Advanced Spectral and Spatial Techniques for Hyperspectral Image Analysis and Classification

a dissertation presented
by
Nicola Falco

in partial fulfilment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Information and Communication Technologies -
Telecommunication Area
at the
Department of Information Engineering and Computer Science
University of Trento
Trento, Italy

Electrical and Computer Engineering
at the
Faculty of Electrical and Computer Engineering
University of Iceland
Reykjavik, Iceland

February 2015

Thesis Advisers:

Prof. Lorenzo Bruzzone

Prof. Jón Atli Benediktsson

Thesis Committee:

Prof. José M. Bioucas Dias

Prof. Gustau Camps-Valls

Prof. Enrico Magli

Don't believe what your eyes are telling you.
All they show is limitation.
Look with your understanding,
find out what you already know,
and you'll see the way to fly.

Richard Bach

Advanced Spectral and Spatial Techniques for Hyperspectral Image Analysis and Classification

Abstract

Recent advances in sensor technology have led to an increased availability of hyperspectral remote sensing images with high spectral and spatial resolutions. These images are composed by hundreds of contiguous spectral channels, covering a wide spectral range of frequencies, in which each pixel contains a highly detailed representation of the reflectance of the materials present on the ground, and a better characterization in terms of geometrical detail. The burst of informative content conveyed in the hyperspectral images permits an improved characterization of different land coverages. In spite of that, it increases significantly the complexity of the analysis, introducing a series of challenges that need to be addressed, such as the computational complexity and resources required.

This dissertation aims at defining novel strategies for the analysis and classification of hyperspectral remote sensing images, placing the focal point on the investigation and optimisation techniques for the extraction and integration of spectral and spatial information. In the first part of the thesis, a thorough study on the analysis of the spectral information contained in the hyperspectral images is presented. Though, independent component analysis (ICA) has been widely used to address several tasks in the remote sensing field, such as feature reduction, spectral unmixing and classification, its employment in extracting class-discriminant information remains a research topic open to further investigation. To this extend, a profound study on the performances of different ICA algorithms is performed, highlighting their strengths and weaknesses in the hyperspectral image classification task. Based on this study, a novel approach for feature reduction is proposed, where the use of ICA is optimised for the extraction of class-specific information. In the second part of the thesis, the spatial information is exploited by employing operators from the mathematical morphology framework. Morphological operators, such as attribute profiles and their multi-channel and multi-attribute extensions, are proved to be effective in the modelling of the spatial information, dealing, however, with issues such as the high feature dimensionality, the high intrinsic information redundancy and the a-priori need for parameter tuning in filtering, which are still open. Addressing the first two issues, the reduced attribute profiles are introduced, in this thesis, as an optimised version of the morphological attribute profiles, with the property to compress all the meaningful geometrical information into a few features. Regarding the filter parameter tuning issue, an innovative strategy for automatic threshold selection is proposed. Inspired by the concept of granulometry, the proposed approach defines a novel granulometric characteristic function, which provides information on the image decomposition according to a given measure. The approach exploits the tree representation of an image, allowing us to avoid additional filtering steps prior to the threshold selection, making the process computationally effective.

The outcome of this dissertation advances the state-of-the-art by proposing novel methodologies for accurate hyperspectral image classification, where the results obtained by extensive experimentation on various real hyperspectral data sets confirmed their effectiveness. Concluding the thesis, insightful and concrete remarks to the aforementioned issues are discussed.

Framsæknar aðferðir sem byggja á upplýsingum í rófi og rúmi fyrir greiningu og flokkun mynda af mjög hárri vídd

Ágrip

Nýlegar framfarir í skynjaratækni hafa leitt til þess að nú er í meira mæli en áður hægt að afla fjarkönnunarmynda af afar hárri vídd (e. hyperspectral images) með mikilli upplausn, bæði í rófi og rúmi (e. spectral and spatial). Myndirnar eru samansettar af hundruðum samliggjandi rófrása sem ná yfir vítt tíðniróf og sérhver myndeining geymir í smáatriðum upplýsingar um endurkast efna sem eru á yfirborði jarðar auk nákvæmra upplýsinga um rúmfræðileg atriði. Hið mikla magn upplýsinga sem geymdar eru í myndum af hárri vídd leyfa betri framsetningu af þeim fjölbreytilegu gerðum landslags sem myndað er með fjarkönnunartækni. Þrátt fyrir þetta, er notkun á svona gögnum flókin í greiningu og huga þarf sérstaklega að ýmsum vandamálum, t.d. flóknari útreikningum og öflugum vélbúnaði sem þarf til úrvinnslunnar.

Í Þessari ritgerð er leitast við setja fram nýjar aðferðir til greiningar og flokkunar fjarkönnunarmynda af mjög hárri vídd. Sérstök áhersla er lögð á rannsóknir og tækni við bestun til að draga fram róf- og rúmupplýsingar og þá síðan heildar upplýsingarnar. Í fyrri hluta ritgerðarinnar er kynnt ítarleg rannsókn á greiningu rófupplýsinga í fjarkönnunarmyndum af mjög hárri vídd. Þótt óháð þáttagreining (ICA – e. Independent Component Analysis) hafi víða verið notuð til að vinna ýmis verkefni í fjarkönnun, t.d. við víddarfækkun, afblöndun rófs og flokkun, þá er notkun ICA til að draga fram sundurgreiningarupplýsingar fyrir flokkun ekki vel þekkt og er hún enn viðfangsefni rannsókna. Af þessum sökum er hér gerð ítarleg rannsókn á frammistöðu nokkurra mismunandi ICA algríma, þar sem sérstaklega eru skoðaðir bæði styrkleikar og veikleikar algrímanna til flokkunar mynda af mjög hárri vídd. Í framhaldi af framangreindri rannsókn er ný víddarfækkunaraðferð sett fram. Í þessari aðferð er notkun ICA bestuð til að draga fram upplýsingar um einstaka flokka.

Í seinni hluta ritgerðarinnar eru rúmupplýsingar notaðar með upplýsingum sem fást með notkun virkja stærðfræðilegrar formfræði. Formfræðilegir virkjar eins og auðkennaprófílar og útvíkkanir þeirra fyrir margar rásir og mörg auðkenni hafa reynst mjög öflugir til að gera líkön sem byggja á rúmfræðilegum upplýsingum. Hins vegar hefur reynst erfitt að leysa vandamál með þessari aðferð, þar sem fjöldi vídda verður oft mikill, umfremd (e. redundandcy) verður í gögnunum og nauðsynlegt er að skilgreina stika fyrirfram. Í ritgerðinni er unnið að lausn fyrstu tveggja vandamálanna.

Víddafækkun auðkennaprófíla er kynnt sem bestuð útgáfa formfræðilegra auðkennaprófíla með þeim eiginleika að öllum helstu flatarupplýsingum er þjappað inn í auðkenni. Þá er ný aðferð sem byggir á sjálfvirkri þröskuldun sett fram til að ákvarða síunarstika. Hún byggir á hugmyndinni um að nota sífellt stækkandi gildi á síunarstikum (e. granulometry), og er skilgreint sérstakt einkennisfall sem gefur upplýsingar um uppskiptingu myndarinnar samkvæmt gefinni mælingu. Nýja aðferðin notast við trjáframsetningu myndarinnar og losar okkur við viðbótarsíunarskref áður en valið á þröskuldunum fer fram. Það að losna við síunarskrefin gerir ferlið reiknilega hagkvæmt.

Niðurstöður þessarar ritgerðar eru framlag til stöðu þekkingar á fræðasviðinu þar sem kynntar eru nýjar aðferðir fyrir nákvæma flokkun myndgagna af mjög hárri vídd. Niðurstöðurnar, sem fengust með umfangsmiklum tilraunum á margs konar raunverulegum gögnum af mjög hárri vídd, staðfestu gildi aðferðanna. Í lok ritgerðarinnar eru dregin saman og rædd atriði um framangreindar aðferðir og niðurstöður.

Acknowledgments

I would like to sincerely thank my advisers Jón Atli Benediktsson and Lorenzo Bruzzone, for their guidance, support, and most important, for their friendship, which gave me the possibility to learn and grow personally and professionally during my doctoral studies.

I would like to thanks my colleagues at the VRII at Háskóli Íslands, Behnood, Eysteinn, Frosti, Gabriele, Pedram and Prashanth, for their friendship and support in the academic world and every-day life. A special thanks goes to Jakob for all the advices on my studies, the always interesting discussions, for helping in several occasions, and most importantly, for letting me win at bowling. I wish to thank all the RSLab members at University of Trento, Ana-Maria, Begum, Claudia, Francesca, Davide, Leonardo, Massimo for sharing this experience and for all the time passed together. Special thanks go to Carlo and Sicong, for the reciprocal support during the research period.

I also wish to thank prof. José M. Bioucas Dias, prof. Gustau Camps-Valls and prof. Enrico Magli, for being part of the committee and for their comments and suggestions regarding the manuscript.

In these three years, I met many people and friends that made this period very special. A big thanks goes to Anna and Fabio, Atli Steinn, Gro, Hannes, Morgane, Pauline and Þorbjörg, for making my icelandic social life a bit more active. Going back to Italy, I must thanks my old friends Cippo, Enrico, Mitch, Mike, Preca, Simone, Toni for their company and for being always there when I was going back to Italy.

I want to thank Kyriaki, that with her enthusiasm and positivity gave me the strength to overtake any obstacle. Finally, a thought goes to my family, which has always been with me either when I was hundred or thousands miles far away from home, for supporting me and believing in me; this achievement would not have been possible without you.

Nicola
February 2015

Contents

List of Figures

List of Tables

Part I

Overview and Background

1
Introduction

This Chapter introduces the dissertation, providing an overview on the remote sensing field, focusing on the hyperspectral images and the challenges related to their analysis. The objectives and the contributions are then described.

1.1 Overview on Remote Sensing

The innate human desire to explore and understand the intangible pushes the boundaries of the scientific and technical limits, and is what made remote sensing the field of science of today. Aristotle, in De Anima, exposes the nature of light as a state of actual transparency in a potentially transparent medium and thus represents the necessary condition for vision. Eighteen hundred years after him, Leonardo da Vinci sets in detail the principles underlying the "camera obscura", while Isaac Newton, in 1666, using a prism proves that the light could be dispersed into a spectrum of colours, and using a second prism, the color could be re-combined into white light, giving birth to the science and art of "drawing with light", broadly known as "photography". Not long after, the first photograph in history of humanity was taken by Niepce (1827), while Gaspard-Félix Tournachon (Nadar) took in 1858 the first aerial photograph from a balloon from an altitude of 1,200 feet over Paris. New methods and technologies for sensing of the Earth's surface going beyond the traditional black and white aerial photograph, required a new, more comprehensive term to be established. The term *remote sensing* came to fill in this gap, initially introduced in 1960. *Remote Sensing* (RS) is the field of science that includes all those activities necessary for the observation, acquisition and interpretation of information related to objects, events, phenomena or any other item under investigation, without making physical contact with the object, event, or phenomenon under investigation. Since the launch of the first satellite for space exploration (Sputnik-1) in the late fifties, advances in the satellite technology burst, offering a multitude of spaceborne and airborne platforms with on-board sensors able to detect a great number of heterogeneous sources of information, for the study not only of distant celestial objects but also for the Earth Observation (EO).

Remote sensing systems collect data by detecting the energy that is reflected from an object or area under investigation. Considering the electromagnetic radiation as the principal physical carrier of information, a main differentiation of remote sensing systems is based on the typology of the source of energy exploited. Depending on whether these systems measure the radiation that is naturally available, or the energy used to illuminate the target under investigation is emitted by the sensor, are defined as passive or active, respectively. Passive sensors rely on the energy provided by the Sun, which is either reflected, or absorbed and then re-emitted from the Earth's surface. While the reflected energy (e.g., visible radiation) is available only when the Sun illuminates the Earth, the emitted energy (e.g., thermal infrared radiation) can be detected at any time, as long as the amount of energy is large enough to be recorded. Examples of the most popular passive sensors are cameras, scanning sensors and microwave radiometers. Active sensors instead, emit the energy required to illuminate the target under investigation, and then detect the backscattered radiation. Examples of broadly used active systems are the RAdio Detection And Ranging (RADAR) and Light Detection And Ranging (LiDAR). In this case, being the sensor the source of radiation, the data acquisition can be performed at any time.

Figure 1.1: Honoré Daumier (French, 1808-1879). Nadar Élevant la Photographie à la Hauteur de l'Art, May 25, 1862. Brooklyn Museum photograph, 2004.

The vast variety of available sensors, which provide data either in image or signal formats, allows to tackle a large number of applications with remarkable advantages. In general, each family of sensors is characterised by properties such as spatial, spectral, radiometrical and temporal resolutions, which are strictly related to their physical implementation resulting more or less suitable for a precise application. This entails the development of advanced techniques for data processing and interpretation that are sensor and application dependent. Space exploration is the RS domain that leads by far the technological advances, providing important know-how also for the Earth monitoring and for its understanding as a celestial object. Another main application is related to the environmental monitoring, where remote sensing techniques are used for studying human activities, such as urban planning, agriculture land usage, and natural phenomena, such as damage assessment due to earthquakes or floods, eruptions, climate change (e.g. glaciers), deforestation. Protected areas with fragile ecosystems can be studied by means of non-invasive remote sensing-based monitoring, without carrying any risk of environmental damage, replacing in this way costly field campaigns. Other important applications include meteorology, national security and natural resource management. The dissemination of remote sensing data is another important topic and is strictly connected to geographic information systems (GIS). Such platform allows remote sensing data obtained by different sources to be combined in order to make the information readily understandable to the final users.

1.2 Introduction to Hyperspectral Images

Earth remote sensing includes data collection on the environment, geology, climate, and other characteristics of the Earth by means of sensors positioned in the air or in Earth orbit. An important distinction between the systems broadly used to this end, refers to the coverage of electromagnetic spectrum. Focusing on passive optical systems, the sensor acquires data as in image format, detecting a portion of the electromagnetic radiation reflected from the Earth's surface in a range of wavelengths that includes the visible, near-infrared and short-wavelength infrared regions of the electromagnetic spectrum.

The sensor system, for instance the scanner, is composed by detectors that scan the scene and store the radiance detected as a quantised sample of the continuous data stream, forming a pixel characterised by a digital number, DN. To create multi-channel images that show specific portions of the EM field, the detected beam is split into different spectral components by inserting a system of spectral filters and optical components (e.g., prism, grating). For a more detailed review on sensor systems and different typology of scanners, please refer to [87, 91, 92]. According to the characteristics of the scanner, sensor systems are distinguished by their different resolutions, which also define the characteristics of the acquired images. The minimum size of an object that the sensor is able to distinguish from the ground represents the spatial resolution, and depends on the altitude of the sensor and its angle of view (i.e., the angle subtended by the sensor), which is defined in terms of Instantaneous Field Of View (IFOV). In digital imaging, the resolution is limited by the pixel size. The spectral resolution is the minimum wavelength at which the instrument is sensitive, while the radiometric resolution is defined as the minimum energy able to be detected by the sensing system. The intrinsic radiometric resolution of a sensor depends on the detector's signal to noise ratio. In a digital image, the radiometric resolution is limited by the number of discrete quantisation levels used to digitise the continuous intensity value. Considering a three-dimensional space (x, y, λ), where x and y are spatial coordinates and λ the spectral coordinate, each pixel is the integral of the radiance in a small volume (cube). The minimum value obtained by the integral represents the radiometric resolution, whereas the spatial resolution is represented by the size of a cube in the plane (x, y). The spectral resolution is the minimum bandwidth on which the measured radiation is integrated (Figure 1.2). Although the acquisition system could detects signals with high resolutions, it counts on various critical points due to physical constraints and instrumental limitations. Indeed, the acquisition of the images is usually affected by the sensor's noise, bad pixel location and atmospheric contribution, requiring different levels of pre-processing in order to ensure the image quality in terms of spectral, spatial and radiometric accuracy [87] and make the data available for further analysis. According to criteria that include spectral range, spectral and spatial resolutions and number of bands, the acquired images are identified as panchromatic, multispectral and hyperspectral. Panchromatic images are mono-channel data, which spatial resolution is maximised with a consequently minimisation of the spectral resolution. In such images, the high geometrical detail permits objects on the ground to be represented in detail, however, the information of the target's spectral characteristic results poor, meaning that objects of different nature can be represented in the same range of pixel values, making their discrimination difficult to achieve. In multispectral images, the augmented spectral dimension, which is represented by a few wide spectral

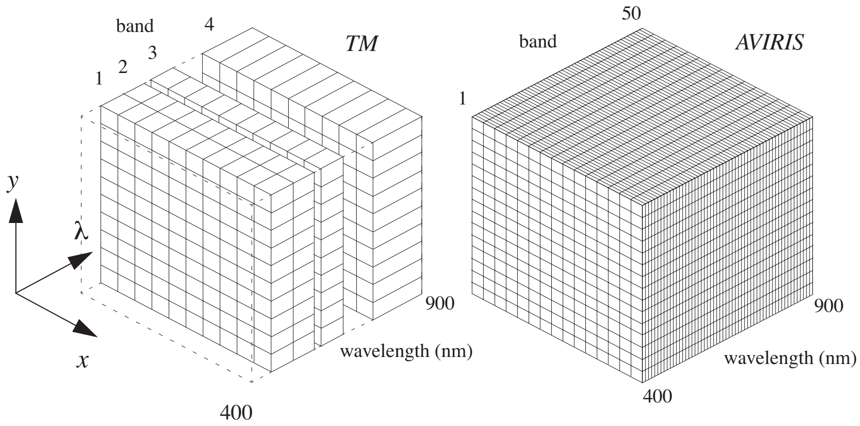


Figure 1.2: Comparison of the spatial and spectral sampling of the Landsat TM and AVIRIS in the VNIR spectral range. Each cell represents the spatial-spectral integration region of one pixel. The sampling of the spectral dimension operated by the Landsat TM results incomplete with relatively broad spectral bands, while the spectral sampling in the case of AVIRIS is relatively continuous over the VNIR range [92].

channels that cover wide portions of the electromagnetic spectrum, provides useful information on the nature of the targets and facilitates their discrimination and classification. In hyperspectral images, the spectral resolution is further improved, where the spectral information is maximised, providing data characterised by hundreds of narrow and contiguous spectral-channels. Consequently, each pixel can be represented as a vector in which a given value corresponds to the radiation at a given spectral band. The high dimensionality of this vector intrinsically provides a finer representation of the spectral signature of the target, leading to a better discrimination among different materials with respect to multispectral images, which are characterised by only few spectral channels (Figure 1.2). Moreover, recent technological advances in sensor technology have led to the development of a new generation of hyperspectral sensors able to provide images with improved spatial resolution. For instance, an image acquired by Hyperion sensors (mounted on EO-1 satellite) has a spatial resolution of 30 m, while ROSIS-3 (airborne spectrometer) can provide images with a spatial resolution of 1.7 m if the acquisition is taken at the altitude of 3 km. CASI-1500 can provide a data cube of 144 spectral bands with a spectral resolution of 1.25 m. From these few examples, we can see that the contextual information, becomes an important source of information that can be exploited for distinguishing different objects on the ground. Due to these

Table 1.1: Technical characteristics of some hyperspectral sensors developed over last years [30].

Sensor	Manufacturer	Platform	No. of bands	Spectral resolution	Spectral range
Hyperion	NASA GSFC	Satellite	220	10nm	0.4-2.5 µm
MODIS	NASA	Satellite	36	40nm	0.4-14.3 µm
CHRIS Proba	ESA	Satellite	up to 63	1.25nm	0.415-1.05 µm
AVIRIS	NASA JPL	Aerial	224	10nm	0.4-2.5 µm
HYDICE	Naval Research Lab	Aerial	210	7.6nm	0.4-2.5 µm
PROBE-1	Earth Search Science	Aerial	128	12nm	0.4-2.45 µm
CASI 550	ITRES Research Ltd	Aerial	288	1.9nm	0.4-1 µm
CASI 1500	ITRES Research Ltd	Aerial	288	2.5nm	0.4-1.05 µm
SASI 600	ITRES Research Ltd	Aerial	100	15nm	0.95-2.45 µm
TASI 600	ITRES Research Ltd	Aerial	64	250nm	8-11.5 µm
HyMap	Intergrated Spectronics	Aerial	125	17nm	0.4-2.5 µm
ROSIS-3	DLR	Aerial	115	4nm	0.43-0.85 µm
EPS-H	GER Corporation	Aerial	133	0.67nm	0.43-12.5 µm
EPS-A	GER Corporation	Aerial	31	23nm	0.43-12.5 µm
DAIS 7915	GER Corporation	Aerial	79	15nm	0.43-12.3 µm
AISA Eagle	Spectral Imaging	Aerial	244	2.3nm	0.4-0.97 µm
AISA Eaglet	Spectral Imaging	Aerial	200	-	0.4-1.0 µm
AISA Hawk	Spectral Imaging	Aerial	320	8.5nm	0.97-2.45 µm
AISA Dual	Spectral Imaging	Aerial	500	2.9nm	0.4-2.45 µm
MIVIS	Daedalus	Aerial	102	20nm	0.43-12.7 µm
AVNIR	OKSI	Aerial	60	10nm	0.43-1.03 µm

properties, hyperspectral images have been widely exploited in different applications, ranging from forestry management, pollution detection and mineral exploration. Table 1.1 provides a summary of the most commonly used sensors usually mounted on aircraft or spacecraft, reporting the principal spectral characteristics.

1.3 Hyperspectral Image Classification: Challenges

The Earth Observation domain entails numerous open research issues to overcome, ranging from the hardware technology itself to the higher level data analysis algorithms for the remote sensing image understanding. Focusing on the later, remote sensing image classification emerges as one of the major challenges. Image classification refers to the process of identifying the diverse objects, materials or items of interest with common properties that are grouped

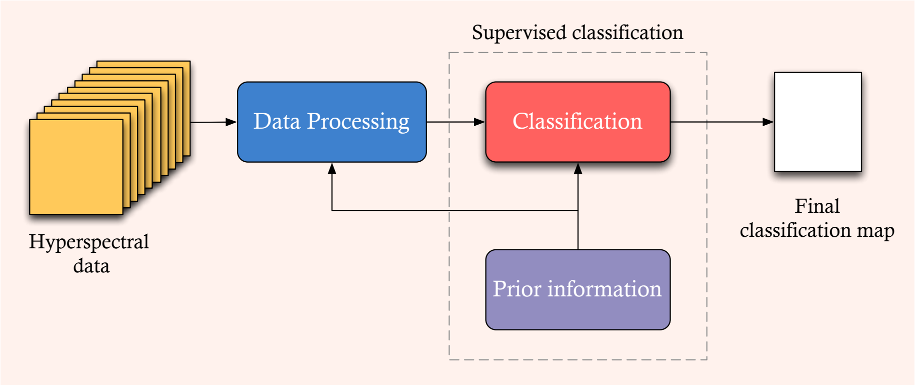


Figure 1.3: General scheme of a supervised image classification approach. Available prior information can be used in both the classification stage and the data processing stage.

into the so-called "classes" of coverage present on the ground of the investigated area of interest. Product of this process is a thematic map, where pixels are characterised by a given label, usually represented by a colour or symbol, used to uniquely identify the items within a class. A general scheme of image classification is illustrated in Figure 1.3, in which available information can be exploited in both the data processing and the classification stage. If on the one hand the burst of informative content conveyed in hyperspectral images, represented by both high spectral and spatial resolutions, provides the base for obtaining high accuracy in the identification of different land-covers, on the other hand it introduces a number of challenges that need to be efficiently addressed.

First, the high dimensionality of the data causes a variety of issues in hyperspectral classification referred to in literature as the "curse of dimensionality". The high dimensionality, which is represented by the spectral dimension, makes the analysis computationally expensive, limiting the exploitation of traditional classification approaches, usually employed in multispectral image analysis. In the context of supervised classification, in which labelled samples are used in the classification process, the ratio between the number of available training samples (which is usually small) and the spectral dimension (which is high), affects the generalization capability of the classifier. In general, it has been observed that, beyond a certain point, the inclusion of additional features, while keeping the number of training samples constant, leads to a decrease of both the accuracy and the generalization of the classification process.

In the machine learning domain, this behaviour is known as the Hughes phenomenon (named after Gordon F. Hughes) [54].

Second, the increase of the spatial resolution in the new generation of spectrometers introduces other important issues in the analysis and classification of hyperspectral images. The high geometrical detail of the scene leads to the presence of objects that are composed by several spatial correlated pixels, resulting in an increase of the intraclass variability [13]. The aforementioned phenomenon decreases the effectiveness of the analysis when only the spectral information is considered, enforcing the need of strategies that integrate the analysis of both spectral and contextual domains in order to maximize the exploitation of the information combined in these images.

1.4 Objectives of this Dissertation

The research work presented in this dissertation aims at investigating and defining novel techniques for the analysis and supervised classification of remote sensing hyperspectral images. In particular, the focus is on the investigation and optimisation of strategies, based on the use of independent component analysis (ICA) and morphological operators, for the extraction and integration of both spectral and spatial information contained in hyperspectral images. In recent studies, ICA proved its effectiveness in extracting useful information to address the hyperspectral image classification task. However, many issues related to the computational cost and how to effectively extract class-specific information, need to be further investigated. Morphological operators, such as morphological attribute profiles and their multi-channel and multi-attribute extensions, proved to be effective in modelling the spatial characteristics. However, issues such as parameter tuning in filtering, need to be addressed, in order to obtain a reliable and representative image decomposition. The high dimensionality of the profiles, which leads to a high intrinsic information redundancy and thus to the Hughes phenomenon, is still an open issue.

Aiming at overcoming the aforementioned issues and limitations that affect the analysis of hyperspectral image classification, the following objectives are defined:

- to deeply investigate the behaviour and performance, in terms of computational cost and execution time, as well as classification accuracy, of the most widely used ICA algorithms in the remote sensing field, under different experimental set-ups.

- to develop a novel strategies to limit the Hughes phenomenon for hyperspectral image classification by exploiting ICA.
- to design an innovative technique for spatial information extraction by using morphological attribute profiles, while addressing the information redundancy issue.
- to move towards a fully automatic approach to the selection of filtering parameters used for the computation of attribute profiles.
- to define a methodology that integrates both spectral and spatial information within a classification scheme.

1.5 Organization of the Dissertation

This dissertation is organised as follows:

Part I provides an introduction to the remote sensing field and the context in which the dissertation is developed.
Chapter 1 introduces the remote sensing field, providing a description of both the challenges and the objectives addressed in this thesis.
Chapter 2 presents an overview of the state-of-the-art in spatial and spectral information extraction domains. Moreover, it provides the theoretical background on independent component analysis and morphological operators.

Part II includes the strategies developed for spectral information extraction based on the exploitation of ICA.
Chapter 3 presents a thorough study on the performances of different independent component analysis algorithms for the extraction of class-discriminant information in remote sensing hyperspectral image classification.
Chapter 4 describes a novel feature reduction technique based on ICA, whose aim is to extract subsets of class-specific independent components for the hyperspectral image classification.

Part III presents the contributions of this dissertation on spectral-spatial analysis for hyperspectral image classification.
Chapter 5 introduces the novel concept of reduced attribute profiles as an optimised version of the morphological attribute profiles. This Chapter provides a solution to both the high dimensionality and the information redundancy issues that affect the morphological attribute profiles.
Chapter 6 presents a new methodology that combines the findings in Chapter 4 and Chapter 5, fusing the spectral and spatial information for hyperspectral image classification.
Chapter 7 introduces a step towards a fully automated procedure for building the attribute profiles, presenting a novel automatic strategy for threshold selection.

Finally, Chapter 8 concludes this dissertation remarking its most important findings, and discussing on the most prominent future research directions.

2
Background and Related Work

This Chapter provides an overview on the most widely used spectral and spatial techniques developed over the last years in pattern recognition, machine learning and image processing, for the analysis of hyperspectral images. Then, the theoretical background on both independent component analysis and morphological operators is provided.

2.1 Introduction

Image classification in hyperspectral remote sensing images is a complex task that employs a number of processes aiming at addressing the challenging issues that emerge from the nature of the hyperspectral images. Considering the spectral domain, each single pixel is considered as an independent entity of information. The high dimensionality makes the analysis computationally expensive, while the Hughes' phenomenon (curse of dimensionality) [54] arises when the ratio between the number of available training samples and the number of spectral channels is small. This affects the generalization capability of the classifier. Most studies in the current literature address the curse of dimensionality issue by exploiting feature extraction and feature selection techniques, aiming at decreasing the dimensionality of the feature space by retaining the most useful information. Other issues arise when hyperspectral images with improved spatial resolution, where the scene is characterized by objects composed by groups of pixels highly correlated, are considered. The improved detail increases the complexity of the image, adding a certain spectral variability to the pixels that belong to the same object or class The complexity increases by increasing of the spatial resolution. In this context, approaches based only the spectral information result less effective, providing classification maps with high uncertainty, especially for those classes with limited number of samples. Therefore, in order to minimise the uncertainty of the classification, the contextual information should be extracted and included in the analysis. In this Chapter, an overview of the most widely used techniques for dimensionality reduction and spatial information extraction approaches is presented, focusing on independent component analysis and mathematical morphology, which are the techniques that will be considered in this dissertation.

2.2 Related Work

2.2.1 Overview on Dimensionality Reduction Approaches

High-dimensional data sets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments. One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are "important" for understanding the underlying phenomena of interest. Dimensionality reduction can be seen as the process of deriving a set of degrees of freedom, which can be used to reproduce most of the variability of a data set. Dimensionality reduction has a long history as an approach to data visualisation, and for extracting key low dimensional features. Apart from teaching us about the data, dimensionality reduction can lead us to better models for inference. It can be divided into two major components, the feature selection and the feature extraction.

In the feature selection approach, the selection of a subset of the original feature set is usually obtained according to the evaluation of a fitness function followed by a search strategy. A number of statistical distance measures [43, 87] such as divergence, Bhattacharyya distance, Jeffries-Matusita (JM) distance and mutual information [49], are used to assess the separability and / or the mutual dependency among class distributions based on the available training set. Once the criterion function is chosen, a search strategy is needed in order to identify the features that better fulfil the criterion function. Considering the high spectral dimension of hyperspectral data (usually around hundreds of spectral channels) an exhaustive search strategy would result to be not feasible from the computational perspective. Sub-optimal strategies are broadly used, such as the Sequential Backward Selection (SBS) [71] and the Sequential Forward Selection (SFS) [107] methods. The first method performs a top-down search where the final feature subset is built up by starting from the complete set of features, while the second method applies a bottom-up search strategy, in which the starting point is an empty set. Both the methods are affected by the so-called nesting effect [58, 86]. In case of SBS technique, the discarded features cannot be selected again and added to the subset while in the case of SFS the selected features cannot be discarded in a second moment. The Sequential Forward Floating Selection (SFFS) and the Sequential Backward Floating Selection (SBFS) [86] methods were proposed to overcome the nesting effect. The steepest ascent and the fast constrained [93] algorithms, in which the feature selection problem is represented by a multi-dimensional binary space, are effective strategies that have shown better results compared to SFFS technique, even if the required computation time is slightly higher. Furthermore, heuristic search algorithms based on the evolutionary concept of natural selection, such as Genetic Algorithms (GAs) [46], are also used in several fields as well as in hyperspectral image analysis, where multi-objective fitness function can be used to find useful spatially invariant features for image classification [14].

Based on the task to be accomplished, i.e., compression, target detection, identification of endmembers and classification, several feature extraction techniques have been developed, ranging from supervised to unsupervised approaches. Supervised techniques based on discriminant analysis [43] have been widely used for the extraction of class-discriminative feature. Discriminant Analysis Feature Extraction (DAFE) [43] reduces the dimensionality optimizing the Fisher‘s ratio. The criterion of class separability is usually formulated by using within-class, between-class, and mixture-scatter matrices. The main advantage is that the approach is distribution-free. A drawback of this method is that if the difference in the class-mean vectors is small, the features chosen are not reliable. If one mean vector is very different from the others, its class will eclipse the others in the computation of the between-class covariance matrix, making the feature extraction process less effective. Decision Boundaries Feature Extraction (DBFE) [64], which is based on the definition of discriminantly redundant features and discriminant informative features, is able to predict the minimum dimension of the feature subset able to achieve the same classification accuracy as in the original feature space and find the necessary feature vectors. Both the DAFE and DBFE demand a large number of training samples for a high-dimensional space, since the computation of the class-statistical parameters is performed at full dimensionality. Considering the case of a limited number of training samples, Projection Pursuit (PP) [59] was proposed in order to avoid the computation at full dimensionality, which is done in a lower-dimensional subspace. The method achieves the dimensionality reduction by optimizing a projection index, which is the minimum Bhattacharyya distance among the classes, taking into consideration first-order and second-order statistics. Non-parametric Weighted Feature Extraction (NWFE) [63] was proposed as a trade-off between the advantages and limitations of the DAFE and DBFE techniques. The method weights every sample to compute the local means and defines new non-parametric between-class and within-class scatter matrices to get more features. However, these techniques are affected by a higher computational load, making the all feature extraction process considerably slow.

Among unsupervised approaches, which do not take advantage of prior information, the Karhunen-Loève transform [43] (also known as Principal Component Analysis, PCA), is one of the most widely used approach. It is able to concentrate the most significant part of the information in a new feature space composed by few principal components. The analysis of eigenvalues is used to determine the significance of principal components (PCs), so that the dimensional reduction is achieved by selecting the PCs according to the magnitude of their eigenvalue, achieving excellent data compression [68] and a good representation in terms of minimum mean square error. Similar approaches are the Maximum Noise Fraction (MNF) [47] and the Noise-Adjusted Principal Component (NAPC) [65], which aim at identifying the projection that maximizes the signal to noise ratio. Independent Component Analysis (ICA) [24, 56] is a well-know unsupervised source separation process that aims at identifying a linear transformation that minimizes the statistical dependence between its components by only considering the observation of their mixture signals. Non-linear versions of the aforementioned approaches obtained through kernel methods [95], such as kernel PCA [39, 90], kernel MNF [80] and kernel ICA [3], have also been developed and proposed in the literature for the analysis of hyperspectral data. Such methods handle non-linearities by mapping the data into high dimensional feature space via the kernel function and then performing a linear analysis in that space. The main drawback of such methods is the higher computational cost required with respect to the linear versions.

2.2.2 Overview on Spatial Information Extraction

In order to minimize the uncertainty of the classification, the information related to the spatial context needs to be included in the analysis. In the last years, methods based on spectral and spatial analysis have been developed to address such issue. Image segmentation, which is a procedure of partitioning of the image into homogeneous regions, has been widely explored in hyperspectral image analysis for the inclusion of spatial information. Several techniques can be found in the recent literature based on different strategies, for instance, watershed [100], minimum spanning forest [101] and multinomial logistic regression [10]. A different strategy is to employ advanced classifiers that combine apposite kernels for the spectral and spatial information into multi-kernel learning [102] and composite kernels [67] strategies. Such approaches, however, rely on features that are extracted ad hoc before the kernel computation. Approaches based on graph theory within a statistic framework, such as Markov random fields (MRFs), have also been exploited and optimized [66, 77], since the standard definition of the neighbour system in high-dimensional context makes the problem computationally intractable. A simple but effective way to include spatial information is the extraction of spatial features by applying filters (e.g, morphological operators, Gabor filters, wavelets decompositions) to the spectral bands. The obtained features are then used to enrich the input space that is used to learn the classifier. Recently, several promising methods have been developed as part of the mathematical morphology, which is a framework for the analysis of spatial structures based on set theory, lattice algebra, and integral geometry. These methods have been used in retrieving and modelling contextual information (e.g., geometry, shape, and edges) in particular for hyperspectral images [15, 40]. Morphological operators, such as attribute profiles (APs) [26], have been successfully exploited in the RS domain to include the spatial information in the data analysis, in tasks such as land-cover classification [27] and change detection [35]. APs provide a multi-level decomposition of the original image, which is obtained by applying a severe thinning / thickening filtering [11] on connected regions. APs are an interesting tool as they extract contextual information according to specific attributes, i.e., measurements that can be performed on a connected region. APs have many advantages: a) Different attributes can be defined, providing a variety of different image decompositions; b) Attributes can be measurements that are not related to the geometry of the region (e.g., standard deviation); c) The filtering is performed on connected regions, while the geometrical detail of the unfiltered regions is fully preserved. This high flexibility renders the APs a powerful tool for extracting complementary spatial information of the structures in the scene.

2.3 Independent Component Analysis

The high dimensionality of hyperspectral data can provide a better characterization of the spectral behaviour of different land-covers, however the redundancy of information should be detected and discarded in order to improve the discriminant analysis. In general, in pre-processing steps, a PCA transformation is applied to the data in order to reduce the dimensionality and obtain a better representation of the whole dataset with a smaller signal to noise ratio. Due to the nature of this orthogonal transformation, the approach results to be not class discriminant, obtaining a new feature space in which, usually, only the first few components are considered, neglecting possible information. Moreover, in the case of non-Gaussian processes, as the class distributions are in hyperspectral data, the variance may not be the quantity of interest. Based on higher order statistics, ICA could be used as a feature extraction approach for extracting the most representative components from hyperspectal images.

ICA is a well known unsupervised blind source separation technique, extensively used in several fields, aimed at finding statistically independent components (ICs) by only considering the observation of mixture signals. The problem of blind separation has been widely investigated in various field such as biomedical signal analysis and processing, e.g., in electroencephalography (EEG), in electrocardiography (ECG), in electromyography (EMG), in magnetoencephalography (MEG) and in electronystagmography (ENG) [60, 70, 82, 98, 104]. ICA-based methods are also applied to geophysical data processing, data mining, speech enhancement, image recognition and wireless communications [23]. During the most recent years, ICA has also received attention in the hyperspectral remote sensing data analysis, in particular for feature reduction [106], spatial unmixing [79], and classification [29, 32, 81, 105].

2.3.1 The Linear Mixing Model

In this section, we provide an introduction to the theoretical background on ICA. Let us consider n mixtures of random variables $x_1, x_2, \ldots, x_n$ which are defined as a linear combination of n random variables $s_1, s_2, \ldots, s_n$. The mixing model can be written as:

$$x_i = a_{i,1}s_1 + a_{i,2}s_2 + \ldots + a_{i,n}s_n \quad i = 1, \ldots, n. \tag{2.1}$$

In terms of random vectors, the model can be rewritten as:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.2}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ is the observed vector, $\mathbf{A}$ is the unknown mixing matrix with element $a_{ij}, i, j = 1, \ldots, n$ (which are real coefficients) and $\mathbf{s} = [s_1, s_2, \ldots, s_n]^T$ is the unknown source vector. By estimating the unmixing matrix of $\mathbf{A}$, called $\mathbf{W}$, the $\mathbf{s}$ vector that represents the independent components (ICs) is obtained by:

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{2.3}$$

The estimation of the ICA model is possible if the following assumptions and restrictions are satisfied: 1) the sources are statistically independent; 2) the independent components must have a non-Gaussian distribution; 3) the unknown mixing matrix $\mathbf{A}$ is assumed square and full rank. Under these conditions, the ICA model can be rewritten as:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \simeq \mathbf{s}, \tag{2.4}$$

where $\mathbf{W} \simeq \mathbf{A}^{-1}$. The problem can be solved by estimating $\mathbf{W}$ to obtain $\mathbf{y}$ that represents the best possible approximation of $\mathbf{s}$. Nevertheless, since $\mathbf{W}$ and $\mathbf{s}$ are unknown in the ICA model, three ambiguities necessarily hold:

1. The variances (energies) of the independent components cannot be determined. That is because any scalar multiplier in one of the sources s_i could always be canceled by dividing the corresponding column $\mathbf{a}_i$ of $\mathbf{A}$ by the same scalar.

2. For similar reasons, also the order of the independent components cannot be ranked.

3. The sign cannot be determined. This means that dark and bright regions may have the same meaning, which is not critical in most applications.

In the remote sensing literature, many ICA algorithms based on the maximization of different criteria can be found. Among them, the most used approaches are FastICA [55], JADE [16] and Infomax [5]. A theoretical definition of these algorithms will be provided in Chapter 3, where a detailed comparison among them is presented.

2.4 Morphological Operators

Mathematical Morphology (MM) is a well-established framework built upon set theory, lattice algebra and integral geometry, whose operators are exploited for the investigation of spatial features (i.e., geometry, shape, edges) of geometrical structures present in an image [94, 96]. Many operators have been presented in the literature and most of them are defined for binary and greyscale images. Dilation and erosion are the basic morphological operators. They are based on a moving window (or kernel), called structuring element (SE). Let us consider an object in the image as a connected region, which is a flat area where the pixels have the same value. In general, dilation causes objects to dilate or grow in size, whereas erosion causes objects to shrink. The effect of the filtering, i.e., the way objects dilate or shrink, depends upon the choice of the SE (shape and size). By combining dilation and erosion we obtain the closing and opening operators. Those operators are used to remove objects that cannot contain the SE, while preserving objects with a similar shape as the SE. However, a distortion of these objects that remains after the filtering is introduced, with a consequent loss of information related to the geometrical characteristics of the objects. This issue can be solved by the introduction of closing and opening by reconstruction, which are based on geodesic transformations and permit the preservation of the geometrical characteristics of the objects that are not removed. A further advancement was made by the introduction of morphological profiles (MP), which is a stack of filtered images obtained by a sequential application of a morphological filter by reconstruction with the SE increasing in size at each step. In general, a single application of a morphological operator is not enough for representing all the objects within the scene. The MP provides a multi-scale decomposition of the image, which goal is to obtain a better representation of the scene by taking into account that objects can appear at different scale. The reader is referred to [94, 96] for a complete background on morphological operators, and to [83] for the definition of MP. All the aforementioned operators are based on the use of a SE, making the filtering highly dependent on the shape of the used SE. A different approach was introduced in [11] with attribute filters, where the morphological transformation is attribute-based, removing the constraint of choosing a particular shape of the SE. Consequently, the effect of the filtering is not shape-dependent any more, whereas, it is adaptive to the considered region and its surrounding. In a similar ways as for the morphological filters, it is possible for the attribute filters to build a multi-scale representation of the images, i.e., morphological attribute profiles (APs) [26].

2.4.1 Attribute Filters and Tree Representations

A two-dimensional gray-scale image I, which can be defined as a mapping from the image domain $E \subseteq$ into $\mathbb{Z}$, can be fully represented as a set of connected components $\mathcal{C}$, defining a partition π_i of E. The way $\mathcal{C}$ is defined leads to different partitions. If we consider a connected operator ψ, by definition it will operate on I only by merging the connected components of the given set $\mathcal{C}$ [88]. Thus, the result of the filtering will be a new partition π_ψ that is coarser (i.e., containing less regions) than the initial one, meaning that for each pixel $p \in E$, $\pi_I(p) \subseteq \pi_{\psi(I)}(p)$ [78, Chapter 7]. The coarseness of the partition generated by a connected operator is determined by a parameter λ (i.e., a size-related filter parameter). Given two instances of the same connected operator with different filtering parameters, ψ_{λ_i} and ψ_{λ_j}, which we denote for simplicity as ψ_i and ψ_j, respectively, there is an ordering relation between the resulting partitions: $\pi_{\psi_i} \subseteq \pi_{\psi_j}$ given $\lambda_i \leq \lambda_j$. Among the different types of connected operators, attribute filters (AFs) are largely diffused. AFs filter connected components in $\mathcal{C}$ according to an attribute $\mathcal{A}$ that is computed on each component. In particular, the value of an attribute $\mathcal{A}$ is evaluated on each connected component in $\mathcal{C}$ and this measure is compared with a reference threshold λ in a binary predicate T_λ (e.g., $T_\lambda := \mathcal{A} \geq \lambda$). In general terms, if the predicate is true the component is maintained otherwise it is removed. According to the attribute considered, different filtering effects can be obtained leading to a simplification of the image. These effects are driven by characteristics such as the regions' scale, shape or contrast. Indeed, the high flexibility of the attribute filter relies on their capability in modelling the spatial information based on any measure that could be computed on a connected component, ranging from measures that are purely geometric (e.g. area, length of the perimeter, image moments, shape factors), to textural ones (e.g. range, standard deviation, entropy), and more.

Connected operators, such as attribute filter, can be implemented relying on representations of an image as a tree (e.g., the min- and max-tree) [89], which is a data structure that can represent each connected region as a node at different grey-levels. Splitting the transformation process in three distinct phases, the tree data representation is able to increase the filtering efficiency. In the first phase, the tree structure is created, where the connected components are identified and the hierarchical structure between nodes is defined. In the second phase, the criterion is evaluated at each node, preserving the nodes that satisfy a given binary predicate T, and removing the others. The final phase is the image restitution, where the pruned tree is converted back to the image. Attribute filters are among those filters that can be easily implemented on tree representations since they natively work on connected components (conversely to connected filters based on structuring elements). According to the way the set of connected components $\mathcal{C}$ is defined, different tree representations of the same image and hence different filters are obtained. A *max-tree* representation is obtained by considering the upper level set $\mathcal{U}(f) = \{X : X \in \mathcal{C}([f \geq \lambda]), \lambda \in \mathbb{Z}\}$. By pruning the max-tree, an anti-extensive filter is obtained (i.e., bright regions will be removed), thus, if the operator is also idempotent and increasing, it leads to an opening. Analogously, a *min-tree* representation and an attribute closing operator are obtained by considering the lower level set $\mathcal{L}(f) = \{X : X \in \mathcal{C}([f \leq \lambda]), \lambda \in \mathbb{Z}\}$, in which the connected components are defined according to a decreasing ordering relation. A different tree representation is given by the *inclusion tree* (or *tree of shapes*) in which the components are defined by a saturation operator that fills holes in components. A hole in a region $X \in \mathcal{C}$ is defined as a component that is completely surrounded by X. The inclusion tree is constructed by progressively saturating the image starting from its regional extrema (i.e., local maxima and minima in the image) until reaching only a single component fully covering E. The inclusion tree can equivalently be obtained by merging the upper and the lower level sets of an image [17]. The sequence of inclusions induced by saturation determines the components in the tree and their links defining the hierarchy. Since the saturation operator is contrast invariant (i.e., bright and dark regions will be treated the same), the filters operating on this tree will be self-dual (quasi self-dual in the case of discrete images).

2.4.2 Attribute Profiles

Let I be a digital grey-scale image and $\mathbb{Z}^n$ $(n = 2$, i.e., 2D images) its definition domain. A morphological transformation, ψ, is a mapping from a subset, E, of the image domain, I, to the same definition domain, E, with $\psi(I) \rightarrow \mathbb{Z}^n$. A profile $\Pi(I)$ is defines as a sequential filtering performed by considering a family of increasing criteria $T = \{T_\lambda : \lambda = 0, ..., L\}$, with $T_0 = \mathit{true}\, \forall C \subseteq E$, where

λ is a set of reference scalar values used in the filtering and C is a connected region in the image. Following this definition and considering a max- and a min-tree, the attribute opening profile, Π_{γ^T}, and the attribute closing profile, Π_{φ^T}, can be defined as follows:

$$\Pi_{\gamma^T}(I) = \left\{ \Pi_{\gamma^{T_\lambda}} : \Pi_{\gamma^{T_\lambda}} = \gamma^{T_\lambda}(I),\ \forall \lambda \in [0, \ldots, L] \right\} \tag{2.5}$$

$$\Pi_{\varphi^T}(I) = \left\{ \Pi_{\varphi^{T_\lambda}} : \Pi_{\varphi^{T_\lambda}} = \varphi^{T_\lambda}(I),\ \forall \lambda \in [0, \ldots, L] \right\}, \tag{2.6}$$

where $\varphi^{T\lambda}$ and $\gamma^{T\lambda}$ represent a morphological attribute closing and attribute opening, respectively. The attribute profile, $\Pi(I)$, is obtained by concatenating the opening and closing profiles as follows:

$$\Pi(I) = \left\{ \Pi^{-}_{\varphi^T}(I), I, \Pi_{\gamma^T}(I) \right\}, \tag{2.7}$$

where $I = \Pi_{\varphi^{T_0}} = \Pi_{\gamma^{T_0}}$ correspond to the original grey-scale I, and $\Pi^{-}_{\varphi^T}(I)$ represents the $\Pi_{\varphi^T}(I)$ in reverse order. It can be seen that the profile results in a vector of $2L + 1$ images.

Another important operator that is extensively used in this work is the so-called differential attribute profiles (DAP), $\Delta(I)$. It is obtained by computing the derivative of the AP, and it shows the residual of the progressive filtering, i.e., the connected regions that have been filtered between two adjacent levels of the AP, and their relative grey values. The DAP can be defined as follows:

$$\Delta(I) = \left\{ \Delta_{\varphi^T}(I), \Delta_{\gamma^T}(I) \right\}. \tag{2.8}$$

In this case, the obtained profile is represented by a vector of $2l$ images. A concept that worth mentioning is the possibility to have non-increasing criteria, which leads to more general definitions of opening and closing, with φ^{T_λ} and γ^{T_λ} denoting the thickening and the thinning profiles, respectively.

Analogously, when considering the contrast invariant operator ρ, which is based on the inclusion tree, the profile Π_ρ, named self-dual attribute profile (SDAP) [18, 28], can be obtained:

$$\Pi_\rho(I) = \left\{ \Pi_{\rho^{T_\lambda}} : \Pi_{\rho^{T_\lambda}} = \rho^{T_\lambda}(I),\ \forall \lambda \in [0, \ldots, L] \right\} \tag{2.9}$$

with $\Pi_{\rho^{T_0}}(I) = I$.

2.4.3 Extension to Multi-Channel and Multi-Attribute

Morphological operators are in general non-linear connected transformations computed on an ordered set of values. This means that any their extension to multivariate values is an ill-posed problem. The usual strategy is to apply the operator to each channel separately and fuse or create a stack of the obtained profiles. However, in the case of hyperspectral images, which feature space has a high dimensionality, this strategy becomes unattainable. In [7], a morphological operator was applied to a sub-space of the original data obtained by using PCA, and only the first most informative principal components (PCs) were considered. The concatenation of each obtained MPs resulted in a new structure, called extended morphological profile (EMP).

Analogously, the same procedure can be adopted for the APs case [27] and SDAP. Let I be a multi-channel data composed of r features. The extended morphological attribute profiles (EAP) is defined as the concatenation of the AP built on each feature f:

$$EAP(I) = \Big\{\Pi(f_1), \Pi(f_2), ..., \Pi(f_r)\Big\}. \tag{2.10}$$

A further extension, which is based on the flexibility of the AP in considering any possible measure applicable to a connected region as criterion, is the concatenation of the EAPs obtained by different attributes, which results in the definition of the extended multi-attribute profile (EMAP) [27]:

$$EMAP(I) = \Big\{EAP(I)_{a_1}, EAP(I)_{a_2}, ..., EAP(I)_{a_q}\Big\}, \tag{2.11}$$

where a_i represents the i-th given attribute, with $i = 1, 2, ..., q$. When the EMAP is built, a multiple presence of the original feature f is included in the profile. This is avoided by including them once only in the first EAP and not include them at all in the later EAPs.

Part II

Spectral Information Analysis

3 Analysis of ICA Algorithms

This Chapter presents a thorough study on the performances of different Independent Component Analysis (ICA) algorithms for the extraction of class-discriminant information in remote sensing hyperspectral image classification. The analysis aims to address a number of important issues regarding the use of ICA in the RS domain. Three scenarios are considered and the performances of the ICA algorithms are evaluated and compared against each other, in order to reach the final goal of identifying the most suitable approach to the analysis of hyperspectral images in supervised classification.

3.1 Introduction

ICA is a well known unsupervised blind source separation technique, extensively used in several fields, aimed at finding statistically independent components (ICs) by only considering the observation of mixture signals. When applied to hyperspectral data, ICA extracts the source components that generate the mixed signal measured by the sensor and the independent components refer to the different classes presented in the scene. Several algorithms have been proposed in the literature for implementing ICA based on the maximization of different criteria. Different algorithms provide diverse feature sets for classification. However, only a limited number of studies addresses the comparative performance of these algorithms. The available studies are in most cases related to biomedical signals analysis [25, 52, 62, 103] yielding results that are not consistent in terms of the most efficient ICA algorithm. All the review papers in the aforementioned domain agree on the identification of the three most prominent algorithms, that are Infomax [5], FastICA [55] and JADE [16]. However, an in-depth comparative study that addresses simultaneously fundamental questions on the properties and the efficiency of ICA implementations for the analysis of hyperspectral remote sensing images is still missing.

In the literature, a common approach is to apply ICA after dimensionality reduction, which is usually carried out by PCA. This approach is applied in [61, 81], where PCA is performed firstly and then the ICA is applied to the most important principal components with the accumulative variance of 99% and 98.58%, while the remaining components are discarded. In other studies [29, 105, 108], JADE and FastICA are used to extract subsets of ICs by exploiting the PCA phase implemented in the algorithms for dimensionality reduction. PCA aims to globally decorrelate the data and maximize the variance. The main limitation of PCA is that it is based on using the global second order statistics for the whole image. Consequently, the sensitivity to critical classes composed of a small number of pixels is reduced [22]. It is also well known that the criterion for retaining a certain number of components based on the calculation of the accumulated sum of eigenvalues is not an effective measure in terms of class discriminant, as demonstrated in [21]. Thus, PCA should not be used as a pre-processing tool for classification purposes [85]. Note that, according to the studies conducted in [56], ICA results obtained after PCA are in general not sufficient to estimate the ICs, since after the use of PCA only information on a subset of orthogonal components is available. In general, some weak ICs may be hidden in the dimensionality reduction process. An attempt to identify a better pre-processing approach than PCA is performed in [32], where a Noise-adjusted Principal Components (NAPC) is used for dimensionality reduction. The obtained results show that the principal components from the NAPC can better maintain the object information in the original data than those from PCA, allowing the ICA to provide better object classification.

The aim of this work is twofold; first, the identification of an effective strategy for the extraction of class-discriminant features with ICA. In the analysis different supervised feature extraction and selection approaches to dimensionality reduction (DR), which are investigated as pre-processing before applying ICA, are considered. Second, the addressing the lack in the literature of an extended comparative study on the three most frequently used implementations of ICA in the broader field of signal processing: Infomax, FastICA and JADE, aiming at assessing the most efficient and reliable methodology to follow when employing the ICA technique for accurate and cost efficient classification of hyperspectral images. Importantly the computational cost is assessed in relation to the number of samples used for the source estimation.

3.2 Independent Component Analysis (ICA)

In this work, three different implementations of ICA are investigated for feature extraction. In particular, the analysis focus on the Infomax, FastICA and the JADE algorithms, which are briefly introduced in the next subsection. As mentioned previously, the scope of this study is to present a complete comparison among the most widely used ICA algorithms in the remote sensing field. For the sake of scientific concreteness, the exploitation of more recent implementations of ICA that are used in the broader signal processing field is attempted. To the best of author's knowledge, one of the most recent implementation of ICA stated to outperform FastICA is RobustICA [109]. This method is presented in the next section. However, since the computational cost was excessively high, the method is evaluated in only one experiment and the results are discussed in the corresponding section.

An important issue that characterizes ICA transformation is the non prioritization of the ICs. Accordingly, multiple ICA applications result in different IC sets, which are diverse both in the order of appearance and in the content, thus making a performance comparison inconsistent. This behaviour is caused by the fact that ICA uses random vectors as initial projections. Wang and Chang addressed this problem in [106] proposing an initialization algorithm in conjunction with the virtual dimensionality (VD) [21] to generate an appropriate set of initial projections. The algorithm was designed for FastICA. However, in order to exploit the original setup without modifying the algorithms, the identity matrix of size $n \times n$ has been chosen as a common initialization for the ICA transformation. It is possible that in some cases the identity matrix gives worse results than a random initialization in terms of convergence time. The advantage in using a constant initialization is the consistency of the obtained components and their ordering. In this Section, a briefly introduction of the ICA algorithms considered in this study is provided.

3.2.1 Infomax

Infomax [5] is based on the minimization of the mutual information between the input and output of a neural network with non-linear units. The mutual information of a pair of random variables x and y can be defined as:

$$I(x;y) = H(x) - H(x \mid y), \tag{3.1}$$

where $H(x \mid y)$ is the conditional entropy defined as:

$$H(x \mid y) = H(x,y) - H(y). \tag{3.2}$$

Considering the entropy as a measurement of uncertainty and the mutual information as a measurement of the dependency between random variables, the matrix $\mathbf{W}$ is determined so that the mutual information among the components of the transformed vector $\mathbf{y}_i$ is minimized. The convergence is quite slow since the inverse matrix has to be computed at each iteration.

The algorithm's implementation used in this work is a part of the *EEGLAB* package [31]. The algorithm performs ICA decomposition using the logistic infomax ICA algorithm developed in [5] with a natural gradient feature as defined by Amary, Cichocki and Yang [2]. The algorithm performs a sphering (whitening) of the data in order to increase the convergence rate. This means that the unmixing matrix that is processed becomes

$$\mathbf{W} = \textit{weights matrix} \cdot \textit{sphere matrix}. \tag{3.3}$$

3.2.2 FastICA

The FastICA algorithm proposed in [55] is a very efficient and robust method for ICA. It exploits the negentropy J, which is a measurement of non-Gaussianity that gives a measure of the distance from normality. It is defined as:

$$J(\mathbf{y}) = H(\mathbf{y}_{Gaussian}) - H(\mathbf{y}), \tag{3.4}$$

with $\mathbf{y}$ being a random vector, $H(\mathbf{y})$ the entropy of $\mathbf{y}$ and $H(\mathbf{y}_{Gaussian})$ the entropy of a Gaussian random vector with the covariance matrix equal to the one of $\mathbf{y}$. Negentropy is always nonnegative and is zero only in case of Gaussian distribution. Because of the complexity of (3.4), the following moment-based approximation has been introduced [56]:

$$J(y) \propto [E\{(G(y)\} - E\{G(v)\}]^2, \tag{3.5}$$

where y is a standardized non-Gaussian variable, v is a standardized Gaussian variable and G is a non-quadratic function. The learning rule for FastICA is based on a fixed-point iteration scheme [55] that has been found to be considerably faster than using gradient descent methods for solving ICA. Before the FastICA algorithm can be applied, the input vector data should be centered and whitened. The scheme finds the maximum of the non-Gaussianity of $\mathbf{w}^T\mathbf{x}$. The basic fixed-point iteration for the estimation and decorrelation of one single independent component is:

$$\begin{aligned} \mathbf{w}_{i+1} &\leftarrow E\{\mathbf{x}g(\mathbf{w}_i^T\mathbf{x})\} - E\{\acute{g}(\mathbf{w}_i^T\mathbf{x})\}\mathbf{w}_i \\ \mathbf{w}_{i+1} &\leftarrow \mathbf{w}_{i+1} - \sum_{j=1}^{i}(\mathbf{w}_{i+1}^T\mathbf{w}_j)\mathbf{w}_j, \end{aligned} \tag{3.6}$$

where $g(u)$ is a non-quadratic function that represents the derivative of the non-quadratic function G in (3.5). The algorithm converges when the old and new values of $\mathbf{w}$ (where $\mathbf{w}$ represents one row of $\mathbf{W}$), point in the same direction. The FastICA algorithm can be used to perform projection pursuit as well, thus providing a general-purpose data analysis method that can be used both in an exploratory fashion and for the estimation of independent components (or sources). The algorithm can estimate the ICs in two different ways: 1) deflationary orthogonalization, which is shown in (3.6), 2) symmetric orthogonalization, which is shown in (3.7). The first approach performs orthogonalization using the Gram-Schmidt method, estimating the ICs one by one, while the second approach estimates all the ICs in parallel.

In our experiments, the second approach is used mainly for two reasons: 1) to avoid the cumulative error in the estimation, and 2) to estimate the ICs by a parallel computation, thus making the algorithm faster. In this case, the basic fixed-point iteration in FastICA with symmetric orthogonalization is as follows:

$$\begin{aligned} \mathbf{w}_{i+1} &\leftarrow E\{\mathbf{x}g(\mathbf{w}_i^T\mathbf{x})\} - E\{\acute{g}(\mathbf{w}_i^T\mathbf{x})\}\mathbf{w}_i \\ \mathbf{W} &\leftarrow (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}}\mathbf{W} \quad \text{with} \quad \mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_m)^T. \end{aligned} \tag{3.7}$$

3.2.3 JADE

The Joint Approximate Diagonalization of Eigenmatrices (JADE) [16] is a widely used and parameter-free implementation of ICA. In the pre-processing, a whitening transformation is performed on the mixtures, which makes the original components uncorrelated and thus independent in terms of second order statistics, and the unmixing matrix **W** orthogonal. The approach exploits the concept of cumulant tensor, which can be seen as a generalization of the covariance matrix. Let us consider the whitened unmixing matrix **W** and the cumulant tensor $\mathbf{F}(\mathbf{M})$, which is a linear symmetric operator. We can define an eigenmatrix **M** such that

$$\mathbf{F}(\mathbf{M}) = \lambda \mathbf{M}, \tag{3.8}$$

where every eigenmatrix has the form $\mathbf{M} = \mathbf{w}_n \mathbf{w}_n^T$, where $\mathbf{w}_n$ is a row of the unmixing matrix **W**. Thus, knowing the eigenmatrix of the tensor, it is easy to obtain the independent components. The main problem is that the eigenvalues are not distinct, and thus, the matrices cannot be uniquely defined. Considering that **F** is a linear combination in the form $\mathbf{w}_n \mathbf{w}_n^T$, it can be observed that the matrix **W** diagonalizes $\mathbf{F}(\mathbf{M})$ for any **M**. This means that it is important to choose a set of n different matrices $\mathbf{M}_i$ that makes the matrices $\mathbf{W}\mathbf{F}(\mathbf{M}_i)\mathbf{W}^T$ as diagonal as possible. The diagonality can be measured as the sum of squares of diagonal elements and is defined as:

$$J_{JADE}(\mathbf{W}) = \sum_i \|diag(\mathbf{W}\mathbf{F}(\mathbf{M}_i)\mathbf{W}^T)\|^2. \tag{3.9}$$

One method of join approximate diagonalization of the $\mathbf{F}(\mathbf{M}_i)$ is to maximize J_{JADE}.

3.2.4 RobustICA

RobustICA [109] is a recent method for deflationary ICA, in which the kurtosis is the general contrast function to be optimized. The method performs the optimization by a computationally efficient technique based on an optimal step size (adaption coefficient). The technique computes algebraically (i.e., without iterations) the step size globally optimizing the kurtosis in the search direction at each extracting vector update. In the derivation of the algorithm, no-simplifying assumptions concerning specific type of sources (real or complex, circular or noncircular, sub-Gaussian or super-Gaussian) are involved. The method presents a number of advantages with significant practical impact when compared to other kurtosis-based algorithms such as the original FastICA and its variants:

- Pre-whitening is not required, so that the performance limitations it imposes can be avoided and the sequential extraction (deflation) can be carried out, e.g., via linear regression.
- Sub-Gaussian or super-Gaussian sources can be extracted in the order specified by the user if the Gaussianity character of the sources is known in advance.
- The optimal step-size technique provides some robustness to the presence of saddle points and spurious local extrema in the contrast function.
- In the experimental analysis performed in [109], the method shows a very low computational cost measured in terms of source extraction quality versus number of operations, even without pre-whitening.

For further details about the implementation, it is suggested referring to [109].

3.3 Design of Experiments and Investigations

The analysis presented in this Chapter aims at identifying which ICA implementation provides better results in terms of classification accuracy and computational cost. This is studied in three scenarios:

- *Low-dimensional space:* This represents the most common scenario in remote sensing image analysis, where ICA is exploited. In general a small subset of features is obtained by performing dimensionality reduction on a high-dimensional feature space. The ICs are then extracted by processing the reduced subset. In the analysis the use of a number of feature extraction and feature selection methods used for dimensionality reduction was considered. The obtained results are compared against the general case in which PCA is exploited for feature reduction. The goal is to analyse and compare the performance of the three ICA algorithms applied to different subsets of features, identifying which pair of ICA algorithm and feature reduction technique gives the best classification accuracy.

- *High-dimensional space:* The performance of ICA is evaluated by considering the entire data set. The obtained feature space is then reduced by selecting the most informative features by exploiting a supervised feature selection algorithm. These features are then used in classification. The aim is to investigate the effectiveness of the ICA algorithms in extracting useful independent components directly from the original feature space, without initially projecting the data into a smaller subspace.

- *Spatial down-sampling:* In this scenario the ICA is applied to subsets of image samples obtained by spatially down-sampling the original image. The goal is to investigate how the performance of the ICA is affected by decreasing the number of samples used for the source estimation, and thus if it is possible to achieve classification accuracies that are similar to those obtained by using the entire data set. The exploitation of a reduced spatial subset would also positively affect the computational time of the ICA.

In the analysis, based on the above scenarios three experiments are designed as described below.

3.3.1 Experiment I: Low-Dimensional Space

Hyperspectral images are usually pre-processed by reducing the feature space in order to decrease the computational cost, discard redundant information and mitigate the noise contribution. Regarding the use of ICA for feature reduction, the most common strategy in remote sensing image analysis is to apply the PCA technique to the original image followed by the ICA. PCA is used to extract the high-variance components while filtering out the low-variance components. It is worth noting that the use of PCA is encouraged by the fact that it is implemented in the ICA as part of the algorithm for whitening purposes (see Section 3.2), where the user can decide to perform the dimensionality reduction by choosing the number of components to be retained. However, the PCA transformation provides a subset of components that after selection does not preserve class-separability. This also affects the independent components. In this experiment a different strategy is proposed and investigated. The aim is to provide a reduced feature set where the class-information is preserved and used as input to the ICA, avoiding the use of the PCA-based reduction approach. To this purpose, considering the context of supervised classification, the dimensionality reduction is performed by exploiting three supervised feature selection and extraction techniques, namely the Steepest Ascent (SA) search algorithm (in which the Jeffries-Matusita distance is used as the criterion function in feature selection), the Local Fisher Discriminant Analysis (LFDA), and the Non-parametric Weighted Feature Extraction (NWFE). The strategy adopted in the experiment consists of three steps: a) dimensionality reduction; b) application of the ICA to the obtained feature subset; c) evaluation in terms of classification accuracy of the effectiveness of the extracted ICs in discriminating the classes. The procedure is repeated for every ICA algorithm, considering different subsets of the retained components, starting from a minimum of 5 components up to 40 components. For simplicity, the different strategies are referred as *DR-approach-ICA*, where *DR-approach* is one of the feature extraction/selection techniques aforementioned (e.g., in the case of NWFE the strategy would be NWFE-ICA). The background information on the feature extraction and feature selection approaches that are used in this work is provided in Section 3.4.2.

3.3.2 Experiment II: High-Dimensional Space

Experiment II aims at investigating the effectiveness of the independent components obtained by considering the entire original hyperspectral data set, without performing any feature reduction (which reduces both redundancy and noise but may introduce information loss). The strategy adopted in the experiment is defined as follows: a) ICA is applied to the entire data set and all the components are retained; b) the most informative components are selected by applying the SA feature selection algorithm; c) the effectiveness of the subset is assessed in terms of classification accuracy. Also, in this case we take advantage of the training samples in order to select the best independent components. JADE's computation load is extremely high when the dimensionality of the feature space becomes large. This is due to the fact the JADE implementation has to estimate the initial vector of n eigenmatrices whose dimension is $n \times n$ (see Section 3.2.3), where n is the number of the sources to estimate. When n increases, the size of the initial projection increases as the cube of n, requiring the availability of a significant quantity of physical memory. For these reasons, JADE is not employed in this experiment and in general it should be avoided when a high-dimensional space is considered.

3.3.3 Experiment III: Spatial Down-Sampling

The third experiment aims at investigating the effectiveness of the ICA in extracting informative components when applied to a down-sampled data set (i.e., only a portion of the total number of pixels is analysed). The analysis consists of seven sub-experiments. In the first three sub-experiments the sampling rate is decreased by three different integer factors: 2, 3, 4. In the last four sub-experiments, different sizes of training samples are considered. The experiment has been conducted considering both scenarios 1 and 2, i.e., low-dimensional space and high-dimensional space, respectively. However, the results obtained from the analysis of the Botswana and Hekla data sets in high-dimensional space are very poor, especially when Infomax is used. Thus, for this scenario, only the results obtained by using FastICA performed on the Salinas data set are reported.

3.4 Experimental Setup

3.4.1 ICA parameter tuning

In the experimental analysis, an implementation of each ICA algorithm based on MATLAB (© The MathWorks, Inc.) scripting language is used.

Infomax

As mentioned in the Section 3.2, the initial weight matrix is initialized as an identity matrix. The training stops when the weight-change goes below the predefined threshold value, which is set by default at 10^{-6} when $n < 33$ and 10^{-7} otherwise, or after a maximum number of ICA training steps of 512.

FastICA

Different parameters need to be tuned. The non-quadratic function $g(u)$ is set as $\tanh(au)$ with $a = 1$, which is proven a good approximation of negentropy [56]. In order to avoid a random initialization, an identity matrix of size $n \times n$ is given in input as initial guess. In this work, the symmetric ortogonalization is chosen for the reasons explained in Subsection 3.2.2. The algorithm stops when the convergence is reached, meaning that the weight-change is less than 10^{-4}, or when the maximum number of iterations, which is set at 1000, is reached.

JADE

this technique is parameter free, i.e., no tuning is needed. The only experimental parameter that can be tuned is related to the stopping criterion, which is thresholded at 10^{-6} by default.

RobustICA

the method requires the tuning of few parameters. Two different approaches of deflation are possible: 1) via ortogonalization, 2) via linear regression. In this work deflationary ortogonalization is used. The threshold for statistical-significant termination test is set at 10^{-4}, while 1000 is the maximum number of possible iterations for each extracted source.

3.4.2 Feature Reduction

This section provides a briefly introduction to the feature reduction techniques used in this work.

Steepest Ascent (SA) Feature Selection

The supervised feature selection is based on the sub-optimal Steepest Ascent search algorithm using as criterion function the Jeffries-Matusita distance. The strategy is based on the search for constrained local extremes in a discrete binary space. More information can be found in [93].

Local Fisher Discriminant Analysis (LFDA)

It is a linear supervised dimensionality reduction method. It combines the ideas of Fisher's Discriminant Analysis (FDA) [41] and Locality Preserving Projections (LPP) [51]: between-class separability is maximized while within-class local structure is preserved. LFDA has an analytic form of the embedding matrix and the solution can be easily computed just by solving a generalized eigenvalue problem. Therefore, LFDA is scalable to large data sets and computationally reliable. More information can be found in [99].

Nonparametric Weighted Feature Extraction (NWFE)

the NWFE algorithm [63] takes advantage of the desirable characteristics of DAFE and Nonparametric Discriminant analysis (NDA) [44], while avoiding their shortcomings. DAFE is fast and easy to apply, but it is able to extract only $L-1$ features, with L the number of classes. This limitation reduces the performance particularly when the difference in mean values of classes is small. NDA focuses on training samples near the required decision boundary, but it does not perform well when either the covariance matrices of the classes are not equal. The main idea of NWFE is to assign different weights to every sample to compute the weighted means and to define new nonparametric between-class and within-class scatter matrices to obtain more than $L - 1$ features.

Principal Component Analysis (PCA)

it is one of the most widely exploited unsupervised approaches in feature reduction. The basic idea of PCA is to find the linearly transformed components that provide the maximum amount of variance possible. Usually the first few components account for a large proportion of the total variance of data and are used to reduce the dimensionality of the original data. However, all components are needed to accurately reproduce the correlation coefficients within the original image. PCA is an unsupervised technique and as such does not include label information for the data.

3.4.3 Classification

Support vector machine

In the experimental analysis, a support vector machine (SVM) classifier is employed for classification purposes, using a Radial Basis Function (RBF) kernel. The algorithm exploited is the LIBSVM [20] library developed for MATLAB®. The one-against-one multi-class strategy is used. The regularization parameter C and the kernel parameter γ are estimated by exploiting a grid-search using a 10-fold cross-validation. This means that the training set is first divided into 10 subsets of equal size, and then each subset is tested using the classifier trained on the remaining 9 subsets. In order to identify the best parameters, exponentially growing sequences of C and γ are considered. In particular, $C = \{10^{-2}, 10^{-1}, ..., 10^{4}\}$ and $\gamma = \{2^{-3}, 2^{-2}, ..., 2^{4}\}$.

Random forest

In the experiment II, the classification results obtained by using SVM are compared to the ones obtained by using Random Forest (RF) [12, 57] classifier.

This experiment is an exploratory one, in order to validated the role of the classifier in the entire process. Having seen that, the pattern of the classifiers behaviours follow similar trends, without loss of generality the entire experimental design is conducted using the SVM classifier. On the contrary of SVM, random forest does not require any parameter tuning.

3.4.4 Data Sets' Description

The experimental analysis is carried out on three real hyperspectral data sets characterized by different spatial and spectral resolutions. For each data set, the training samples and the test samples are generated in such way that the two sets results mutually exclusive (i.e., no shared samples between the two sets).

Salinas: The data set is described in Appendix A.1. For Salinas data set, the training set used in the experiments is made up of 15% randomly selected samples from each class.

Hekla: The data set is described in Appendix A.2. In the case of Hekla data set, the training set is generated by a random selection of 50 samples from each class.

Botswana: The data set is described in see Appendix A.3. In the case of Botswana, the training set is generated by random selection of 20% of samples from each class.

3.5 Experimental Results and Discussion

In this section, the results of the experiments described in Section 3.3 are presented and discussed in depth. For each experiment, the performance is reported in terms of classification accuracy, kappa coefficient and the computational time required for the convergence of the ICA. The individual ICA algorithms are compared and analysed using the three performance measures. For a better understanding of the obtained results, the overall classification accuracies are given in percentage (%), while the comparison between accuracies is given in percentage points (*pp*), which are simply the arithmetic difference of two percentages.

3.5.1 Experiment I: Low-Dimensional Space

The results of the analysis conducted in experiment I are depicted in Figure 3.1. The taxonomy is based on the different DR approaches, showing for each of them the behaviour of the classification accuracy for the three ICA algorithms considering different subsets of components. Table 3.1 reports the best results obtained by the ICA algorithms for each strategy, showing the number of retained components, the overall accuracy (OA), the kappa coefficient (k) and the computational time (CPU time).

For the Salinas data set (Figures 3.1, left column), the strategy LFDA-ICA obtains the best classification accuracy. In this case, all the ICA algorithms perform similarly in terms of classification accuracy and number of components required. JADE algorithm outperforms the others in terms of overall accuracy, which reaches a score of 95.48%. However, in terms of computational cost, FastICA requires about 30% less with respect to JADE and 60% less CPU time with respect to Infomax. NWFE-ICA appears to be the second best strategy. In this case, JADE and FastICA provide very similar trends, obtaining as highest OA 94.99% and 95.07%, respectively, while the performance of Infomax are strongly affected when increasing the number if ICs. All the ICA algorithms obtain the highest accuracy with 20 components. In terms of computational cost, JADE requires about 20% less CPU time than FastICA and about 75% less than Infomax. In the case of the PCA-ICA strategy, the maximum accuracy obtained by JADE is 95.10% (25 ICs). This requires an higher CPU time than FastICA and Infomax, which provided as highest OA 94.28% and 94.62%, respectively (20 ICs). However, considering the global trend, JADE provides the best classification accuracies with respect to FastICA and Infomax. Also in the case of SA-ICA, JADE provides in general a better global trend with respect to the other two ICAs. In terms of computational time, JADE and FastICA require equal CPU time, while Infomax results to be the slowest.

Considering the results obtained for Hekla data set (see Figures 3.1, second column), the NWFE-ICA approach achieves the best classification accuracy compared to all the other strategies. All the three ICAs provide the best performance when 5 components are retained. The obtained OAs are very close to each other, however, in terms of computational cost, JADE requires about 60% less than FastICA and more than 95% less than Infomax. The LFDA-ICA strategy provides the best OA when 10 components are considered, with an accuracy slightly higher compared to the one obtained with the PCA-ICA. On the other hand, considering different subsets of components, PCA-ICA shows

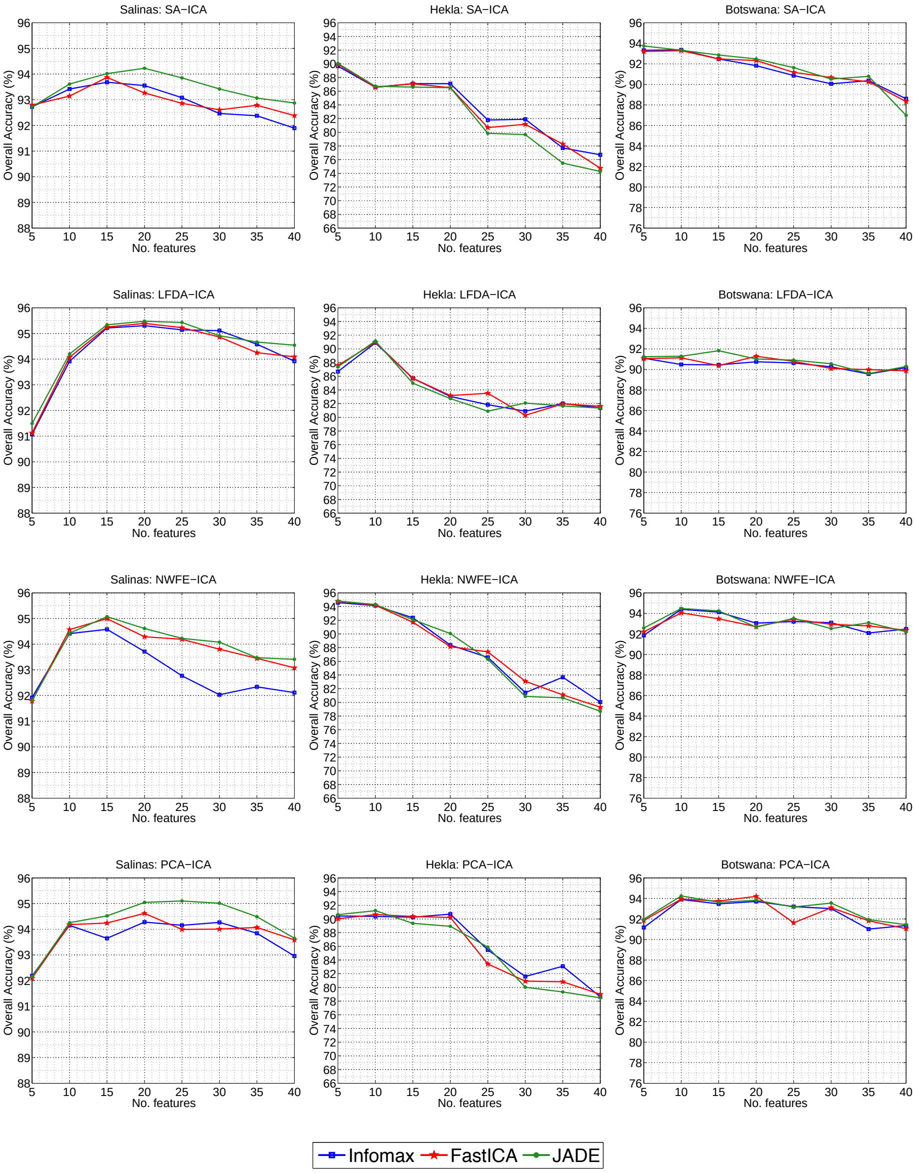


Figure 3.1: Experiment I: comparison of the overall classification accuracy obtained by Infomax, FastICA and JADE for different DR strategies (SA-ICA, LFDA-ICA, NWFE-ICA, PCA-ICA) and different number of features: (left column) Salinas data set; (middle column) Hekla data set; (right column) Botswana data set.

a better trend than the LFDA-ICA. In terms of computational time, JADE and FastICA require similar computational time, which is about 80% less than Infomax. The SA-ICA strategy provided similar (and in some cases better) results with respect to the LFDA-ICA even thought it is the approach that gave the lowest maximum classification accuracy. Each ICA algorithm reached its best performance with 5 components. JADE and FastICA obtained very close results in terms of both OA and CPU time, while Infomax provided a slightly lower OA requiring a much higher computational time.

In the analysis of the Botswana data set (Figures 3.1, third column), the best accuracies are obtained when applying NWFE-ICA and PCA-ICA. The former provided the highest accuracy when 10 components were considered, where JADE and Infomax provided a slightly higher accuracy than FastICA. However, the computational time required by JADE is about 50% less than FastICA and 90% less than Infomax. A similar analysis can be done for both PCA-ICA and SA-ICA strategies. LFDA-ICA approach provided the lowest classification accuracy with respect to the other strategies. Also in this case, the three ICA algorithms obtain very similar classification accuracies, while the computational time required by JADE is about 50% and 60% smaller than that required by the FastICA and the Infomax, respectively.

In general, the use of feature extraction algorithms for pre-processing achieves higher classification accuracies than to using feature selection. The reason might be that a better minimization of the noise contribution is achieved when the feature extraction algorithms are employed. Considering the best results reported in the Table 3.1 (highlighted in gray), for each data set JADE achieved accuracies that are slightly higher than those of the other ICA algorithms. The highest improvement was achieved in case of Botswana, where JADE outperformed Infomax improving the OA by 1.384*pp* In terms of computational time required to achieve the best classification accuracy, the JADE's performance is comparable to the one obtained by FastICA. Infomax resulted in general the worst performing technique, both from the computational time and the classification accuracy points of view.

In this experiment, another technique that was recently proposed in the neuroscience field, RobustICA [109], has been used. In [109], RobustICA is presented and compared to the kurtosis-based FastICA, considering the deflationary orthogonalization (i.e., the components are extracted one by one). The study is applied to the biomedical problem of atrial activity (AA) extrac-

Table 3.1: Classification results obtained in Experiment I (Figure 3.1). Only the best results are reported. Classification results obtained on the original spectral channels are given for comparison. ”No. feat.” denotes the number of feature retained, ”OA (%)” denotes percentage overall accuracy, ”k” indicates the kappa coefficient and ”Time” gives the computational time in seconds.

			SA-ICA			LFDA-ICA		
		Spectr.	Infomax	FastICA	JADE	Infomax	FastICA	JADE
Salinas	No. feat.	204	15	15	20	20	20	20
	OA (%)	94.55	93.68	93.87	94.23	95.30	95.39	**95.48**
	k	0.91	0.93	0.93	0.94	0.90	0.95	0.95
	Time (s)	-	12.70	7.20	7.22	17.17	7.10	10.56
Hekla	No. feat.	157	5	5	5	10	10	10
	OA (%)	93.89	89.65	89.93	90.06	90.88	90.96	91.14
	k	0.80	0.88	0.88	0.86	0.90	0.90	0.90
	Time (s)	-	44.71	2.38	1.62	34.90	3.40	3.95
Botswana	No. feat.	145	10	10	5	15	20	15
	OA (%)	93.42	93.32	93.28	93.74	90.44	91.28	91.82
	k	0.93	0.93	0.93	0.93	0.90	0.91	0.91
	Time (s)	-	26.90	2.50	0.49	47.50	35.08	18.02
			NWFE-ICA			PCA-ICA		
			Infomax	FastICA	JADE	Infomax	FastICA	JADE
Salinas	No. feat.		15	15	15	20	20	25
	OA (%)		94.58	94.99	95.07	94.28	94.62	95.10
	k		0.94	0.94	0.95	0.93	0.93	0.95
	Time (s)		14.33	4.52	3.57	17.54	7.91	29.70
Hekla	No. feat.		5	5	5	10	5	10
	OA (%)		94.57	94.79	**94.81**	90.64	90.43	91.21
	k		0.94	0.94	0.94	0.89	0.89	0.90
	Time (s)		26.91	1.42	0.58	32.96	1.87	3.85
Botswana	No. feat.		10	10	10	10	10	10
	OA (%)		94.43	94.00	**94.47**	93.89	93.93	94.24
	k		0.94	0.94	0.94	0.93	0.94	0.94
	Time (s)		42.28	9.51	3.01	39.58	4.65	4.45

tion in atrial fibrillation (AF) electrocardiograms (ECGs). In that context, RobustICA was claimed to be more efficient than FastICA in providing better sources with a lower computational cost. In [74] RobustICA was compared to JADE for ECG artefacts removal from Electroencephalogram (EEG) signals. Also in this case, RobustICA was preferred than JADE. Even if the scope of

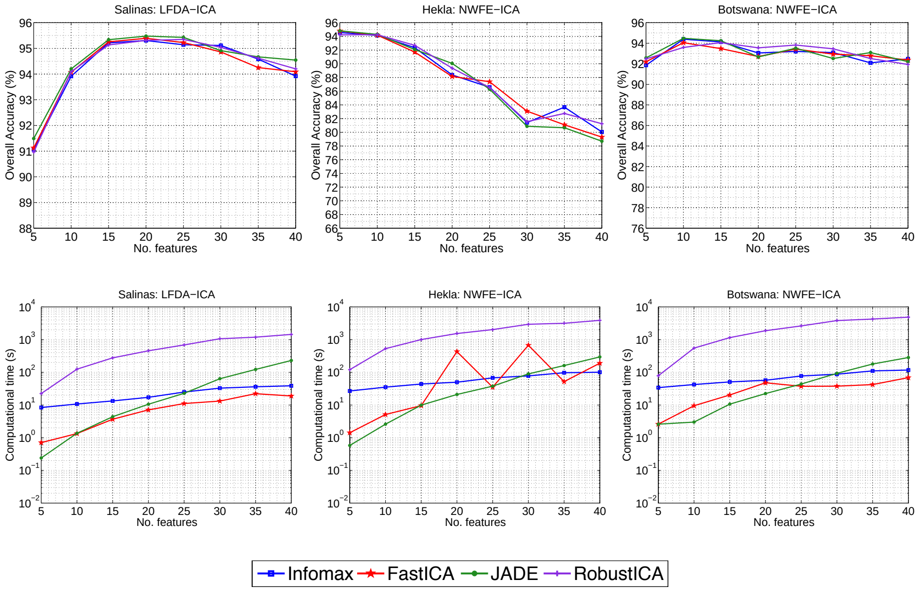


Figure 3.2: Experiment I: comparison of the overall classification accuracy and computational cost obtained by Infomax, FastICA, JADE and RobustICA versus the number of features, considering the best DR strategies: (first column) LFDA-ICA for Salinas data set, (second column) NWEFE-ICA for Hekla data set; (third column) NWFE-ICA for Botswana data set.

the study is not to exploit all the existing implementations but only the most widely used in the remote sensing field, an exploratory experiment was carried out using the aforementioned implementation. Taking into account these results, the algorithm is tested and compared to the best cases (i.e., LFDA-ICA for Salinas, NWFE-ICA for Hekla and Botswana). Figure 3.2 shows the comparison between the ICA algorithms, while the best obtained results are reported in Table 3.2 From the experimental analysis, it can be seen that in terms of classification accuracies, the performance of RobustICA is in line with the ones obtained by FastICA, Infomax and JADE approaches, providing the best accuracy among the other ICA algorithms in case of Botswana. However, the required computational cost is much higher, resulting in a extremely slow computational time (especially in case of Botswana), which seems to increase linearly with the number of extracted components. This methods have been used in neuroscience field but never in remote sensing field. However, the nature and the properties of the hyperspectral images are different from signals analyzed in neuroscience field. Considering the obtained results and because

Table 3.2: Classification results obtained in Experiment I considering RobustICA algorithm and the best DR strategies (Figure 3.2). Only the best results are reported. ”No. feat.” indicates the number of feature retained, ”OA (%)” denotes percentage overall accuracy, ”k” gives the kappa coefficient and ”Time” gives the computational time in seconds.

		Infomax	FastICA	JADE	RobustICA
Salinas: LFDA-ICA	No. feat.	20	20	20	25
	OA (%)	95.30	95.39	95.48	95.35
	k	0.90	0.95	0.95	0.95
	Time (s)	17.17	7.10	10.56	688.74
Hekla: NWFE-ICA	No. feat.	5	5	5	10
	OA (%)	94.57	94.79	94.81	94.25
	k	0.94	0.94	94	0.94
	Time (s)	26.91	1.42	0.58	529.01
Botswana: NWFE-ICA	No. feat.	10	10	10	15
	OA (%)	94.43	94.00	94.47	94.59
	k	0.94	0.94	0.94	0.94
	Time (s)	42.28	9.51	3.01	1151.5

of the significant computational cost required even when feature reduction is performed in pre-processing, the use of these techniques does not seem appropriate for hyperspectral images. Thus it will not be considered further in the experimental analysis.

3.5.2 Experiment II: High-Dimensional Space

Following the design of the experiment I, also in this case the overall accuracies are entirely depicted in Figure 3.3, while the best results are reported in Table 3.3. This Table shows for each chosen algorithm (in this case only Infomax and FastICA), the number of retained components, the overall accuracy (OA), the kappa coefficient (k) and the computational time, which is the estimation cost of the ICA on the entire original data set. Both of the results obtained by using SVM and RF classifiers are reported. From the analysis of the results obtained by using SVM, it can be noticed that FastICA outperforms Infomax for all the three data sets, increasing the OA by 2.164*pp* in the case of Salinas and by 1.334*pp* in the case of Botswana. In the case of Hekla, the results are quite similar, showing a little improvement of 0.56*pp* for FastICA. The kappa coefficient follows a similar trend as well. FastICA requires

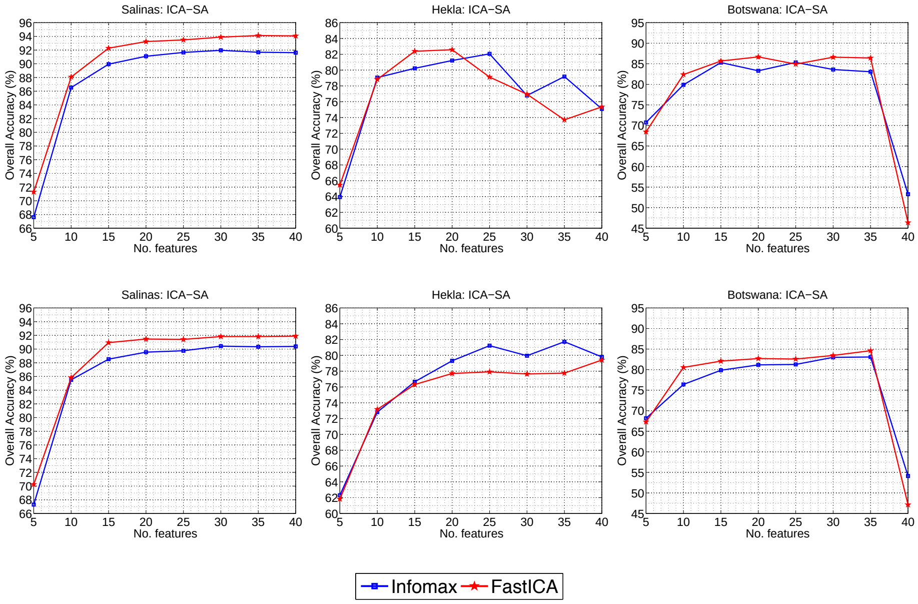


Figure 3.3: Experiment II: comparison of the overall classification accuracy obtained by Infomax and FastICA versus the number of features by using (top row) SVM, and (bottom row) RF, for Salinas, Hekla and Botswana data sets.

a CPU time, which is one order of magnitude less with respect to Infomax, confirming the superiority of the fixed-point algorithm. The use of the entire data set, without performing dimensionality reduction, assures that there is no information loss in the process, as it may happen when any feature extraction/selection technique is used. Considering the case of Salinas, the selection of 35 ICs provided an OA of 94.12%, which is quite close to the one obtained in the first experiment by the SA-ICA approach (94.23%) (see in Table 3.1). However by applying ICA to the entire data set there is no noise reduction, the extracted components carry noise which affects the final classification results. This becomes more evident in the experimental results obtained for the Hekla and Botswana data sets, in which the accuracies are quite low with respect to the ones obtained in the first experiment.

Considering the results obtained by using RF, the overall accuracies are in general lower than the ones of SVM. However, the obtained results follow a similar trend of the ones obtained with SVM in case of Salinas and Botswana, where FastICA outperform Infomax, while in case of Hekla, Infomax is the one that obtained the highest overall accuracy.

Table 3.3: Classification results obtained in Experiment II. Only the best results are reported. "No. feat." indicates the number of feature retained, "OA (%)" denotes percentage overall accuracy, "k" gives the kappa coefficient and "Time" gives the computational time in seconds.

		SVM		RF	
		Infomax	FastICA	Infomax	FastICA
Salinas	No. feat.	30	35	30	40
	OA (%)	91.96	94.12	90.43	91.89
	k	0.91	0.93	0.89	0.91
	Time(s)	2253.16	281.22	2253.16	281.22
Hekla	No. feat.	25	20	35	40
	OA (%)	82.05	82.58	81.73	79.41
	k	0.80	0.80	0.79	0.77
	Time(s)	3694.57	2506.31	3694.57	2506.31
Botswana	No. feat.	25	20	35	35
	OA (%)	85.33	86.67	83.06	84.60
	k	0.84	0.86	0.82	0.83
	Time(s)	3028.29	337.19	3028.29	337.19

3.5.3 Experiment III: Spatial Down-Sampling

In this experiment the performances of the ICA algorithms are investigated and compared when a spatial down-sampling of the ICA's input data is performed. The analysis presents the results obtained by considering two scenarios.

Low-dimensional space

In this experiment, the three ICA algorithms were tested on the three different data sets taking into account only the *DR-approach-ICA* strategies that gave the best results in terms of accuracies in the experiment I (see Table 3.1), i.e., the LFDA-ICA in case of Salinas data set, and the NWFE-ICA in the case of both Hekla and Botswana data sets. Table 3.4 reports the number of samples employed in the experiment. *None (all samples)* denotes the case in which the entire image is considered. This coincides with the results obtained in experiment I, and they are reported here for comparison. *Ds2x, Ds3x, Ds4x* denote a decrease of the sampling rate of a factor 2, 3 and 4, respectively. *Train 1* denotes the initial training sets, i.e., 15% of the ground truth samples in the case of Salinas, 50 samples for each class in case of Hekla, and 20% of the ground truth in

Table 3.4: Experiment III: description of the data set considered in terms of numbers of samples.

Downsampling	Salinas	Hekla	Botswana
None (all samples)	111104	336000	377856
Ds2x	55552	168000	188928
Ds3x	37035	112000	125952
Ds4x	27776	84000	94464
Train 1	8112	600	644
Train 2	5403	480	482
Train 3	2697	360	323
Train 4	1075	120	161

case of Botswana. *Train 2* denotes 10% of the ground truth in case of Salinas, 40 samples for each class in the case of Hekla, and 15% of the ground truth samples in the case of Botswana. *Train 3* denotes 5% of the ground truth in case of Salinas, 30 samples for each class in the case of Hekla, and 10% of the ground truth samples in the case of Botswana. *Train 4* denotes 2% of the ground truth in case of Salinas, 10 samples for each class in the case of Hekla, and 5% of the ground truth samples in the case of Botswana. For an easier interpretation of the global behaviours of the ICA algorithms, the results are reported as graphs in Figures 3.4-3.6, which show the OAs and the computational time obtained when different down-sampling factors and training samples are considered. For a more exhaustive analysis of the effect of the reduction of the training samples on the performance of the ICA algorithms, two different approaches are considered. The first approach takes into account a real life situation, where the same number of training samples for the classifier and the ICA is considered, In this case, the observed variation of the classification accuracy trend is caused of the degradation of both the ICs and the effectiveness of the classifier. In the second approach, the number of the training samples in input to the ICA varies, while the one in input at the classifier (which coincides with the original training set (i.e., *Train. 1*)), remains the same. Even if this strategy is unusual, it permits us to evaluate the real effect of the reduction of the training samples on the performance of the ICA. For each of the Figures 3.4, 3.5 and 3.6, the top row reports the results obtained considering the first approach, the middle row reports the results obtained by the second approach and the

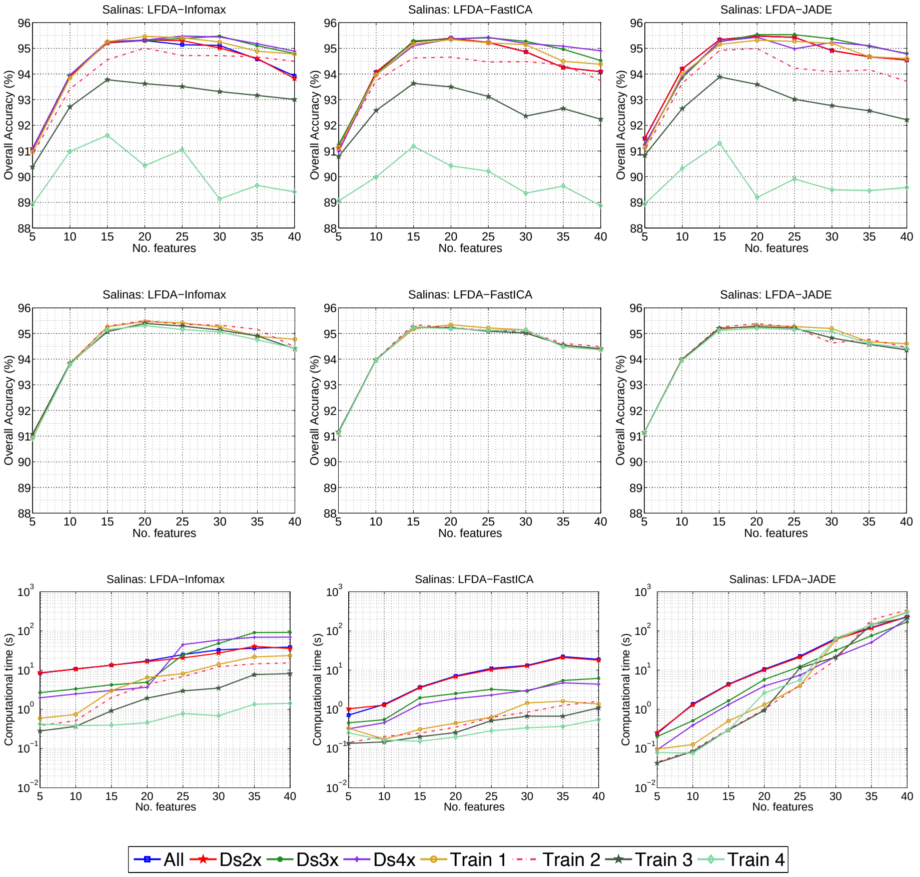


Figure 3.4: Experiment III in low dimensional scenario: comparison of the overall classification accuracy provided by (first column) Infomax, (second column) FastICA, (third column) JADE, for different number of samples on Salinas data set. Top row shows the results obtained by exploiting the first approach (i.e, the same number of training samples are given as input to both the ICA and the classifier), while the middle row shows the ones obtained by exploiting the second approach (i.e, the number of the training samples given as input to the ICA varies, while the one in input to the classifier remains the same). The bottom row shows the computational time related to the first approach.

bottom row shows the log-lin plots of the computational times (plotted in logarithmic scale on the y-axis) for each of the sub-experiment reported in the top row. The analysis is done for different numbers of components (plotted in linear scale on the x-axis) retrieved each time.

Focusing on the Salinas data set, the performances of each ICA technique, obtained by applying a down-sampling of a factor 2, are very close to the ones obtained by using all the samples. Similarly, the computational times required

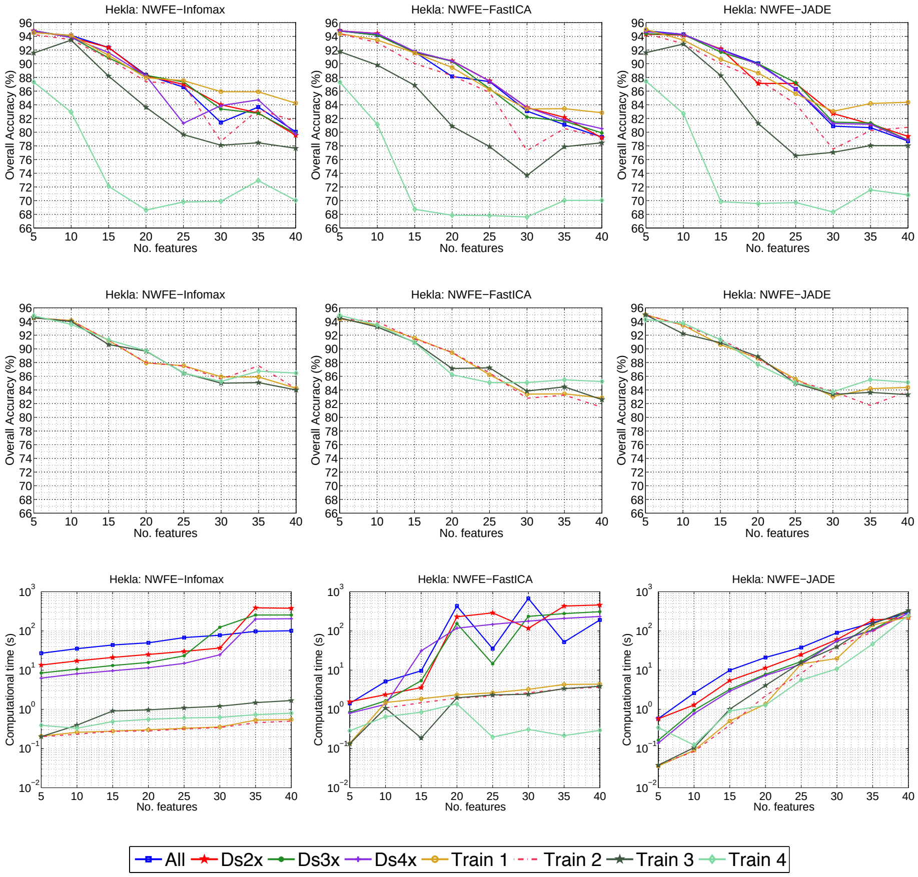


Figure 3.5: Experiment III in low dimensional scenario: comparison of the overall classification accuracy provided by (first column) Infomax, (second column) FastICA, (third column) JADE, for different number of samples on Hekla data set. Top row shows the results obtained by exploiting the first approach (i.e, the same number of training samples are given as input to both the ICA and the classifier), while the middle row shows the ones obtained by exploiting the second approach (i.e, the number of the training samples given as input to the ICA varies, while the one in input to the classifier remains the same). The bottom row shows the computational time related to the first approach.

for the convergence are similar. Improvements in the overall accuracy, and especially in the computational time, are more evident when higher factors are considered. In particular, when Infomax is used, the best OA (95.48%) is obtained when the sampling rate is decreased by a factor 3 and 4. While the OAs are quite similar, the computational times improve with respect to the case in which all the samples are used. For example the computational time decreased to 6.47 *s* by using only the training samples, achieving an OA of 95.46%. Con-

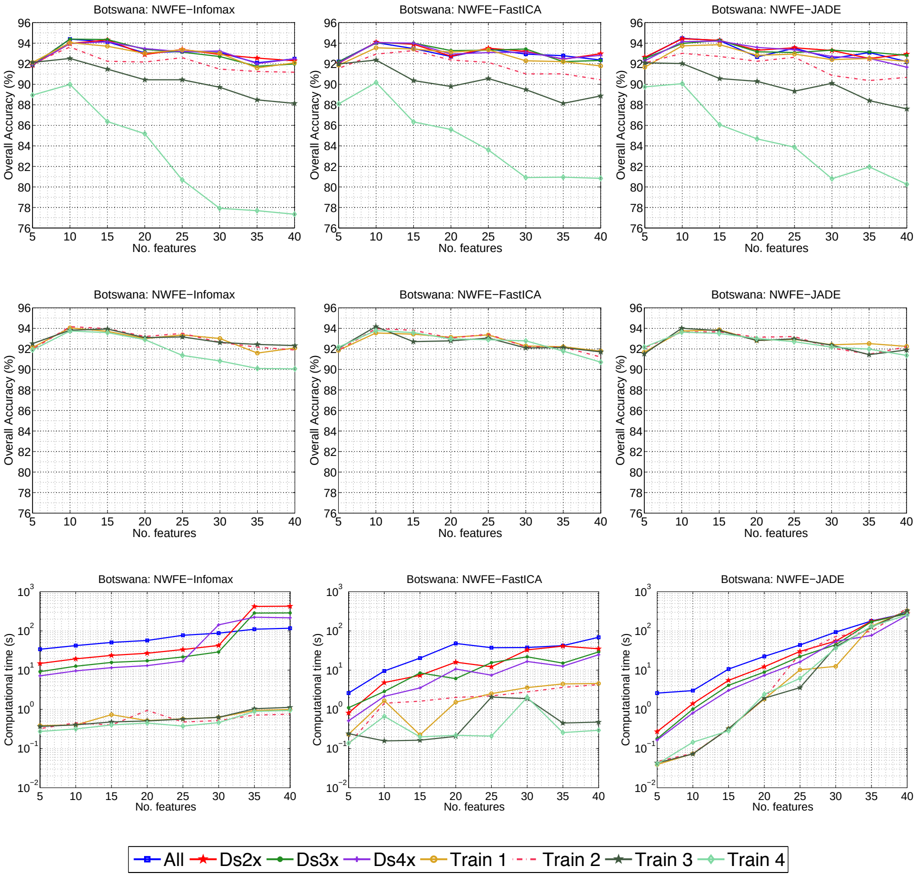


Figure 3.6: Experiment III in low dimensional scenario: comparison of the overall classification accuracy provided by (first column) Infomax, (second column) FastICA, (third column) JADE, for different number of samples on Botswana data set. Top row shows the results obtained by exploiting the first approach (i.e, the same number of training samples are given as input to both the ICA and the classifier), while the middle row shows the ones obtained by exploiting the second approach (i.e, the number of the training samples given as input to the ICA varies, while the one in input to the classifier remains the same). The bottom row shows the computational time related to the first approach.

sidering FastICA, the best OA (95.43%) was achieved by applying a down-sampling of a factor 4, while the obtained computational time was 2.31 *s*. For JADE, the highest OA (95.53%) is obtained by reducing the sample rate by a factor 3, halving the computational time (5.74 *s*) with respect to the case when the entire data set is used (10.56 *s*).

Analysing the results obtained for the Hekla data set, the down-sampling reduces the computational times. For Infomax, 6.25 *s* is the computational time

Table 3.5: Classification results obtained in Experiment III considering the high dimensional scenario. The results are related to Salinas data set by using FastICA. "No. feat." indicates the number of feature retained, "OA (%)" denotes percentage overall accuracy, "k" gives the kappa coefficient and "Time" gives the computational time in seconds.

	Ds2x (55552 samples)		Ds3x (37035 samples)		Ds4x (27776 samples)		Training samples (8112 samples)	
No. feat.	OA (%)	k	OA (%)	k	OA (%)	k	OA (%)	k
5	69.25	0.65	65.98	0.62	65.36	0.61	62.48	0.57
10	87.53	0.86	85.32	0.84	85.22	0.84	84.13	0.82
15	92.43	0.92	89.57	0.88	90.39	0.89	90.45	0.89
20	93.55	0.93	90.91	0.90	91.35	0.90	91.57	0.91
25	93.72	0.93	92.17	0.91	92.06	0.91	**91.59**	0.91
30	93.78	0.93	92.01	0.91	**92.63**	0.91	91.27	0.90
35	**94.21**	0.94	**92.62**	0.92	91.34	0.90	90.57	0.90
40	93.97	0.93	91.17	0.90	90.89	0.90	90.15	0.90
Time (s)	102.84		84.57		208.88		161.46	

reached with a down-sampling of a factor 4, which is a much lower compared to the 26.91 *s* obtained in the experiment I. While time improved, the OA resulted 94.83%. The classification performances obtained by using the FastICA remain the same for all the sub-experiments, while the computational time decreased from 1.42 *s* (obtained in the experiment I) to 0.80 *s*. The JADE algorithm improved the computational time, which decreased from 0.58 *s* to 0.036 *s* when using only the training samples, achieving an OA of 95.00%, which is slightly higher than the previous OA of 94.81%.

For Botswana data set, the down-sampling does not improve the performances of JADE, which achieved a slightly lower OA (94.28 % with 15 components) than the one obtained in the experiment I (94.47% with 10 components), with a computational time that slightly increased to 4.1 *s*. A different behaviour can be noticed for Infomax and FastICA. Both of them decreased the computational cost, which was reduced by 77% (from 42.28 *s* to 9.57 *s*) for Infomax, and by 50% (from 9.51 *s* to 4.81 *s*) for FastICA, without decreasing the classification accuracies (from 94.43% to 94.51% for Infomax, and from 94.00% to 94.05% for FastICA).

Comparing the only variation in the number of training samples, it can be seen that a decrease of the number of training samples in the first approach (top row) strongly affects the classification accuracy. The analysis performed by varying only the number of training samples given as input to the ICA (middle row), keeping constant the number of the training samples in input to the classifier, shows that the quality of the extracted ICs is not affected, providing similar trend of classification accuracy for a different number of training samples. This result points out that a variation in the number of the training samples affects more incisively the classifier rather than the ICA performance. In terms of computational time, decreasing the number of training samples coincides in general with a decrease of the computational cost. This is more evident in the case of Infomax and FastICA, while in case of JADE the computational cost increases with the number of retained components.

In general, it can be stated that when dimensionality reduction is performed, down-sampling as a pre-processing approach can contribute to the improvement of the computational time of the ICA algorithms without decreasing the overall classification accuracy. This finding is significant, especially when the computational time is an important aspect of the analysis, as is the case of the analysis of hyperspectral images.

High-dimensional space

The results reported in Table 3.5 show the performances of ICA when a spatial down-sampling of the image is performed before applying ICA in the high-dimensional space scenario are considered. The convergence capability of the ICA is strongly affected by the decreased number of input samples when applied to the entire image, while it fails to converge when very few input samples are considered. For this reason only the results obtained by using the first five data sets described in Table 3.4 are considered. For similar reasons, only the results obtained by applying FastICA on Salinas data set are reported. Table 3.5 reports the OA and the k coefficient for different subsets of features, while the computational time is referred to the total time requested for the extraction of all the ICs (i.e., 204 ICs). It can be seen that performing a down-sampling of factor a 2 (meaning that half of the total number of samples are discarded), the obtained overall classification accuracy (94.21%) is quite similar to the one achieved by considering all the samples (94.12%, see Table 3.3), whereas the computational time for extracting the ICs decreases by $63,4\%$. However, when the down-sampling factor increases, the performance of the ICA decreases, i.e., the classification accuracy decreases while the required computational time to converge increases.

3.6 Conclusions

In this Chapter, a detailed comparison among three widely used ICA algorithms (i.e., Infomax, FastICA and JADE) for hyperspectral image classification was presented. The analysis took into account different scenarios in order to compare and identify the best strategy for extracting class-discriminant components based on the use of ICA. The ICA algorithms were tested in both low and high dimensional spaces.

In the first scenario, ICA algorithms were tested performing dimensionality reduction with alternative strategies rather than the PCA (which usually is implemented in conjunction with the ICA algorithms and used for retaining a certain number of components). Supervised feature selection/extraction techniques have been exploited and compared to the case in which PCA is used. The results of the analysis pointed out that the exploitation of prior information in feature extraction methods for dimensionality reduction allows ICA algorithms to provide better feature sets which led to more accurate classifications. In this scenario, which is the most common in the analysis of hyperspectral images, JADE was the ICA approach that provided the best performance in terms of classification accuracy, while it provided results comparable to the ones obtained by FastICA in terms of computational time (in many cases it was faster). Infomax resulted in general to be the worst in terms of both computational time and classification accuracy.

The second scenario was aimed at investigating the performance of ICA when the entire data set is considered. Using the entire data set, without applying any dimensionality reduction, assures that no information is lost before performing the ICA. The analysis in this scenario showed that FastICA outperforms Infomax both in terms of computational time and classification accuracy. In this case, JADE could not be exploited since it requires a massive computational load when the number of estimated components becomes high. When the entire data set is considered, there is theoretically no information loss. However, the full set of selected components is more noisy, thus affecting the classification results.

The third scenario showed that the reduction of the number of samples on which applying ICA can in general improve the ICA convergence speed, without decreasing the classification accuracies. The approach is more effective in low dimensional spaces, where there are no issues with the Hughes' phenomenon, especially when the number of training samples given as input to both the ICA algorithm and the classifier are chosen properly. Indeed, the experiments showed that SVM is more affected by a decrease of the number of the training samples than the ICA, which can provide "good" ICs even when few samples are exploited for the transformation. This observation becomes very important in applications for which the computational time and the number of available samples are crucial aspects. Consequently, the inclusion of the analysis of prior information in computational efficient strategies should foster the development of new ICA-based methodologies for the analysis of large hyperspectral remote sensing images.

Part of this Chapter is published in:

N. Falco, L. Bruzzone, and J. A. Benediktsson. A Comparative Study of Different ICA Algorithms for Hyperspectral Image Analysis. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2013 5th Workshop on*, Gainesville, Florida, 2013 [34].

N. Falco, J. A. Benediktsson, and L. Bruzzone. A Study on the Effectiveness of Different Independent Component Analysis Algorithms for Hyperspectral Image Classification. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7(6):2183–2199, June 2014. ISSN 1939-1404. doi: 10.1109/JSTARS.2014.2329792 [36].

4
Feature Reduction Based on ICA

This Chapter presents a novel feature dimensionality reduction strategy based on ICA applied on hyperspectral images. An optimized methodology aiming at extracting subsets of class-informative independent components for hyperspectral supervised classification is proposed and discussed. The selection of the most representative components is assured by the minimization of the reconstruction error, which is computed on the training samples used for the supervised classification.

4.1 Introduction

The goal of the spectral analysis is to extract informative features that can be used in the classification task. However, many studies have shown that not all the spectral space is needed for a good representation of the image. On the contrary, a part of the spectral space contains information that is noisy and redundant. In the context of hyperspectral image classification, the analysis of the entire spectral space is a difficult task due to several factors. One of the main issues is related to the high computational load that is required for the analysis of high spectral dimensionality data. Another issues is that the ratio between the number of available training samples (usually low) and the spectral dimension (usually high) is small, affecting the generalization capability of the classifier [54] (this is known as the Hughes phenomenon). Feature reduction techniques are usually adopted in order to extract a sub-space of informative features based on different criteria, while discarding all the rest. The experimental analysis conducted in Section 3.3.1 shown that when PCA is used prior to ICA for dimensionality reduction, it provides a sub-set of components that, in general, does not preserve class-separability, affecting the independent components. This was also demonstrated in other studies [22, 32]. In this Chapter, an optimized feature reduction approach, which exploits the properties of ICA aiming at extracting class-informative components for supervised classification purposes, is proposed. ICA analysis is optimized to address the supervised classification task based on the use of prior information provided by training samples, which are available in the supervised context. In Section 3.3.3, it was shown that the reduction of the number of samples used as input to an ICA algorithm can, in general, improve the ICA convergence speed, without affecting significantly the classification results. That was noticed in particular in a low-dimensional scenario (i.e., dimensionality reduction was performed prior to ICA), whereas, when the dimensionality reduction was not considered, the decrease of the number of training samples used as input to the ICA was affecting negatively the performance of the classifier. In that case, the issue was to extract class-discriminant features by exploiting all the training samples in a high dimensional space.

Aiming at finding an optimized approach that effectively exploits the information extracted by ICA, in the proposed approach, the ICA is separately applied to each class in a high-dimensional space (meaning that no dimensionality reduction is applied prior ICA), extracting sets of ICs that are strictly dependent on the training samples of each single class. The idea is to extract ICs that are suitable to represent each specific class. After the ICA decomposition, the reconstruction error is evaluated in order to identify the best ICs in terms of class representation. The reconstruction error is, thus, exploited to address the issue related to the non-prioritization of the extracted ICs, i.e., multiple applications of ICA provide different IC sets, which are diverse both in the order of appearance and in content. The final sub-set is then optimized by applying a feature selection technique based on genetic algorithm approach. Based on our previous study [36], FastICA resulted the technique that provided the best performance in extracting the whole source matrix, requiring less computational resources with respect to JADE and Infomax. Therefore, FastICA is chosen here as the applied ICA decomposition technique.

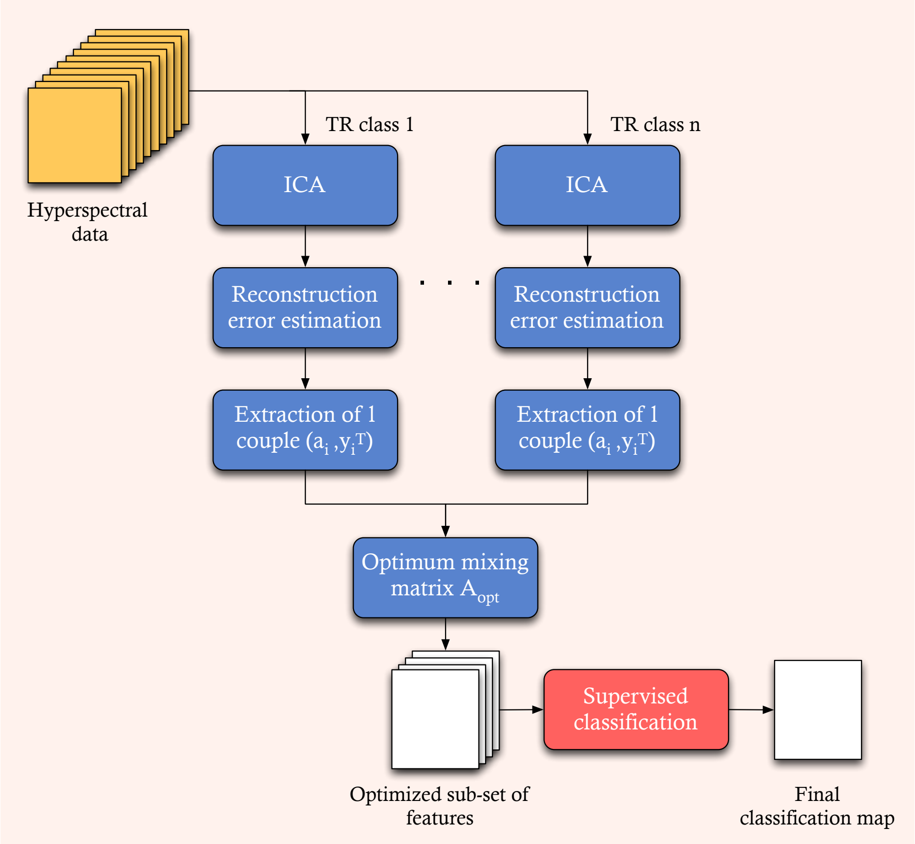

Hyperspectral data
TR class 1
TR class n
ICA
ICA
Reconstruction error estimation
Reconstruction error estimation
Extraction of 1 couple (a_i, y_i^T)
Extraction of 1 couple (a_i, y_i^T)
Optimum mixing matrix A_{opt}
Supervised classification
Optimized sub-set of features
Final classification map

Figure 4.1: General scheme of the proposed technique for feature reduction based on ICA.

4.2 Proposed Technique for Feature Reduction

In this Section, the proposed approach is presented. Figure 4.1 shows the general scheme. Let $\mathbf{X}$ be the observed data, represented by a $m \times p$ matrix, with m spectral channels and p pixels, whose elements $[\mathbf{x}_1, \ldots, \mathbf{x}_m]^T$ are the mixtures of the observed data, the linear mixing model adopted for hyperspectral images can be rewritten as:

$$\mathbf{X} = \mathbf{AS} = \sum_{i=1}^{m} \mathbf{a}_i \mathbf{s}_i^T, \tag{4.1}$$

where $\mathbf{A}$ is an $m \times m$ matrix and represents the unknown mixing matrix with elements $[\mathbf{a}_1, \ldots, \mathbf{a}_m]$ and $\mathbf{S}$ is an $m \times p$ matrix whose elements are the unknown sources $[\mathbf{s}_1, \ldots, \mathbf{s}_m]^T$. The proposed algorithm consists of the following steps:

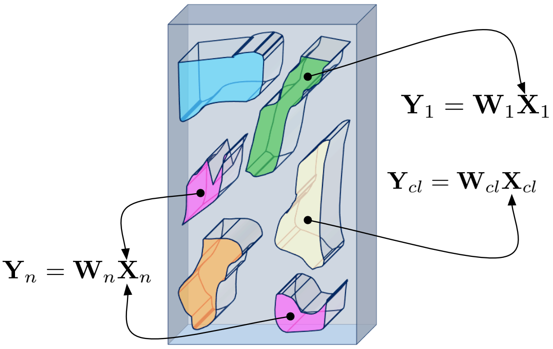

$\mathbf{Y}_1 = \mathbf{W}_1\mathbf{X}_1$
$\mathbf{Y}_{cl} = \mathbf{W}_{cl}\mathbf{X}_{cl}$
$\mathbf{Y}_n = \mathbf{W}_n\mathbf{X}_n$

Figure 4.2: Clustering based on the training samples. A full size vector of ICs is extracted from each cluster separately.

4.2.1 Extraction of Class-Specific ICs

n clusters representing the n classes of interest are extracted from the data set. Each cluster $\mathbf{X}_{cl}$, where $cl = 1, \ldots, n$, coincides with the training samples of each class. For each of them, the unmixing matrix $\mathbf{W}_{cl}$ and the independent components $\mathbf{Y}_{cl}$ are estimated by using FastICA, as shown in Figure 4.2.

4.2.2 Evaluation of the Reconstruction Error

The reconstruction error provides a measure of the class information associated with a single component and is used to rank the extracted ICs. The estimation of the reconstruction error is obtained by computing the Frobenius norm, denoted by $||.||_F^2$, between the original data set and the back projection of the obtained ICs. It is mathematically defined as follows:

$$E_{cl} = ||\mathbf{X}_{cl} - \mathbf{A}_{cl}Y_{cl}||_F^2 = \left|\left|\mathbf{X}_{cl} - \sum_{i=1}^{m} \mathbf{a}_i\mathbf{y}_i^T\right|\right|_F^2, \qquad (4.2)$$

with $\mathbf{A}_{cl} = \mathbf{W}_{cl}^{-1}$. Here, $\mathbf{a}_i$ is a column vector of the mixing matrix $\mathbf{A}_{cl}$, which represents the spectral signature related to the class, and $\mathbf{y}_i$ is a row vector of the estimated source matrix $\mathbf{Y}_{cl}$. Considering the relation in (4.2), the m pairs $(\mathbf{a}_i, \mathbf{y}_i)$ are ranked based on their relative contribution, where high contribution means low reconstruction error. The ranking is assessed by applying the following iterative procedure, which identifies the l-th couple that minimizes

Algorithm 1 Algorithm for the ranking of the couples $\mathbf{a}_i, \mathbf{y}_i^T$ based on the reconstruction error.

```
$\mathbf{X} \leftarrow \mathbf{X}_{cl}$
for $j \leftarrow 1$ to $l$ do
    for $i \leftarrow 1$ to $m$ do
        $E_i = \|\mathbf{X} - \mathbf{a}_i \mathbf{y}_i^T\|_F^2$
    end for
    $i_{opt} = \arg\min_i \mathbf{E}(i)$
    $E_{opt} = \min \mathbf{E}(i)$
    $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{a}_{i_{opt}} \mathbf{y}_{i_{opt}}^T$
    $\mathbf{a}_{i_{opt}} \leftarrow \infty$
    $\mathbf{y}_{i_{opt}} \leftarrow \infty$
    $idx_j = i_{opt}$
    $E\_vec_j = E_{opt}$
end for
return idx, E_vec
```

the reconstruction error:

$$idx = \arg\min_i err(i) = \{i \mid \min_i \left\|\mathbf{X}_l - \mathbf{a}_i \mathbf{y}_i^T\right\|_F^2\}, \tag{4.3}$$

$$\mathbf{X}_{l+1} \leftarrow \mathbf{X}_l - \mathbf{a}_{idx} \mathbf{y}_{idx}^T, \tag{4.4}$$

with $i = 1, \ldots, m$. Here, idx represents the index of the chosen l-th couple at the l-th iteration. $\mathbf{X}_l$ is initialized as equal to $\mathbf{X}_{cl}$, and updated at each iteration by subtracting the contribution provided by $\mathbf{a}_{idx} \mathbf{y}_{idx}^T$ identified at the previous iteration as shown in (4.4). The procedure requires the tuning of the parameter l, which represents the number of couples to retain after the ranking. The algorithm for the computation of the reconstruction error and the l indices for a single class is shown in Algorithm 1.

4.2.3 Identification of the Optimal Mixing Matrix

From the previous step, for each class cl, a matrix $\mathbf{A}'_{cl}$ composed of the best elements $[\mathbf{a}_1, \ldots, \mathbf{a}_l]$ is defined, where l is the total number of couples retained for a given class. The optimal mixing matrix is represented by $\mathbf{A}_{opt} = [\mathbf{A}'_1, \mathbf{A}'_2, \ldots, \mathbf{A}'_n]$. The obtained $\mathbf{A}_{opt}$ is an $m \times (nl)$ matrix. Based on the choice

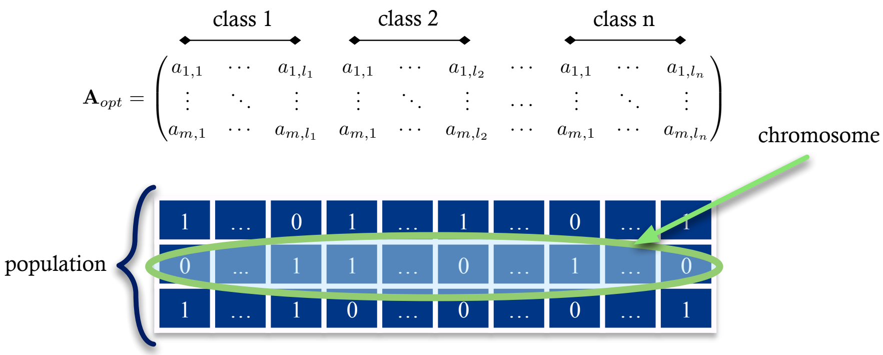


Figure 4.3: Selection based on genetic algorithm approach applied to the A_{opt}.

of l, the matrix $\mathbf{A}_{opt}$ can have quite a high dimensionality. A further selection based on GAs is performed on the elements of the matrix $\mathbf{A}_{opt}$, where a chromosome identifies which column is selected for the transformation (see Figure 4.3). The fitness function is evaluated on the transformation $\mathbf{Y}_{GA} = \mathbf{W}_{GA}\mathbf{X}$, where $\mathbf{Y}_{GA}$ are the final sub-set of ICs, $\mathbf{W}_{GA} = \mathbf{A}_{GA}^{-1}$ is the new unmixing matrix derived from the reduced version of $\mathbf{A}_{opt}$ and $\mathbf{X}$ is the original observed image. In our algorithms, the computation of the unmixing matrix, which can lead to an underdetermined system, is done by using the Moore-Penrose pseudoinverse.

It is worth mentioning that the analysis of each class is independent of the other. This allows the computation of the ICA and the estimation of the reconstruction error to be performed in a parallel distributed system. In this way, the computational time of the ICA for the entire data set is significantly decreased and can be approximated to be similar to the computational time of a single class ICA. The computation in parallel fashion can be also adopted for computing the fitness function for each population in order to optimize the selection based on GA.

4.3 Experimental Setup

4.3.1 FastICA Tuning

FastICA is not a parameter-free approach. In our experiments, the non-quadratic function $g(u)$, which represents the derivative of the non-quadratic function G, is set as $\tanh(au)$ with $a = 1$. This choice provides a good approximation of negentropy, as proven in [56]. Here, symmetric orthogonalization is chosen since in our analysis every feature extracted has the same importance and its computation results faster (see Section 3.2.2). Other parameters are related to the stopping criterion. The algorithm stops when the convergence is reached, meaning that the weight change has to be less than 10^{-4}, or the maximum number of iterations (which is set at 1000), is reached. One more parameter is the guess for the initial projection. In order to make the performance comparison consistent, the identity matrix of size $n \times n$ is chosen for initialization.

4.3.2 Genetic Algorithm Tuning

A search strategy based on GA is employed to reduce the size of A_{opt} by selecting the most representative column vectors a_i. In this study the classification accuracy obtained by the SVM classifier with the Radial Basis Function (RBF) kernel is considered as a fitness function to be maximized. However, other measures could be integrated as fitness function. Since the kernel parameter estimation is computationally expensive, the estimation is performed once for each population using 5-fold cross-validation. The selection strategy is based on Stochastic Universal Sampling (SUS) [4], where sigma scaling [76] is employed in order to avoid premature convergence. The parameters of the GA, such as crossover rate, mutation rate and population size, are determined empirically through a set of preliminary experiments. In this work, a uniform crossover is used, with a crossover rate of 0.80 and a mutation rate of 0.01. The length of a chromosome is computed as nl, where n is the number of classes of a specific data set and l is the chosen number of (a_i, y_i^T) couples that minimize the reconstruction error. The search criterion stops when 50 generations are computed.

4.3.3 Classification

For classification purposes, an SVM classifier [20] is exploited considering a Radial Basis Function (RBF) kernel. The algorithm employs the one-against-one multi-class strategy. For the estimation of the regularization parameter, C, and the kernel parameter, γ, cross-validation based on the grid-search approach is performed. In particular, an exponentially growing sequences of C and γ are considered, with $C = \{10^{-2}, 10^{-1}, \ldots, 10^{4}\}$ and $\gamma = \{2^{-3}, 2^{-2}, \ldots, 2^{4}\}$. Each classification result in Section 4.4 is obtained by using a 10-fold cross-validation, i.e., that the training set is split into 10 sets, where 9 of them are used for training the model and the one left is used for validation. In this way the choice of the parameters results unbiased.

4.3.4 Data Sets' Description

The experimental analysis is carried out on four real hyperspectral data sets characterized by different spatial and spectral resolutions. For each data set, the training samples and the test samples are generated in such way that the two sets results mutually exclusive (i.e., no shared samples between the two sets).

Pavia University: The data set is described in Appendix A.4. Information about the training set used in the experiments are also provided.

Pavia Center: The data set is described in Appendix A.5. Information about the training set used in the experiments are also provided.

Salinas: The data set is described in Appendix A.1. In case of Salinas data set, the training set employed in the experiments is made up of 15% randomly selected samples from each class.

Hekla: The data set is described in Appendix A.2. In case of Hekla, the training set is generated by a random selection of 50 samples from each class.

4.4 Experimental Results and Discussion

In this section, the feature dimensionality reduction approach based on ICA is tested and the obtained results on the four data sets are shown. Aiming at providing a qualitative analysis of the presented approach, the effectiveness in extracting class-informative features is assessed in terms of classification accuracies and kappa coefficients. For a better understanding of the obtained results, the overall classification accuracies are given in percentage (%), while the comparison between accuracies is given in percentage points (*pp*), which are simply the arithmetic difference of two percentages. For each data set, the behaviour of the proposed approach is tested for a different choice of the parameter l, which indicates the number of the retained best couples (a_i, y_i^T) that minimize the reconstruction error. The parameter l is sets as $l = 1, 2, 3, 4$. The

Table 4.1: Classification of the four data sets by employing the proposed ICA-based feature reduction approach. "No. feat." denotes the number of features selected based on the reconstruction error, "No. feat. GA" denotes the number of features after the GA selection, "OA (%)" indicates the percentage overall accuracies and "k" indicates the kappa coefficients. Classification results obtained by exploiting the original spectral bands and by using the PCA-ICA strategy are given for comparison.

				Proposed approach			
		Spectr.	PCA-ICA	$l = 1$	$l = 2$	$l = 3$	$l = 4$
Pavia University	No. feat.	103	12	9	18	27	36
	No. feat. GA	-	-	8	10	10	12
	OA (%)	77.91	82.55	79.11	85.25	86.25	**87.69**
	k	0.72	0.78	0.73	0.80	0.82	**0.84**
Pavia Center	No. feat.	102	8	9	18	27	36
	No. feat. GA	-	-	6	12	17	18
	OA (%)	97.35	97.86	97.98	98.17	**98.57**	98.50
	k	0.96	0.97	0.97	0.97	**0.98**	0.98
Salinas	No. feat.	204	20	16	32	48	64
	No. feat. GA	-	-	13	17	26	29
	OA (%)	94.57	94.62	93.71	**95.30**	95.17	94.93
	k	0.91	0.93	0.93	**0.95**	0.95	0.94
Hekla	No. feat.	157	5	12	24	36	48
	No. feat. GA	-	-	7	12	20	26
	OA (%)	93.89	90.40	91.97	94.47	**96.28**	96.00
	k	0.91	0.88	0.91	0.93	**0.95**	0.95

proposed approach is then compared to the spectral case, where all the spectral bands are used as input to the classifier, and to the common strategy based on ICA for feature reduction (i.e., PCA is used as dimensionality reduction prior to ICA). In the last case, the shown results represent the best case obtained by varying the number of components retained from 2 to the total number of available spectral channels. The numerical results are reported in Table 4.1, while Figures 4.4 - 4.7 show the best ICA subsets extracted for each data sets. In Figures 4.8 and 4.9, the classification maps obtained by using the proposed approach are shown and compared to the map obtained by using the original spectral channels and to the best case of PCA-ICA approach. By comparing the obtained results, one can see that the proposed approach is able to provide representative subsets. While this is less evident for Pavia Center and Salinas, where the classes are already well represented and separated in the spectral domain, for the Pavia University and Hekla data sets, the effectiveness of the proposed approach becomes clearer. In those cases, significantly higher classification accuracies are achieved for the proposed approach as compared to the spectral and the common strategy cases. In particular, for Pavia University, the best classification accuracy is achieved with $l = 4$, obtaining after the selection a subset of 12 components, with a sharp improvement of 9.78 *pp* compared to the spectral case, and of 5.15 *pp* compared to the best case of PCA-ICA (which is obtained by extracting 12 features). In the case of Hekla, the best classification accuracy is achieved with $l = 3$, obtaining after the selection a subset of 20 components. In the Hekla case, the classification accuracy is improved of 5.88 *pp* compared to the best case of PCA-ICA, and of 2.39 *pp* compared to the spectral case. In the case of Pavia Center and Salinas, the best classification accuracies are achieved with $l = 3$ and $l = 2$, respectively, retaining after the selection a subset of 17 components with a slight improvement respect to both the spectral case and PCA-ICA.

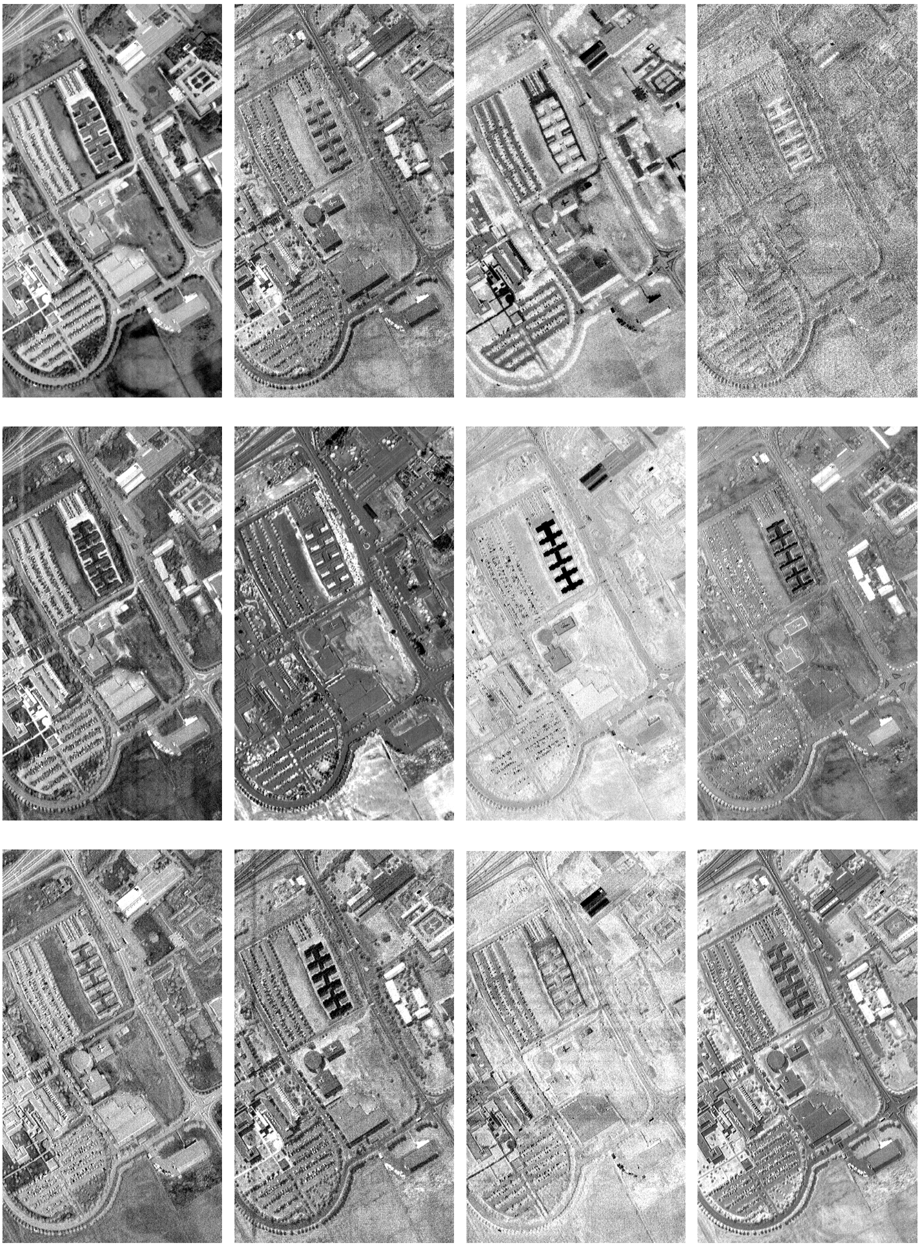


Figure 4.4: Subset of ICs extracted for Pavia University.

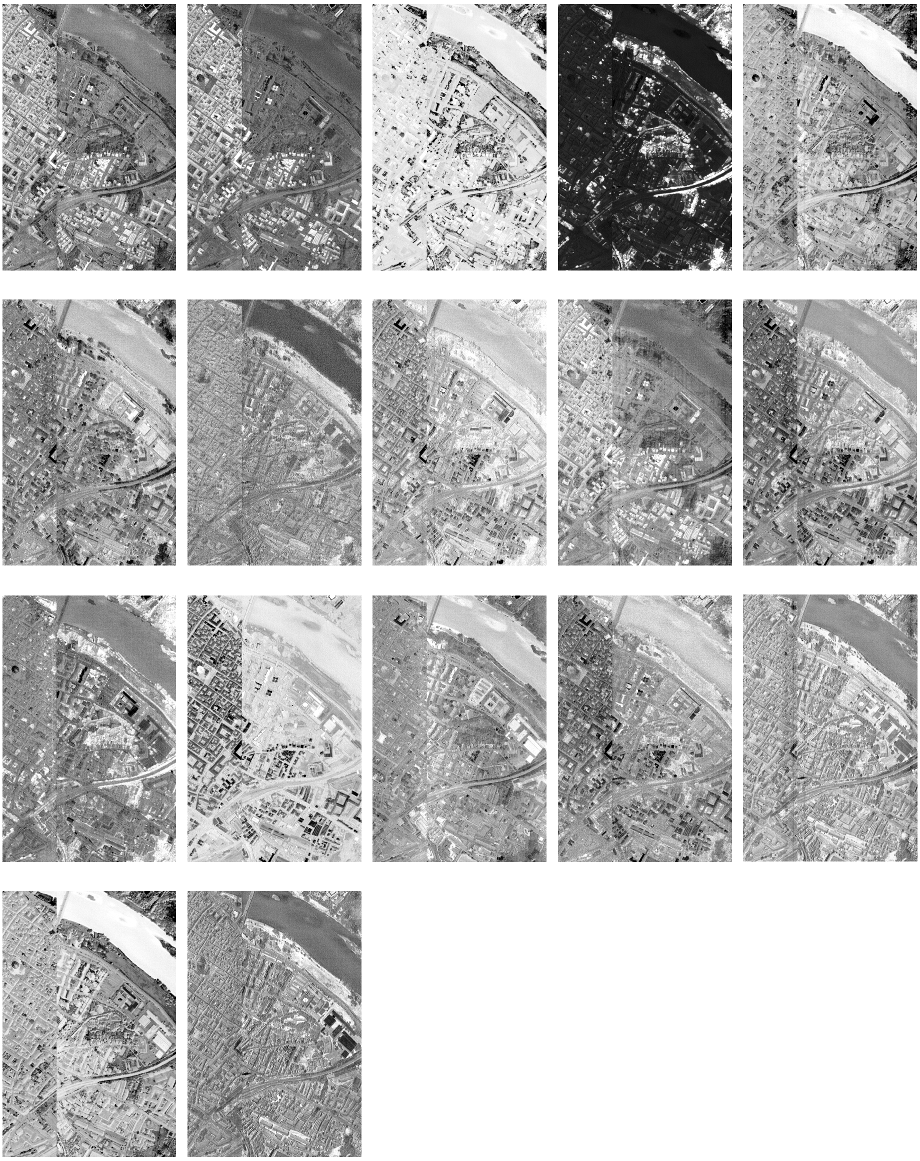


Figure 4.5: Subset of ICs extracted for Pavia Center.

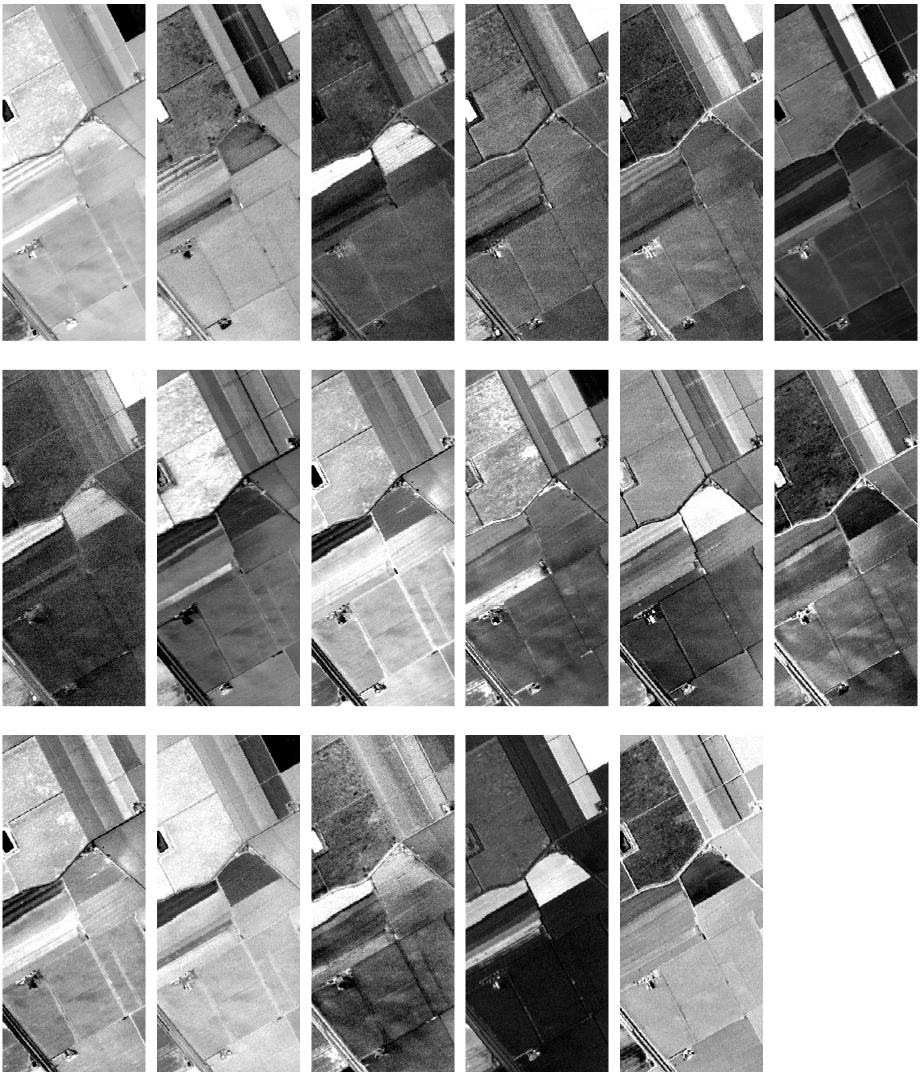


Figure 4.6: Subset of ICs extracted for Salinas.

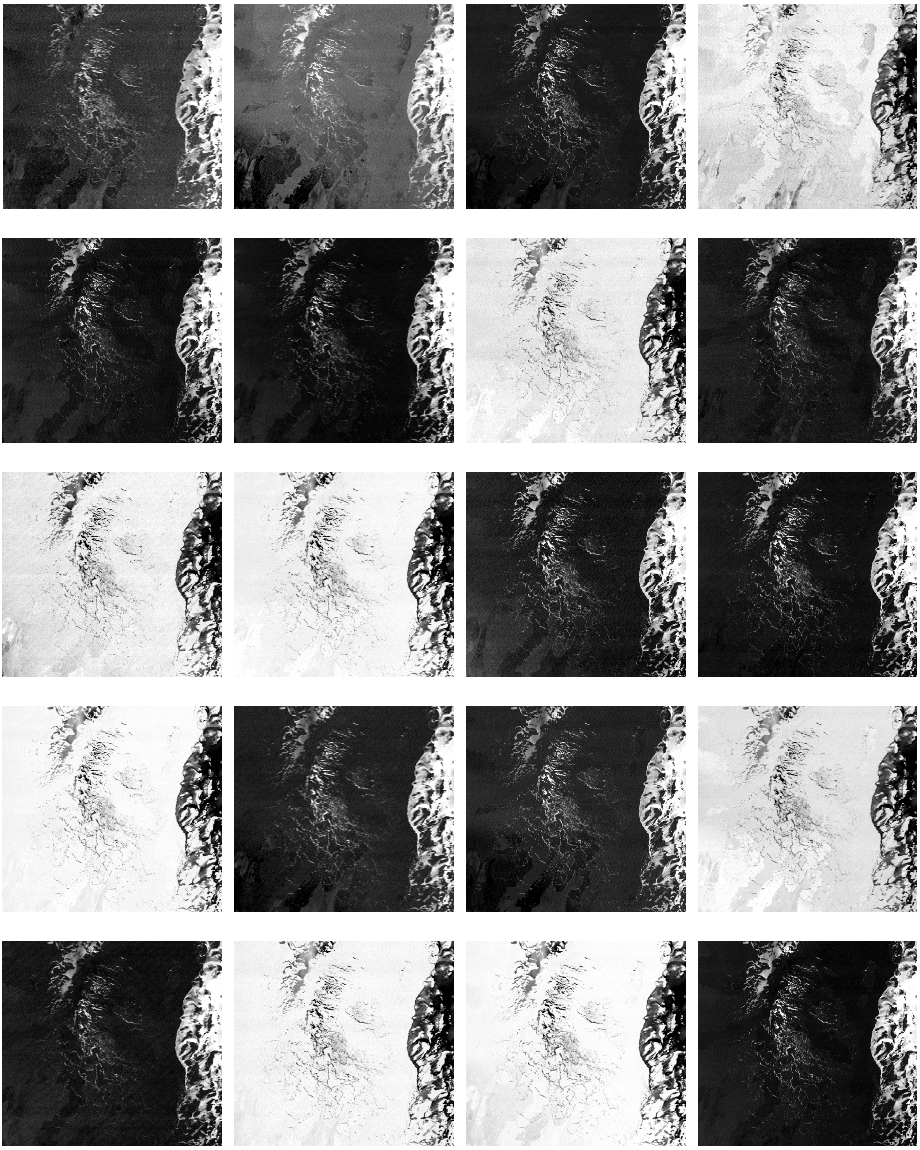
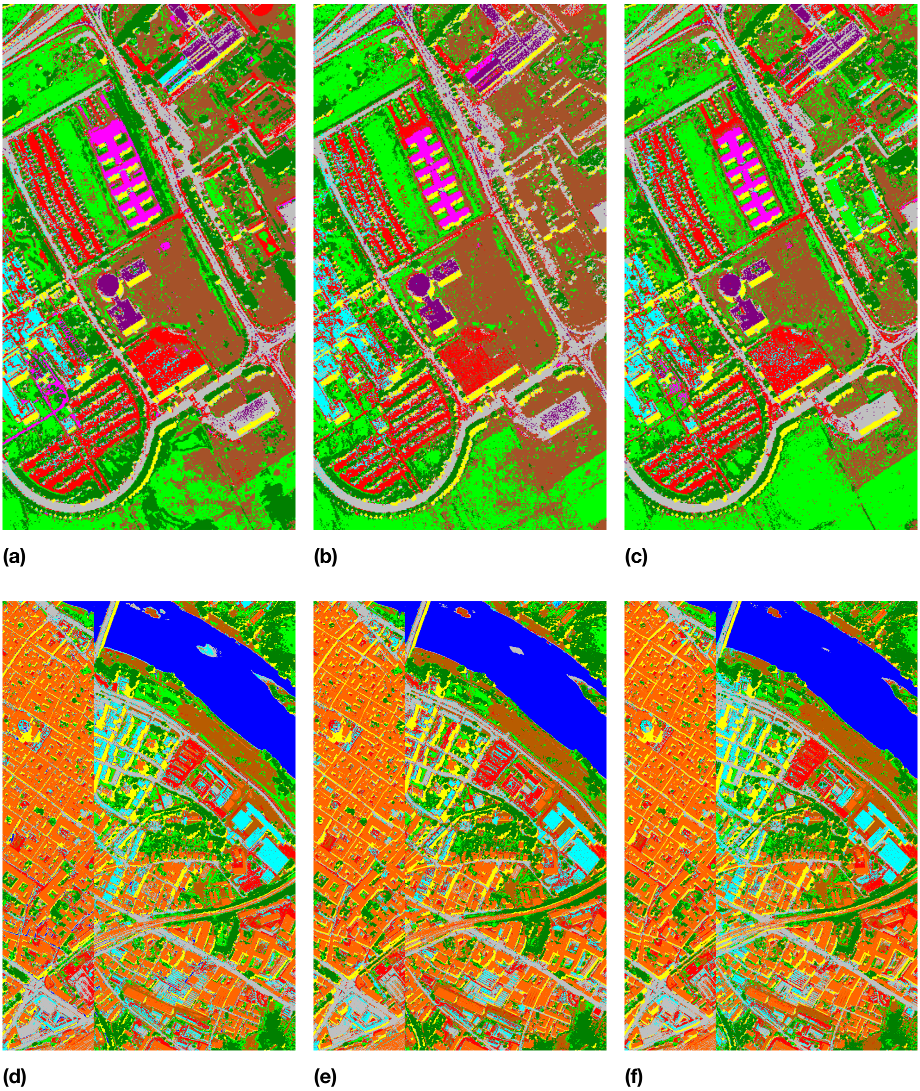


Figure 4.7: Subset of ICs extracted for Hekla.



(a)

(b)

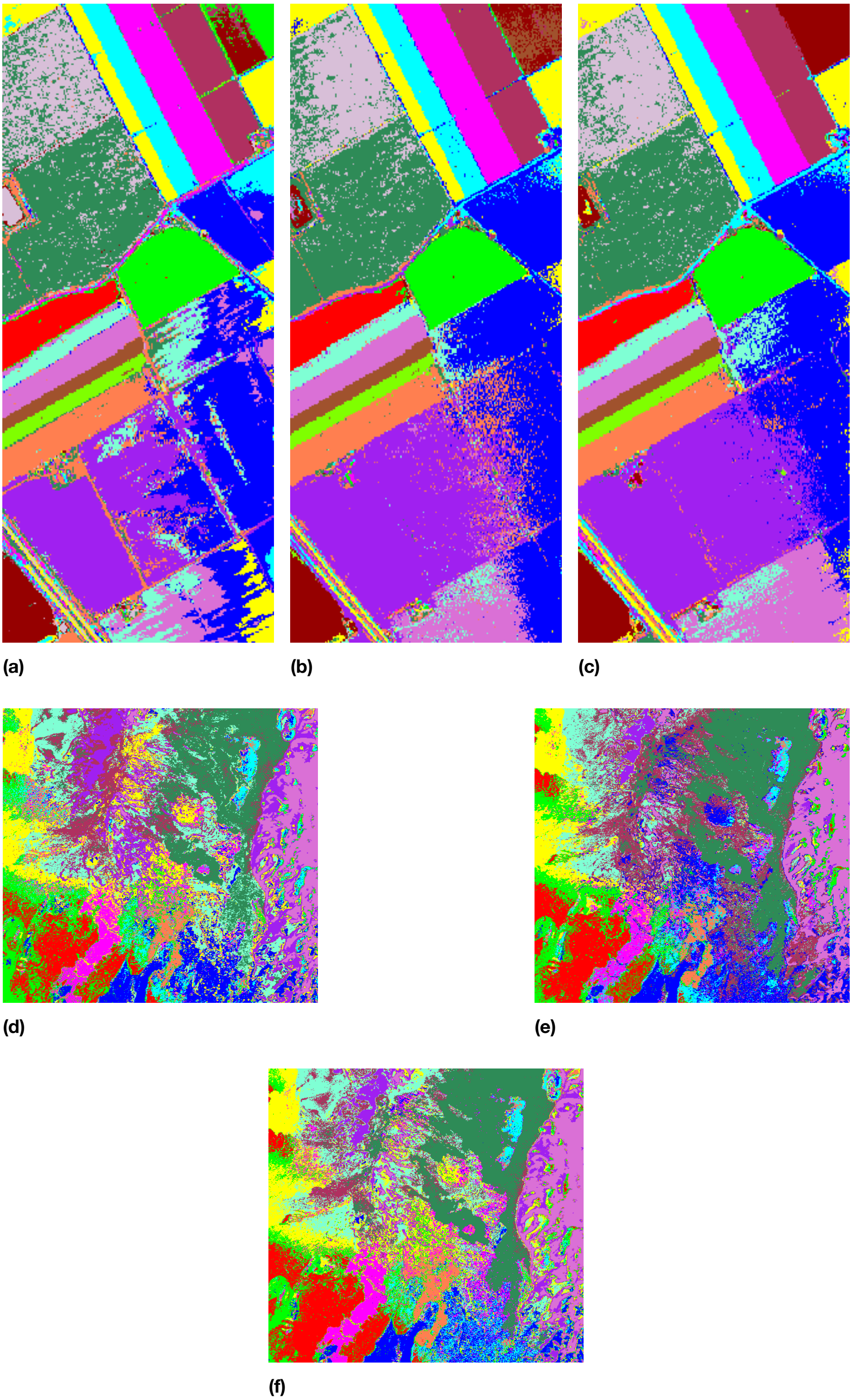
(c)

(d)

(e)

(f)

Figure 4.8: Classification maps of Pavia University (top row) and Pavia Center (bottom row): (a)(d) Spectral case; (b)(e) PCA-ICA; (c)(f) Proposed technique.



(a) (b) (c)

(d) (e)

(f)

Figure 4.9: Classification maps of Salinas (top row) and Hekla (middle and bottom row): (a)(d) Spectral case; (b)(e) PCA-ICA; (c)(f) Proposed technique.

4.5 Conclusions

This Chapter presented a feature dimensionality reduction technique based on ICA suitable for supervised hyperspectral image classification. The goal of this study was to extract class-informative features where the use of ICA was optimized for its application in a high-dimensional scenario (i.e., no dimensionality reduction was performed prior ICA). The reconstruction error computed on the training samples of each single class (defined for the classification stage) was used as estimation of the class-information content. In particular, the retrieving of class-information was assured by solving an optimization problem based on the minimization of the reconstruction error of the ICs extracted from each specific class. The searching strategy was further optimized by employing a genetic algorithm-based approach, which led to an additional reduction. The obtained results shows that an appropriate use of ICA can bring prominent improvements in selecting the most representative components from hyperspectral images, providing improved results in classification.

Part of this Chapter is published in:

N. Falco, L. Bruzzone, and J. A. Benediktsson. An ICA based approach to hyperspectral image feature reduction. In *Geoscience and Remote Sensing Symposium, 2014. IGARSS 2014. IEEE International*, pages 3470–3473. IEEE, July 2014. ISBN 978-1-4799-5775-0. doi: 10.1109/IGARSS.2014.6947229 [37].

N. Falco, J. A. Benediktsson, and L. Bruzzone. Spectral and Spatial Classification of Hyperspectral Images Based on ICA and Reduced Morphological Attribute Profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 2015. (accepted) [38].

Part III

Integration of Spatial Information

5

Reduced Attribute Profiles for the Analysis of Spatial Information

This Chapter presents an optimized version of the morphological attribute profiles, namely reduced attribute profiles, focusing on addressing the issue related to the high dimensionality, which leads to an highly intrinsic information redundancy, for a better representation of the spatial information.

5.1 Introduction

The new generation of hyperspectral sensors are able to provide images with an improved spatial resolution. The spatial context information, which is represented by the neighbourhood pixel system, becomes an important information source for distinguishing different objects on the ground. However, the exploitation of this information source increases the complexity of the classification process. Recent improvements in mathematical morphology have provided new techniques, such as Attribute Profiles (APs) [26], able to extract contextual information of the investigated regions at different scale-level. APs are an extension of Morphological Profiles (MPs), developed by Pesaresi and Benediktsson [83], made up by concatenating the results obtained by applying more severe attribute filters, which operate on connected regions that compose the image. The concatenation of APs gives Extended Attribute Profiles (EAPs) [27]. The high flexibility of APs and EAPs in extracting different information from the regions made them a powerful tool for modelling the spatial information in remote sensing images. However, one of the main drawbacks of these techniques is the need to select the optimal range of values related to the family of criteria adopted by each filter step, making it difficult to identify if prior knowledge of the investigated area is not available. This fact induces another important issue that is related to the high dimensionality of the profiles. In order to be able to characterize all the regions in the scene, the values used as threshold should cover a wide range. Depending on the scene, this can result in very long profiles in which the geometrical information related to several regions become redundant. The high dimensionality of an AP leads to a large number of features, which increases even more when Extended Attribute Profiles (EAPs) and Extended Multi-Attribute Profiles (EMAPs) [27] are considered, resulting in the Hughes phenomenon.

In this work, a novel strategy for extracting spatial information from hyperspectral images based on Differential Attribute Profiles (DAPs) is proposed, which was shown to be suitable in extracting and characterizing structures both in VHR image analysis [26] and in change detection [35]. A DAP is computed as the derivative of the AP, showing at each level the residual between two adjacent levels of the AP. From the analysis of the multilevel behaviour of the DAP, it is possible to extract geometrical features corresponding to the structures within the scene at different scales. Such information is used to compress the APs in few features, obtaining the *reduced attribute profiles* (rAPs). The aim of this work is to exploit the potential of the APs in the classification task, while decreasing both the high dimensionality and the redundancy that affects the original APs.

5.2 Proposed Reduced Attribute Profiles

Aiming at minimizing the intraclass variability due to the increase in geometrical detail [13] and the uncertainty of the classification, the information related to the spatial context needs to be included in the analysis. In this study, mathematical morphology-based techniques are used for the extraction of the spatial information. According to Section 2.4.3, it is easy to understand that EAPs and EMAPs provide a rich multi-level description of the scene. However, the dimensionality of the feature space increases when EAP is considered as input to a classification stage. The situation is even more challenging when an EMAP is considered. Another important issue is related to the relevant presence of redundant information within the AP, which can be seen from the high sparsity that characterizes the DAP. This is due to the way in which the profile is built. Regions that are not considered in the filtering are preserved at each scale and the same information is propagated along the profile.

Inspired by research made on Differential Morphological Profile (DMPs) in [1, 83], where the DMPs were used for the definition of segmentation techniques, an optimized version of the AP representation is here presented. By exploits the most informative geometrical features extracted by the DAPs, the issue related to the redundancy that affects the APs can be addressed. The proposed solution aims at fusing the information contained in the AP by identifying the best level of representation for each region present in the scene. This is possible by analysing the DAP and by using the corresponding AP's values in order to compress the AP into two single features, one for each thickening and thinning profile. The following steps describe the proposed technique related to a single attribute case, while Figure 5.1 shows the general scheme of the method.

5.2.1 Attribute Profiles and Region Extraction

The first step is to obtain the AP for each single feature. The critical phase in building the AP is the choice of the λ values that are used as references for the filtering phase. An optimal choice of the range values is the one that provides a proper representation of the regions present in the scene, which is highly scene and attribute dependent, and it is usually based on prior information of the scene. This issue is considered in Chapter 7. The proposed method relies on the analysis of the regions that are filtered at each level of the AP. Aiming at extracting those regions, the DAP is exploited. The DAP is obtained by differentiating the AP (see Section 2.4.2), representing the residual of the AP, where each level of the DAP shows the regions that have been filtered between two adjacent levels of the AP, in terms of grey-level values. This characteristic allows the identification of connected regions related to each grey-value of the DAP, with the advantage of preserving the geometrical shape without any loss in terms of detail. From this step, thinning (closing) and thickening (opening) profiles are analysed separately due to the different information that they provide.

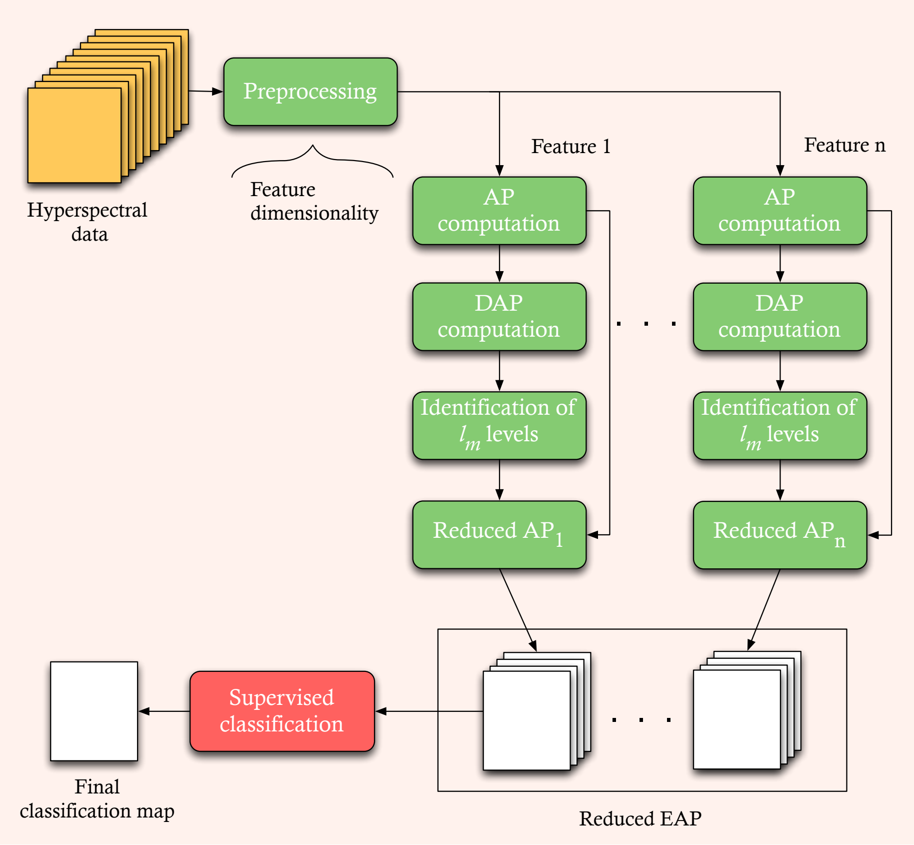


Figure 5.1: General scheme of the procedure to obtain the reduce AP.

5.2.2 Identification of the Representative Levels

This step aims at finding the level where a given connected region is well represented in terms of homogeneity. Let us consider the case of an increasing attribute, where the size of the filtered regions increases when the criterion value λ increases. As general behaviour, a given region grows starting from the first level, where few pixels are considered, and increases in size at each filtering step by merging with the surrounding regions, reaching, after a certain number of morphological transformations, the level in which the structural meaning of the region is partially or totally lost. In order to identify that level, a homogeneity measure, which is computed on the connected region taking into account the original image pixel values, is defined and analysed along the profile. For a given connected region C, the homogeneity measure H is computed as follows:

$$H(C) = P(C) \times S(C), \tag{5.1}$$

where $P(\cdot)$ is the size in pixels of the connected region and $S(\cdot)$ is the standard deviation computed on the pixels within the connected region considering the original values. The joint use of the two parameters ensures that a region selected as meaningful will be as spectrally homogeneous and large as possible. Consequently, the goal is to identify the level where the homogeneity of a connected region changes drastically, and consider as the meaningful level L_m the one the precedes this effect. This is obtained by searching for those two adjacent levels, whose difference in $H(C)$ intensity is maximum:

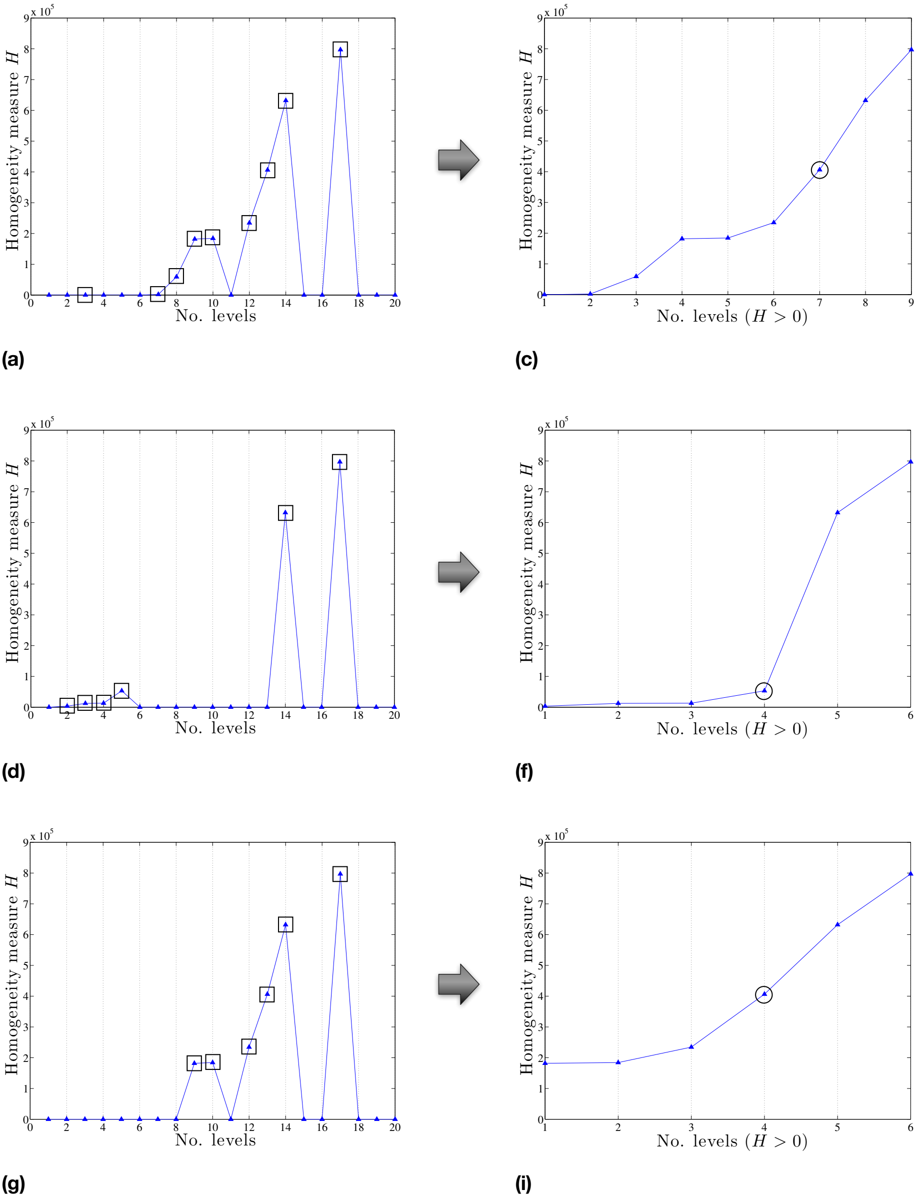
$$L_m(C) = \arg\max_L \{H(C_{L+1}) - H(C_L)\} \tag{5.2}$$

with C_L and C_{L+1} defined as follows:

$$C_L : \{p \mid p \in C \text{ at level } L\} \tag{5.3}$$

$$C_{L+1} : \{p \mid p \in C \text{ at level } L+1\} \tag{5.4}$$

where $C_L \subseteq C_{L+1}$. This implies that C_{L+1} could be the result of the merging of more connected regions, which are compared to the C_{L+1} separately. Figure 5.2 shows three examples of possible behaviours of a homogeneity measure computed for an increasing criterion (e.g., *the diagonal of the bounding box that encloses a given region*). Considering the non-hierarchical nature of the DAP, the levels that have zero values for a given region are not considered in the analysis since, at those levels of the AP, the region is not affected by the filtering (see Figure 5.2 where squares indicate the considered levels and circles indicate the meaningful levels). The computation of L_m is based on the assumption that $H(C)$ is monotone increasing (after discarding the zero-value levels). When non-increasing attributes are considered, the initial assumption does not hold. To overcome this issue, the H profile computed for each extracted region is sorted in terms of size of regions in such a way the new H profile has a similar behaviour to the one of an increasing criterion. After this, the procedure illustrated above can be applied to the modified H. This solution allows the analysis of the homogeneity to be performed without losing the information provided by the attribute, which is intrinsic in the shape of the extracted regions.



(a) (c) (d) (f) (g) (i)

Figure 5.2: Examples of homogeneity measure $H(C)$ for an increasing criterion (e.g., *diagonal of the bounding box*) computed on a given region C (left column) considering 20 filtering values. The levels with zero intensity are not considered in the analysis since the region under investigation is not affected by the filtering. The circles indicate the levels L_m that are chosen (right column), and precede the maximum change in terms of H intensity.

5.2.3 Fusion of the Attribute Profiles

In this step, the geometrical information contained in each profile (i.e., thinning and thickening of the AP), is fused into two images, whose connected regions Cs are associated to the values of the AP at the scale level denoted by L_m. The reduced AP, is thus defined as:

$$r\Pi(I) = \left\{ r\Pi_{\varphi^T}(I), I, r\Pi_{\gamma^T}(I) \right\}. \tag{5.5}$$

The obtained feature space has a size of three feature types that combine the most informative geometrical information, according to the homogeneity measure.

5.2.4 Extension to Multi-Channel and Multi-Attribute

The rAP is directly extracted from the original AP, resulting affected by the same limitations when multi-channel data are considered. The extension to hyperspectral data analysis is obtained by applying the morphological analysis to a subset of features identified by applying feature dimensionality reduction techniques. The same concepts of multi-channel and multi-attribute introduced for APs [27] can be applied to the rAP, obtaining the reduced EAP (rEAP) and the reduced EMAP (rEMAP). In this case, the dimension of the future space of a rEAP is calculated as $(r3)$, where r corresponds to the number of features processed in the analysis. For the rEMAP, the feature space size corresponds to $(2rq + r)$, with q the number of the considered attributes. It is worth noting that the feature space size does not depend on the number of filtering thresholds L, as it is for the original EMAP, which size corresponds to $(2Lrq + r)$. This gives the possibility, if necessary, to increase the range of family criteria, T, for a better identification of the regions that compose the scenes (which leads to a more representative decomposition of the image) without incurring a consequent increasing the final dimension of the final feature space.

Figure 5.3 shows an example of the pipeline to obtain a single reduced AP, where in a) the attribute profile is built considering a set of L thresholds λ; b) the differential attribute profile is derived; c) for each connected region C, the $H(C)$ is evaluated and the maximum change identified, giving the level L_m represented by the map of levels; d) the reduced attribute profile is obtained by mapping the original profile into a single feature (i.e., one for closing and one for opening), according to the map of levels. When the concatenation

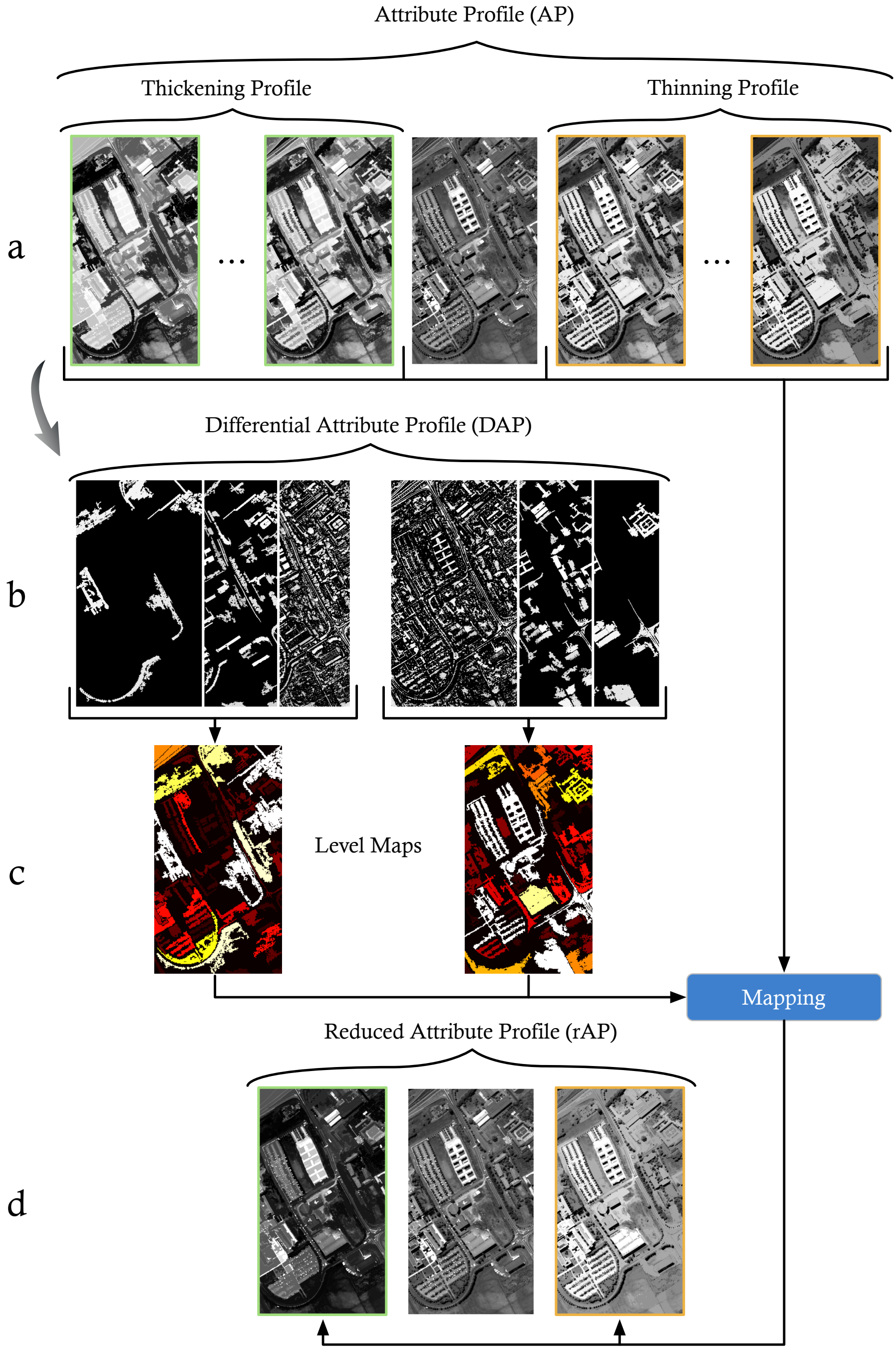


Figure 5.3: Pipeline of the processing required to obtain a reduced AP. *a*) the attribute profile is built considering a set of L thresholds λ; *b*) the differential attribute profile is derived; *c*) for each connected region C, the $H(C)$ is evaluated and the maximum change identified, giving the level L_m represented by the map of levels; *d*) the reduced attribute profile is obtained by mapping the original profile into a single feature (i.e., one for closing and one for opening), according to the map of levels.

of different rEAPs lead to a high-dimensional vector, a fusion process [29] is preferable. The evaluation of the multi-attribute information is performed by fusing the outcome of the classification obtained by each single rEAP. More specifically, the fusion strategy considered in this work assigns a pixel to a class according to the majority voting strategy. However, in the case of a tie in votes for two or more class labels, majority voting cannot be exploited. In this case, for each class label for which a tie is observed, the average class-accuracy obtained by the classifiers in agreement on the same class label is computed and considered for comparison. The final decision is made according to the classifiers that obtained the highest averaged classification accuracy.

5.3 Experimental Setup

Aiming at comparing the classification performance of the proposed optimization to the original AP, the work presented in [29] is considered as state-of-the-art. Therefore, the experimental setup used in this work correspond to the one in [29]. The experimental analysis is performed considering the data set of Pavia University (described in Appendix A.4). According to [29], dimensionality reduction is performed by PCA, retaining the first four components that contain more than 99% of the total variance.

In this dissertation, four attribute are considered as measures for modelling the contextual information:

area This is an increasing attribute that models the scale according to the cardinality of the considered regions. For this attribute, the following set of thresholds is considered: $\lambda_a = [100, 500, 1000, 5000]$.

diagonal of the bounding box This is an increasing attribute and it models the extension of the region. For this attribute, the following set of thresholds is considered: $\lambda_d = [10, 25, 50, 100]$.

moment of inertia[53] This is an non-increasing attribute. It is represented by the first moment of Hu and represents a measure of the elongation of a region, which is independent on the scale and orientation of the regions. For this attribute, the following set of thresholds is considered: $\lambda_i = [0.2, 0.3, 0.4, 0.5]$.

standard deviation This is an increasing attribute. It models the homogeneity of the regions, taking into account the gray-level values of their pixels. It

Table 5.1: Comparison between the classification performance of the original EAPs and reduced EAPs for the Pavia University data set. For each attribute, the Table reports the percentage overall accuracies ”OA(%)”, and the kappa coefficients ”*k*. Classification results obtained by employing the original spectral bands and by using the four PCs exploited to built the EAP and rEAP are given for comparison.

			Original EAPs				
	Spectr.	PCA	EAP$_a$	EAP$_d$	EAP$_s$	EAP$_i$	EMAP
No. feat.	103	4	36	36	36	36	132
OA (%)	77.91	72.88	**90.00**	85.42	**86.56**	69.80	77.81
k	0.72	0.65	0.87	0.81	0.82	0.63	71.08
			Reduced EAPs - (proposed approach)				
	Spectr.	PCA	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$	rEMAP
No. feat.	103	4	**12**	**12**	**12**	**12**	**36**
OA (%)	77.91	72.88	88.44	**87.20**	85.01	**80.59**	**90.95**
k	0.72	0.65	0.86	0.84	0.81	0.78	0.88

does not rely on the scale or shape of the regions as the other previous attributes. For this attribute, the following set of thresholds is considered: $\lambda_s = [20, 30, 40, 50]$.

5.3.1 Classification

For classification purposes an SVM classifier [20] was exploited considering a Radial Basis Function (RBF) kernel. The algorithm employs the one-against-one multi-class strategy. For the estimation of the regularization parameter, C, and the kernel parameter, γ, a 10-fold cross-validation based on the grid-search approach is performed considering an exponentially growing sequences of C and γ, where $C = \{10^{-2}, 10^{-1}, ..., 10^{4}\}$ and $\gamma = \{2^{-3}, 2^{-2}, ..., 2^{4}\}$.

5.4 Experimental Results and Discussion

Table 5.1 shows the results obtained by using the proposed approach and the original APs. The best results based on the comparison between the two techniques (i.e., AP versus rAP) are reported in bold. In particular, the classification performance obtained by the rEAPs is consistent with the state-of-the-art, obtaining similar classification accuracies in case of the *area, diagonal of the bounding box* and *standard deviation* attributes. However, in the case of the *inertia*, the rEAP$_s$ provides an improvement of 10.79 *pp* with respect to the original EPAs. Following the same strategy as in [29], the EMAP and rEMAP are obtained by concatenating all the EAPs and rEAPs, respectively, to obtain a unique vector of features. Also in this case, the reduced version of the EMAP outperforms the original EMAP with an improvement of 13.14 *pp*. One can notices the increase of the Hughes' effect when the original EMAP is used, whereas, in the case of rEMAP the multi-attribute information is better exploited as demonstrated by the classification accuracies. It is worth noting that, rEAPs and rEMAPs required only 12 and 36 features (three time less than the original EAPs and EMAPs), respectively, to provide results comparable to state-of-the-art accuracies.

5.5 Conclusions

In this Chapter, a novel strategy for extracting spatial information from hyperspectral images based on the analysis of the morphological DAPs has been proposed. The approach presented in this work aimed at reducing both the dimensionality and the redundancy by extracting the most informative spatial information from an AP. This has been done by analysing the multi-scale behaviour of the DAP, which permits us to extract geometrical features corresponding to the structures within the scene at different scales. The proposed approach consisted in two stages. The first stage was to characterize all the regions with a homogeneity measure and use this measure for identifying the level L_m of scale that best represents a given region. The second step was to fuse the geometrical information of the extracted regions into a single map considering their level L_m previously identified. The proposed method reduces the dimensionality of the AP to a space composed by three feature types, two related to the reduced thickening and thinning profile and one to the original single tone image. The experimental analysis, which has been carried out on an hyperspectral image on the area of the University of Pavia, in Italy, shows the effectiveness and the potentialities of the proposed strategy. By applying our the method to the four extracted principal components, the obtained reduced EAPs were characterized by a feature space of just 12 features, which gave remarkably better classification accuracies when compared to both the PCA and the whole data set, showing the importance of including spatial features in the analysis. In the comparison with the original EAPs, which were composed by 36 features, the presented approach provided classification accuracies consistent to the state of the art while achieved up to 10% improvement for the inertia attribute case, employing three times less of features. This approach permits the compression of the most informative geometrical information, according to the measure of homogeneity, taking into consideration the multi-scale transformation performed by the AP. The final reduced AP is characterized by a feature space that accounts for three features types, the reduced thickening and thinning profiles and the original image. Thus, the dimension of the feature space of the reduced EAP is calculated as $(n * 3)$, where n is the number of the components used in the analysis. The low dimensionality of the produced feature set gives the possibility to the user to enhance the feature space with additional features that provide complementary class-discriminant information, which could be useful for the final classification.

Part of this Chapter is published in:

N. Falco, J. A. Benediktsson, and L. Bruzzone. Extraction of spatial features in hyperspectral images based on the analysis of differential attribute profiles. In L. Bruzzone, editor, *Remote Sensing*, volume 8892, page 88920O. International Society for Optics and Photonics, Oct. 2013. doi: 10.1117/12.2029199 [33].

N. Falco, J. A. Benediktsson, and L. Bruzzone. Spectral and Spatial Classification of Hyperspectral Images Based on ICA and Reduced Morphological Attribute Profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 2015. (accepted) [38].

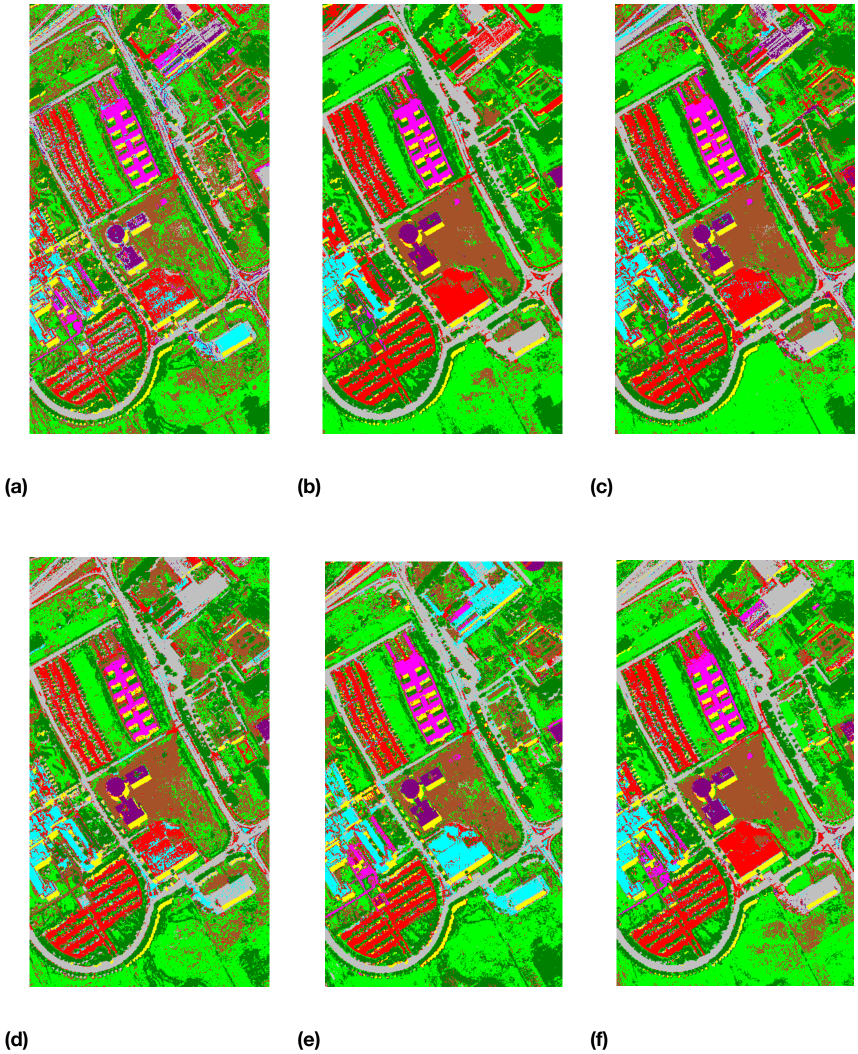


Figure 5.4: Classification maps obtained for Pavia University by exploiting: (a) PCA (4 comp.); (b) rEAP$_a$; (c) rEAP$_d$; (d) rEAP$_s$; (e) rEAP$_i$; (f) rEMAP.

6
Spectral and Spatial Information Integration

This Chapter presents a new technique that combines spectral and spatial information for supervised hyperspectral image classification. The feature reduction based on independent component analysis introduced in Chapter 4 is the main core of the spectral analysis, where the exploitation of prior information coupled to the evaluation of the reconstruction error assures the identification of the best class-informative subset of independent components. Reduced Attribute Profiles (rAPs), introduced in Chapter 5 and designed to address the issues of information redundancy that affects the common morphological APs, are then employed for the modelling and fusion of the contextual information.

6.1 Introduction

In this Chapter, the previous works presented in Chapter 4 and in 5 are extended, proposing a novel method to supervised classification based on both spectral and spatial information analysis. Considering the most recent studies, where APs are exploited, the spectral analysis is usually relegated to the identification of few PCA components, which are then exploited for building the

APs, EAPs and EMAPs, while supervised feature extraction techniques (e.g., DBFE, NWFE) are eventually employed in order to reduce the dimensionality of such huge vectors. Indeed, when a large range of filtering thresholds is considered, the dimension of the feature space of the obtained profile increases resulting in a very large number of features and, thus, in the Hughes phenomenon. In the literature, the issue was investigated by considering many approaches. In [6], the high dimensionality of the morphological profiles was reduced by exploiting feature extraction and feature selection techniques prior to classification, which is a strategy that has also been widely exploited in recent studies [8, 45, 72] on spectral-spatial classification using attribute profiles, where a chain composed by different feature extraction approaches where used to extract the final subset . A compact representation of the morphological profiles, called Morphological Characteristic (MC), was obtained in [83] by analysing the Differential Morphological Profile (DMP) to identify if the underlying region of each pixel is darker or brighter than its surroundings. In [84], an extension of the MC was presented, where the characteristics of scale, saliency, and level of the DMP are identified by a 3-D index for each pixel in the image. A strategy based on a sparse classifier and SUnSAL (Sparse Unmixing by variable Splitting and Augmented Lagrangian) [9] for the analysis of the entire EMAP, was presented in [97].

In this study, the spectral analysis becomes a fundamental part, which aims at extracting the optimal subset of class-informative features. To this purpose, the feature reduction technique based on ICA, introduced in Chapter 4. In the approach, the analysis is performed for each specific class, where the selection of the most representative components relies on the minimization of the reconstruction error computed on the training samples employed for the supervised classification. The spatial analysis is then performed by extracting spatial features based on mathematical morphology. To this purpose, the reduced APs (rAPs), introduced in Chapter 5, are obtained by evaluating the contextual information for each connected region by identifying the best level of representation, according to a homogeneous measure. Such analysis permits the contextual information to be preserved, and at the same time, to address the dimensionality issue, which leads to a highly intrinsic information redundancy. Figure 6.1 shows the general schema of the proposed technique.

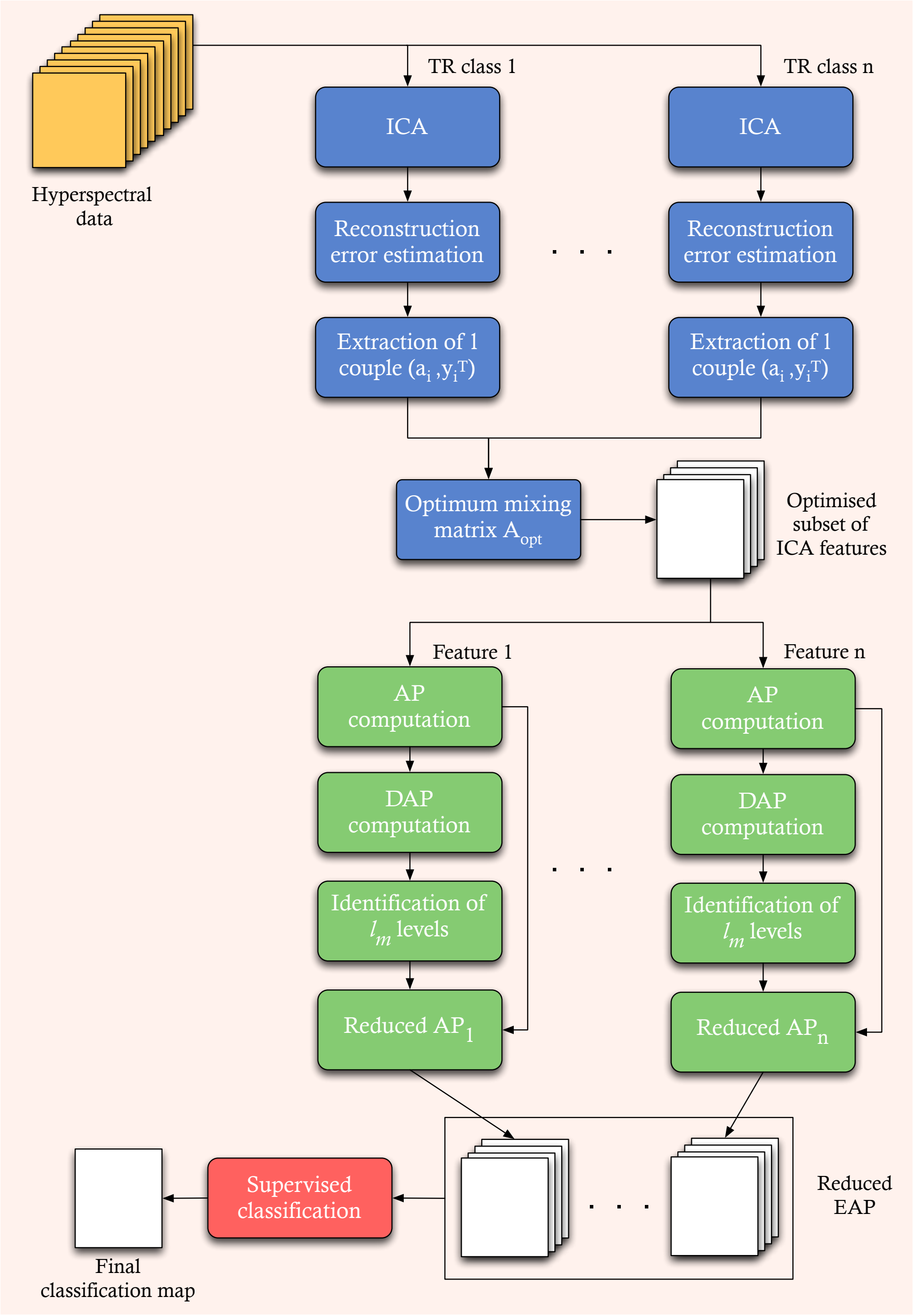

Hyperspectral data
TR class 1
TR class n
ICA
ICA
Reconstruction error estimation
Reconstruction error estimation
. . .
Extraction of 1 couple (a_i, y_i^T)
Extraction of 1 couple (a_i, y_i^T)
Optimum mixing matrix A_{opt}
Optimised subset of ICA features
Feature 1
Feature n
AP computation
AP computation
DAP computation
DAP computation
. . .
Identification of l_m levels
Identification of l_m levels
Reduced AP_1
Reduced AP_n
Final classification map
Supervised classification
. . .
Reduced EAP

Figure 6.1: General scheme of the proposed technique for spectral and spatial information integration for supervised classification.

Table 6.1: Family of increasing criteria employed in Experiments 1 and 2 for each attribute. The number of thresholds is indicated in brackets.

	Experiment 1		Experiment 2	
Area	$[100, 500, 1000, 5000]$	(4)	$[500 : 500 : 5000]$	(10)
Std. dev.	$[20, 30, 40, 50]$	(4)	$[20 : 5 : 50]$	(7)
Diagonal	$[100, 200, 400, 600]$	(4)	$[50 : 50 : 600]$	(12)
Inertia	$[0.2, 0.3, 0.4, 0.5]$	(4)	$[0.2 : 0.05 : 0.5]$	(7)

6.2 Design of Experiments and Investigations

In these experiments, the ICA-based scheme is employed for the extraction of class-representative components, which are then used for building the rAPs. In particular, the feature subsets that provided the best classification accuracy in Chapter 4 are used as input for building the rAPs. This gives the possibility for further comparison between the spectral based approach and the spectral and spatial based approach.

In the analysis, four attributes are considered for the modelling of the spatial information, such as *area* (a), *diagonal of the bounding box* (d), *standard deviation* (s) and *moment of inertia* (i). The proposed method is based on a region extraction process, where a better filtering of the scene would lead to the extraction of regions that would not be identified otherwise. In order to test the performances on different ranges of thresholds, two experiments are set up, where two families of increasing criteria are considered. Experiment 1 exploits the set of values that is usually employed in the literature, while in Experiment 2, the number of thresholds is increased, giving a thicker image decomposition. Table 6.1 shows the rage of thresholds used for building the profiles. It is important to note that an increase of the number of thresholds does not cause an increase of the dimension of the feature space of the rAPs, and thus, of the rEAPs.

6.3 Experimental Setup

For the experimental setup related to the ICA-based approach for feature dimensionality reduction, which includes FastICA and GA tuning, please refer too Section 4.3.

6.3.1 Classification

An SVM classifier [20] with a Radial Basis Function (RBF) kernel is exploited for classification purposes. The algorithm employs the one-against-one multi-class strategy. For the estimation of the regularization parameter, C, and the kernel parameter, γ, 10-fold cross-validation based on the grid-search approach is performed. In particular, an exponentially growing sequences of C and γ are considered, with $C = \{10^{-2}, 10^{-1}, ..., 10^{4}\}$ and $\gamma = \{2^{-3}, 2^{-2}, ..., 2^{4}\}$.

6.3.2 Data Sets' Description

The analysis is carried out on the IC subsets obtained by the four real hyperspectral data sets used in Chapter 4. For each data set, the training samples and the test samples are generated in such way that the two sets results mutually exclusive (i.e., no shared samples between the two sets).

Pavia University: The data set is described in Appendix A.4. Information about the training set used in the experiments are also provided.

Pavia Center: The data set is described in Appendix A.5. Information about the training set used in the experiments are also provided.

Salinas: The data set is described in Appendix A.1. In case of Salinas data set, the training set employed in the experiments is made up of 15% randomly selected samples from each class.

Hekla: The data set is described in Appendix A.2. In case of Hekla, the training set is generated by a random selection of 50 samples from each class.

6.4 Experimental Results and Discussion

This section presents the experimental results of the proposed technique for the integration of spectral and spatial information for classification obtained on the four data sets. Table 6.2 reports all the classification results obtained in Experiments 1 and 2 for each data set, while Figures 6.2 - 6.5 show the classification maps of the best cases (represented in bold in Table 6.2). The results ob-

Table 6.2: Classification results obtained by exploiting the proposed technique for spectral and spatial classification. For each data set, the Table reports the percentage overall accuracies "OA(%)" and the kappa coefficients "*k*" obtained in Experiments 1 and 2 for each attribute. The number of features exploited are given in parentheses.

	Pavia University (12 ICs + 24 spatial features)							
	Experiment 1				Experiment 2			
	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$
OA (%)	94.05	94.90	87.50	83.36	**95.60**	**95.92**	**89.56**	**84.03**
k	0.92	0.93	0.84	0.78	**0.94**	**0.94**	**0.86**	**0.79**
	Pavia Center (17 ICs + 34 spatial features)							
	Experiment 1				Experiment 2			
	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$
OA (%)	99.11	**99.12**	**98.69**	97.83	**99.12**	99.08	97.79	**98.22**
k	0.99	**0.99**	**0.98**	0.97	**0.99**	0.99	0.97	**0.97**
	Salinas (17 ICs + 34 spatial features)							
	Experiment 1				Experiment 2			
	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$
OA (%)	99.14	97.17	95.43	89.19	**99.51**	**97.62**	**95.47**	**90.59**
k	0.99	0.97	0.95	0.88	**0.99**	**0.97**	**0.95**	**0.89**
	Hekla (20 ICs + 40 spatial features)							
	Experiment 1				Experiment 2			
	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$	rEAP$_a$	rEAP$_d$	rEAP$_s$	rEAP$_i$
OA (%)	98.61	98.76	97.27	90.34	**98.87**	**99.14**	**97.84**	**95.16**
k	0.98	0.99	0.97	0.89	**0.99**	**0.99**	**0.98**	**0.94**

tained in Experiments 1 and 2 confirms the fact that the inclusion of spatial information provides a general improvement in the classification accuracies with respect to the case where only spectral information (i.e., the ICs) is considered (see Table 4.1, Section 4.4). In particular, in the case of Pavia University data set, the rEAPs are built starting from the 12 ICs selected by applying the ICA-based feature reduction approach, obtaining profiles that include 36 features. In this case, the attributes *area* and *diagonal* provided the best results obtaining a maximum improvement of 8.23 *pp*. In the case of the Pavia Center data set, the rEAPs are built on 17 ICs obtaining a final vector of 51 features. From the analysis, it can be seen that a good classification accuracy can be achieved

Table 6.3: Classification results obtained by exploiting the rEMAPs for Experiment 1 and 2. "OA (%)" denotes the percentage overall accuracies and " *k*" indicates the kappa coefficients.

		Pavia University	Pavia Center	Salinas	Hekla
Experiment 1	OA (%)	94.59	**99.28**	98.73	98.78
	k	0.93	**0.99**	0.98	0.99
Experiment 2	OA (%)	**96.28**	99.11	**98.95**	**99.08**
	k	**0.95**	0.99	**0.99**	**0.99**

by exploiting the spectral information (see the spectral case in Table 4.1, Section 4.4). However, a slight improvement can be obtained by employing spatial information. Also in this case, the attributes *area* and *diagonal* provided the best accuracies. In the case of the Salinas data set, as for the Pavia Center case, the rEAPs are composed by 51 features including 17 ICs. In this case, the attribute *area* obtained the best classification accuracy with an improvement of 4 *pp* with respect to the only spectral case. In case of the Hekla data set, 20 ICs were extracted, which are used to build 60-feature rEAPs. In this case, the attributes *area* and *diagonal* and *standard deviation* provided an improvement with respect to the spectral case. The best classification accuracy was obtained by using the attribute *diagonal* giving an increase of 3 *pp*. The attributes *area* and *diagonal* are the ones that provided better classification accuracies, while *inertia* resulted in a worse classification accuracy. This is probably due to the fact that the identification of a proper range of thresholds is not trivial, especially for non-increasing criteria, where this is less intuitive with respect to the increasing criteria. Such issue is considered in the next Chapter, and it will be considered in our future research to provide an automatic approach that would be independent of the attribute and the image considered. By comparing the results obtained in Experiments 1 and 2, one can see that a larger range of thresholds leads in general to more representative rEAPs. Table 6.2 shows in bold the best classification accuracies based on the comparison between the two cases (i.e., Experiment 1 versus Experiment 2).

A further experiment is based on the fusion of the information provided by each rEAP to obtained the rEMAP. The strategy adopted for the multi-attribute analysis is based on the fusion of the classification results obtained by the rEAPs (see Section 5.2.4). This choice is justified by the fact that this solution is more robust than using a unique stack of features, while the dimensionality of the problem remains low with a consequently advantage in terms of computational cost. In general, the employment of the fusion strategy provides classification results (Table 6.3) that are quantitatively and qualitatively similar, and in some cases better, than the best case obtained by employing a single rEAP. This is also proved by the classification maps of the rEMAPs (Figures 6.2e, 6.3e, 6.4e and 6.5e) where each class is spatially better represented (i.e., less noisy) with respect to the single rEAP case.

The results obtained in this study by exploiting spectral, spatial information and their combination, can be compared to the ones obtained in other recent works [8, 45, 72, 73], where the use of attribute profiles were exploited and combined with supervised feature extraction techniques in order to reduce the final dimension of the profile and discard the redundant information. In particular, the proposed technique for spectral and spatial analysis outperformed the approaches presented in [72, 73] in terms of accuracies. Here, the Pavia University data set was used for testing by exploiting the attributes *standard deviation* and *area*. Furthermore, supervised feature extraction techniques were employed to both provide the initial feature subset and reduce the final dimension of the profile space. The proposed method outperforms also the approaches considered in [45] for the Pavia Center data set. In particular, by comparing the results obtained in the presented study, one can see that by employing the presented ICA-based approach we are able to reach higher overall accuracy with respect to the spectral-spatial case in [45]. In particular, the rEMAP is able to achieve a higher accuracy than for the case in [45] in which supervised feature extraction techniques are exploited for dimensionality reduction of the final profile. In the case of the Pavia University data set, the proposed approach obtained higher accuracies compared to the case in which the original AP are used, while it provided very close (and in some cases higher) accuracies to the cases in which feature extraction techniques are exploited. In terms of accuracies, the proposed approach outperforms also the strategy adopted in [8] considering the case in which standard training set is exploited for the classification. The reason for such comparison is to prove the effectiveness of both the ICA-based approach in extracting class-informative features and the reduced APs in providing subsets of spatial features in which the redundant information is discarded. Moreover, the comparison proves that by optimizing the information extraction, the inclusion of additional process steps in the classification chain, such as the multiple use of supervised feature extraction techniques, can be avoided.

6.5 Conclusions

In this Chapter, a novel technique for spectral and spatial supervised classification of hyperspectral images is presented. The technique aims at optimizing the usage of ICA in class-informative feature extraction, while minimizing the disadvantages in the use of APs and its extensions (i.e., EAPs and EMAPs), such as the information redundancy, which limits the classification capabilities. In particular, class-informative features were extracted by exploiting a novel dimensionality reduction strategy based on ICA, where the use of ICA was optimized for its application in a high-dimensional scenario (i.e., no dimensionality reduction was performed prior ICA). The retrieving of class-information was assured by solving an optimization problem based on the minimization of the reconstruction error of the ICs extracted for each specific class. The reconstruction error, computed on the training samples defined for the classification stage, was used as estimation of the class-information content and exploited to rank the extracted ICs. The spatial information was, then, extracted from the identified subset of ICs by employing the reduced Attribute Profiles (rAPs). The rAP is an optimized version of the well known morphological morphological Attribute Profiles (APs). In rAP the geometrical information related to the filtered regions are adaptively selected based on a homogeneous measure. The extraction is based on the multi-level analysis of DAP, which shows filtered regions between adjacent levels of the AP. By using this approach, it was possible to fuse the geometrical information into few features. The method was tested on four real hyperspectral images, which were different in spectral / spatial resolutions and content. The obtained results showed the effectiveness of the proposed technique in extracting spectral and spatial features providing higher or similar accuracies when compared to state of the art.

Part of this Chapter is going to appear in:
N. Falco, J. A. Benediktsson, and L. Bruzzone. Spectral and Spatial Classification of Hyperspectral Images Based on ICA and Reduced Morphological Attribute Profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 2015. (accepted) [38].

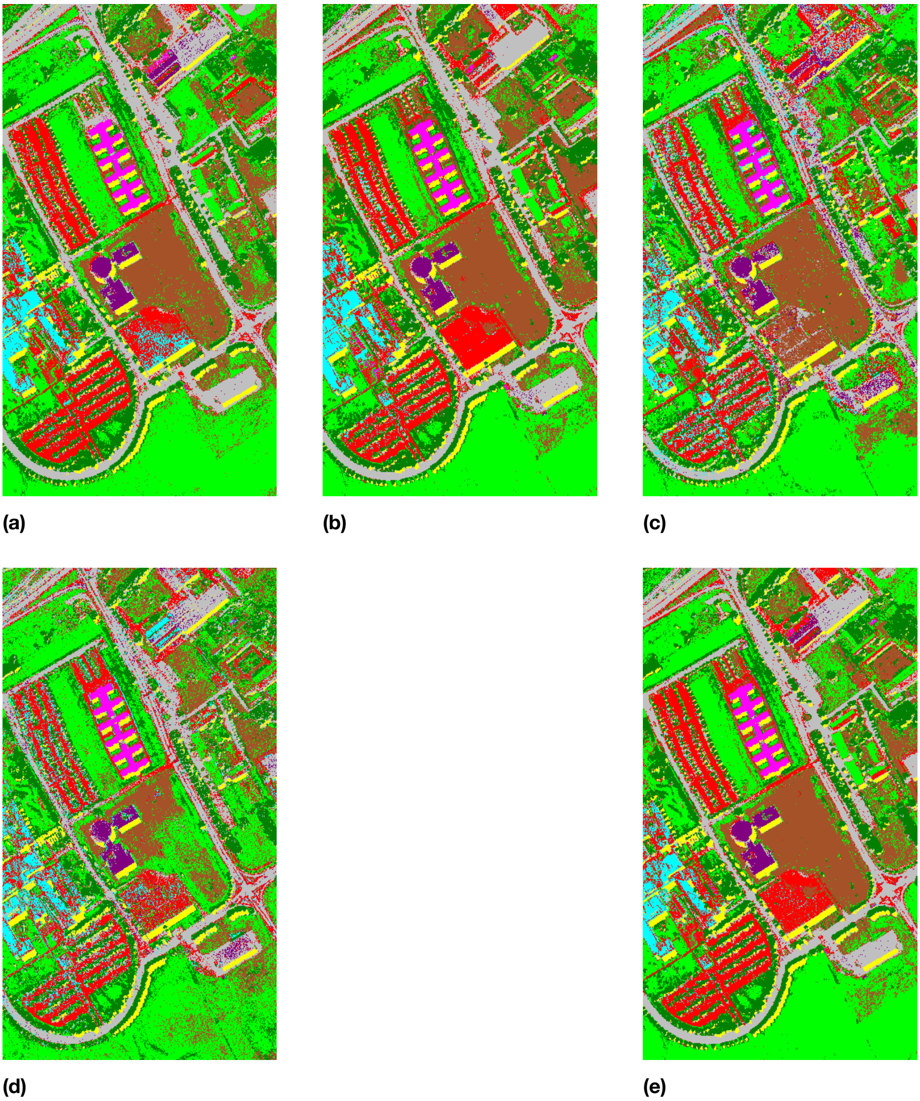
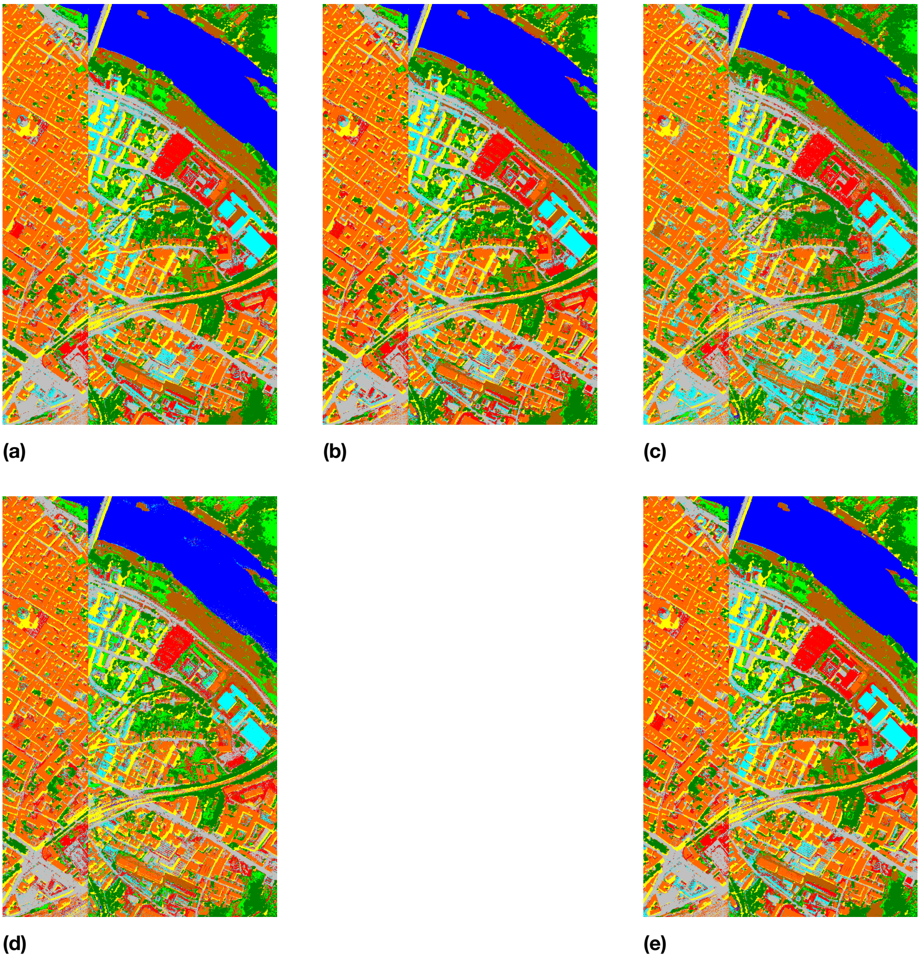


Figure 6.2: Classification maps of Pavia University: (a) rEAP$_a$; (b) rEAP$_d$; (c) rEAP$_s$; (d) rEAP$_i$; (e) rEMAP.



(a)

(b)

(c)

(d)

(e)

Figure 6.3: Classification maps of Pavia Center: (a)rEAP$_a$; (b) rEAP$_d$; (c) rEAP$_s$; (d) rEAP$_i$; (e) rEMAP.

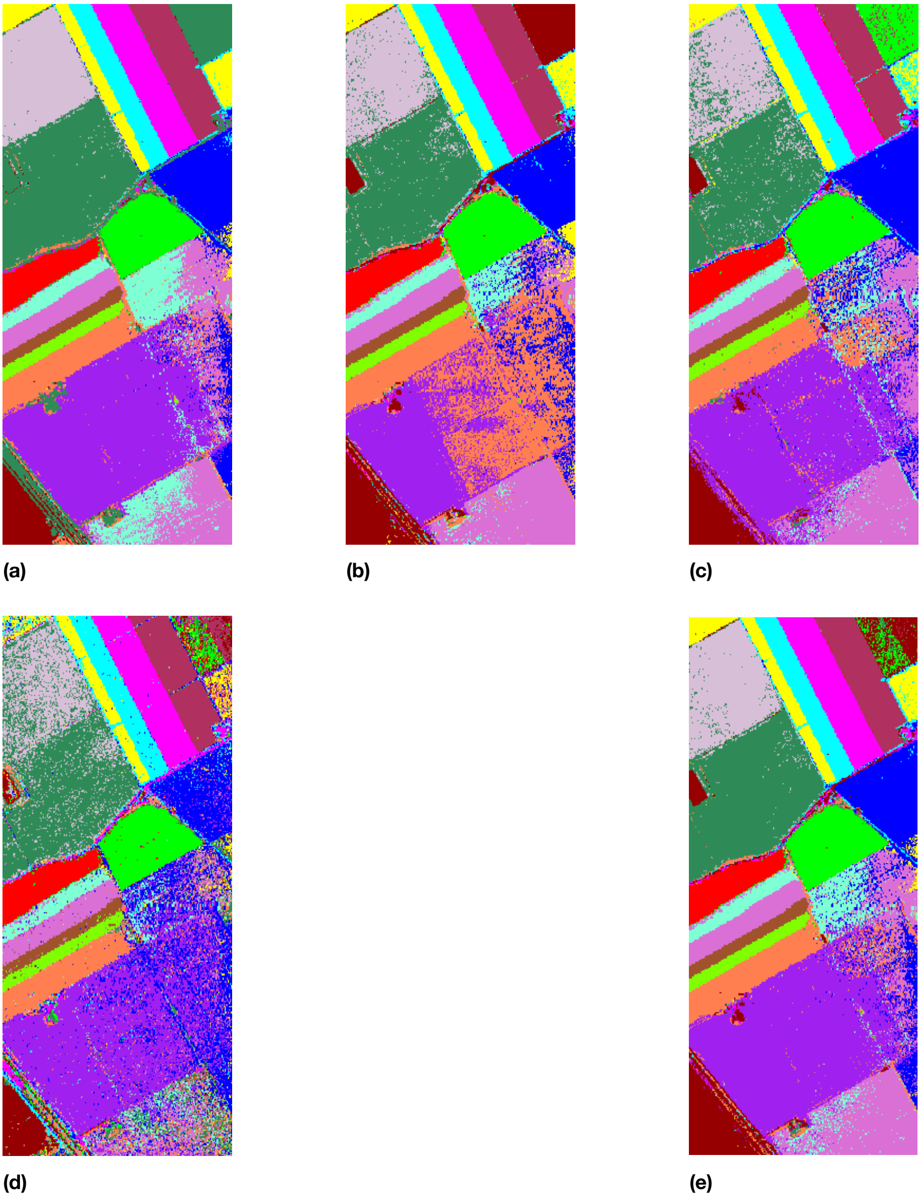


Figure 6.4: Classification maps of Salinas: (a) rEAP$_a$; (b) rEAP$_d$; (c) rEAP$_s$; (d) rEAP$_i$; (e) rEMAP.

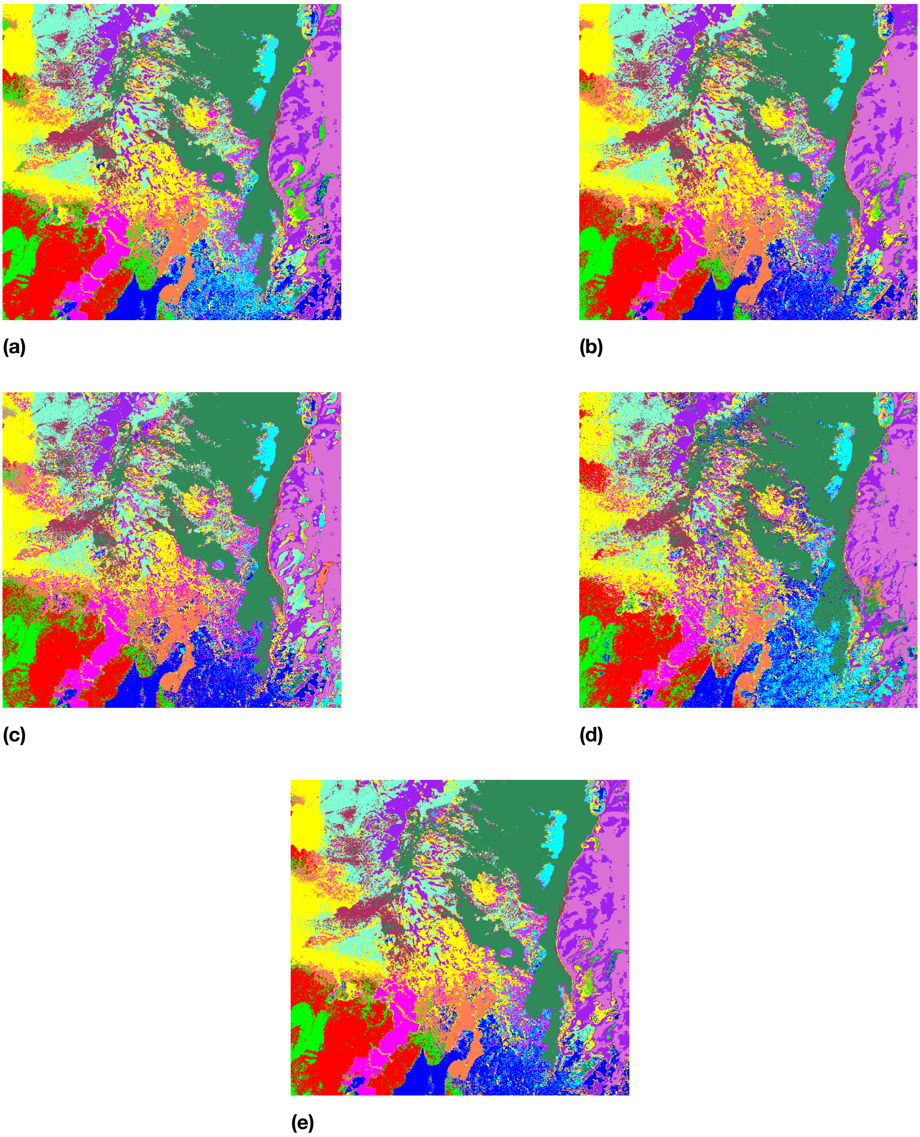

(a)
(b)
(c)
(d)
(e)

Figure 6.5: Classification maps of Hekla: (a) rEAP$_a$; (b) rEAP$_d$; (c) rEAP$_s$; (d) rEAP$_i$; (e) rEMAP.

7
Automatic Threshold Selection for Profiles of Attribute Filters

This Chapter proposes a novel technique for the automatic selection of the filters' parameter, aiming at addressing the issue related to the choice of a proper set of filtering thresholds used to build representative attribute profiles. The technique is based on the concept of granulometric characteristic function, which provides information on the image decomposition according to a given measure, and exploits the tree representation of an image, which allows us to avoid filtering steps usually required prior the threshold selection, making the process computationally effective.

7.1 Introduction

Profiles obtained by the sequential application of morphological attribute filters are very powerful and flexible operators, able to richly extract information on the spatial arrangement and characteristics of the objects in a scene. However, the image decomposition relies on the selection of thresholds, which should be tuned in order to provide a profile that is both representative (i.e., it contains salient structures in the image) and non-redundant (i.e., objects are present only in one or few levels of the profile). The selection of an appropriate range of thresholds remains one of the main open issues. In the literature, few attempts have been done to solve the issue related to the selection of the threshold values. The common approach is based on field-knowledge of the scene, where the values are manually selected by a visual analysis of the scene under consideration [35]. In [69] the set of threshold was derived after a preliminary classification and clustering of the input image, while, in [73], considering a supervised classification scenario, the thresholds for the specific attribute of standard deviation were based on statistics of the available training samples. In [42] the filter thresholds were chosen based on the analysis of a granulometric curve (i.e., a curve related to the size distribution of the structures in the image [96]). In particular, the thresholds selected are those whose granulometric curve best approximates the one obtained by considering a large set of thresholds. The main drawback that all the aforementioned approaches have is the need of computing a large number of filtering steps (potentially with all possible thresholds) in order to be able to identify those thresholds that are significant. Consequently the computational cost and the memory to store an unidentified number of filtered images need to be considered.

In this work, a novel strategy for the automatic selection of the thresholds based on the concept of *granulometric characteristic functions* (GCFs) is proposed. The GCF can be seen as an extension of the concept of granulometry. Considering a series of morphological opening operations, a granulometric curve is computed as the sum of the pixel values of each image result of an opening versus the threshold reference [96]. By duality, an anti-granulometric curve is derived by the closing operator. Here, a GCF is defined as a measure that is computed on the tree representation of the input image, showing the evolution of a characteristic measure for increasing values of the thresholds (i.e., increasingly coarser filters). In this framework, different measures (e.g., related to the gray-levels, number of pixels, etc.) can be considered, leading to the definition of different GCFs. The proposed technique exploits the tree representation of an image; in particular, for gray-scale images, the corresponding tree representation provides useful a-priori information (i.e., prior to the filtering) related to the actual range of the attribute values. Due to this, the computation of the GCF can be performed directly from the tree, without requiring any filtering of the analysed image. Aiming at showing the flexibility of the proposed approach based on the analysis of GCFs, three measures for computing the GCF are considered for testing. Similarly to [42], in the proposed approach, the set of selected thresholds correspond to the one that best approximates the GCF computed on the full set of thresholds. By approximat-

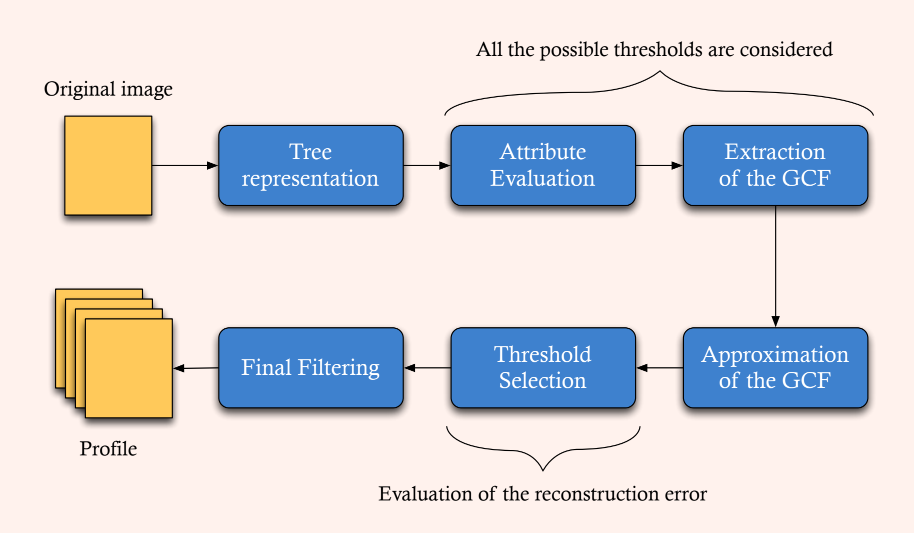


Figure 7.1: General scheme for the proposed automatic threshold selection technique.

ing the GCF curve, it is assumed that the distribution of a given measure along the profile can be extracted and approximated by using the subset of selected thresholds. In [42], such a strategy was applied on a conventional granulometry curve, requiring to explicitly filter the images accordingly with an initial set of thresholds, which was manually defined prior to the filtering. The advantages of the proposed approach are that the initial range corresponds to the set of the all possible thresholds, thus it does not require any initial selection, and the GCF curve is derived from the tree structure, without involving any filtering step. To identify the thresholds to build the profiles, the method employs a piecewise linear regression model [75]. The approach is tested on Pavia data set, considering three different tree representations, min-, max- and inclusion trees.

7.2 Proposed Automatic Threshold Selection

The proposed technique is based on descriptive functions computed on a profile, namely *granulometric characteristic functions* (GCFs), which extract the trend of a given property that characterize the effect of the image decomposition along the tree. The selection of representative thresholds relies on the approximation of the GCFs aiming at preserving the distribution in order to extract those thresholds that provide a significant change in the effect of the filtering. Figure 7.1 shows the general schema of the proposed technique.

7.2.1 Measure and Granulometric Characteristic Function

A *granulometric characteristic function* (GCF) is defined as a function returning a measure $\mathcal{M}$ which is computed on a profile:

$$\mathrm{GCF}(\Pi_\psi(f)) = \{\mathcal{M}(\psi_i)\}_{i=1}^{L}. \tag{7.1}$$

Thus, if $\mathcal{M} : f \to \mathbb{R}$, $\mathrm{GCF}(\Pi_\psi(f))$ leads to L scalar values (one for each image in the profile). GCF is inspired by the granulometry curve; in spite of this, it can be extended to other characteristics, additionally to the sum of gray-levels (as it is conventionally considered in the granulometry). As standard granulometric curves show the interaction of the size of the image structures with the filters when the filter parameter varies, so GCFs provide information on the effect of subsequent filtering with respect to some characteristics of the image. In this work, the definitions of three measures and the relative GCFs are proposed:

Sum of gray-level values As for the conventional granulometry, this measure provides information related to the effect of the filtering with respect to the changes in terms of gray-levels that are produced in the image.

$$\mathrm{GCF}_{val}(\Pi_\psi(f)) = \left\{\sum |f - \psi_i(f)|\right\}_{i=1}^{L}. \tag{7.2}$$

Number of changed pixels Another possible measure is the number of pixels that change gray-value at different filtering. In this case, the GCF is sensitive to changes in the spatial extent of the regions rather than in gray-levels.

$$\mathrm{GCF}_{pix}(\Pi_\psi(f)) = \left\{\mathrm{card}[f(p) \neq \psi_i(f)(p)], \forall p \in E\right\}_{i=1}^{L}, \tag{7.3}$$

with card[·] the cardinality of a set.

Number of changed regions This GCF shows the number of connected components that are affected by each filter and it is a topological measure invariant to the spatial extent and gray-level variations induced by the filterings.

$$\mathrm{GCF}_{reg}(\Pi_{\psi}(f)) = \left\{\mathrm{card}[\mathcal{C}(f)] - \mathrm{card}[\mathcal{C}(\psi_i(f))]\right\}_{i=1}^{L}. \qquad (7.4)$$

According to the definitions reported above, the GCFs are monotonic increasing functions, since the measures that are considered increase for progressively coarser filters. Clearly, other measures can be considered for the definition of different GCFs, if the interest lies in investigating the effects of the filtering with respect to other image characteristics.

7.2.2 Threshold Selection

The problem to address is the selection, among the set $\bar{\Lambda} = \{\lambda_i\}_{i=1}^{L}$ of all possible values of λs, of a subset $\hat{\Lambda} = \{\hat{\lambda}_i\}_{i=1}^{\hat{L}}$ with $\hat{L} \ll L$. The full set $\bar{\Lambda}$ is extremely scene dependent and can potentially be very large, making the problem of selecting the subset $\hat{\Lambda}$ more complicated to realize since the full set is not readily accessible. A possible strategy for the selection relies on the computation of a profile by considering a relatively large number of λs (considering all of them in real scenarios is impractical) and prune the profile by selecting some of filtered images and related filter parameters so defining $\hat{\Lambda}$. However, such an approach is limited by the need of generating the filtered images in order to perform the selection and by the lack of guarantee that all possible threshold are considered for selection. The method exploits the GCFs, defined in Section 7.2.1, in order to select those values λs that lead to "significant" changes in the effect of the filters (as measured by the considered GCF). This approach is not new since considering the granulometric curve for estimating values of λ that generate salient filtering images has been already proposed in [42]. However, this work take advantage of the exploitation of the tree representation of the image (augmented with the values of the attributes for each node) prior to any filtering. In particular, each node, which maps a region of spatially connected pixels in the image, gives information related to the value of attributes, gray-level and number of pixels. This allows us to know all possible values of λ (i.e., to know exactly the full set $\bar{\Lambda}$) and compute the GCFs before any filtering. Similarly to [42], in the proposed approach, the set $\hat{\Lambda}$ of the selected thresholds correspond to the one that best approximates the GCF computed on the $\bar{\Lambda}$ thresholds. The main assumption is that by approximating the GCF curve, the distribution of the measure $\mathcal{M}$ that underlies the GCF can be extracted and approximated by using the selected $\hat{L}$ thresholds. The approximation of the GCF curve is achieved by means of piecewise linear regression [75], in which the independent variable (i.e., the chosen measure) is partitioned into intervals and approximated with separate line segments that fit each interval. The boundaries between the segments are identified by breakpoints, for which projections over the x-axis correspond to the set of selected thresholds $\hat{\Lambda}$. The automatic selection of the thresholds relies on the analysis of the reconstruction error computed between the GCF obtained by considering all the possible threshold values and the approximation of the GCF.

7.3 Experimental Setup

The experimental analysis is performed on the first principal component extracted from Pavia University. In particular, the min-, max- and inclusion trees representations are computed on the PC and used to derive the three GCFs. Here, the attribute *area* is considered for building the attribute and self-dual attribute profiles. However, according to the definition of the GCF, the method can be computed considering any attribute. The set of thresholds for the attribute *area* is composed of $L = 2076$, $L = 1963$ and $L = 2508$ unique values for the min-tree, the max-tree and the inclusion tree, respectively.

7.4 Experimental Results and Discussion

In the proposed approach, the number of segments used for the approximation are identified by minimizing the reconstruction error (i.e., the mean absolute error, MAE) between the GCF and its approximation. Due to the nature of the employed regression model, the first and the last breakpoints correspond to the first threshold (i.e., equal to 0, resulting in the original input image) and to the last threshold (resulting in all pixels having the same gray-scale value), respectively, which do not provide useful information. For this reason, these two breakpoints are discarded. Accordingly, the number of thresholds equals to the number of segments minus one. An example of reconstruction error estimation, computed for all the tree representations, is shown in Figure 7.2. There, the reconstruction error has been computed for $\hat{L} = 1, \ldots, 15$ segments. Considering the obtained trend of the reconstruction errors, the attribute closing profile, the attribute opening profiles and the self-dual attribute profile are built by using the first 4 selected thresholds for each GCF, i.e., considering the first 5 segments. Figure 7.3 shows the regression analysis performed to approximate the three derived GCFs. In the figure, the red circles represent the real GCFs (computed with all the thresholds in $\bar{\Lambda}$), the green line denotes the approximation based on 5 segments, and the black circles identify the breakpoints (i.e., the $\hat{\Lambda}$). One can see that the breakpoints are different from one to another GCF, providing different sets $\hat{\Lambda}$ of thresholds, meaning that a different measure has a different distribution along the profile. Figures 7.4 - 7.12 show the attribute closing, the attribute opening and the self-dual attribute profiles, obtained by selecting the 4 thresholds, which correspond to the breakpoints without the extremities. From the obtained results, it is possible to notice how the considered GCFs are able to model the contextual information according to the chosen measures, while the proposed approach for threshold selection is able to identify those λs values that better characterize the main changes (i.e., changes in slope) in the the distribution of the original GCFs.

7.5 Conclusions

In this Chapter, a novel technique for the automatic selection of thresholds based on a new concept of granulometric characteristic functions, is presented. Granulometric characteristic functions are derived from the hierarchical representations of the image, and computed considering a measure of interest. By exploiting the tree (i.e., min-, max-, inclusion- trees) representations, filtering steps prior the selection of the threshold set become unnecessary, making the approach computational efficient. Three GCFs are defined based on different measures, such as, the sum of the gray-level values (i.e., based on the conventional granulometry), the number of changed pixels and the number of changed regions. In addition to the standard granulometry, which is related to the volume (sum of the gray values) of variations, GCFs derived from the other measures show the effects of the decomposition in terms of spatial extent (i.e., how large are the areas that got changed). At this point, a piecewise linear regression model is employed to approximate the GCF. An algorithm identifies the number and position of the breakpoints that minimise the reconstruction error of the GCF, providing the best approximation of the GCF. The meaningful thresholds are then derived from the obtained breakpoints. The effectiveness of the proposed approach is assessed by a qualitative analysis of the obtained APs and SDAPs built on Pavia University data set. According to the chosen attribute (i.e., area) and measures, the obtained image decompositions present effective multi-level characterizations of the original input scene, providing profiles that are representative of and non redundant. The selection of thresholds that provides similar information is avoided due to the characteristic of the regression model, which inserts a breakpoint where the curve presents a change in slope, which corresponds to a significant change in the image decomposition.

Part of this Chapter is submitted in:

G. Cavallaro, N. Falco, M. Dalla Mura, L. Bruzzone, and J. A. Benediktsson. Automatic Threshold Selection for Profiles of Attribute Filters Based on Granulometric Characteristic Functions. In *Proc. of 12th International Symposium on Mathematical Morphology (ISMM 2015)*, 2015. (accepted) [19].

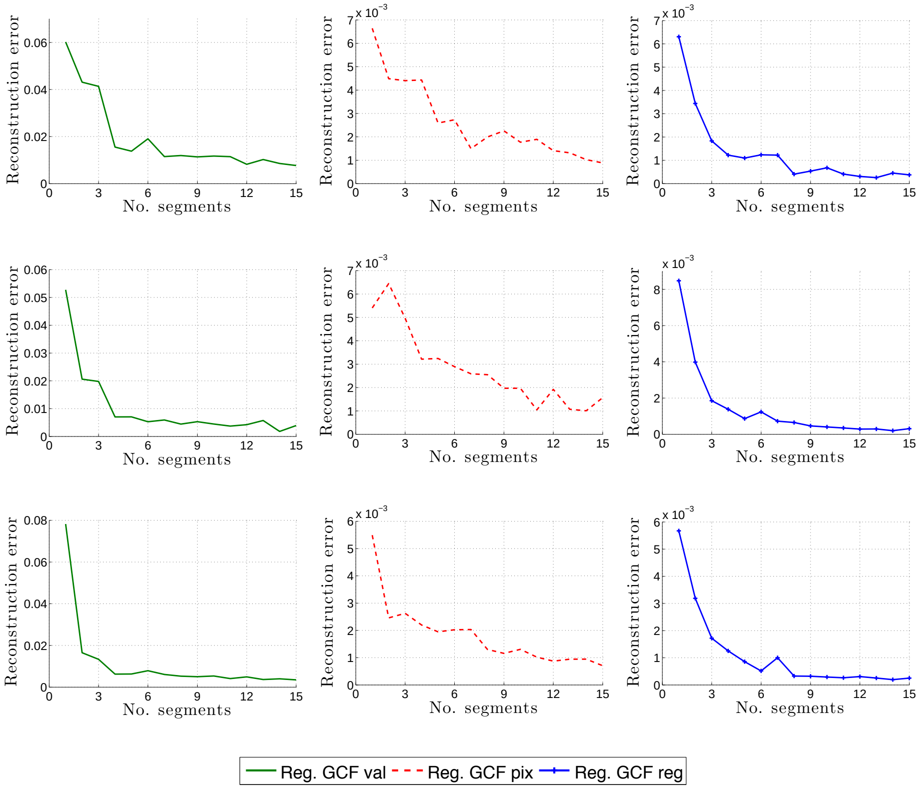
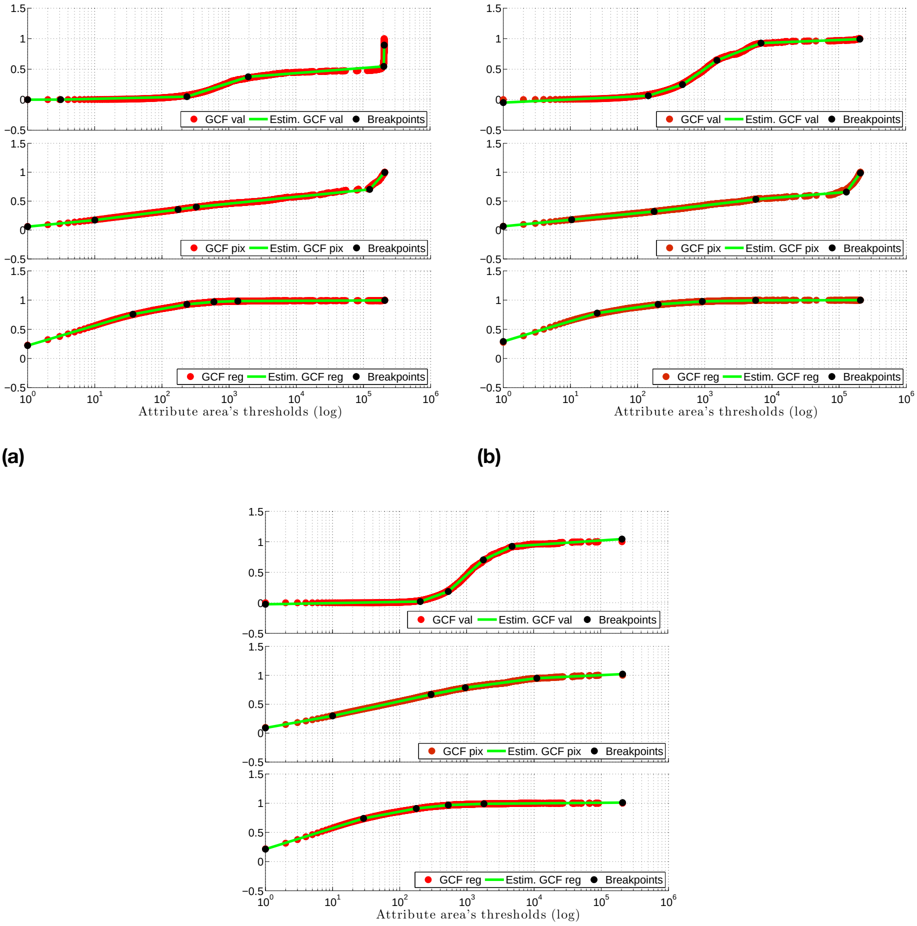


Figure 7.2: Evaluation of the reconstruction error computed for the estimation of the GCFs derived by min-tree (top line), max-tree (middle line) and inclusion tree (bottom line) for Pavia University. The GCF$_{val}$ is denoted in green, the GCF$_{pix}$ is denoted in red, and the GCF$_{reg}$ is denoted in blue.



(a)

(b)

(c)

Figure 7.3: GCFs derived by (a) min-tree, (b) max-tree and (c) inclusion tree. For each GCF (red circles), the estimated curve (green line) and the breakpoints (black circles), which are used to derive the thresholds, are shown.

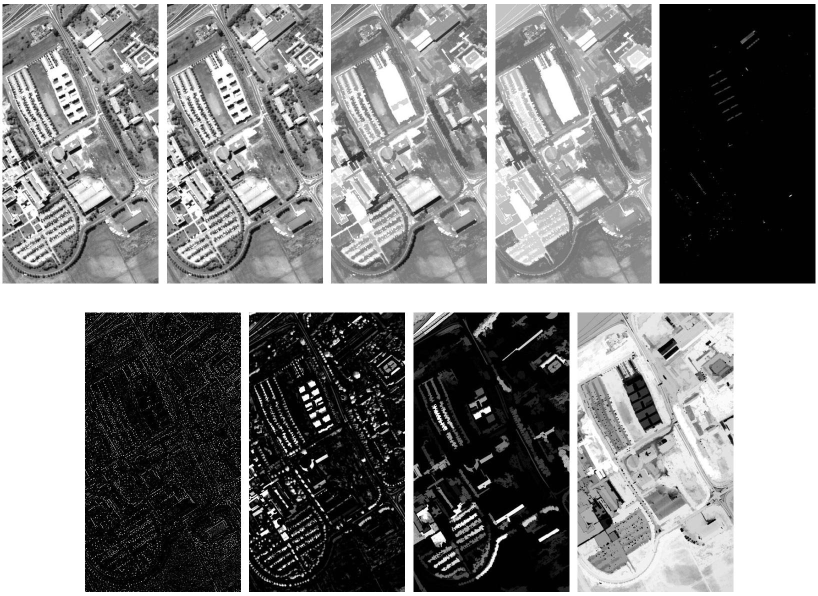


Figure 7.4: Π_φ computed on Pavia based on GCF$_{val}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

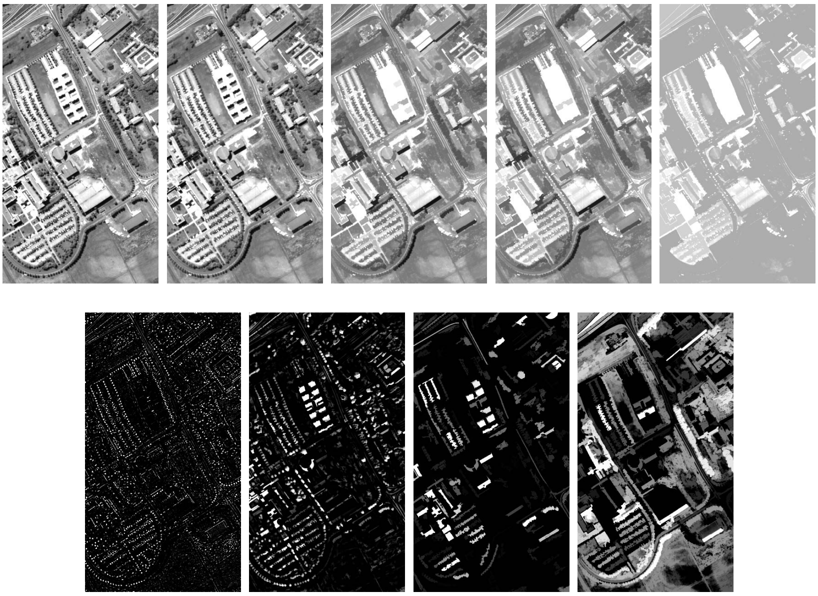


Figure 7.5: Π_φ computed on Pavia based on GCF$_{pix}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

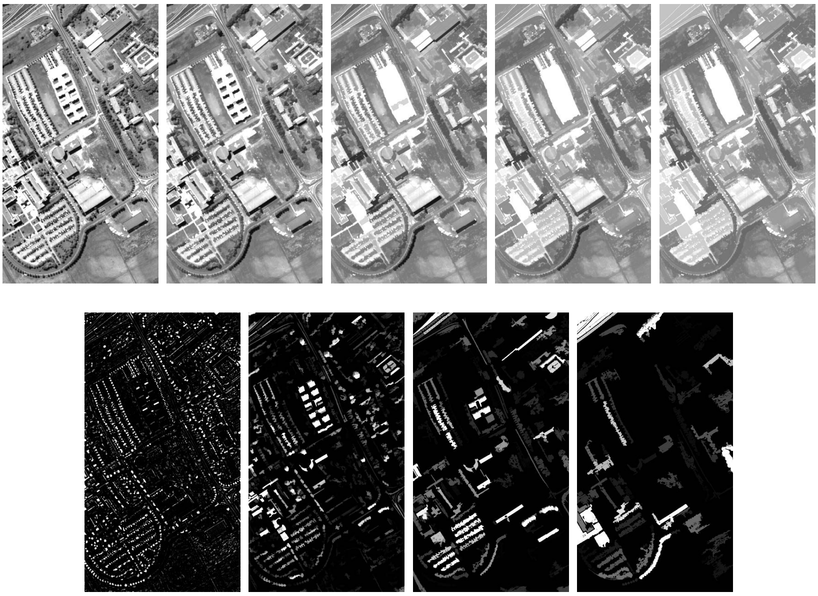


Figure 7.6: Π_{φ} computed on Pavia based on GCF$_{reg}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

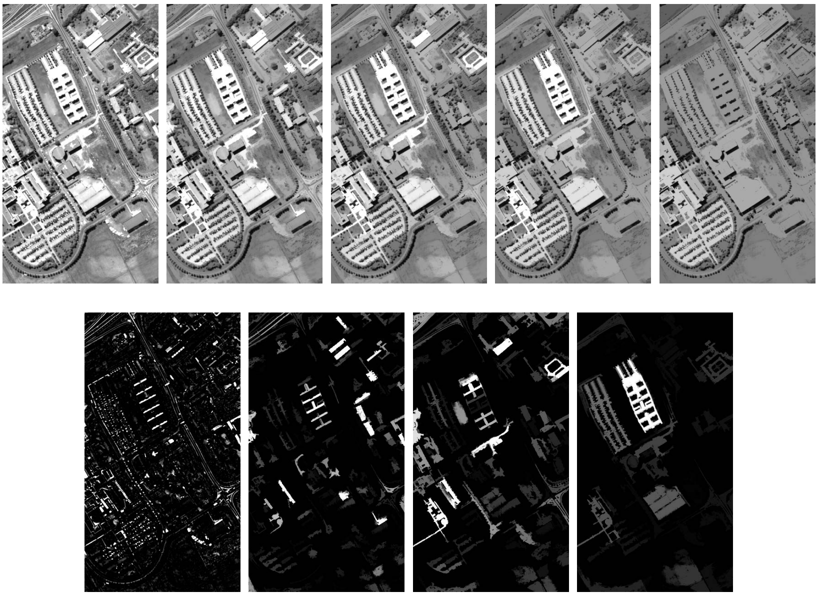


Figure 7.7: Π_{γ} computed on Pavia based on GCF$_{val}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

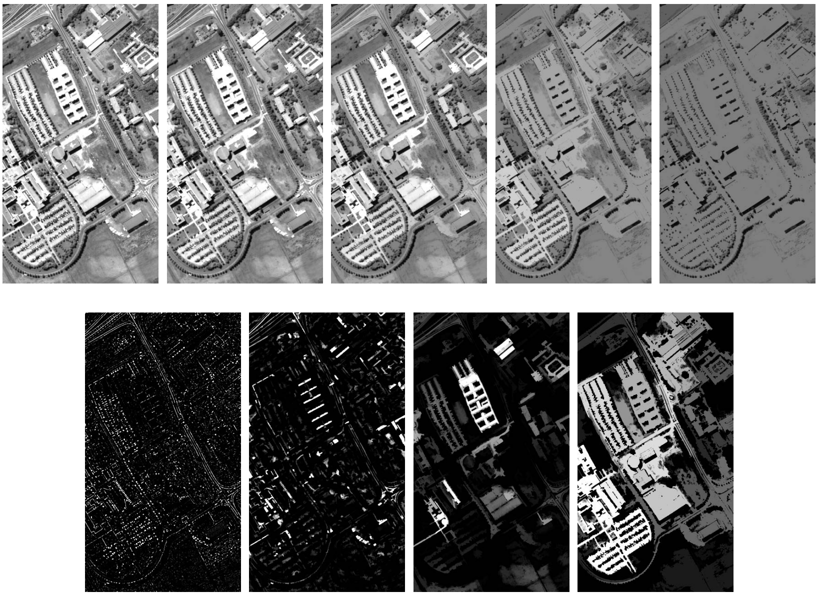


Figure 7.8: Π_γ computed on Pavia based on GCF$_{pix}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

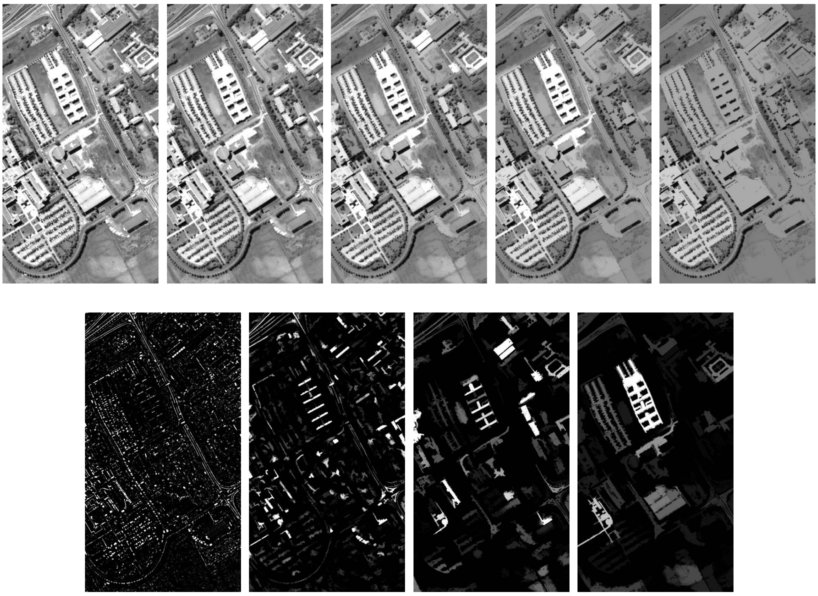


Figure 7.9: Π_γ computed on Pavia based on GCF$_{reg}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

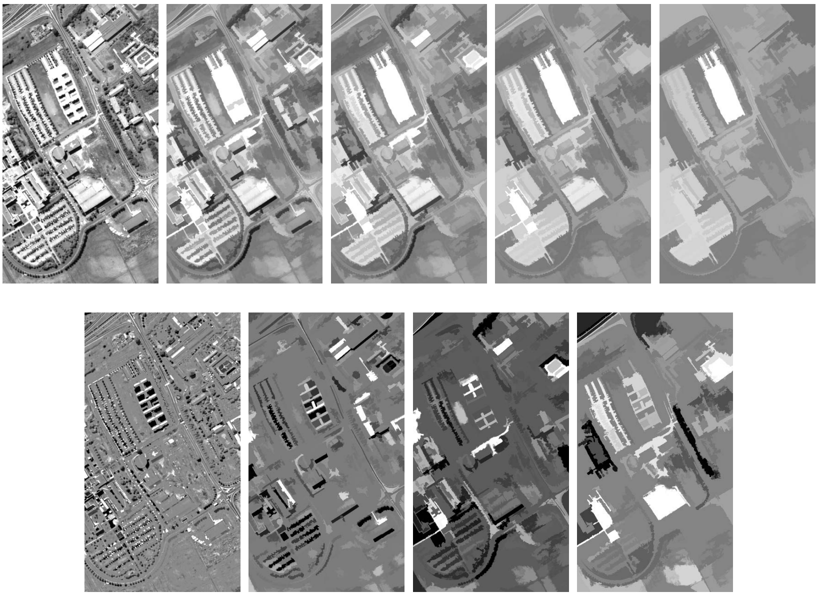


Figure 7.10: Π_ρ computed on Pavia based on GCF$_{val}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

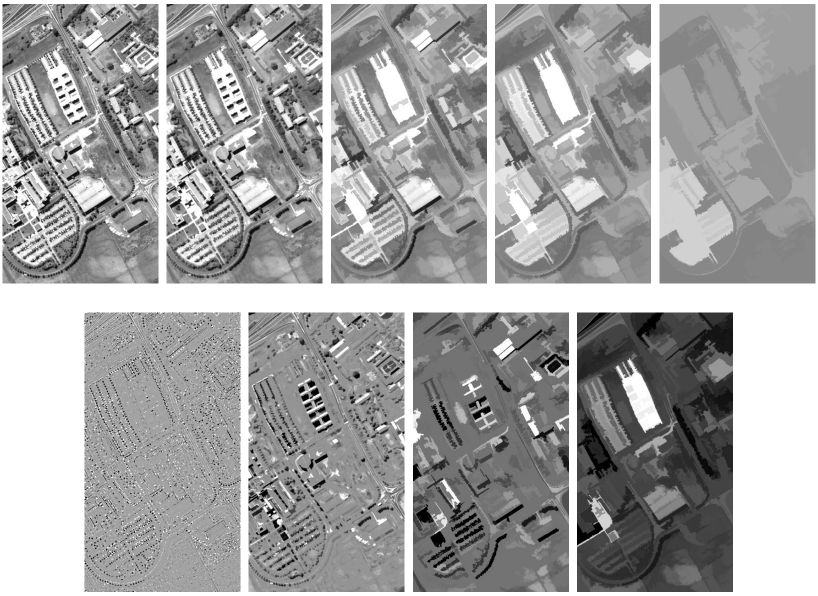


Figure 7.11: Π_ρ computed on Pavia based on GCF$_{pix}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

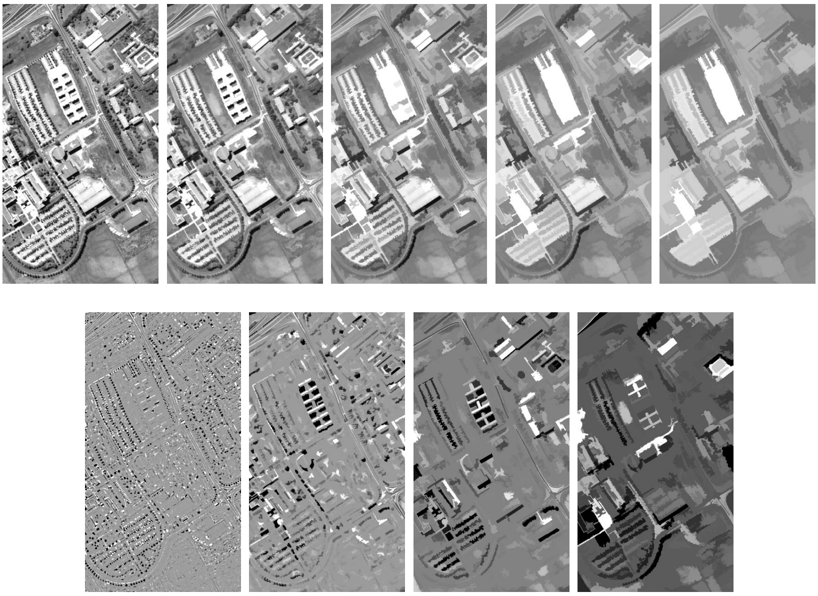


Figure 7.12: Π_ρ computed on Pavia based on GCF$_{reg}$. Thresholds' value is increasing from left to right, with the first column coinciding with ψ^{T_0} (i.e., the original image).

8
Conclusions

This Chapter concludes the thesis presenting a general discussion on the work and the obtained results, reviewing the main contributions. Finally, promising directions for future work developments are presented.

The latest advances in the remote sensing field have contributed significantly at the broad availability of high quality hyperspectral image data. Accordingly, the development of efficient and robust algorithms for the analysis of these data is a very important topics in the remote sensing field. Coupling the high spectral resolution to an increasing spatial resolution of the latest sensors, hyperspectral images are unarguably congruous for classification and land-cover mapping. The classification problem, aiming at detecting and identifying the different land-covers that characterize a given geographical area of interest, is a complex process that involves different procedures whose aim is to extract and analyse all the useful spectral and spatial information that hyperspectral images contain. This thesis aimed at developing methodologies for classification of hyperspectral images that accurately detect and identify the various types of land-covers.

8.1 Contribution of this Dissertation

A first step (see Chapter 3) towards this direction is performed by means of a detailed comparison among the three most broadly used ICA algorithms for hyperspectral image classification, i.e., Infomax, FastICA and JADE. This is an essential step in order to acquire profound insights on the exact assumptions and trade-offs that each implementation encompasses as well as the advantages on the hyperspectral image analysis. The results pointed out that the exploitation of prior information in feature extraction approaches used for dimensionality reduction prior to ICA, allows the extraction of better sets of independent components, leading to more accurate classifications. Infomax resulted in general to be the worst in terms of both computational time and classification accuracy. JADE was the implementation with the best performance in terms of classification accuracy, while in terms of computational time it is comparable to the ones obtained by FastICA being in many cases faster. Similarly, when scaling up to higher dimensionality and without applying any dimensionality reduction prior to ICA, ensuring in this way the preservation of information, FastICA outperformed Infomax, in both computational time and classification accuracy. JADE, on the contrary, required a massive computational load and thus is not adequate for this type of analysis. Information lossless schemas imply full consideration of the data, resulting thus more noisy and negatively affecting the classification accuracy. In Chapter 3, ICA performances were also tested against the number of input samples. This pointed out that decreasing the number of input samples increases significantly the convergence speed, while maintaining the classification accuracies. The approach was more effective in lower dimensionality spaces, where Hughes' phenomenon is not critical. Unlike supervised classification algorithms (i.e., SVM), ICA was proven to be negligibly affected by the reduction of the input sample size and could still provide "good" ICs even when few samples were exploited. This observation is of essential importance in applications, for which the computational time and the number of available samples are crucial issues.

Based on the previous findings, in Chapter 4, a novel ICA-based feature reduction approach was specially designed to retrieve class-informative features in a high-dimensionality scenario, i.e., where no prior dimensionality reduction is applied. The selection of the ICs subset was decided upon the minimization of a criterion function based on the reconstruction error measured for the ICs extracted from each specific class. A genetic algorithm based approach was employed for the selection of the final subset. The results obtained confirm that an appropriate use of ICA can bring prominent improvements in selecting the most representative components, leading to significantly higher classification accuracies, while maintaining the computational cost and time of the ICA extraction low. In particular, the application of the ICA on the entire Salinas and the Hekla data sets (see Table 3.3), where the ICs were selected by employing the SA feature selection approach, resulted in poor classification OAs, 91.96% and 82.05%, respectively; while, considering feature reduction step prior to ICA (see Table 3.1) the OAs were 95.48% and 94.81%, respectively. The proposed schema obtained slightly higher OAs (95.30% and 96.28%, respectively). However, it introduced an automatic approach that non only isolates the most informative features without any supervision, but also identifies the optimum number of the components to keep (see Table 4.1).

Having analysed the exploitation of the spectral information that hyperspectral images can provide, the analysis was extended to the spatial information domain. The first step toward this direction was the definition of a novel strategy for extracting spatial information based on a optimized version of attribute profiles (AP) (presented in Chapter 5), aimed at reducing both the dimensionality and the redundancy of the information that characterises the AP. The algorithm considered, built upon multi-scale analysis of the DAP behaviour, resulted in the extraction of geometrical features that correspond to meaningful structures in the scene at different scales. According to homogeneity criteria, the original AP was compressed, fusing the most informative geometrical information into few features. The emerged *reduced AP*'s feature space accounts for three features types, the reduced thickening and thinning profiles as well as the original image. Compared against to the original APs, the reduced APs achieved comparable or higher classification accuracies, while using only few features (i.e., in the presented case was one third of the number of feature of the original AP). It is worth noting that, in contrast to the original AP, the number of thresholds used in the filtering process does not affect the final number of features that compose the reduced AP. This property brings important advantages in the cases of multi-attribute and multi-channel analysis, where the use of reduced EAPs and reduced EMAPs for modelling the spatial context limits the Hughes phenomenon.

In Chapter 6, the aforementioned findings from Chapters 4 and 5, were explored for defining a methodology which takes advantage of the utmost of the available information fusing both spectral and spatial features in the classification task. The most representative features, as identified by the feature selection approach based on ICA (Chapter 4), were processed by a spatial analysis algorithm. A sharp improvement of the classification accuracies with respect to the spectral features used alone and most importantly against the current state-of-the-art approaches was achieved. Moreover, compared to the spectral-spatial approaches based on AP and proposed in the state-of-the-art, the obtained results proved that the inclusion of additional process steps in the classification chain, such as the multiple use of supervised feature extraction techniques, can be avoided when an optimization of the extraction of both spectral and spatial information is performed.

In Chapter 7, the issue related to the selection of the optimum range of filtering thresholds that provides representative and non redundant profiles, was address by presenting an automatic selection approach. The algorithm was based on the new concept of granulometric characteristic functions, derived from hierarchical representations of the image, and computed considering a measure of interest. In contrast with the current state-of-the-art, the presented approach exploits the tree, i.e., min-, max-, inclusion tree representations of the image, permitting to identify the complete range of available thresholds and sequentially to select a subset without applying any filtering. This breakthrough methodology advances the level of independence of the algorithm, making the approach high computational and memory efficient. these properties that are very important when large images are considered in the analysis.

8.2 Future Research Developments

In this dissertation, innovative methodologies are presented, which significantly improve the state-of-the-art in analysing and extracting information from hyperspectral images. These methodologies considered both spectral and spatial information, focusing on supervised image classification. The experiments carried out, pointed out on a series of potential improvements that are promising directions for future research.

- Part II: *a*) The proposed feature dimensionality strategy, based on ICA could be improved by defining an automatic estimation of the parameter l, which represents the number of ICs to be retained. The choice of l should not necessarily be the same for each class. This improvement would allows us to obtain a fully automatic and parameter-free approach. *b*) The exploitation of the genetic algorithms could be improved in terms of computational efficiency, if a different fitness function (e.g., Jeffries-Matusita distance) is evaluated, replacing the SVM classifier that increases the computational load due to the cross-validation procedure. *c*) Considering the advances in parallel computing, kernel-based feature extraction (e.g., kernel ICA) implementations, which require high computational effort, could also be investigated.

- Part III: *a*) The employment of an automatic threshold selection strategy in the spectral-spatial classification approach, is a straight forward step in order to make the entire procedure fully automatic. This would provide the context necessary for the validation of the proposed technique in providing informative profiles to address the hyperspectral image classification task. *b*) The tree structures (i.e., min-, max- and inclusion trees) have proven their efficiency in implementing attribute operators, avoiding any filtering step. Such a strategy should be included in the extraction of the reduced AP, in order to further decrease the computational cost. Moreover, the new defined *granulometric characteristic curves* and homogeneity measures could be jointly used for a better characterisation of each connected component. *c*) Even if in this dissertation, we focused on the support vector machine classifier, different classifiers could be investigated as well. In particular, strategies based on multiple classifiers, whose effectiveness is proven in terms of both stability and performance with respect to the conventional classifiers, could be integrated in the proposed spectral-spatial classification to improve the stability of the classification accuracy.

A
Data Sets Description

This Appendix gives a brief introduction to the hyperspectral data sets used in the experimental analysis and to the sensors used for their acquisition. The descriptions of the classes of interest, the training and the test sets, used for the accuracy assessment, are also provided.

A.1 Salinas Valley, California (Salinas)

Salinas[1] data set has been acquired over Salinas Valley, California, in 1998. The acquisition has been done by using the AVIRIS sensor (see Appendix A.6.1). The original data set is composed of 224 bands with a spectral range between 0.43 µm and 2.5 µm. The image has a size of 512 × 217 pixels with a spatial resolution of 3.7 m. In this study, the corrected data set is considered by discarding the 20 water absorption bands: [108-112], [154-167], 224. The ground reference data contains 16 classes of interest (described in Table A.1). A false color composition of the data set and the reference map are shown in Figures A.1a and A.1b, respectively.

[1] Available on-line through the Grupo de Inteligencia Computacional from the Basque University (EPV/EHU): `http://www.ehu.es/ccwintco/index.php?title=Home`.

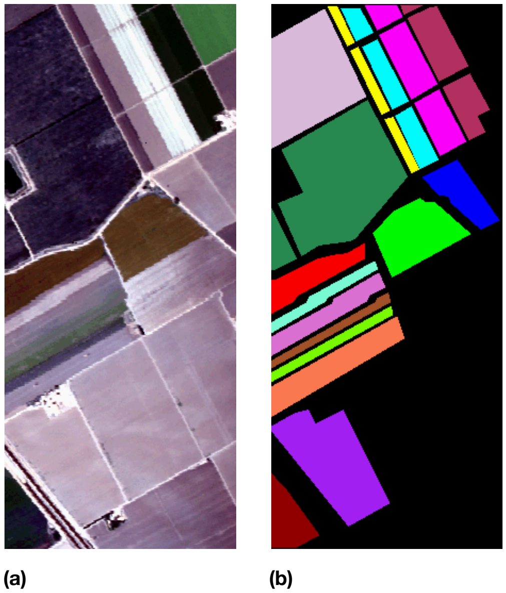


Figure A.1: Salinas data set description: (a) false color image; (b) reference map.

Table A.1: Classes and numbers of training / test samples for Salinas data set.

Salinas Data Set			
No.	Class	Training	Test
1	Broccoli green weeds 1	301	1708
2	Broccoli green weeds 2	558	3168
3	Fallow	296	1680
4	Fallow rough plow	209	1185
5	Fallow smooth	401	2277
6	Stubble	593	3366
7	Celery	536	3043
8	Grapes untrained	1690	9581
9	Soil vineyard develop	930	5273
10	Corn senesced green weeds	491	2787
11	Lettuce romaine 4 weeks	160	908
12	Lettuce romaine 5 weeks	289	1638
13	Lettuce romaine 6 weeks	137	779
14	Lettuce romaine 7 weeks	160	910
15	Vineyard untrained	1090	6178
16	Vineyard vertical trellis	271	1536

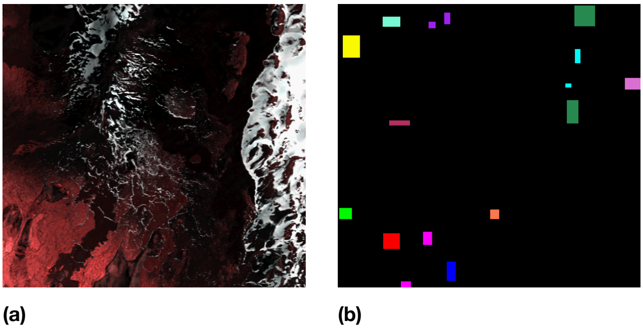


Figure A.2: Hekla data set description: (a) false color image; (b) reference map.

Table A.2: Classes and numbers of training / test samples for Hekla data set.

Hekla Data Set			
No.	Class	Training	Test
1	Andesite lava moss cover	50	973
2	Scoria	50	500
3	Hyperclatite formation	50	634
4	Andesite lava 1980 III	50	1446
5	Rhyolite	50	354
6	Andesite lava 1980 I	50	658
7	Andesite lava 1991 II	50	360
8	Andesite lava 1991 I	50	2689
9	Firn and glacier ice	50	408
10	Andesite lava 1970	50	292
11	Lava with Tephra and Scoria	50	650
12	Snow	50	663

A.2 Hekla volcano, Iceland (Hekla)

Hekla[2] data set was collected in June 17, 1991 on the active Hekla volcano, which is located in south-central Iceland, by the AVIRIS sensor. Due to the failure of the near-infrared spectrometer (spectrometer 4) during the data acquisition, 64 channels appeared blank. After discarding the noisy and the blank channels, the final data set included 157 spectral channels. The image has dimensions of 600 × 560 pixels with a geometric resolution of 20 m. It shows mainly lava flows from different eruptions and older hyaloclastites (formed

[2] Available from the University of Iceland upon request.

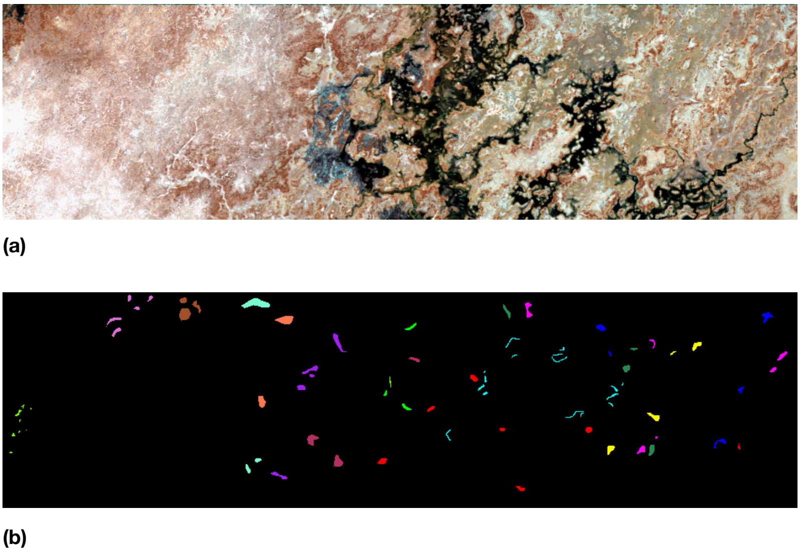


Figure A.3: Botswana data set description: (a) false color image; (b) reference map.

during subglacial eruptions). The ground reference data contains 12 classes of interests, which are described in Table A.2. Figures A.2a and A.2b show a false color composition of the image and the reference map, respectively.

A.3 Okavango Delta, Botswana (Botswana)

Botswana[3] data set was collected over the Okavango Delta, Botswana, in May 31, 2001 by the Hyperion sensor (see Appendix A.6.2). The acquired image is characterized by 242 bands covering the 0.4-2.5 µm portion of the spectrum with a spectral resolution of 10 nm. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10-55], [82-97], [102-119], [134-164], [187-220]. The image shows an area of 256 × 1476 pixels with a spatial resolution of 30 m. The ground reference data represent 14 land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta. The data set and the reference map are shown in Figures A.3a and A.3b, respectively, while Table A.3 provides information related to the classes. More information about the data set can be found in [50].

[3] Available on-line through the Center for Space Research at the University of Texas at Austin `http://www.csr.utexas.edu/hyperspectral/codes.html`.

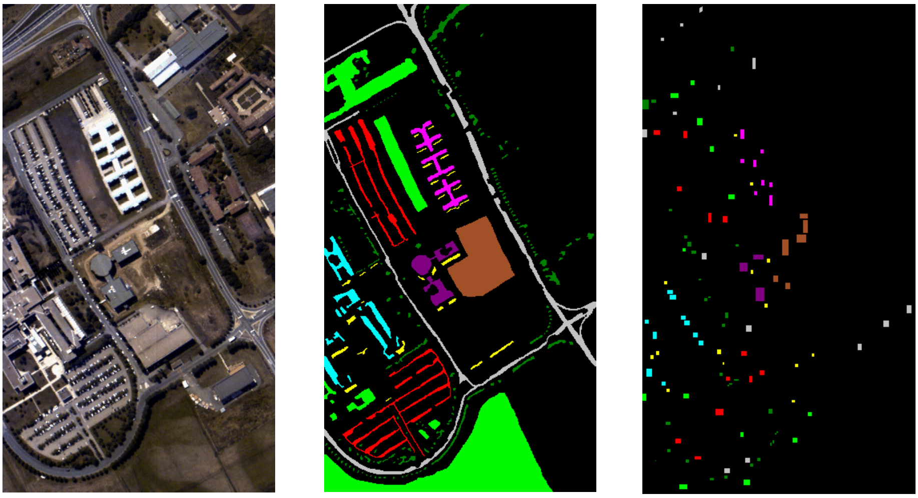
Table A.3: Classes and numbers of training / test samples for Botswana data set.

Botswana Data Set			
No.	Class	Training	Test
1	Water	54	216
2	Hippo grass	20	81
3	Floodplain grasses1	50	201
4	Floodplain grasses2	43	172
5	Reeds1	53	216
6	Riparian	53	216
7	Firescar2	51	208
8	Island interior	40	163
9	Acacia woodlands	62	252
10	Acacia shrubland	49	199
11	Acacia grasslands	61	244
12	Short mopane	36	145
13	Mixed mopane	53	215
14	Exposed soil	19	76

A.4 Pavia, university area, Italy (Pavia University)

Pavia University[4] data set was acquired by the optical airborne sensor ROSIS-03 (see Appendix A.6.3) over the university area of the city of Pavia, Italy. The image is composed by 103 spectral channels with a spectral range between 0.43 µm and 0.86 µm. The image shows an area of 610 × 340 pixels with a spatial resolution of 1.3 m per pixel. In the data set, nine classes of interest are considered, namely: Asphalt, meadow, gravel, trees, metal sheets,bare soil, bitumen, self-blocking bricks and shadows. The data set and the reference map are shown in Figure A.4a and A.4b, respectively, while the class information are reported in Table A.4.

[4] Provided by Prof. Paolo Gamba from the Telecommunications and Remote Sensing Laboratory, University of Pavia.



(a) (b) (c)

Figure A.4: Pavia University data set description: (a) true color image; (b) reference map; (c) training samples.

Table A.4: Classes and numbers of training / test samples for Pavia University data sets.

Pavia University			
No.	Class	Training	Test
1	Asphalt	548	6631
2	Meadow	540	18646
3	Gravel	392	2099
4	Trees	524	3064
5	Metal sheets	265	1345
16	Bare soil	532	5029
7	Bitumen	375	1330
8	Self-blocking bricks	514	3682
9	Shadows	231	947

A.5 Pavia, central area, Italy (Pavia Center)

Pavia Center[5], as for the previous data set, was acquired by the ROSIS-3 sensor during a flight campaign over Pavia. In this case, the data set is composed by 102 spectral bands, with a scene of 1096 × 715 pixels. Nine classes of interest

[5] Provided by Prof. Paolo Gamba from the Telecommunications and Remote Sensing Laboratory, University of Pavia.

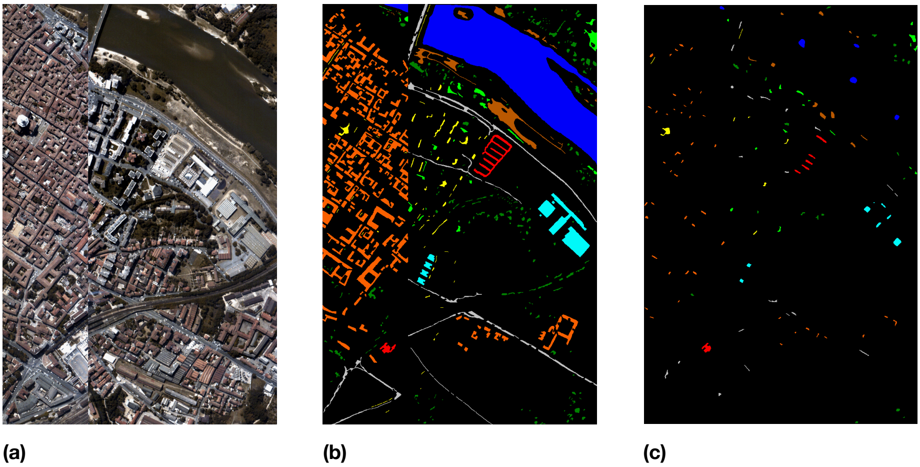

(a)
(b)
(c)

Figure A.5: Pavia Center data set description: (a) true color image; (b) reference map; (c) training samples.

Table A.5: Classes and numbers of training / test samples for Pavia Center data sets.

Pavia Center			
No.	Class	Training	Test
1	Water	824	65147
2	Trees	820	6778
3	Meadow	824	2266
4	Self-blocking bricks	808	1891
5	Bare soil	820	5764
6	Asphalt	816	8432
7	Bitumen	808	6479
8	Tiles	1260	41566
9	Shadows	476	2387

are considered, namely: Water, trees, meadow, self-blocking bricks, bare soil, asphalt, bitumen, tiles and shadows. the data set and the related reference map are shown in Figure A.5a and A.5b, respectively, while the class information is reported in Table A.5.

A.6 Hyperspectral Sensors

A.6.1 AVIRIS

The AVIRIS (Airborne Visible/Infrared Imaging Spectrometer)[6] [48] was developed by NASA Jet Propulsion Laboratory (JPL), providing images since 1987. The sensor system, composed of four spectrometers, measures a portion of the solar reflectance spectrum that covers the wavelength span from 0.4 µm to 2.5 µm, through 224 contiguous spectral channels with 10 nm intervals. The image acquisition is performed from areal platforms that can fly at different altitudes, which determine the pixel size and swath width of the acquired image.

A.6.2 Hyperion

Hyperion[7] is a one of the three primary instruments on board EO-1 NASA's spacecraft, which has been launched in 2000. The sensor acquires images composed by 220 spectral channels, covering a wavelength span from 0.4 µm to 2.5 µm with a spectral resolution of 10 nm. The maximum acquired area per image is of 7.5 km by 100 km with a spatial resolution of 30 m.

A.6.3 ROSIS-03

The ROSIS-03 (Reflective Optics System Imaging Spectrometer)[8] was developed jointly by Daimler-Chrysler Aerospace AG, the GKSS Research Centre and the German Aerospace Center, DLR. The sensor has been lunched for the first time in 1992 and operates from areal platforms. The sensor covers a portion of the reflectance spectrum from 0.43 µm to 0.86 µm, providing an image cube with 115 spectral channels with a spectral interval of 4 nm. The spatial resolution varies with the flying altitude. With an altitude of 3 km the pixel size is 1.7×1.7 m^2.

[6] http://aviris.jpl.nasa.gov/aviris

[7] http://eo1.usgs.gov/sensors/hyperion

[8] messtec.dlr.de/link-80-en

B
Accuracy Assessment

This Appendix provides the notions of overall accuracy and kappa coefficient, used in this dissertation for the accuracy assessment.

In this dissertation, the accuracy assessment is based on the analysis of the confusion matrix computed on the classification results. From the confusion matrix useful parameters that indicate how good the obtained classification is are derived. The parameters that are used are the overall accuracy and the kappa coefficient.

Overall Accuracy (OA) The overall accuracy (OA) represents the number (or percentage) of pixels that are correctly classified. Considering a total of C classes, the OA is mathematically defined as the ratio between the total number of corrected pixels for each class, N_i (which are summed along the major diagonal), divided by the total number of referenced pixels that are being tested, T:

$$OA = \frac{\sum_i^C N_i}{T}. \tag{B.1}$$

Kappa Coefficient (κ) The kappa coefficient (or kappa statistic) provides a measure of overall classification quality by comparing the agreement against the one expected by chance. It is mathematically defined as follows:

$$k = \frac{m_o - m_c}{1 - m_c}, \tag{B.2}$$

where m_o represents the proportion of correct agreement in the test set, and m_c is the proportion of agreement that is expected by chance. The possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement).

Bibliography

[1] H. G. Akcay and S. Aksoy. Automatic Detection of Geospatial Objects Using Multiple Hierarchical Segmentations. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(7):2097–2111, 2008. ISSN 01962892. doi: 10.1109/TGRS.2008.916644.

[2] S. Amari, A. Cichocki, and H. H. Yang. A New Learning Algorithm for Blind Signal Separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763, 1996.

[3] F. R. Bach and M. I. Jordan. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3:1–48, 2002. ISSN 0003-6951. doi: 10.1162/153244303768966085.

[4] J. E. Baker. Reducing bias and inefficiency in the selection algorithm. In *Proceedings of the second international conference on genetic algorithms*, pages 14–21, 1987.

[5] A. J. Bell and T. J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, Nov. 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.6.1129.

[6] J. A. Benediktsson, M. Pesaresi, and K. Arnason. Classification and Feature Extraction for Remote Sensing Images From Urban Areas Based on Morphological Transformations. *Geoscience and Remote Sensing, IEEE Transactions on*, 41(9):1940–1949, Sept. 2003.

[7] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson. Classification of Hyperspectral Data From Urban Areas Based on Extended Morphological Profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(3):480–491, Mar. 2005.

[8] S. Bernabe, P. R. Marpu, A. Plaza, M. D. Mura, and J. A. Benediktsson. Spectral–Spatial Classification of Multispectral Images Using Kernel Feature Space Representation. *Geoscience and Remote Sensing Letters, IEEE*, 11(1):288–292, Jan. 2014. ISSN 1545-598X. doi: 10.1109/LGRS.2013.2256336.

[9] J. M. Bioucas-Dias and M. A. T. Figueiredo. Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, pages 1–4. IEEE, June 2010. ISBN 978-1-4244-8906-0. doi: 10.1109/WHISPERS.2010.5594963.

[10] J. S. Borges, J. M. Bioucas-Dias, and A. R. S. Marcal. Bayesian hyperspectral image segmentation with discriminative class learning. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(6):2151–2164, June 2011. ISSN 0196-2892. doi: 10.1109/TGRS.2010.2097268.

[11] E. J. Breen and R. Jones. Attribute Openings, Thinnings, and Granulometries. *Computer Vision and Image Understanding*, 64(3):377–389, Nov. 1996. ISSN 10773142. doi: 10.1006/cviu.1996.0066.

[12] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324.

[13] L. Bruzzone and B. Demir. A Review of Modern Approaches to Classification of Remote Sensing Data. In I. Manakos and M. Braun, editors, *Land Use and Land Cover Mapping in Europe*, volume 18 of *Remote Sensing and Digital Image Processing*, pages 127–143. Springer Netherlands, 2014. ISBN 978-94-007-7968-6. doi: 10.1007/978-94-007-7969-3.

[14] L. Bruzzone and C. Persello. A Novel Approach to the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images With Improved Generalization Capability. *Geoscience and Remote Sensing, IEEE Transactions on*, 47(9):3180–3191, Sept. 2009. ISSN 0196-2892. doi: 10.1109/TGRS.2009.2019636.

[15] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson. Advances in Hyperspectral Image Classification: Earth monitoring with statistical learning methods. *Signal Processing Magazine, IEEE*, 31(1):45–54, 2014.

[16] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *Radar and Signal Processing, IEE Proceedings F*, 140(6):362–370, 1993. ISSN 0956-375X. URL http://perso.telecom-paristech.fr/~cardoso/guidesepsou.html.

[17] V. Caselles and P. Monasse. *Geometric description of images as topographic maps*. Springer Berlin Heidelberg, 2010. ISBN 9783642046100. doi: 10.1007/978-3-642-04611-7_1.

[18] G. Cavallaro, M. D. Mura, J. A. Benediktsson, and L. Bruzzone. A comparison of self-dual attribute profiles based on different filter rules for classification. In *Geoscience and Remote Sensing Symposium, 2014. IGARSS 2014. IEEE International*, pages 1265–1268, Quebec City, QC, 2014. ISBN 978-1-4799-5775-0. doi: 10.1109/IGARSS.2014.6946663.

[19] G. Cavallaro, N. Falco, M. Dalla Mura, L. Bruzzone, and J. A. Benediktsson. Automatic Threshold Selection for Profiles of Attribute Filters Based on Granulometric Characteristic Functions. In *Proc. of 12th International Symposium on Mathematical Morphology (ISMM 2015)*, 2015. (accepted).

[20] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, Apr. 2011. ISSN 21576904. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[21] C.-I. Chang and Q. Du. Estimation of Number of Spectrally Distinct Signal Sources in Hyperspectral Imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(3):608–619, 2004.

[22] A. Cheriyadat and L. M. Bruce. Why principal component analysis is not an appropriate feature extraction method for hyperspectral data. In *Geoscience and Remote Sensing Symposium, 2003. IGARSS 2003. IEEE International*, pages 3420–3422, 2003. ISBN 0-7803-7929-2. doi: 10.1109/IGARSS.2003.1294808.

[23] A. Cichocki and S.-i. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, Inc., 2002. ISBN 9780471607915.

[24] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287 – 314, 1994. ISSN 0165-1684.

[25] N. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun. Comparison of blind source separation algorithms for fmri using a new matlab toolbox: GIFT. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages 401–404, 2005. ISBN 0780388747.

[26] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone. Morphological Attribute Profiles for the Analysis of Very High Resolution Images. *Geoscience and Remote Sensing, IEEE Transactions on*, 48 (10):3747–3762, 2010. ISSN 01962892. doi: 10.1109/TGRS.2010.2048116.

[27] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone. Extended profiles with morphological attribute filters for the analysis of hyperspectral data. *International Journal of Remote Sensing*, 31(22): 5975–5991, Dec. 2010. ISSN 0143-1161. doi: 10.1080/01431161.2010.512425.

[28] M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone. Self-dual attribute profiles for the analysis of remote sensing images. In G. K. Soille, Pierre and Pesaresi, Martino and Ouzounis, editor, *Mathematical Morphology and Its Applications to Image and Signal Processing*, pages 320–330. Springer Berlin Heidelberg, 2011.

[29] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone. Classification of Hyperspectral Images by Using Extended Morphological Attribute Profiles and Independent Component Analysis. *Geoscience and Remote Sensing Letters, IEEE*, 8(3):542–546, 2011. ISSN 1545-598X. doi: 10.1109/LGRS.2010.2091253.

[30] M. Dalponte, L. Bruzzone, L. Vescovo, and D. Gianelle. The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas. *Remote Sensing of Environment*, 113(11): 2345–2355, 2009. ISSN 00344257. doi: 10.1016/j.rse.2009.06.013.

[31] A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component

analysis, Mar. 2004. ISSN 0165-0270. URL http://sccn.ucsd.edu/eeglab/.

[32] Q. Du, I. Kopriva, and H. Szu. Independent-component analysis for hyperspectral remote sensing imagery classification. *Optical Engineering*, 45(1):017008–1 – 017008–13, Jan. 2006. ISSN 0091-3286. doi: 10.1117/1.2151172.

[33] N. Falco, J. A. Benediktsson, and L. Bruzzone. Extraction of spatial features in hyperspectral images based on the analysis of differential attribute profiles. In L. Bruzzone, editor, *Remote Sensing*, volume 8892, page 88920O. International Society for Optics and Photonics, Oct. 2013. doi: 10.1117/12.2029199.

[34] N. Falco, L. Bruzzone, and J. A. Benediktsson. A Comparative Study of Different ICA Algorithms for Hyperspectral Image Analysis. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2013 5th Workshop on*, Gainesville, Florida, 2013.

[35] N. Falco, M. Dalla Mura, F. Bovolo, J. A. Benediktsson, and L. Bruzzone. Change Detection in VHR Images Based on Morphological Attribute Profiles. *Geoscience and Remote Sensing Letters, IEEE*, 10(3):636–640, May 2013. ISSN 1545-598X. doi: 10.1109/LGRS.2012.2222340.

[36] N. Falco, J. A. Benediktsson, and L. Bruzzone. A Study on the Effectiveness of Different Independent Component Analysis Algorithms for Hyperspectral Image Classification. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7(6):2183–2199, June 2014. ISSN 1939-1404. doi: 10.1109/JSTARS.2014.2329792.

[37] N. Falco, L. Bruzzone, and J. A. Benediktsson. An ICA based approach to hyperspectral image feature reduction. In *Geoscience and Remote Sensing Symposium, 2014. IGARSS 2014. IEEE International*, pages 3470–3473. IEEE, July 2014. ISBN 978-1-4799-5775-0. doi: 10.1109/IGARSS.2014.6947229.

[38] N. Falco, J. A. Benediktsson, and L. Bruzzone. Spectral and Spatial Classification of Hyperspectral Images Based on ICA and Reduced Morphological Attribute Profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 2015. (accepted).

[39] M. Fauvel, J. Chanussot, and J. A. Benediktsson. KERNEL PRINCIPAL COMPONENT ANALYSIS FOR FEATURE REDUCTION IN HYPERSPECTRALE IMAGES ANALYSIS. In *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*, volume 1, pages 238–241, 2006. ISBN 1424404134.

[40] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton. Advances in Spectral–Spatial Classification of Hyperspectral Images. *Proceedings of the IEEE*, 101(3):652–675, Mar. 2013. ISSN 0018-9219. doi: 10.1109/JPROC.2012.2197589.

[41] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. ISSN 00034800. doi: 10.1111/j.1469-1809.1936.tb02137.x.

[42] G. Franchi and J. Angulo. Comparative Study on Morphological Principal Component Analysis of Hyperspectral Images. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2014 6th Workshop on*, pages 1–4, 2014.

[43] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Computer science and scientific computing. Academic Press, Inc., second edition, 1990. ISBN 0122698517. doi: 10.1144/GSL.SP.2003.213.01.01.

[44] K. Fukunaga and J. M. Mantock. Nonparametric discriminant analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-5 (6):671–678, 1983.

[45] P. Ghamisi, J. A. Benediktsson, and J. R. Sveinsson. Automatic Spectral–Spatial Classification Framework Based on Attribute Profiles and Supervised Feature Extraction. *Geoscience and Remote Sensing, IEEE Transactions on*, pages 1–12, 2013.

[46] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, 1989. ISBN 0201157675.

[47] A. A. Green, M. Berman, P. Switzer, and M. D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *Geoscience and Remote Sensing, IEEE Transactions on*, 26(1):65–74, 1988. ISSN 01962892. doi: 10.1109/36.3001.

[48] 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, M. R. Olah, and O. Williams. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, 65(3):227–248, 1998.

[49] B. Guo, R. I. Damper, S. R. Gunn, and J. D. B. Nelson. A fast separability-based feature selection method for high-dimensional remotely-sensed image classification. *Pattern Recognition*, 41(5):1653–1662, 2008.

[50] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(3):492–501, 2005.

[51] X. He and P. Niyogi. Locality preserving projections. *Neural Information Processing Systems*, 16:153, 2004.

[52] K. E. Hild, G. Alleva, S. Nagarajan, and S. Comani. Performance comparison of six independent components analysis algorithms for fetal signal extraction from real fMCG data. *Physics in Medicine and Biology*, 52 (2):449–62, Jan. 2007. ISSN 0031-9155. doi: 10.1088/0031-9155/52/2/010.

[53] M.-K. Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187, Feb. 1962. ISSN 0096-1000. doi: 10.1109/TIT.1962.1057692.

[54] G. F. Hughes. On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, 14(1):55–63, 1968. ISSN 00189448. doi: 10.1109/TIT.1968.1054102.

[55] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3): 626–34, Jan. 1999. ISSN 1045-9227. doi: 10.1109/72.761722. URL `http://research.ics.aalto.fi/ica/fastica/`.

[56] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., 2001. ISBN 047140540X.

[57] A. Jaiantilal. Classification and Regression by randomForest-matlab, 2009. URL `http://code.google.com/p/randomforest-matlab`.

[58] A. K. Jain and D. Zongker. Feature selection: evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997. ISSN 01628828. doi: 10.1109/34.574797.

[59] L. O. Jimenez and D. A. Landgrebe. Hyperspectral data analysis and supervised feature reduction via projection pursuit. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(6):2653–2667, 1999. ISSN 01962892. doi: 10.1109/36.803413.

[60] T. P. Jung, S. Makeig, C. Humphries, T. W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–78, Mar. 2000. ISSN 0048-5772.

[61] C. Jutten, S. Moussaoui, and F. Schmidt. How to Apply ICA on Actual Data ? Example of Mars Hyperspectral Image Analysis. In *Digital Signal Processing, 2007 15th International Conference on*, pages 3–12. IEEE, July 2007. ISBN 1-4244-0881-4. doi: 10.1109/ICDSP.2007.4288502.

[62] A. Kachenoura, L. Albera, L. Senhadji, and P. Comon. ICA: a potential tool for BCI systems. *IEEE Signal Processing Megazine*, pages 57–68, 2008. doi: 10.1109/MSP.2007.909530.

[63] B.-C. Kuo and D. A. Landgrebe. Nonparametric weighted feature extraction for classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(5):1096–1105, 2004. ISSN 01962892. doi: 10.1109/TGRS.2004.825578.

[64] C. Lee and D. A. Landgrebe. Feature extraction based on decision boundaries. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(4):388–400, 1993. ISSN 01628828. doi: 10.1109/34.206958.

[65] J. Lee, A. Woodyatt, and M. Berman. Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform. *Geoscience and Remote Sensing, IEEE Transactions on*, 28(3):295–304, May 1990. ISSN 01962892. doi: 10.1109/36.54356.

[66] J. Li, J. M. Bioucas-Dias, and A. Plaza. Spectral-Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields. *Geoscience and Remote Sensing,*

IEEE Transactions on, 50(3):809–823, 2012. ISSN 0196-2892. doi: 10.1109/TGRS.2011.2162649.

[67] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson. Generalized composite kernel framework for hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 51 (9):4816–4829, 2013. ISSN 01962892. doi: 10.1109/TGRS.2012.2230268.

[68] S. Lim, K. Sohn, and C. Lee. Principal component analysis for compression of hyperspectral images. In *Geoscience and Remote Sensing Symposium, 2001. IGARSS'01. IEEE 2001 International*, volume 1, pages 97–99, 2001. ISBN 0780370317.

[69] Z. Mahmood, G. Thoonen, and P. Scheunders. Automatic threshold selection for morphological attribute profiles. In *Geoscience and Remote Sensing Symposium, 2012. IGARSS 2012. IEEE International*, pages 4946–4949, 2012. ISBN 9781467311595.

[70] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski. Independent Component Analysis of Electroencephalographic Data. In *Advances in Neural Information Processing Systems*, volume 8, pages 145–151, 1996.

[71] T. Marill and D. M. Green. On the effectiveness of receptors in recognition systems. *Information Theory, IEEE Transactions on*, 9(1):11–17, 1963. ISSN 00189448. doi: 10.1109/TIT.1963.1057810.

[72] P. R. Marpu, M. Pedergnana, M. D. Mura, S. Peeters, J. A. Benediktsson, and L. Bruzzone. Classification of hyperspectral data using extended attribute profiles based on supervised and unsupervised feature extraction techniques. *International Journal of Image and Data Fusion*, 3(3): 269–298, Sept. 2012. ISSN 1947-9832. doi: 10.1080/19479832.2012.702687.

[73] P. R. Marpu, M. Pedergnana, M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone. Automatic Generation of Standard Deviation Attribute Profiles for Spectral–Spatial Classification of Remote Sensing Data. *Geoscience and Remote Sensing Letters, IEEE*, 10(2):293–297, Mar. 2013. ISSN 1545-598X. doi: 10.1109/LGRS.2012.2203784.

[74] V. Matic, W. Deburchgraeve, and S. Van Huffel. Comparison of ICA algorithms for ECG artifact removal from EEG signals. In *Proc. of the 4th Annual symposium of the IEEE-EMBS Benelux Chapter (IEEE-EMBS).*, pages 2–5, 2009.

[75] V. E. McZgee and W. T. Carleton. Piecewise Regression. *Journal of the American Statistical Association*, 65(331):1109–1124, 1970. doi: 10.1080/01621459.1970.10481147.

[76] M. Mitchell. *An introduction to genetic algorithms*. MIT Press, 1998.

[77] G. Moser, S. B. Serpico, and J. A. Benediktsson. Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE*, 101:631–651, 2013. ISSN 00189219. doi: 10.1109/JPROC.2012.2211551.

[78] L. Najman and H. Talbot. *Mathematical Morphology: From Theory to Applications*. ISTE-Wiley, June 2010. doi: 10.1002/9781118600788.

[79] J. M. P. Nascimento and J. M. B. Dias. Does independent component analysis play a role in unmixing hyperspectral data? *Geoscience and Remote Sensing, IEEE Transactions on*, 43(1):175–187, Jan. 2005. ISSN 0196-2892. doi: 10.1109/TGRS.2004.839806.

[80] A. A. Nielsen. Kernel maximum autocorrelation factor and minimum noise fraction transformations. *Image Processing, IEEE Transactions on*, 20(3):612–24, Mar. 2011. ISSN 1941-0042. doi: 10.1109/TIP.2010.2076296.

[81] J. A. Palmason, J. A. Benediktsson, J. R. Sveinsson, and J. Chanussot. Classification of hyperspectral data from urban areas using morpholgical preprocessing and independent component analysis. In *Geoscience and Remote Sensing Symposium, 2005. IGARSS 2005. IEEE International*, volume 1, pages 176–179. IEEE, 2005. ISBN 0-7803-9050-4. doi: 10.1109/IGARSS.2005.1526133.

[82] S. D. Parmar, H. K. Patel, and J. S. Sahambi. Separation performance of ICA algorithms on FECG and MECG signals contaminated by noise. *Applied Computing*, 3285:184–190, 2004.

[83] M. Pesaresi and J. A. Benediktsson. A new approach for the morphological segmentation of high-resolution satellite imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 39(2):309–320, 2001. ISSN 01962892. doi: 10.1109/36.905239.

[84] M. Pesaresi, G. K. Ouzounis, and L. Gueguen. A new compact representation of morphological profiles: report on first massive VHR image processing at the JRC. In S. S. Shen and P. E. Lewis, editors, *SPIE Defense, Security, and Sensing*, volume 8390, pages 839025–839025–6. International Society for Optics and Photonics, May 2012. doi: 10.1117/12.920291.

[85] S. Prasad and L. M. Bruce. Limitations of Principal Components Analysis for Hyperspectral Target Recognition. *Geoscience and Remote Sensing Letters, IEEE*, 5(4):625–629, 2008.

[86] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994. ISSN 01678655. doi: 10.1016/0167-8655(94)90127-9.

[87] J. A. Richards and X. Jia. *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag Berlin Heidelberg, Berlin Heidelberg, 4th edition, 2006. ISBN 3540251286. doi: 10.1007/978-3-642-30062-2.

[88] P. Salembier and J. Serra. Flat zones filtering, connected operators, and filters by reconstruction. *Image Processing, IEEE Transactions on*, 4(8): 1153–1160, 1995. ISSN 10577149. doi: 10.1109/83.403422.

[89] P. Salembier, A. Oliveras, and L. Garrido. Antiextensive connected operators for image and sequence processing. *Image Processing, IEEE Transactions on*, 7(4):555–570, 1998. ISSN 10577149. doi: 10.1109/83.663500.

[90] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *Artificial Neural Networks—ICANN'97*, pages 583–588. Springer, 1997.

[91] J. R. Schott. *Remote Sensing: The Image Chain Approach*. Oxford University Press, 2007. ISBN 0195178173.

[92] R. A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Academic Press, Inc., 2007. ISBN 0123694078.

[93] S. B. Serpico and L. Bruzzone. A new search algorithm for feature selection in hyperspectral remote sensing images. *Geoscience and Remote Sensing, IEEE Transactions on*, 39(7):1360–1367, 2001. ISSN 01962892. doi: 10.1109/36.934069.

[94] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, Inc., London, June 1982.

[95] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*, volume 47. Cambridge University Press, 2004. ISBN 0521813972. doi: 10.2277.

[96] P. Soille. *Morphological Image Analysis*. Springer Berlin Heidelberg, Berlin, Heidelberg, second edition, 2004. ISBN 978-3-642-07696-1. doi: 10.1007/978-3-662-05088-0.

[97] B. Song, M. Dalla Mura, P. Li, A. J. Plaza, J. M. Bioucas-Dias, J. A. Benediktsson, and J. Chanussot. Remotely Sensed Image Classification Using Sparse Representations of Morphological Attribute Profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(8): 5122–5136, Aug. 2014. ISSN 0196-2892. doi: 10.1109/TGRS.2013.2286953.

[98] A. Subasi and M. Ismail Gursoy. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, 37(12):8659–8666, Dec. 2010. ISSN 09574174. doi: 10.1016/j.eswa.2010.06.065.

[99] M. Sugiyama. Local Fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 905–912, New York, New York, USA, 2006. ACM Press. ISBN 1595933832. doi: 10.1145/1143844.1143958.

[100] Y. Tarabalka, J. Chanussot, J. Benediktsson, J. Angulo, and M. Fauvel. Segmentation and Classification of Hyperspectral Data using Watershed. *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, 3:III – 652–III – 655, July 2008. doi: 10.1109/IGARSS.2008.4779432.

[101] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(5):1267–1279, 2010.

[102] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski. Learning relevant image features with multiple-kernel classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(10):3780–3791, 2010. ISSN 01962892. doi: 10.1109/TGRS.2010.2049496.

[103] K. Vanderperren, M. De Vos, J. R. Ramautar, N. Novitskiy, M. Mennes, S. Assecondi, B. Vanrumste, P. Stiers, B. R. H. Van den Bergh, J. Wagemans, L. Lagae, S. Sunaert, and S. Van Huffel. Removal of BCG artifacts from EEG recordings inside the MR scanner: a comparison of methodological and validation-related aspects. *NeuroImage*, 50(3):920–34, Apr. 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.01.010.

[104] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *Biomedical Engineering, IEEE Transactions on*, 47(5):589–93, May 2000. ISSN 0018-9294. doi: 10.1109/10.841330.

[105] A. Villa, J. A. Benediktsson, J. Chanussot, and C. Jutten. Hyperspectral Image Classification With Independent Component Discriminant Analysis. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(6): 4865–4876, 2011. ISSN 01962892.

[106] J. Wang and C.-I. Chang. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. *Geoscience and Remote Sensing, IEEE Transactions on*, 44(6):1586–1600, 2006. ISSN 01962892. doi: 10.1109/TGRS.2005.863297.

[107] A. W. Whitney. A Direct Method of Nonparametric Measurement Selection. *Computers, IEEE Transactions on*, C-20(9):1100–1103, 1971. ISSN 00189340. doi: 10.1109/T-C.1971.223410.

[108] J. Xia, P. Du, X. He, and J. Chanussot. Hyperspectral Remote Sensing Image Classification Based on Rotation Forest. *Geoscience and Remote Sensing Letters, IEEE*, 11(1):239–243, 2014.

[109] V. Zarzoso and P. Comon. Robust Independent Component Analysis by Iterative Maximization of the Kurtosis Contrast With. *Neural Networks, IEEE Transactions on,* 21(2):248–261, 2010.