thesis, Delft University of Technology, 2001. Pages: 44

[120] P. J. Toivanen, J. Ansamäki, J. P. S Parkkinen, and J. Mielikäinen. Edge detection in multispectral images using the self-organizing map. *Pattern Recognition Letters*, 24(16):2987–2994, 2003. Pages: 16

[121] P.W. Trahanias and A.N. Venetsanopoulos. Color edge detection using vector order statistics. *IEEE Transactions on Image Processing*, 2:259–264, 1993. Pages: 16

[122] W. Tu and S. Sun. Transferable discriminative dimensionality reduction. In *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, pages 865–868. IEEE,

2011. Pages: 51, 54

[123] J. van de Weijer, T. Gevers, and A.D. Bagdanov. Boosting color saliency in image feature detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):150 –156, 2006. ISSN 0162-8828. Pages: 17

[124] T. Vo-Dinh. A hyperspectral imaging system for in vivo optical diagnostics. *Engineering in Medicine and Biology Magazine, IEEE*, 23(5):40–49, 2004. Pages: 4, 5, 38

[125] P. Von Buenau, F.C. Meinecke, F. Kiraly, and K.R. Müller. Finding stationary subspaces in multivariate time series. 2009. Pages: 7, 50, 51, 57

[126] K. N. Walker, T. F. Cootes, and C. J. Taylor. Locating salient object features. In *Proceedings of the British Machine Vision Conference*, pages 557–566, 1998. Pages: 17

[127] T.D. Wang, J.M. Crawford, M.S. Feld, Y. Wang, I. Itzkan, and J. Van Dam. In vivo identification of colonic dysplasia using fluorescence endoscopic imaging. *Gastrointestinal endoscopy*, 49(4):447–455, 1999. Pages: 4, 5, 38

[128] CL Wilson and MD Garris. Handprinted character database 3. Technical report, National Institute of Standards and Technology, 1992. Pages: 70

[129] S. J. Winawer, R. H. Fletcher, L. Miller, F. Godlee, M. H. Stolar, C. D. Mulrow, S. H. Woolf, S. N. Glick, T. G. Ganiats, J. H. Bond, L. Rosen, J. G. Zapka, S. J. Olsen, F. M. Giardiello, J. E. Sisk, R. Van Antwerp, C. Brown-Davis, D. A. Marciniak, and R. J. Mayer. Colorectal cancer screening: clinical guidelines and rationale. *Gastroenterology*, 112(2):594–642[Erratum, Gastroenterology 1997;112:1060, 1998;114:625.], 1997. Pages: 38

[130] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson. Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Trans. on Geoscience & Remote Sensing*, 47(8):2973–2987, 2009. Pages: 74

[131] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. *IEEE Trans. Systems, Man, and Cybernetics*, pages –, 2010. Pages: 74

[132] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson. Segmentation and classification of hyperspectral images using watershed transformation. *Patt.Recogn.*, 43(7):2367–2379, 2010. Pages: 74

[133] S. Di Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33:116–125, 1986. Pages: 16, 21, 32

[134] M. Zhu and A.M. Martinez. Subclass discriminant analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(8):1274–1286, 2006. Pages: 64

[135] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. Pages: 8, 9

[136] W. G. Zoller, I. P. Arlart, P. Merckle, and G. Gottschalk. Rationale diagnostik gastrointestinaler tumoren. *Chirur Gastroenterologie*, 16:70–74, 2000. Pages: 38

SUMMARY

Spectral imaging has been extensively applied in many fields, including agriculture, environmental monitoring, biomedical diagnostics, etc. Thanks to the advances in sensor technology, spectral imaging systems nowadays provide finer and finer spectral resolution needed to characterize the spectral properties of materials. The high spectral resolution, however, raises an issue as the difference in spectral information between two adjacent wavelength bands is typically very small. As a result, much of the data in a scene seems to be redundant. However, critical information is embedded that often can be used to identify materials. Therefore, finding appropriate approaches for visualizing and analyzing this rich source of information remains challenging for research in spectral imaging.

This thesis proposes a new method for data visualization using edge detection. Images captured at different wavelength bands often provide different edge information. Therefore, edge detection can be employed as a visualization tool since it allows for a rough localization of objects in the image. Because edges are rare events in an image, the proposed method first recasts the problem of edge detection into detecting events that have a small probability in feature space constructed by the spatial gradient magnitudes from all spectral channels. The edge strength of a pixel is defined by the confidence value that the spectral measurements at this pixel are from an event with low probability. We demonstrate that the method is especially useful for cases in which objects are embedded in background clutter or just appear in a few bands.

One of the main concerns in classifying spectral imaging is the small sample size problem. The number of labeled samples for training is often insufficient due to the fact that collecting labeled data is an expensive and time consuming process. This thesis exploits two recent approaches in machine learning to deal with this situation, called transfer learning and semi-supervised learning.

Another issue which often arises in spectral imaging is that data domains shift. In remote sensing, data collected from the same location but at different time points often exhibit remarkable differences between their spectral characteristics. In multispectral endoscopy for example, different types of cancer exhibit different spectral signatures. To make it possible to transfer knowledge learnt from one or several source domains to

a new target domain, transfer learning techniques can be used. This thesis specifically addresses the questions "When, what, and how to transfer".

To address the question "When to transfer", we proposed a method that selects suitable training data sets for a given test set. The selection scheme is based on a similarity measurement between data domains to avoid deterioration from irrelevant source domains. Experimental results demonstrated that classification is already significantly improved when a few data sets that are presumably similar to a given test set are selected for training instead of using all available data sets.

To address the question "What and how to transfer", we proposed a method that aims at finding a transformation of the original feature space such that the source and target distributions are matched. The method generalizes a well-known Fisher feature extraction method to the domain shift problem by learning invariant features across different domains. Different from classical Fisher feature extraction, the proposed method not only minimizes the within-class scatters but also minimizes the difference in distributions between different domains. Therefore, the constructed subspace reduces the drift in distributions among different domains and at the same time preserves the discriminants across classes. We demonstrated, on both artificial and real data, that learning invariant features with respect to the domains is essential to overcome the domain shift obstacle.

Different from transfer learning, semi-supervised learning copes with small sample size problems by leveraging unlabeled data from the same classification problem. This thesis presents two semi-supervised methods: (i) to select optimal training samples as the center points of the clusters resulting from clustering all data (training and test data), and (ii) to extend the representation set by unlabeled data when using a dissimilarity representation. The key advantage of the dissimilarity representation is that it provides a way to embed knowledge about the structural information of data into powerful feature-based statistical approaches. Up to now, the representation or prototype set has been usually selected from the training data, limiting the different aspects that can be captured, especially when the training data set is small. Based on the fact that it is not necessary to know the labels of the samples used in the representation set, we show that the test set can be included in the dissimilarity representation yielding a richer description of the objects of interest, and thus substantially improves the classification performance, especially within small sample size problems.

Concluding, the resolution of spectral imaging is vastly increasing, both in the frequency range as well as spatially, which requires novel analysis methodologies that can cope with the increase in resolutions. This thesis proposes to enrich the spectral data to facilitate visualizing spectral events. To deal with classification tasks based on spectral data, a number of novel learning methodologies are introduced that are particularly geared to have few labeled data.

SAMENVATTING

Spectrale beeldvorming wordt in veel disciplines, zoals landbouw, milieu en medische diagnostiek, op grote schaal toegepast. Dankzij ontwikkelingen in sensortechnologie, kunnen spectrale beeldsystemen tegenwoordig voor de zeer fijne resoluties zorgen die nodig zijn om de spectrale eigenschappen van materialen te beschrijven. Deze fijne resolutie brengt ook een probleem met zich mee, aangezien de verschillen tussen twee naastgelegen golflengtebanden vaak zeer klein zijn. Dit zorgt ervoor dat veel data, die cruciaal is om de materialen te identificeren, overbodig kan lijken, terwijl er essentiÈle informatie in zit die gebruikt kan worden voor visualisatie en analyse. Het vinden van passende methoden om deze rijke bron van informatie te visualiseren, en te analyseren, blijft daarom een van de belangrijkste uitdagingen in de spectrale beeldanalyse.

Dit proefschrift introduceert een methode voor datavisualisatie die gebruikt maakt van de detectie van randen. Spectrale beelden kunnen in verschillende kanalen informatie over verschillende randen bevatten; de detectie van randen is een belangrijke methode omdat het voor globale lokalisatie van objecten in het beeld kan zorgen. Omdat randen in een beeld schaars zijn, wordt het probleem getransformeerd in een probleem van detectie van gebeurtenissen die een kleine kansdichtheid hebben in een ruimte die is geconstrueerd uit de spatiale gradiÈnt magnitudes van alle spectrale kanalen. De randsterkte van een bepaalde pixel wordt bepaald door de kans dat de spectrale metingen in deze pixel door een onwaarschijnlijke gebeurtenis veroorzaakt worden. Wij laten zien dat deze methode bijzonder geschikt is voor gevallen waarin de objecten verborgen zijn in de achtergrond, of alleen in een paar banden zichtbaar zijn.

Een van de belangrijkste problemen in spectrale beeldanalyse is dat het aantal gelabelde voorbeelden voor het trainen van een classifier vaak onvoldoende is; dit komt doordat het verzamelen van gelabelde data duur en tijdrovend is. Dit probleem staat bekend als het "small sample size" probleem. Dit proefschrift benut twee recente machine learning methodes die kunnen omgaan met deze situatie, namelijk leren met overdracht en semi-gesuperviseerd leren.

Een ander veelvoorkomend probleem in spectrale beeldanalyse is het verschuiven van de datadomeinen, het zogenaamde domain shift probleem. Bijvoorbeeld: data die op dezelfde locatie, maar op verschillende tijdstippen, verzameld is, kan opvallende verschillen vertonen in de spectrale kenmerken. In multispectrale endoscopie kunnen verschillende soorten kanker verschillende spectrale signaturen hebben. Om de kennis uit het ene domein naar een ander domein over te dragen, kunnen technieken die leren met overdracht gebruikt worden. Dit proefschrift gaat specifiek in op de vragen wanneer moet er overdacht plaatsvinden, wat moet er overgedragen worden, en hoe moet deze overdracht plaatsvinden?

Om de vraag wanneer er overdracht moet plaatsvinden te beantwoorden, hebben wij een methode geẸntroduceerd die geschikte trainingsdatasets selecteert, gegeven een bepaalde testdataset. De selectie is gebaseerd op de mate van overeenkomst tussen de twee datadomeinen. Dit helpt om de situatie te voorkomen waarin dat irrelevante brondomeinen de classificatieresultaten verslechteren. Experimentele resultaten laten zien dat de classificatieresultaten significant beter zijn wanneer slechts enkele datasets die op de testset lijken geselecteerd worden om mee te trainen.

Om de vragen te beantwoorden wat er overgedragen moet worden en hoe deze overdracht moet plaatsvinden, hebben wij een methode geẸntroduceerd die een transformatie van de originele feature space probeert te vinden, zodanig dat de distributies in bron- en doeldomeinen overeen komen. Deze methode generaliseert een bekende Fisher-feature-extractiemethode naar het probleem van het verschuiven van domeinen, door features te leren die invariant zijn over verschillende domeinen. Onze methode minimaliseert zowel de intraklasse spreiding als het verschil in distributies tussen verschillende domeinen, terwijl de klassieke Fisher kenmerkextractie alleen de intraklasse spreiding minimaliseert. De geconstrueerde subruimte reduceert zo het verschil in verdelingen, maar behoudt tegelijkertijd de discriminerende informatie tussen de klassen. Wij hebben zowel op kunstmatige als op echte data laten zien dat het leren van invariante kenmerken essentieel is om het domain-shift-probleem de baas te worden.

Semi-gesuperviseerd leren pakt het small -sample -size probleem aan door ongelabelde data uit hetzelfde classificatieprobleem te gebruiken. In dit proefschrift worden twee semigesuperviseerde methoden gepresenteerd: (i) om de optimale trainpunten te selecteren als middelpunten van de clusters die verkregen worden door clustering van alle data (zowel trainings- als testdata), en (ii) om de representatieset uit te breiden met behulp van ongelabelde data wanneer gebruik wordt gemaakt van ongelijkheidsrepresentatie. Het belangrijkste voordeel van de ongelijkheidsaanpak is dat het een manier is om informatie over de structuur van de data geschikt te maken voor krachtige statistische methodes. Tot recent werd de representatieset uit de trainingsdata gekozen, wat beperkend is voor het aantal aspecten van de data dat beschreven kan worden, vooral bij kleine trainingsdatasets. Omdat de labels van de representatieset niet noodzakelijk zijn, kunnen we laten zien dat het gebruiken van de testset binnen de representatieset een betere beschrijving oplevert van de objecten, en aanzienlijk de classificatieresultaten verbetert, vooral in small sample size gevallen.

De resolutie van spectrale beeldvorming blijft toenemen, zowel in frequentiebereik als in ruimtelijk bereik, met als gevolg dat er nieuwe technieken nodig zijn die met deze toename om kunnen gaan. In proefschrift wordt voorgesteld om de spectrale data te verrijken met als doel om de visualisatie van spectrale gebeurtenissen mogelijk te maken. Het classificatieprobleem van spectrale data wordt benaderd met een aantal nieuwe methoden die vooral gericht zijn op problemen met kleine hoeveelheden gelabelde data.

ACKNOWLEDGEMENTS

More than four years of pursuing a Ph.D. at Delft University of Technology (TUDelft) has been an unforgettable time in my life with full of challenges and enlightenment. Many people have helped/contributed directly and indirectly to my work and made this thesis possible. I am deeply grateful for all that I have received throughout these years.

First and foremost, I would like to thank Bob Duin, my supervisor and my mentor, for his continuous support and guidance. Dear Bob, you have always kept the balance between guiding me and giving me freedom in my research. Your valuable suggestions and encouragement have given me confidence and kept me motivated. In my project, I spent half of the time at TUDelft, Delft, the Netherlands and half of the time at the Carinthian Tech Research (CTR) company, Villach, Austria. You were always there whenever I needed your advices. Even when I was in Villach, you still managed regular weekly telephone meetings with me to discuss about my work and to update me on the works of other members in our PRLab group. That made me feel like I was still in Delft and was always part of our group. It is truly my great pleasure to be your student, Bob!

I am grateful to Marco Loog, my co-supervisor, who has actively contributed to my work. Dear Marco, meetings with you are always fruitful and also bring lots of fun. Your critical remarks have helped me improve myself greatly, especially in defining research questions and presenting the research line. I often intended to do a lot of simulations/experiments, some of those might have taken days to finish. You kept advising me to balance between experimental and theoretical/writing work. I have become more careful in setting up experiments and spent more time analyzing the results. I have indeed learnt a lot from you, Marco.

I would like to thank Marcel Reinders, my promotor, for the invaluable comments, questions, and suggestions that led to significant improvement of this thesis. Thank you also for going through my propositions.

My sincere thanks come to Raimund Leitner for his guidance and assistance during the time I worked at CTR. Being there gave me the chance to get acquainted with hyperspectral imaging equipments and practical applications. Exciting chats and social activities with Gerald, Martin, and Thomas made my life in Villach more joyful. I appreciate the CTR for financially supporting my Ph.D. project through the Austrian COMET funding program.

I deeply thank all (former) members of the Pattern Recognition & Bioinformatics group for maintaining not only a scientifically open atmosphere where I can ask numerous questions without hesitation but also a friendly and nutritious environment, which certainly increased my productivity. David has been always willing to answer my questions and clearly explain to me Prtools and DDtools, especially at the beginning of my PhD. I thank Olga for her collaborations and Veronika for translating the summary of my thesis into Dutch. I have also enjoyed many interesting discussions with Pavel, Emile, Wan-Jui, Laurens, Carmen, Diana, Yenisel, Jesse, Konstantin, Serguei, Lu, Fei-Fei, Vikas, Gorkem, and Yuanhao. Saskia was always kindly supporting me with all the administrative work even when I was in Villach. Robbert provided me with technical supports on the server and computer softwares.

Sharing the office with Yan and Alessandro was really a great thing. I had not only lots of conversations ranging from work to life with you in the office but also invitations to your places for Chinese (many thanks to Tianmu as well) and Italian cuisines. Tianmu and Yan, I will never forget the birthday ’co-celebration’ with Yan that you invited me to. This made me feel warm in the cold Delft winter. Also I like your special birthday gift very much. Alessandro, thank you also for generously helping me with the move from Delft to Groningen. Inge and Bob have invited us multiple times to their place and organized museum and walking trips in The Hague for which I really appreciated. I will remember all our group dinners, PRBorrels, coffee-talks, and other activities.

A warm ’thank you’ to my too-many-to-name Vietnamese friends with whom I shared unforgettable cheerful time at all the places I have been in the last years (Delft, Vienna, and Groningen). Particularly, thanks to *anh* Trung, *anh* Vu, Nghi, and Duong for the great hospitality whenever I need accommodation for a short stay in Delft; to *anh* Dat & *chi* Nhung, Ha & Thuy, *anh* Tuan Anh & Tinh for being there in our critical time.

I am grateful to my grandmother, my grandmother in law, my parents, my mother in law, and my ’big’ family who have been encouraging and supporting me throughout every moment.

This thesis is dedicated to my beloved wife, Minh Anh, who is always by my side and shares my every up and down. During my Ph.D., I (we) had traveled a lot by trains at weekends between Villach and Vienna ($\sim$ 4.5 hours), and between Delft and Groningen ($\sim$ 3.5 hours). To me, they were the trains of ’happiness’ as I knew you were there waiting for me (and preparing delicious dinners). Being with you, I am happy at every place. Finally, thank you, my lovely son − Viet Tung, for your arrival during the mean time. You bring so much joy and happiness to us!

CURRICULUM VITAE

Cuong Viet Dinh (In Vietnamese: Đinh Việt Cường) was born on January 18, 1983 in Vinhphuc, Vietnam. He received a Bachelor in Computer Science from Vietnam National University, Hanoi in 2005 and an M.Eng in Electronics and Computer Engineering from Korea University, Seoul, Korea in 2007. During his Master study, he focused on video indexing and retrieval techniques putting emphasis on text detection in video. In June, 2007, he came back to Vietnam and worked at the Research and Development Department, LG Electronics Vietnam company.

From September 2008 to December 2012, he was a Ph.D. student working on a collaborated project between the Pattern Recognition Laboratory, Delft University of Technology, Delft, The Netherlands and Carinthian Tech Research (CTR) company, Villach, Autria. This resulted in the presented dissertation entitled “Learning from Weakly Representative Data and Applications in Spectral Image Analysis". From January 2013, he is a post-doc at the Netherlands Cancer Institute (NKI), Amsterdam, The Netherlands.

PUBLICATIONS

Journals

- **Cuong V. Dinh**, Robert P. W. Duin, Ignacio Piqueras-Salazar, and Marco Loog. A Generalized Fisher based Feature Extraction Method for Domain Shift, *Pattern Recognition*, 46(9): 2510-2518, 2013.
- Olga Rajadell, Pedro Garcia-Sevilla, **Cuong V. Dinh**, and Robert P.W. Duin. Improving hyperspectral pixel classification with unsupervised training data selection, *Geoscience and Remote Sensing Letters* (accepted).
- Raimund Leitner, Martin De Biasio, Thomas Arnold, **Cuong V. Dinh**, Marco Loog, Robert P. W. Duin. Multi-spectral video endoscopy system for the detection of cancerous tissue, *Pattern Recognition Letter*, 34(1): 85-93, 2012.
- **Cuong V. Dinh**, Raimund Leitner, Pavel Paclik, Marco Loog, and Robert P. W. Duin. SEDMI: Saliency based edge detection in multispectral images, *Image and Vision Computing*, 29(8): 546-556, 2011.
- Hanjin Ryu, Myunghoon Kim, **Cuong V. Dinh**, Seong Soo Chun, and Sanghoon Sull. Robust Face Tracking based on Region Correspondence and Its Application for Person based Indexing System, *International Journal of Innovative Computing, Information and Control (IJICIC)*, 4(11): 2861-2873, 2008.

Conferences

- **Cuong V. Dinh**, Robert P.W. Duin, Marco Loog. A study on semi-supervised dissimilarity representation, *the 21st International Conference on Pattern Recognition (ICPR)*, Tokyo, Japan, 2012.
- **Cuong V. Dinh**, Marco Loog, Raimund Leitner, Olga Rajadell, Robert P.W. Duin. Training Data Selection for Cancer Detection in Multispectral Endoscopy Images, *the 21st International Conference on Pattern Recognition (ICPR)*, Tokyo, Japan, 2012.
- **Cuong V. Dinh**, Raimund Leitner, R. O. R. Rojas, Marco Loog, and Robert P. W. Duin. A Study of Detecting Cancer Tissues in Multispectral Endoscopy Data, *ICT.OPEN*, 2011.
- Olga Rajadell-Rojas, P. Garcia-Sevilla, **Cuong V. Dinh**, and Robert P. W. Duin. Selection of samples for active labeling in semi-supervised hyperspectral pixel classification, *Image and Signal Processing for Remote Sensing XVII*, Prague, Czech Republic, 2011.

- Olga Rajadell-Rojas, **Cuong V. Dinh**, Robert P. W. Duin, and P. Garcia-Sevilla. Semi-supervised hyperspectral pixel classification using interactive labeling, *3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers)*, Lisbon, Portugal, 2011 (**best paper award**).
- **Cuong V. Dinh**, Raimund Leitner, Pavel Paclik, and Robert P. W. Duin. A Clustering Based Method for Edge Detection in Hyperspectral Images, *16th Scandinavian Conference on Image Analysis*, SCIA, Oslo, Norway, 2009.
- **Cuong V. Dinh**, Raimund Leitner, Pavel Paclik, and Robert P. W. Duin. A Method for Edge Detection in Hyperspectral Images, *15th Annual Conference of the Advanced School for Computing and Imaging*, ASCI 2009.
- **Cuong V. Dinh**, Seong Soo Chun, Seungwook Cha, Hanjin Ryu, Sanghoon Sull (2007). An Efficient Method for Text Detection in Video Based on Stroke Width Similarity, *8th Asian Conference on Computer Vision (ACCV 2007)*, Tokyo, Japan, 2007.
- Minh-Anh T. Nguyen, **Cuong V. Dinh**, Tri-Hoai Ngo, Viet-Ha Nguyen. Applying Multi Neural Networks and suitable feature extraction methods in Vietnamese hand-written character recognition, *8th national conference on information technology and communications, pp 37-46*, Hai Phong, Vietnam, August 2005 (in Vietnamese).
- Minh-Anh T. Nguyen, **Cuong V. Dinh**, Tri-Hoai Ngo, Viet-Ha Nguyen. An application of Neural Network in Vietnamese hand-written character recognition, *1st Young Vietnamese Scientist Meeting*, Nha Trang, Vietnam, June 2005.