Spectral and Spatial Classification of Hyperspectral Data

Pedram Ghamisi

A Thesis Presented in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy in Electrical and Computer Engineering at the
University of Iceland
2015

Advisor:
Professor Jón Atli Benediktsson

Thesis Committee:
Professor Jón Atli Benediktsson
Professor Antonio J. Plaza
Professor Magnús Örn Úlfarsson

Opponents:
Professor Melba M. Crawford
Professor David Messinger

Háskóli Íslands / University of Iceland
Verkfræði- og náttúruvísindasvið/ School of Engineering and Natural Sciences
VR-II, Hjardarhaga 2-6, IS-107 Reykjavik, Iceland
Phone + 354 525 4000
verkognatt@hi.is
www.hi.is

ISBN 978-9935-9243-3-9

This thesis is dedicated with gratitude to my parents.

Pedram
31/11/2014

Abstract

Hyperspectral imaging systems have gained a great attention from researchers in the past few years. These systems use sensors, which acquire data mostly from the visible through the middle infrared wavelength ranges and can simultaneously capture hundreds of (narrow) spectral channels from the same area on the surface of the Earth. Thanks to the detailed spectral information provided by hyperspectral sensors, the possibility of accurately discriminating materials of interest with an increased classification accuracy is increased. Furthermore, with respect to advances in hyperspectral imaging systems, the spatial resolution of recently operated sensors is getting finer, which enables analysis of small spatial structures in images.

Without any doubt, classification (or mapping) can be considered as the backbone of most image interpretation in remote sensing. In general, supervised classification approaches classify input data by considering the spectral information (e.g., intensity value of each pixel for grayscale images or intensity vector for RGB or high-dimensional images) of the data to produce a classification map in order to discriminate different classes of interest, by using a set of representative samples for each class, referred to as training samples. This way, by using a combination of training followed by classification, maps are produced from imagery. However, most of the existing classification techniques have been developed for the analysis of multispectral images, and consequently, they are not usually efficient for the classification of hyperspectral images, which can provide a detailed spectral information. This brings up the question whether the currently available classification techniques will be able to handle high-dimensional data.

The main objective of this thesis is the development of efficient spectral-spatial classification approaches in terms of classification accuracies. Beside the importance of classification accuracies, another critical issue for the purpose of hyperspectral image classification is simplicity and speed of the applied approaches. Therefore, in this thesis, a special emphasis is given on proposing robust techniques in terms of classification accuracies as well as being fast. In order to increase the efficiency of the existing techniques and reduce the laborious task of user interaction, a further development of automatic techniques plays a key role in remote sensing data analysis. Such techniques can be used for handling real-time applications such as hazard monitoring and risk management. Three different strategies are considered in the thesis as described below.

In the first strategy, a spectral-spatial classification approach, which is automatic and provides good classification accuracies is proposed. This method is based on integrating a Support Vector Machine (SVM) with Hidden Markov Random Field (HMRF). SVM and HMRF are two powerful approaches for high-dimensional data classification and spatial information extraction, respectively.

In the second strategy, we propose to use adaptive neighborhood systems by considering different approaches based on image segmentation and attribute profiles. These techniques are considered in order to extract spatial information for the purpose of spectral-spatial classification. In order to extract spectral information, SVM and Random Forests are applied due to their good performance in handling high dimensional data with limited number of training samples.

Finally, due to the fact that hyperspectral remote sensors acquire a massive amount of data and obtain many measurements, not knowing which data are relevant for a given problem, the third strategy is using novel feature selection approaches in order to address the *curse of dimensionality* and reduce the redundancy of high dimensional data.

***Index Terms* —** Spectral-spatial classification, Hyperspectral data, hidden Markov random field segmentation, Particle Swarm Optimization (PSO)-based image segmentation, Darwinian PSO (DPSO)-based image segmentation, Fractional Order DPSO (FODPSO)-based image segmentation, Morphological Profile, Attribute Profile, Extended Attribute Profile, Extended Multi-Attribute Profile, Support Vector Machine, Random Forest, Thresholding based image segmentation, Feature selection, Feature Extraction, Principal Component Analysis, Kernel Principal Component Analysis, Independent Component Analysis, Discriminant Analysis Feature Extraction, Decision Boundary Feature Extraction, Nonparametric Weighted Feature Extraction, Genetic Algorithm (GA), Binary Fractional Order Darwinian Particle Swarm Optimization, PSO-based feature selection, FODPSO-based feature selection, GA-based feature selection, Hybridization of GA and PSO feature selection.

Útdráttur

Kerfi sem notuð eru til að taka myndir af gríðarlega hárri vídd hafa notið mikillar athygli rannsakenda á undanförnum árum. Þessi kerfi nota skynjara sem safna gögnum einkum frá sýnilegu yfir miðinnrauða bylgjulengdarsviðið og geta náð samtímis hundruð (þröngra) rófrása fyrir sama svæðið á yfirborði jarðar. Vegna þessara nákvæmu rófupplýsinga sem fást með svona skynjurum er aukinn möguleiki á því að greina á milli þeirra efna á jörðinni sem eru til athugunar hverju sinni, með aukinni flokkunarnákvæmni. Til viðbótar má nefna varðandi þróunina að greinihæfni nýlegra skynjara af gríðarlegri vídd er að verða meiri, sem býður upp á greiningu á litlum rúmfræðilegum hlutum í myndum sem þessir skynjarar gefa.

Það er enginn vafi á því að telja má flokkun (eða kortlagningu) sem hryggjarstykkið í túlkun fjarkönnunarmynda. Almennt eru leiðbeindar flokkunaraðferðir notaðar til að flokka inntaksgögn með því að vinna rófupplýsingar (þ.e. gildi sérhverrar myndeiningar í gráskalamyndum eða gildi vigurs fyrir RGB eða myndir af hárri vídd) gagnanna og búa til flokkunarkort til að aðgreina þá flokka sem eru til skoðunar. Þá eru notuð svokölluð þjálfunargögn eða -sýni sem eru fulltrúar hvers flokks fyrir sig. Á þennan hátt eru kort búin til úr myndum með því að þjálfa fyrst flokkara og beita flokkaranum svo. Hins vegar hafa flestar núverandi flokkunaraðferðir verið þróaðar fyrir greiningur á fjölrása myndum (myndir sem hafa færri en 20 víddir) og henta því ekki endilega fyrir flokkun mynda af gríðarlegri vídd (miklu fleiri víddir en 20) sem bjóða upp á mjög nákvæmar rófupplýsingar. Þetta gefur tilefni til þess að spurt sé hvort hægt sé að beita núverandi flokkunaraðferðum á gögn af mikilli vídd.

Meginverkefni þessarar ritgerðar er að þróa nákvæmar flokkunaraðferðir sem taka tillit til bæði róf- og rúmupplýsinga. Til viðbótar við flokkunarnákvæmni í flokkun mynda af gríðarlegri vídd þarf einnig að hafa í huga hversu einfaldar og hraðvirkar aðferðirnar eru. Af þeim sökum er sérstök áhersla lögð í þessari ritgerð á ónæmar aðferðir sem eru bæði nákvæmar og hraðvirkar. Til að gera fyrirliggjandi aðferðir öflugri og til að minnka hina oft tímafreku gagnvirkni við notendur, eru hér þróaðar sjálfvikrar aðferðir. Slíkar aðferðir má nota í rauntímavinnslu, t.a.m. við eftirlit með umhverfisvá og til að minnka það tjón

sem af getur hlotist. Þrjár meginaðferðir eru rannsakaðar í ritgerðinni.

Fyrst er þróuð sjálfvirk flokkunaraðferð fyrir flokkun á róf- og rúmgögnum. Þessi aðferð byggist á því að tengja saman stoðvigravél (e. Support Vector Machine, SVM) og hulin Markov slemibisvið (Hidden Markov Random Field, HMRF). SVM og HMRF eru tvær öflugar aðferðir sem nota má til flokkunar gagna af hárri vídd og til að draga fram rúmfræðilegar upplýsingar.

Önnur aðferð sem skoðuð er í ritgerðinni gengur út á að þróa kerfi sem notar aðhæft nágrenni með því að velta upp mismunandi leiðum sem byggja á myndbútun (e. image segmentation) og auðkennaprófílum (e. attribute profiles). Þessar aðferðir eru notaðar til að draga fram rúmfræðilegar upplýsingar fyrir róf-rúm flokkun. Til að draga fram rófupplýsingar eru SVM og slembiskógar notaðir vegna þess að báðar þær aðferðir hafa áður sýnt fram á notagildi sitt þegar gögn af hárri vídd eru flokkuð og lítill fjöldi þjálfunarsýna er fyrirliggjandi.

Að lokum má nefna að skynjarar með gríðarlega háa vídd safna bæði mjög miklum gögnum og mörgum mælingum, án þess að ljóst sé hvaða gögn skipta máli fyrir það vandamál sem leysa þarf hverju sinni. Af þessum sökum gengur þriðja aðferðin sem rædd er í ritgerðinni út á þróun nýrrar einkennavalsaðferðar (e. feature selection) sem leitast við að koma í veg fyrir víddarbölvun (e. curse of dimensionality) ásamt því að minnka umfremd í gögnunum.

List of Publications

0.1 Book

1. (Book) J. A. Benediktsson and **P. Ghamisi**, Spectral-Spatial Classification of Hyperspectral Remote Sensing Images, Artech House Publishers, INC, Boston, USA, *in press*.

0.2 Journal papers

1. **P. Ghamisi**, M. S. Couceiro, J. A. Benediktsson and N. M. F. Ferreira, "An Efficient Method for Segmentation of Images Based on Fractional Calculus and Natural Selection," *Expert Systems With Applications*, vol. 39, no. 16, pp. 12407-12417, Nov. 2012.

2. **P. Ghamisi**, M. S. Couceiro, F. M. L. Martins and J. A. Benediktsson, "Multilevel Image Segmentation Based on Fractional-Order Darwinian Particle Swarm Optimization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2382-2394, May 2014.

3. **P. Ghamisi**, M. S. Couceiro, M. Fauvel and J. A. Benediktsson, "Integration of Segmentation Techniques for Classification of Hyperspectral Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 342-346, Jan. 2014.

4. **P. Ghamisi**, J. A. Benediktsson and M. O. Ulfarsson, "Spectral-Spatial Classification of Hyperspectral Images Based on Hidden Markov Random Fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2565-2574, May 2014.

5. **P. Ghamisi**, J. A. Benediktsson and J. R. Sveinsson, "Automatic Spectral-Spatial Classification Framework Based on Attribute Profiles and Supervised Feature Extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 9, pp. 5771-5782, Dec. 2014.

6. **P. Ghamisi**, J. A. Benediktsson, G. Cavallaro and A. Plaza, "Automatic Framework for Spectral–Spatial Classification Based on Supervised Feature Extraction and Morphological Attribute Profiles," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2147 - 2160, Jun. 2014.

7. **P. Ghamisi** and J. A. Benediktsson, "Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization," *IEEE Geoscience and Remote Sensing Letter*, vol. 12, no. 2, pp. 309-313, Jul. 2015.

8. **P. Ghamisi**, M. Dalla Mura and J. A. Benediktsson, "A Survey on Spectral–Spatial Classification Techniques Based on Attribute Profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2335-2353, May 2015.

9. **P. Ghamisi**, M. S. Couceiro and J. A. Benediktsson, "A Novel Feature Selection Approach Based on FODPSO and SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2935-2947, May 2015.

10. **P. Ghamisi**, A. ALi, M. S. Couceiro and J. A. Benediktsson, "A Novel Evolutionary Swarm Fuzzy Clustering Approach for Hyperspectral Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, accepted.*

11. **P. Ghamisi**, J. A. Benediktsson and S. Phinn, "Land-Cover Classification by using both Hyperspectral and LiDAR Data," *International Journal of Image and Data fusion, submitted.*

12. **P. Ghamisi**, G. Cavallaro, Dan Wu, J. A. Benediktsson and A. Plaza, "Fusion of LiDAR and Hyperspectral Data for the Classification of Urban Areas: Case Study in Queensland, Australia," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, submitted.*

0.3 Conference papers

1. **P. Ghamisi**, M. S. Couceiro and J. A. Benediktsson, "Extending the Fractional Order Darwinian Particle Swarm Optimization to Segmentation of Hyperspectral Images," in *Proc. SPIE, Image and Signal Processing for Remote Sensing XVIII*, 2012, pp. 85370F-85370F-11.

2. **P. Ghamisi**, M. S. Couceiro, M. Fauvel, J. A. Benediktsson, "Spectral-Spatial Classification Based on Integrated Segmentation," in *Proc. IEEE IGARSS, 2012*, pp. 1458-1461, 2013.

3. **P. Ghamisi**, Jon Atli Benediktsson, Magnus O. Ulfarsson, "The Spectral Spatial Classification of Hyperspectral Images Based on Hidden Markov Random Field and its Expectation-Maximization," in *Proc. IEEE IGARSS, 2012*, pp. 1107-1110, (**The winner of the IEEE Mikio Takagi student prize 2013 for winning the Student Paper Competition at the conference between almost 70 people**).

4. **P. Ghamisi**, M. S. Couceiro, and J. A. Benediktsson, "Classification of Hyperspectral Images with Binary Fractional Order Darwinian PSO and Random Forests," in *Proc. SPIE, Image and Signal Processing for Remote Sensing XIX*, 2013, pp. 88920S–88920S-8.

5. **P. Ghamisi**, M. S. Couceiro and J. A. Benediktsson, "FODSPO Based Feature Selection for Hyperspectral Remote Sensing Data," *WHISPERS 2014, in press*.

6. **P. Ghamisi**, J. A. Benediktsson, S. Phinn, "Fusion of Hyperspectral and LiDAR Data in Classification of Urban Areas," in *Proc. IEEE IGARSS, 2014*, pp. 181-184, **(Invited paper)**.

7. **P. Ghamisi** and J. A. Benediktsson, "Feature Selection of Hyperspectral Data by Considering the Integration of Genetic Algorithms and Particle Swarm Optimization," in *Proc. SPIE, Image and Signal Processing for Remote Sensing XX*, 2014, pp. 92440J-92440J-6.

8. **P. Ghamisi**, D. Wu, G. Cavallaro, J. A. Benediktsson and S. Phinn, "Am Advanced Classifier for the Joint Use of LiDAR and Hyperspectral Data: A Case Study of Queensland, Australia", in *Proc. IEEE IGARSS, 2015, accepted*.

Acknowledgements

The story of life is quicker than a wink of an eye. I remember the day as if it were yesterday that I commenced my PhD at the University of Iceland. At those days, I put my greatest effort on finishing my study, but after two years or so, I figured out there is no end to education. It is not that you read a book, pass an examination, and finish with education. The whole life, from the moment you are born to the moment you die, is a process of learning.

I would never have been able to get through my dissertation, more than 10 journal papers, a book, a number of proposals as well as the best student paper award without the guidance of my committee members, help from friends, and support from my family and wife. Too often we underestimate the power of a touch, a smile, a kind word, a listening ear, an honest compliment or the smallest act of caring, all of which have the potential to turn a life around. Therefore, this page of my thesis is gratefully devoted to the ones whose supports, kind wishes and positive energies pushed me forward.

Foremost, I would like to express my deepest gratitude to my advisor, Prof. Jon Atli Benediktsson, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I appreciate his vast knowledge and skill in many areas (e.g., pattern recognition, remote sensing, machine learning and so on), and his assistance in writing reports (i.e., papers, scholarship applications and this thesis), which have on occasion made me "GREEN" with envy.

Besides my advisor, I would like to express my sincere appreciation to the rest of my thesis committee: Prof. Antonio Plaza and Prof. Magnus O. Ulfarsson, for their encouragement and insightful comments.

Furthermore, my greatest appreciation goes to my parents for giving birth to me at the first place and supporting me spiritually throughout my life. From the bottom of my heart, I understood that no matter how bad things get, no matter how wrong things go, parents will always be there. I cannot never appreciate them enough for all sacrifices they went through for me.

My sincere thanks also goes to Peyman, Parisa, Mersedeh and Amir. I

would thank you from the bottom of my heart for all the supports, kind wishes and positive energies that you kindly provided me with. I never be able to think ahead and picture my life without you.

A great big, heartfelt thank of mine goes to my nearest and dearest, Shaghayegh whose love, support and smile helped me to get through all difficulties that I have faced during my PhD. No matter there is a long distance in between, being deeply loved by someone gives you strength, while loving someone deeply gives you courage.

I would like to gratefully thank to my friend and colleague, Dr. Behnood Rasti whose friendship gave me strength to overcome the difficulties in my work and life.

My thanks and appreciations also go to my colleagues in developing this work toward my PhD including: Dr. Micael S. Couceiro, Dr. Mathieu Fauvel and Dr. Mauro Dalla Mura as well as my friends in Signal Processing Lab at the University of Iceland including: Jakob, Frosti, Eysteinn, Nicola, Gabriele, Xudong and people who have willingly helped me out with their abilities.

I would like to thank my friends, Abtin, Babak, Farshid, Amin, Amirhossein and others. They were always there cheering me up and stood by me through the good times and bad. So long as the memory of certain beloved friends lives in my heart, I shall say that life is a legend.

Last but not least, thank you Iceland for providing me with this brilliant opportunity and making me a better person than I used to be. I am also very thankful to Icelandic Research Fund for Graduate Students, which financially supported my research.

As a few friendly words with readers, I would say: I hope your dreams take you to the corners of your smiles, to the highest of your hopes, to the windows of your opportunities, and to the most special places your heart has ever known. If you can imagine it, you can achieve it; if you can dream it, you can become it. Don't be frightened of making mistakes. Because if you are making mistakes, then you are making new things, trying new things, learning, living, pushing yourself, changing yourself, changing your world. You're doing things you have never done before, and more im-

portantly, you are doing something. So that's my wish for you, and all of us, and my wish for myself. Make new mistakes. Make glorious, amazing mistakes. Make mistakes nobody's ever made before. Don't freeze, don't stop, don't worry that it isn't good enough, or it isn't perfect, whatever it is: art, or love, or work or family or life. Whatever it is you're scared of doing, do it. Make your mistakes, today and forever. Don't forget that everything will be fine in the end. If it is not fine it is not the end.

Pedram
31/11/2014

Contents

List of Figures

List of Tables

Abbreviations

VHR	Very High Resolution
MRF	Markov Random Field
MAP	Maximum a Posteriori
HMM	Hidden Markov Model
HMRF	Hidden Markov Random Field
SE	Structuring Element
MP	Morphological Profile
EMP	Extended Morphological Profile
DMP	Differential Morphological Profile
AF	Attribute Filter
AP	Attribute Profile
EAP	Extended Attribute Profile
EMAP	Extended Multi-Attribute Profile
EEMAP	Entire Extended Multi-Attribute Profile
SDAP	Self-Dual Attribute Profile
SVM	Support Vector Machine
RF	Random Forest
RBF	Radial Basis Function
Hseg	Hierarchical Segmentation
PC	Principal Component
PCA	Principal Component Analysis
DAFE	Discriminant Analysis Feature Extraction
DBFE	Decision Boundary Feature Extraction

NWFE	Nonparametric Weighted Feature Extraction
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
DPSO	Darwinian Particle Swarm Optimization
FODPSO	Fractional Order Darwinian Particle Swarm Optimization
BFODPSO	Binary Fractional Order Darwinian Particle Swarm Optimization
HGAPSO	Hybridization of Genetic Algorithms and Particle Swarm Optimization
GLCM	Gray-Level Co-occurrence Matrix

CHAPTER 1

Introduction

1.1 Introduction to Hyperspectral Imaging Systems

In the past decade, hyperspectral imaging systems have gained a great attention from researchers. Hyperspectral imaging systems use sensors that mostly operate from the visible through the middle infrared wavelength ranges and can simultaneously capture hundreds of (narrow) spectral channels from the same area on the surface of the Earth. The hyperspectral sensors collect data with pixels that are represented by vectors in which each element is a measurement corresponding to a specific wavelength. The size of each vector is equal to the number of spectral data channels that are collected by the sensor. For hyperspectral images, several hundreds of spectral data channels of the same scene are typically available, while for multispectral images up to ten data channels are usually provided. The detailed spectral information provided by hyperspectral sensors increases the possibility of accurately discriminating materials of interest with an increased classification accuracy. In addition, thanks to advances in hyperspectral technology, the fine spatial resolution of recently operated sensors enables analysis of small spatial structures in images. Several operational imaging systems are currently available providing a large amount of images for various thematic applications, such as:

- *Ecological science*: Hyperspectral images are used to estimate biomass and carbon, biodiversity in dense forest zones and can be used to study land cover changes.
- *Geological science*: It is possible to recover physico-chemical mineral properties such as composition and abundance.
- *Mineralogy*: By using hyperspectral data, not only a wide range of min-

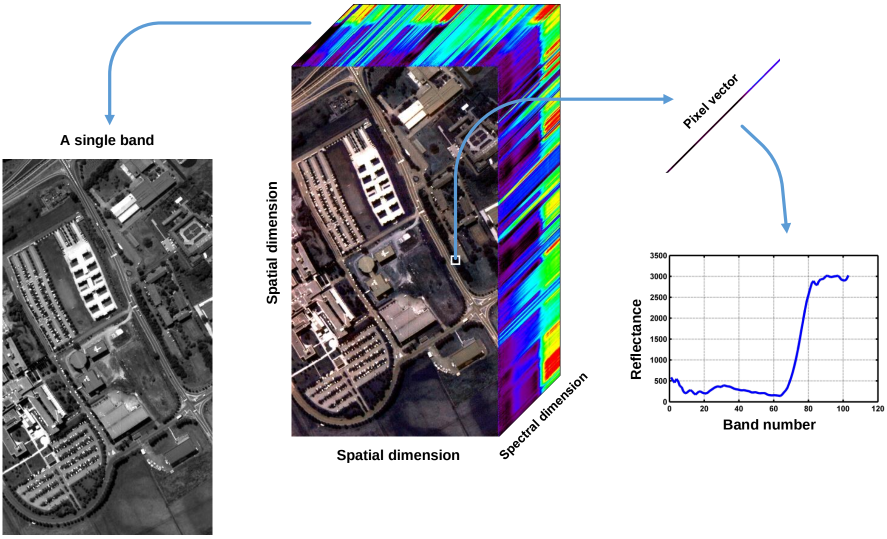


Figure 1.1: An example of a hyperspectral data cube.

erals can be identified but also their relation to the presence of valuable minerals can be understood. Currently, researchers are investigating the effect of oil and gas leakages from pipelines and natural wells on the spectral signatures of vegetation.

- *Hydrological science*: Hyperspectral imagery is taken into account to determine changes in wetland characteristics. Moreover, water quality, estuarine environments and coastal zones can be investigated by using hyperspectral images as well.

- *Precision agriculture*: Hyperspectral data are considered as a powerful tool in order to classify agricultural classes and to extract nitrogen content for the purpose of precision agriculture.

- *Military applications*: The rich spectral-spatial information of hyperspectral data can be also used for target detection. The intrinsic properties of hyperspectral images need to be addressed specifically because conventional algorithms made for multispectral images do not adapt well to the analysis of hyperspectral images.

Hyperspectral images can be considered as a stack of images with different wavelength interval (spectral channels) from the same scene on the surface of the earth. Based on this interpretation, hyperspectral images can be referred to *hyperspectral data cubes*. In other words, each spectral channel represents a gray scale image and all images make a three dimensional hyperspectral cube. Figure 1.1 shows an example of *hyperspectral data cube*. A three dimensional *hyperspectral data cube* consists of $n_1 \times n_2 \times d$ pixels which $n_1 \times n_2$ is the number of pixels in each spectral channel and d represents the number of spectral channels. In greater detail, a hyperspectral image can be introduced from one of the following perspectives:

1. *Spectral perspective (or spectral dimension)*: In this case, a hyperspectral data cube consists of several pixels and each pixel is a vector of d values. Each pixel corresponds to the reflected radiation of the specific region of the earth and has multiple values in spectral bands. This detailed spectral information can be used in order to analyze different materials, precisely. The right image of Figure 1.1 shows a histogram of the one pixel with multiple values for each band in spectral dimension. In this domain, the following points are of importance:

   - In general, vectors of different pixels belonging to a similar material have almost the same values. Different supervised and unsupervised classification techniques are used in order to group the vectors with almost the same characteristic.
   - In general, in each vector, neighborhood pixels in different spectral channels have a strong correlation. Different supervised and unsupervised feature reduction techniques are used in order to reduce the dimensionality of the hyperspectral data cube.

2. *Spatial perspective (or spatial dimension)*: In this case, a hyperspectral data cube consists of d gray scale images with a size of $n_1 \times n_2$. The values of all pixels in the one spectral band make a gray scale image with two dimensions which are spatial and spatial and is shown in Figure 1.1.

   - In the spatial dimension, adjacent pixels quite commonly belong to the same object (in particular for Very High Resolution (VHR) data). This dimension provides valuable information regarding the size and shape of different structures and objects on the earth. There are several ways to extract spatial information (e.g., segmentation) which will be discussed in detail later in this thesis.

The first attempts to analyze hyperspectral images were based on techniques that were developed for multispectral images which only have a few spectral channels, usually less than seven. However, most of the commonly used methods designed for the analysis of gray scale, color or multispectral images are inappropriate and even useless for hyperspectral images. As a matter of fact, with a limited number of available samples, the performance of supervised classification in terms of accuracies will dramatically be downgraded when the number of data channels increases. In addition, the *Hughes* *phenomenon/curse of dimensionality* [1] poses a problem for designing robust statistical estimations. As a result, based on the above characteristics of hyperspectral images, in order to make the most of the rich information provided by the hyperspectral data, the development of new algorithms is required. In the following section, we discuss the specific characteristics of hyperspectral data in more detail. This section is very important for the discussion that follows in the rest of the thesis. In addition, in the introduction part of this thesis, in order to provide some illustrative examples, *Pavia University* hyperspectral data set has been repeatedly used. This data set was captured on the city of Pavia, Italy by the ROSIS-03 (Reflective Optics Spectrographic Imaging System) airborne instrument. The flight over the city of Pavia, Italy, was operated by the Deutschen Zentrum für Luft- und Raumfahrt (DLR, the German Aerospace Agency) within the context of the HySens project, managed and sponsored by the European Union. The ROSIS-03 sensor has 115 data channels with a spectral coverage ranging from 0.43 to 0.86 μm. Twelve channels have been removed due to noise. The remaining 103 spectral channels are processed. The data have been corrected atmospherically, but not geometrically. The spatial resolution is 1.3 m per pixel. The data set covers the Engineering School at the University of Pavia and consists of different classes including: trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow and soil. This data set comprises 640×340 pixels. Figure 1.2 presents a false color image of ROSIS-03 Pavia University data and its corresponding reference samples. These samples are usually obtained by manual labeling of a small number of pixels in an image or based on some field measurements. Thus, the collection of these samples is expensive and time demanding. As a result, the number of available training samples is usually limited, which is a challenging issue in supervised classification.

1.2 High Dimensional Data

Figure 1.3 shows the basic idea of the pixel-wise pattern recognition approach which consists of feature extraction/selection and classification. In pattern recognition, each image pixel is considered as a pattern and its spectrum (a vector of different values of a pixel in different spectral channels) is considered as the initial set of features. Since this set of features is often redundant, feature reduction (feature extraction and/or selection) step is performed aiming at reducing the dimensionality of the feature set (from d_1 dimensions in the original data to d_2 dimensions in a new feature space $d_2 < d_1$) and maximizing separability between classes. The reason why we need to consider this step in hyperspectral data processing will be discussed in subsection 1.2.1. The next step (called classification) refers to partitioning the entire spectral domain into K exhaustive, none overlapping regions, so that every points in this domain is

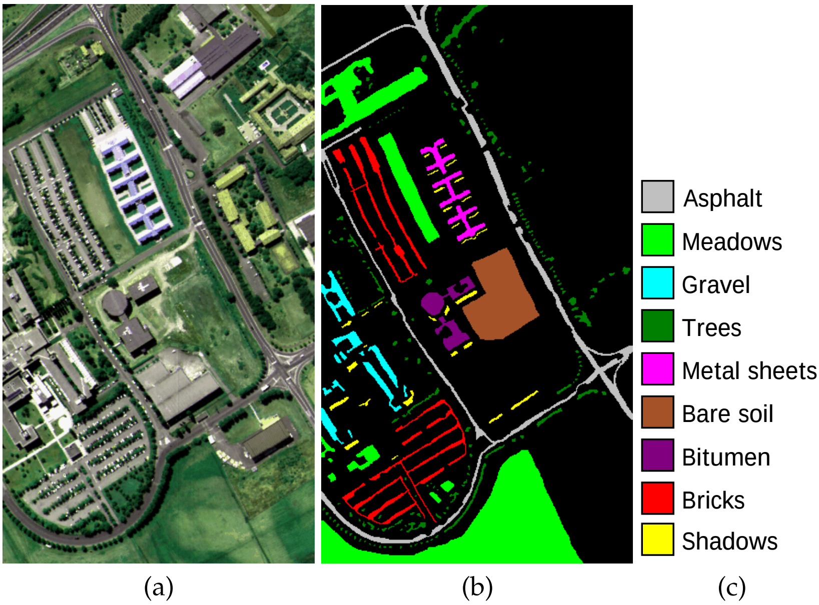


Figure 1.2: ROSIS-03 Pavia University hyperspectral data. (a) Three band false color composite, (b) Reference data and (c) Color code.

uniquely associated with one of the K classes. Once this step is accomplished, each pixel is classified according to its feature set. The output of this step is a one dimension image. The reason why we need to consider this step in hyperspectral data processing will be discussed in subsection 1.2.5.

In the following section, the geometrical and statistical characteristics of hyperspectral data along with the shortcomings of conventional techniques for analyzing of this sort of data have been investigated and possible solutions will be described for each shortcoming.

1.2.1 Geometrical and Statistical Properties of High Dimensional Data and the Need of Feature Reduction

At this point, we are in the era of massive automatic data collection, systematically obtaining many measurements, not knowing which data is appropriate for a problem in hand. The trend of hyperspectral imagery is to record hundreds of spectral channels from the same scene which can characterize chemical composition of different materials and is potentially helpful in analyzing different objects of interest. In the spectral domain, each spectral channel is considered as one dimension and each pixel is represented a point in this do-

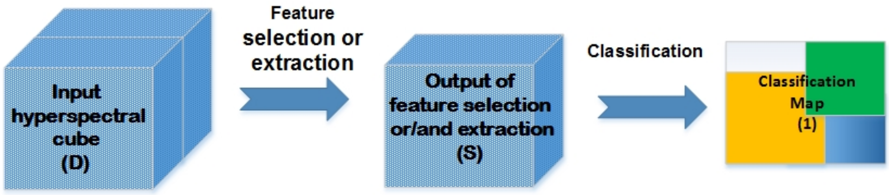

Input hyperspectral cube (D)
Feature selection or extraction
Output of feature selection or/and extraction (S)
Classification
Classification Map (1)

Figure 1.3: The basic chain pixel-wise pattern recognition approaches.

main. By increasing the spectral channels in the spectral domain, theoretical and practical problems may arise and conventional techniques which are applied on multispectral data are no longer appropriate for the processing of high dimensional data. The increased dimensionality of such data is able to improve data information content significantly, but provides a challenge to the conventional techniques for accurate analysis of hyperspectral data. Human experience in three-dimensional (3-D) space misleads one's intuition of geometrical and statistical properties in high-dimensional space [2]. In other words, it is difficult for humans get used to visualizing spaces with higher-dimension than three. Sometimes, this misunderstanding of high dimensional spaces and conventional spaces leads to the wrong choices in terms of data processing. As a result, the main objective of the following sub-sections is to give a brief description of the properties of high dimensional spaces.

A. 1. As dimensionality increases, the volume of a hypercube concentrates in corners [3]

The volume of the hypersphere with radius r and dimension d is computed by [1]

$$V_s(r) = \frac{2r^d}{d} \frac{\pi^{d/2}}{\Gamma(\frac{d}{2})} \tag{1.1}$$

and the volume of a hypercube in $[-r, r]^d$ is calculated by

$$V_c(r) = (2r)^d. \tag{1.2}$$

The fraction of the volume of a hypersphere inscribed in a hypercube is

[1] Reminder: Γ is the *gamma function*, which is an extension of the factorial function, with its argument shifted down by 1, to real and complex numbers. The main property of the gamma function is $x\Gamma(x) = \Gamma(x+1)$. As an example for the gamma function $\Gamma(\frac{5}{2}) = \frac{3}{2}\Gamma(\frac{3}{2}) = \frac{3}{2}\frac{1}{2}\Gamma(\frac{1}{2}) = \frac{3}{4}\sqrt{\pi}$

$$f_{d1} = \frac{V_s(r)}{V_c(r)} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(\frac{d}{2})}. \tag{1.3}$$

The above means that $lim_{d \longrightarrow \infty} f_{d1} = 0$, i.e., as d increases, the volume of the hypercube is increasingly concentrated in the corners.

A. 2. As dimensionality increases, the volume of a hypersphere concentrates in an outside shell [3, 4]

The fraction of the volume in a shell defined by a sphere of radius $r - \epsilon$ inscribed inside a sphere with radius r is

$$f_{d2} = \frac{V_s(r) - V_s(r-\epsilon)}{V_s(r)} = \frac{r^d - (r-\epsilon)^d}{r^d} = 1 - (1 - \frac{\epsilon}{r})^d. \tag{1.4}$$

The above means that $lim_{d \longrightarrow \infty} f_{d2} = 1$, here the volume of a hypersphere is mostly concentrated in an outside shell. In the same way, *it can be proven that the volume of a hyperellipsoid concentrates in an outside shell* [5].

Based on the above-mentioned properties, two important specifications for high-dimensional data can be concluded:

- A high-dimensional space is almost empty, which implies that multivariate data in $\rm I\!R$ is usually in a lower dimensional structure. As a result, high-dimensional data can be projected into a lower subspace without losing considerable information in sense of separability among the different statistical classes.

- Gaussian distributed data have a tendency to concentrate in the tails. In the same way, uniformly distributed data have a tendency to be concentrated in the corners, which makes the density estimation more difficult. In this space, local neighborhoods are almost surely empty which demands the larger band-width of estimation and produces the effect of losing detailed density estimation. For more information and the proof of the claim, please see [5].

A. 3. As dimensionality increases, the diagonals are nearly orthogonal to all coordinate axes [3]

The cosine of the angle between any diagonal vector and a Euclidean coordinate axis is given by:

$$cos(\theta_d) = \pm \frac{1}{\sqrt{d}} \tag{1.5}$$

Here $lim_{d \longrightarrow \infty} cos(\theta_d) = 0$, which implies that the diagonal is more likely to become orthogonal to the Euclidean coordinates in high-dimensional space.

A. 4. The required number of labeled samples for supervised classification increases as the dimensionality increases

As will be discussed later, supervised classification methods classify input data by using a set of representative samples for each class, referred to as *training samples*. Training samples are usually obtained by the manual labeling of a small number of pixels in an image or based on some field measurements.

Fukunaga [6] showed that there is a relation between the required number of training samples and the number of dimensions for different types of classifiers. The required number of training samples is linearly related to the dimensionality for linear classifiers and to the square of the dimensionality for quadratic classifiers. For nonparametric classifiers, it has been shown that the number of required samples exponentially increases as the dimensionality increases.

It is expected that by increasing the dimensionality of data, more information is required in order to detect more classes with more accuracy. At the same time, the aforementioned characteristics show that conventional techniques, which are based on the computation in full dimensional space, may not provide accurate classification results when the number of training samples is not substantial. For instance, while keeping the number of samples constant, after a few features, the classification accuracy actually decreases as the number of features increases [2]. For the purpose of classification, these problems are related to the curse of dimensionality. In [2], Landgrebe shows that too many spectral bands are undesirable from the standpoint of expected classification accuracy. When the number of spectral channels (dimensionality) increases, with a constant number of samples, a higher dimensional set of statistics must be estimated. In other words, although higher spectral dimensions increase the separability of the classes, the accuracy of the statistical estimation decreases.

A. 5. For most high-dimensional data sets, low linear projections have the tendency to be normal (Gaussian), or a combination of normal distributions, as the dimensionality Increases

It has been shown in [7, 8], as the dimensionality tends to infinity, lower dimensional linear projections will approach a normality model with probability approaching one. In this case, normality is regarded as a normal distribution or a combination of normal distributions [5].

Due to the above-mentioned characteristics of high-dimensional spaces, one can easily figure out that the high-dimensional space is completely different from 3-D space. These particular behaviors of high-dimensional data have a significant effect in the context of supervised classification techniques. In order to estimate class parameters, a large number of training samples is needed (which is almost impossible) in order to make a precise estimation. This problem is more severe when dimensionality increases. In a nonparametric approach, in order to the satisfactory estimation of a class density, the number of required training samples is even greater.

It is obvious that a high-dimensional space is almost empty and multivariate data can be represented in a lower dimensional space. Consequently, it is possible to reduce the dimensionality of high-dimensional data without sacrificing significant information and class separability. Based on the difficulties of density estimation in nonparametric approaches, parametric data-analysis techniques may lead to a better performance, where only a limited number of training samples is available to provide the required *a priori* information. As a result, it is desirable to project the high-dimensional data into lower dimensional subspace, where the undesirable effects of high-dimensional geometric characteristics and the so-called curse of dimensionality are decreased.

Each spectral channel characterizes as one dimension in the spectral domain. By increasing the features in the spectral domain, theoretical and practical problems may arise. For instance, while keeping the number of training samples constant, the classification accuracy actually decreases when the number of features becomes large [1]. For the purpose of classification, these problems are related to the curse of dimensionality. In [2], Landgrebe showed that too many spectral bands can be undesirable from the standpoint of expected classification accuracy because the accuracy of the statistical estimation decreases (Hughes phenomenon). The aforementioned issue demonstrates that there is an optimal number of bands for classification accuracy and more features do not necessarily lead to better results. Therefore, use of feature reduction techniques may lead to a better classification accuracy. Figure 1.4 demonstrates that with a limited number of training samples, as the number of features increases, the class separability increases but the accuracy of the statistical estimation decreases. Therefore, by keeping the number of samples constant, after a few features, the classification accuracy actually decreases as the number of features increases. In general, *feature reduction* techniques can be divided into *feature selection* and *feature extraction* techniques.

1.2.2 Feature extraction

Feature extraction can be explained as finding a set of vectors that represents an observation while reducing the dimensionality. Feature extraction is the process of producing a small number of features by combining existing bands. In this line of thought, feature extraction techniques transform the input data linearly or nonlinearly to another domain and extract informative features in the new domain. Feature extraction can be split into two categories; unsupervised and supervised feature extraction where the former is used for the purpose

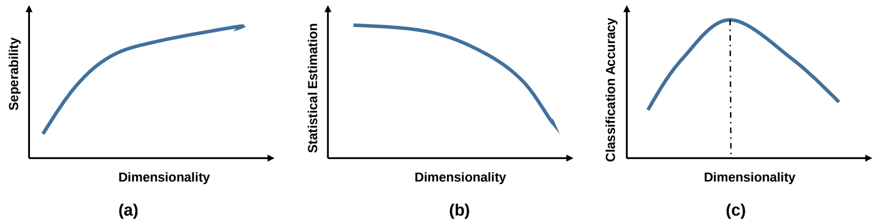


Figure 1.4: Y-axis demonstrates the amount of separability, statistical estimation and classification accuracy in (a), (b) and (c), respectively. X-axis demonstrates dimensionality. With a limited number of training samples, as the number of features increases, the class separability increases but the accuracy of the statistical estimation decreases. In this case, while keeping the number of samples constant, after a few features, the classification accuracy actually decreases as the number of features increases.

of data representation and latter is considered for solving the so-called Hughes phenomenon [1] and reducing the redundancy of data in order to improve classification accuracies. In pattern recognition, it is desirable to extract features which are focused on the discrimination between classes of interest. Although a reduction in dimensionality is of importance, the error rising from the reduction in dimension has to be without sacrificing the discriminative power of classifiers [9]. In continue, a few well-known feature extraction techniques which have been widely used in conjuction with spectral-spatial classifiers will be elaborated.

1.2.2.1 Principal Component Analysis (PCA)

PCA is an unsupervised feature extraction technique. The general aim of PCA is to transform the data into a lower dimensional subspace via a transformation that is optimal in terms of the sum-of-squared error [10]. PCA reduces the dimensionality of a data set with interrelated variables, while retaining as much as possible of the variation in the data set. The dimensionality reduction is obtained by a linear transformation of the data into a new set of variables, the PCs. The PCs are orthogonal to each other and are ordered in such a fashion that the first PC corresponds to the greatest variance, the second component corresponds to the second greatest variance and so on.

Each pixel in a d-bands image can be written as:

$$X_d = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}. \tag{1.6}$$

In order to reduce the dimensionality of the input hyperspectral data, one can estimate the eigenvalues of the covariance matrix as follows:

$$C_{d,d} = \begin{pmatrix} \sigma_{1,1} & \dots & \sigma_{1,d} \\ \vdots & \ddots & \vdots \\ \sigma_{d,1} & \dots & \sigma_{d,d} \end{pmatrix} \tag{1.7}$$

where $\sigma_{i,j}$ is the variance for band i if $i = j$ and otherwise $\sigma_{i,j} = \rho_{ij}\sigma_i\ \sigma_j$ for each pair of different bands where ρ_{ij} is the correlation coefficient between the bands.

The eigenvalues (λ) of the variance-covariance matrix can be calculated as the roots of the characteristic equation as follows:

$$det(C - \lambda I) = 0, \tag{1.8}$$

where *C* is the covariance matrix of the data and *I* is the diagonal identity matrix.

With respect to the eigenvalues, one can calculate the percentage of original variance explained by each PC. This percentage can be estimated with respect to the ratio of each eigenvalue in relation to the sum of the all eigenvalues. In this way, the PCs which contain minimum variance can be eliminated.

The principal component transformation can be expressed as follows:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = \begin{pmatrix} w_{1,1} & \dots & w_{1,d} \\ \vdots & \ddots & \vdots \\ w_{d,1} & \dots & w_{d,d} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}, \tag{1.9}$$

where *Y* is the vector in the mapped space, *W* is the transformation matrix and *X* is the vector of the original data. The column vectors of W are the eigenvectors of *C*, the covariance matrix of the input data, ordered in the same way as the corresponding eigenvalues. These values give information on the relation between the bands and each PC. From these values one can link a main component with a real variable. The eigenvectors can be estimated from the vector - matrix equation for each eigenvalue λ_d as follows:

$$(C - \lambda_d I)w_d = 0, \tag{1.10}$$

where *C* is the covariance matrix, λ_d is the *d*th eigenvalue, *I* is the diagonal identity matrix, and w_d is the *d*th eigenvector.

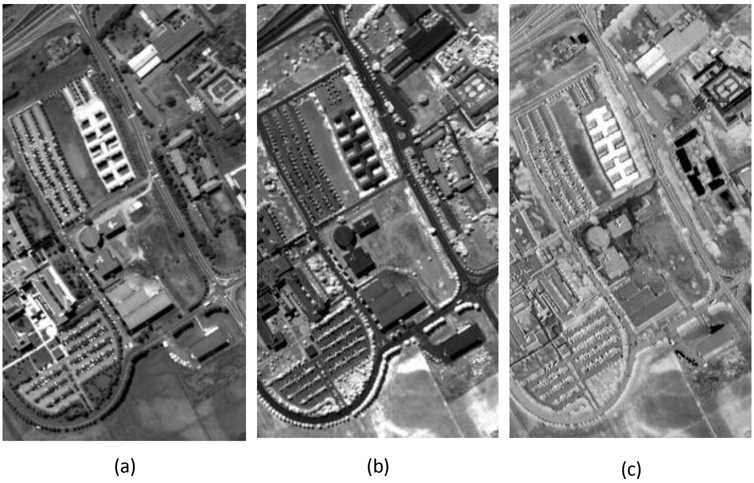


Figure 1.5: The first three obtained PCs for the Pavia University data: a) The first PC, b) the second PC and c) the third PC.

1.2.2.2 Discriminant Analysis Feature Extraction (DAFE)

DAFE is a parametric supervised feature extraction approach. This approach has been extensively used for dimension reduction in classification problems [11]. In this approach, within-class, between-class and mixture scatter matrices are usually considered as the criteria of class separability. The within-class scatter matrix of DAFE (S_w^{DA}) is estimated by:

$$S_w^{DA} = \sum_{i=1}^{K} P_i \Sigma_i, \tag{1.11}$$

where P_i denotes the prior probability of class i where $i = \{1, ..., K\}$ and Σ_i is the class covariance matrix.

The between-class scatter matrix of DAFE (S_b^{DA}) is given by:

$$\begin{aligned} S_b^{DA} &= \sum P_i(m_i - m_0)(m_i - m_0)^T = \\ &\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} P_i P_j (m_i - m_j)(m_i - m_j)^T, \end{aligned} \tag{1.12}$$

where m_i is the class mean for class i. The parameter m_0 is the expected vector of the mixture distribution, which is is given by:

$$m_0 = \sum_{i=1}^{K} P_i m_i. \tag{1.13}$$

In DAFE, The optimal features are extracted by optimizing the Fisher criterion expressed by:

$$J = tr\left[(S_w^{DA})^{-1}(S_b^{DA})\right]. \tag{1.14}$$

The first row of Figure 1.6 demonstrates the first three components of DAFE extracted from the Pavia University data set. DAFE is fast and works well when the distribution of the data is normal (Gaussian), but its concept suffers from the following shortcomings:

1. When the distribution of data is not normal (non-Gaussian), the performance of DAFE will be downgraded and its results will not be promising.
2. When the difference in the mean vectors of the classes is small, the extracted features by DAFE will not be reliable. In the same manner, if one class-mean vector is very different from others, its corresponding class will dramatically influence on the other classes in the sense of computing the between-class covariance matrix [12]. As a consequence, the feature extraction process will be ineffective.
3. DAFE is based on computations at full dimensionality, which demands a huge number of training samples in order to accurately estimate statistics [13].
4. The main shortcoming associated with the concept of DAFE is that this approach is not full rank and its rank at maximum is equal to K-1 where K is the number of classes. In this way, if the rank of the within-class scatter matrix is u, then DAFE only extracts $min(K-1, u)$ features. Since in real situations, the data distribution is complicated, using only K-1 features usually is not sufficient [2].

1.2.2.3 Decision Boundary Feature Extraction (DBFE)

This method was proposed in [14] and relies on extracting informative features from decision boundaries. DBFE considers training samples directly in order to determine the location of the effective decision boundaries, which is the boundary where different classes overlap [15].

For two Gaussian classes, the work flow of DBFE can be summarized as follows [2]:

1. Let m_i and Σ_i be the class mean and class covariance matrix, respectively.

[2] The following steps of DBFE procedure was extracted from [14]

2. Classify the whole bands of the input data by using the training samples.

3. Perform a chi-square threshold test to the correctly classified training samples of each classes and delete outliers. To do so, for the class i, keep X only if the following chi-square threshold test is satisfied:

   $$(X - m_i)^T \Sigma_i^{-1} (X - m_i) < R_{t1}.$$

   Let $(X_1, X_2, ..., X_{L_1})$ be only correctly classified training samples of class w_1 which satisfy the chi-square threshold test and $(Y_1, Y_2, ..., Y_{L_2})$ be only correctly classified training samples of class w_2 which satisfy the chi-square threshold test.

4. Perform a chi-square threshold test of class w_1 to the training samples of class w_2, and keep Y_j only if the following chi-square threshold test is satisfied:

   $$(Y_j - m_1)^T \Sigma_i^{-1} (Y_j - m_1) < R_{t2}.$$

   If the number samples of class w_2 which satisfy the chi-square threshold test is less than L_{min}, keep the L_{min} samples of class w_2 which provide the smallest values.

5. For X_i of class w_1, find the nearest samples of class w_2 kept in step 3.

6. Find the point Po_i where the straight line connecting the pair of samples found in step 5 meets the decision boundary.

7. Find the normal unit vector N_i to the decision boundary at the point Po_i found in step 6 as follows:

   $$N = \nabla h(X) \,|_{X=X_i} = (\Sigma_1^{-1} - \Sigma_2^{-1}) X_i + (\Sigma_1^{-1} m_1 - \Sigma_2^{-1} m_2).$$

8. By repeating steps 5-7 for X_i where $i = \{1, ..., L_1\}$, L_1 unit normal vectors, calculate the estimate of the effective decision boundary feature matrix (Σ_{EDBFM}^1) from class w_1 which can be estimated as follows:

   $$\Sigma_{EDBFM}^1 = \frac{1}{L_1} \sum_i^{L_1} N_i N_i^t.$$

   Repeat steps 3-8 for class w_2.

9. Calculate the final estimation of the effective decision boundary matrix by using the following equation:

   $$\Sigma_{EDBFM} = \frac{1}{2} \left(\Sigma_{EDBFM}^1 + \Sigma_{EDBFM}^2 \right).$$

This work flow can be easily extended for multiclass cases [14]. The second row of Figure 1.6 demonstrates the first three components of DBFE extracted from the Pavia University data set.

Some of main points of using DBFE are listed below:

1. Since DBFE considers directly the classification accuracies rather than other metrics (e.g., statistical distances), it is based on both the mean separation and covariance differences. In this manner, this approach works more efficient than some other feature selectors which are downgraded when there is no mean separation.
2. DBFE is able to efficiently handle the problems of outliers [14].
3. Because DBFE works directly on training samples to determine the location of effective decision boundaries, it demands many training samples. In other words, in a case when we do not have enough training samples, the efficiency of DBFE is downgraded which is not desirable.
4. Since DBFE works directly on training samples to determine the location of effective decision boundaries, this approach can be computationally intensive if the number of training samples is large.
5. Because DBFE works directly on training samples to determine the location of effective decision boundaries, it suffers from the Hughes phenomenon as the number of features increases [15].

1.2.2.4 Nonparametric Weighted Feature Extraction (NWFE)

In order to overcome the limitations of DAFE, NWFE was introduced in [13]. NWFE is a nonparametric supervised feature extraction technique. NWFE is developed based on DAFE by focusing on samples near the eventual decision boundary, rather than considering the same weight for all training samples as with DAFE. The main ideas behind NWFE are to put different weights on different samples in order to compute ′weighted means′ and define new nonparametric within-class and between-class scatter matrices. The main advantages of using NWFE are as follows:

1. NWFE is generally of full rank. This advantage provides the possibility of opting the number of desired features on the opposite way of DAFE which usually can extract *K-1* features (*K* is the number of classes) [2]. In addition, this advantage helps to reduces the issue of singularity [2].
2. The nonparametric nature of between- and within-class scatter matrices in NWFE makes this approach well-suited even for non-normal distributed data.

In NWFE, the nonparametric between-class scatter matrix for K classes is estimated as

$$S_b^{NW} = \sum_{i=1}^{K} \frac{P_i}{K-1} \sum_{j=1_{i\neq j}}^{K} \sum_{l=1}^{n_i} \lambda_l^{(i,j)} (X_l^{(i)} - M_j(X_l^{(i)}))(X_l^{(i)} - M_j(X_l^{(i)}))^T, \tag{1.15}$$

where P_i denotes the prior probability of class i where $i = \{1, ..., K\}$ and $X_l^{(i)}$ is the l-th sample from class i. $\lambda_l^{(i,j)}$ presents scatter matrix weight. n_i is training sample size of class i. $M_j(X_l^{(i)})$ is regarded as the weighted mean of $X_l^{(i)}$ in class j and is given below:

$$M_j(X_l^{(i)}) = \sum_{k=1}^{n_i} W_{lk}^{(i,j)} X_k^j, \tag{1.16}$$

in which $W_{lk}^{(i,j)}$ is the weight for estimating weighted means, which is estimated as follows:

$$W_{lk}^{(i,j)} = \frac{dist(X_l^{(i)}, X_k^{(j)})^{-1}}{\sum_{t=1}^{n_j} dist(X_l^{(i)}, X_k^{(j)})^{-1}}, \tag{1.17}$$

where $dist(a, b)$ means the distance from a to b. As it can be observed, the weight $W_{lk}^{(i,j)}$ is a function of $X_l^{(i)}$ and $X_k^{(j)}$. In this case, if the distance between $X_l^{(i)}$ and $X_k^{(j)}$ is small, the weight $W_{lk}^{(i,j)}$ goes to one, otherwise the weight $W_{lk}^{(i,j)}$ goes to zero.

The scatter matrix weight $\lambda_l^{(i,j)}$ is a function of $X_l^{(i)}$ and $M_j(X_l^{(i)})$. In this case, for class i, if the distance between $X_l^{(i)}$ and $M_j(X_l^{(i)})$ is small, the scatter matrix weight $\lambda_l^{(i,j)}$ goes to one otherwise the scatter matrix weight goes to zero. The scatter matrix weight $\lambda_l^{(i,j)}$ is defined as:

$$\lambda_l^{(i,j)} = \frac{dist(X_l^{(i)}, M_j(X_l^{(i)}))^{-1}}{\sum_{t=1}^{n_j} dist(X_l^{(i)}, M_j(X_l^{(i)}))^{-1}}. \tag{1.18}$$

The nonparametric within-class scatter matrix S_w^{NW} is estimated by

$$S_w^{NW} = \sum_{i=1}^{L} P_i \sum_{l=1}^{n_i} \frac{\lambda_l^{(i,j)}}{n_i} (X_l^{(i)} - M_i(X_l^{(i)}))(X_l^{(i)} - M_i(X_l^{(i)}))^T, \tag{1.19}$$

The optimal features are extracted by optimizing the $(S_w^{NW})^{-1}(S_b^{NW})$. The general work flow of NWFE is as follows [14]:

1. Compute the distances between each pair of sample points and create distance matrix
2. Compute the weight $W_{lk}^{(i,j)}$ by using the distance matrix produced in step 1.
3. Compute the weighted means $M_j(X_l^{(i)})$ by using $W_{lk}^{(i,j)}$.
4. Compute the scatter matrix weight $\lambda_l^{(i,j)}$.
5. Compute S_b^{NW} and S_w^{NW}.
6. Extract features by considering $(S_w^{NW})^{-1}(S_b^{NW})$.

The third row of Figure 1.6 demonstrates the first three components of NWFE extracted from the Pavia University data set. For detailed information regarding NWFE see [13].

1.2.3 Feature Selection

Feature selection perhaps is the most straightforward way to reduce the dimensionality of a data set by simply selecting a subset of features from the set of available features based on a criterion. As an example, imagine one wishes to select the best five bands out of the ten available bands for the classification of a data set with six classes, by using Bhattacharyya distance [6] feature selection technique. To do so, one needs to compute the Bhattacharyya distance between each pair of classes for each subset of size five out of the 10-band data [2]. As output of this procedure, five features which provide the highest Bhattacharyya distance in the feature domain, will be selected. Please note, feature selection techniques do not make any changes on the specification of the input data and just easily select the most informative bands out of the available ones, by considering a criterion. On the contrary, feature extraction performs a linear or nonlinear transformation to the input data, concentrating the information into a smaller number of features. Therefore, feature extraction makes changes on input data via a transformation.

From one point of view, feature selection techniques can be categorized into two categories: unsupervised and supervised. Supervised feature selection techniques aim at finding the most informative features with respect to prior knowledge (e.g., training samples) and lead to better identification and classification of different classes of interest. On the contrary, unsupervised methods are used in order to find distinctive bands when a prior knowledge of

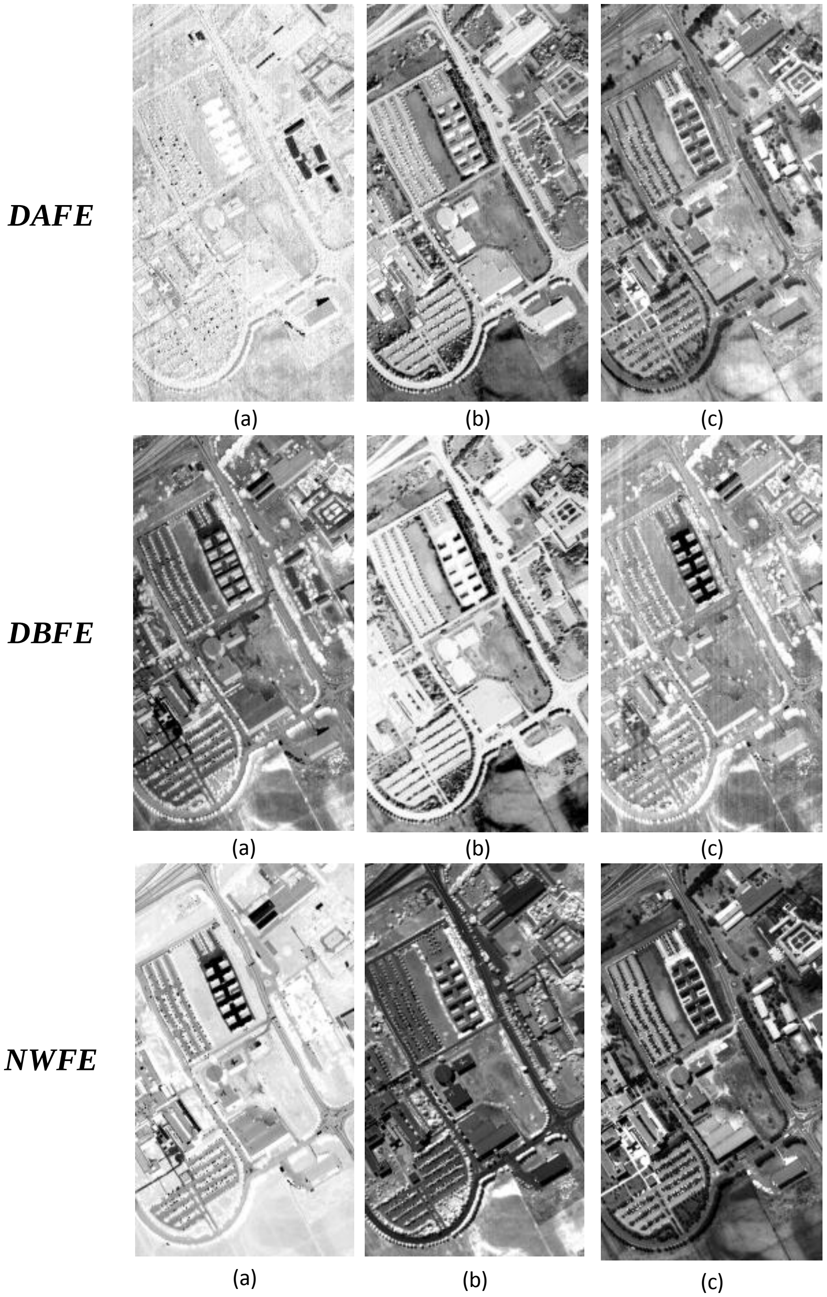


Figure 1.6: The first three components of DAFE, DBFE and NWFE, respectively, for the Pavia University data.

the classes of interest is not available. Information Entropy [16], First Spectral Derivative [17] and Uniform Spectral Spacing [18] can be considered as unsupervised feature selection techniques, while supervised feature selection techniques usually try to find a group of bands achieving the largest class separability. Class separability can be calculated with considering a few approaches such as Divergence [19], Transformed divergence [19], Bhattacharyya distance [6] and Jeffries-Matusita distance [19]. Although conventional feature selection techniques have been used extensively in remote sensing for many years, they suffer from the following shortcomings:

1. Most conventional feature selection techniques are based on the estimation of the second order statistics (e.g., covariance matrix) and because of that, they demand many training samples in order to estimate the statistics accurately [3]. Therefore, in a situation when the number of training samples is limited, the singularity of covariance matrices is possible. In addition, since the existing bands in hyperspectral data usually have some redundancy, the probability of the singularity of the covariance matrix will even increase.

2. In order to select informative bands by using conventional feature selectors, corrupted bands (e.g., water absorption and low SNR bands), are usually pre-removed manually, which is a time-consuming task. In addition, conventional feature selection methods can be computationally demanding since they are based on exhaustive search techniques and they require to calculate all possible alternatives in order to choose the most informative features from the available ones. In this case, in order to select an m features out of a total of n features, these methods must calculate $n!/(n-m)!m!$ alternatives, which is a laborious task and demands a significant amount of high computational memory. In other words, the feature selection techniques are only feasible in relatively low dimensional cases. In this case, as the number of bands increases, the CPU processing time exponentially increases.

In order to address the above-mentioned shortcomings of conventional feature selection techniques, the new trend of feature selection methods is usually based on the use of stochastic and evolutionary optimization techniques (e.g., Genetic Algorithms (GA) and Particle Swarm Optimization (PSO)). The main

[3] Reminder: The mean vector is considered as a first order statistic, since it involves only one variable. On the contrary, covariance matrix is known as a second-order statistic, since it considers the relationship between two variables. In the same way, correlation infers how two variables are related to each other. Higher-order statistics are related to the relationships between more variables. As the order of the statistic increases, the statistical estimation using a limited number of training samples becomes more problematic [2]. This is the reason why we would generally expect that the mean vector can be relatively well estimated with a smaller number of samples compared to the covariance matrix.

reasons behind this trend is that 1) in evolutionary feature selection techniques, there is no need to calculate all possible alternatives in order to find the most informative bands and 2) in evolutionary based feature selection techniques, usually a metric is chosen as fitness function which is not based on the calculation of the second order statistics, and, in this case, the singularity of the covariance matrix is not a problem. Furthermore, despite the conventional feature selection techniques for which the number of required features need to be set by the user, most of evolutionary-based feature selectors are able to automatically select the most informative features in terms of classification accuracy without requiring the number of desired features to be set *a priori*.

In order to make the most of the evolutionary-based feature selection techniques, the use of an efficient metric (fitness function) is vitally important for estimating the capability of different potential solutions. In this case, approaches such as Divergence [19], Transformed divergence [19], Bhattacharyya distance [6] and Jeffries-Matusita distance [19] can be taken into account as the fitness function. However, as mentioned before, these metrics require many training samples in order to estimate statistics accurately. In addition, these metrics are based on the estimation of the second order statistics and when the number of training samples is limited and the input features have a high correlation, the singularity of the covariance matrix downgrades the efficiency of the feature selection step and these metrics cannot lead to a conclusion.

For the purpose of hyperspectral image analysis, Support Vector Machine (SVM) and Random Forests (RF) play an important role since they can handle high dimensional data even if a limited number of training samples is available. In addition, SVM and RF are non-parametric classifiers and in this case, they are suitable for non-Gaussian (non-normal) data sets. Therefore, the output of these classifiers can be chosen as a fitness function. However, it should be noted that either SVM or RF has its own shortcoming when it is considered as the fitness function. For instance, when RF is considered as the fitness function, due to its capability to handle different types of noise, corrupted and noisy bands cannot be eliminated even after high number of iterations. On the contrary, since SVM is more sensible than RF to noise, SVM is able to detect and discard corrupted bands after a few iterations, which can be considered as a privilege for the final classification step. However, SVM needs a time demanding step, cross-validation, in order to tune the hyperplane parameters. In this case, since most of the evolutionary-based feature selection techniques are based on iterative process and producing many potential solutions, SVM needs to be applied many times during the process and because of that, the evolutionary-based feature selection approaches based on the SVM as the fitness function demand a huge amount of CPU processing time, which is not the case for RF. Another alternative would be to use SVM without the cross-validation step and arbitrarily initialize the hyperplane parameters. In this case, the algorithm is not automatic anymore and the obtained results might not be reliable. As a result, a careful choice of the fitness function is very much of importance and we recommend the readers to choose the most appropriate metric based on their problem and data set.

In the literature, there is a huge number of articles related to the use of evolutionary optimization based feature selection techniques. These methods are mostly based on the use of GA and PSO. For example, in [20], the authors developed a SVM classification system which allows the detection of the most distinctive features and the estimation of the SVM parameters (e.g., regularization and kernel parameters) by using a GA. In [21], PSO was considered in order to select the most informative features obtained by morphological profiles for classification. In [22], a method was developed which allows to simultaneously solve problems of clustering, feature detection, and class number estimation in an unsupervised way by considering a PSO.

Below, one of the best-known evolutionary-based feature selection techniques (GA-based feature selection) is discussed in detail. In Chapter 5, a detailed description is also provided for the PSO-based feature selection. In addition, in Chapter 5, based on the shortcomings of GA and PSO, a few advanced evolutionary-based feature selection approaches will be proposed (i.e., the Hybrid GA-PSO (HGAPSO)- and FODPSO-based feature selection methods) [23, 24].

1.2.4 Genetic Algorithm (GA)-based feature selection

GA is inspired by the genetic process of biological organisms. GA consists of several *potential solutions*; called chromosomes or individuals. Each chromosome in a binary GA includes several genes with binary values; 0 and 1, which determine the attributes of each individual. A set of the chromosomes is made up to form a population.

For the purpose of feature selection based on GA, the length of each chromosome should be equal to the number of input features. In this case, the value of each gene, 0 or 1, demonstrates the absence or the presence of the corresponding band, respectively.

The merit of each chromosome is evaluated by using a fitness function. Fitter chromosomes are selected through a selection step as parents for the generation of new chromosomes (offsprings). In that step, two fit chromosomes are selected and combined through a crossover step. Then, mutation is performed on the offsprings in order to increase the randomness of individuals for decreasing the possibility of getting stucked in local optimum [25]. Below, the main steps of the GA are briefly described.

Figure 1.7 shows the general idea of a simple GA. In each generation, first, the fitness values of all the chromosomes in the same population are calculated (e.g., the overall accuracy of SVM on validation samples). Then, the selection

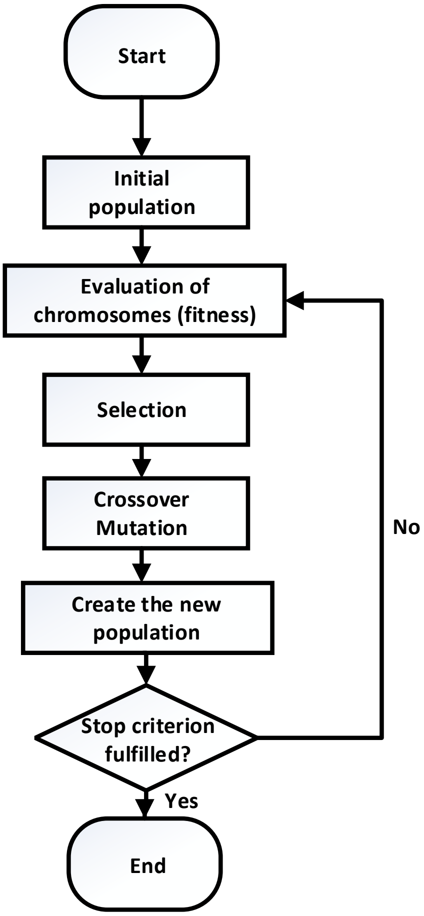

Start
Initial population
Evaluation of chromosomes (fitness)
Selection
Crossover Mutation
Create the new population
Stop criterion fulfilled?
No
Yes
End

Figure 1.7: General idea of the conventional GA.

step is applied. The main idea behind the selection step is to give preference to better individuals (those that have higher fitness value) by allowing them to pass on their specification (genes) to the next generation and prohibit the entrance of worst fit individuals into the next generations. As the generations go on, the chromosomes should get fitter and fitter (i.e., closer and closer to the desired solution). There is a wide variety of techniques for the selection step, but *Tournament selection* [26] and *Roulette Wheel selection* [27] are the most common ones.

Crossover is regarded as the process of taking more than one parent chromosome and producing new offsprings from them. In the crossover step, generally, fitter chromosomes are selected based on their fitness value and recombined with each other in order to produce new chromosomes for the next generation. In this way, once a pair of chromosomes has been selected as parents, crossover can take place to produce offsprings. A crossover probability of 1.0 indicates that all the selected chromosomes are used in reproduction i.e., there are no survivors. However, empirical studies have shown that better results are achieved by a crossover probability of between 0.65 and 0.85, which implies that the probability of a selected chromosome surviving to the next generation unchanged (apart from any changes arising from mutation) [4]. There are a wide variety of methods for performing crossover on the chromosomes, but the most popular ones are one point, two points and uniform crossover [28]. A simple scheme of different crossover methods is shown in Figure 1.8.

Mutation is used along with the crossover operation, in order to increase the randomness of the population and add more diversity on the chromosomes in order to avoid getting trapped in local optimum. Mutation is performed on the chromosomes based on *mutation probability*. Mutation probability (or ratio) is considered as a measure of likeness in which random genes of the chromosome will be flipped into something else (in binary GA, the values are switched from 0 to 1 or 1 to 0). For example if a chromosome is encoded as a binary string of length 100 and if the mutation probability is 0.01, it means that 1 out of the 100 bits (on average) picked at random and switched to another value (0 or 1). A simple scheme of mutation with the probability of 0.1 is shown in Figure 1.9.

The GA is an iterative process, so it is iterated again and again until the stop criterion is met. In this manner, there are different ways which can be considered as a stop criteria. For example, if the difference between the best fitness value and the average of all fitness values in one iteration is less than a predefined threshold value, the process can be terminated. In addition, the number of iterations can be predefined by the user as an another way for terminating the process.

The main shortcoming of the GA is that if a chromosome is not selected, the information contained by that individual is lost. In addition, GA is slow and it demands a high CPU processing time. That problem can get even worse when the problem at hand is complicated. Same as other evolutionary techniques, there is no absolute assurance that GA will be able to find a global optimum.

A detailed description regarding the PSO-, HGAPSO- and FODPSO- based feature selection approaches is provided in Chapter 5.

1.2.5 Conventional spectral classifiers and the importance of considering spatial information

As discussed before, a hyperspectral data cube consists of several pixels and each pixel is a vector of d values which demonstrates the number of spectral

[4] http://www.optiwater.com/optiga/ga.html

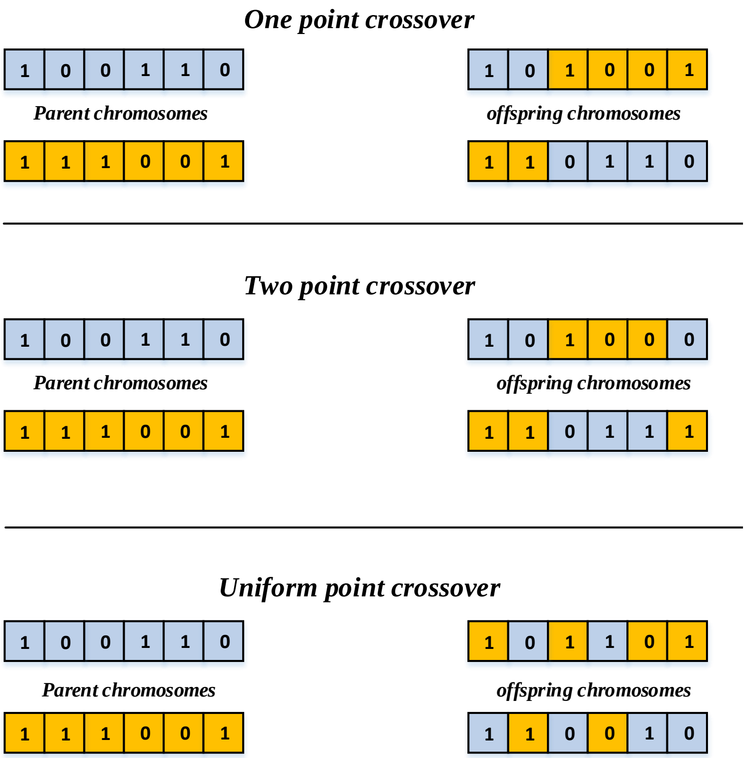


Figure 1.8: A simple scheme of different crossover methods. From top to down: one point, two points and uniform crossover techniques.

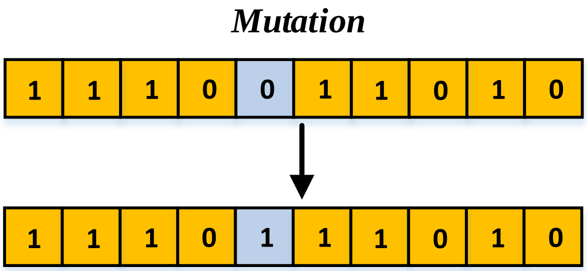


Figure 1.9: A simple scheme of mutation with the probability of 0.1.

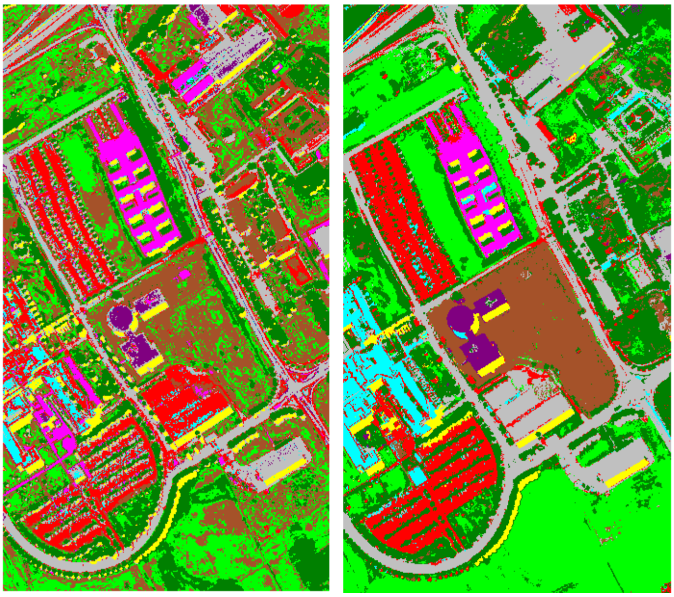


Figure 1.10: An example of classification maps, with (the right image) and without (the left image) considering spatial information. As can be seen, the classification map obtained by considering the both spectral and spatial information is much smoother than the classification map obtained by considering only spectral information. Considering the spatial information can reduce the labeling uncertainty that exists when only spectral information is taken into account, and helps to overcome the salt and pepper appearance of the classification map.

channels. Each pixel corresponds to the reflected radiation of the specific region of the earth and has multiple values in spectral bands. Vectors of different pixels belonging to the similar material with high probability may have almost the same values. Different supervised and unsupervised classification techniques are used in order to group vectors with almost the same spectral characteristic. The procedure of grouping different materials with almost the same spectral characteristics can be considered as the fundamental meaning of image classification. Remote sensing image classifiers try to discriminate different classes of ground cover, for example, from categories such as soil, vegetation, and surface water in a general description of a rural area, to different types of soil, vegetation, and water depth or clarity for a more detailed description.

Hyperspectral imaging instruments are now able to capture hundreds of spectral channels from the same area on the surface of the Earth. By providing very fine spectral resolution with hundreds of (narrow) bands, accurate discrimination of different materials is possible. As a result, hyperspectral data are a valuable source of information for the classifiers. The output of the classification step is a *classification map*. Figure 1.10 (the right image) shows an example of a classification map consisting nine classes, including: trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow and soil.

In follows, a brief description of a few well-known classifiers is provided in order to illustrate the pros and cons of these methods and the reason why spectral and spatial classifiers have been gaining a great attention from different researchers. Broadly speaking, classification techniques can be categorized into two categories; supervised and unsupervised classifiers which can be briefly described as follows:

- Supervised classifiers: These types of methods classify input data by considering its spectral information into a classification map in order to determine classes of interests, by using a set of representative samples for each class, referred to as *training samples*. In order to partition the feature space into decision regions, a set of training samples for each class is used. Training samples are usually obtained by manually labeling a small number of pixels in an image or based on some field measurements. In other words, for a hyperspectral data cube with *d*-bands, which can be represented as a set of n pixel vectors $X = \left\{X_j \in \mathrm{IR^d, j = 1, 2, ..., n}\right\}$, supervised classifiers try to classify the data into a set of classes $\Omega = \{w_1, w_2, ..., w_K\}$.
- Unsupervised classifiers: Another type of classifiers is based on unsupervised classification or clustering. It is referred to as unsupervised because it does not use training samples, and classify the input data only based on an arbitrarily number of initial "cluster centers" which may be user-specified or may be quite arbitrarily selected. During the process, each pixel is associated with one of the cluster centers based upon a similarity criterion. Here, two best-known clustering techniques are briefly described.
    - K-means: This approach [29] is as one of the best-known clustering methods which was introduced by MacQueen. This method starts with a random initial partition of the pixel vectors into candidate clusters and then reassigns these vectors to clusters by reducing the squared error in each iteration, until a convergence criterion is met.
    - ISODATA: This method was firstly introduced in [30] and it follows the same trend with K-means clustering algorithm but with the distinct difference that the former assumes that the number of clusters is known *a priori*, but latter allows for different number of clusters.

Supervised classification techniques play a key role in the analysis of hyperspectral images and a wide variety of applications can be handled by a good classifier, including: land-use and land-cover mapping, crop monitoring, forest applications, urban development, mapping, tracking and risk management.

In the 1990s, neural network approaches attracted many researchers for classifying hyperspectral images [31, 32]. The advantage of using neural network models over the statistical parametric methods are that they are distribution free and thus no prior knowledge about the statistical distribution of classes is needed. A set of weights and non-linearities describe the neural network, and these weights are computed via an iterative training procedure. The main interest of using such approaches considerably increased in the 1990s because of recently proposed feasible training techniques for non-linearly separable data [33]. At this point, the use of neural networks for hyperspectral image classification is limited, primarily due to their algorithmic and training complexity [34] as well as the number of tuning parameters which need to be selected.

RF was first introduced in [35] and it is an ensemble method for classification and regression. Ensemble classifiers get their name from the fact that several classifiers, i.e., an ensemble of classifiers, are trained and their individual results are then combined through a voting process. In order to classify an input vector by RF, the input vector is run down each decision tree (a set of binary decisions) in the forest (the set of all trees). Each tree provides a unit vote for a particular class and the forest chooses the class that has the most votes. For example, if 100 trees are grown and 80 of them predict that a particular pixel is forest and 20 of trees predict it is grass, the final output for that pixel will be forest. Based on studies in [36], the computational complexity of the RF algorithm is $cT\sqrt{MN}\ \log(N)$ where c is a constant, T denotes the number of trees in the forest, M is regarded as the number of variables and N is the number of samples in the data set. It is easy to detect that RF is not computationally intensive but demands a considerable amount of memory since it needs to store an N by T matrix while running. RF does not assume any underlying probability distribution for input data and can provide a good classification result in terms of accuracies, and can handle many variables and a lot of missing data. Another advantage of RF classifier is that it is insensitive to noise in the training labels. In addition, RF provides an unbiased estimate of the test set error as trees are added to the ensemble and finally it does not overfit.

SVMs are another example of supervised classification approach. The general idea behind SVM is to separate training samples belonging to different classes by tracing maximum margin hyperplanes in the space where the samples are mapped [37]. SVMs were originally introduced for solving linear classification problems. However, they can be generalized to non-linear decision functions by considering the so-called kernel trick [38]. A kernel-based SVM is being used to project the pixel vectors into a higher dimensional space and estimate maximum margin hyperplanes in this new space, in order to improve linear separability of data [38]. The sensitivity to the choice of the kernel and regularization parameters can be considered as the most important disadvantages of SVM. The latter is classically overcome by considering cross-validation techniques using training data [39]. The Gaussian radial basis function (RBF) is widely used in remote sensing [38].

The both SVM and RF classification methods are comparable in terms of classification accuracies and have been widely used for the purpose of hyperspectral image classification since they can handle high dimensionality of data with limited number of training samples which is the common issue in remote sensing. However, while both methods are shown to be effective classifiers for non-linear classification problems, SVM requires a computationally demanding parameter tuning in order to achieve optimal results, whereas RF does not require such tuning process and is found to be more robust. In this sense, RF is much faster than SVM and for volumetric data using RF instead of SVM is favorable.

Conventional spectral classifiers consider the hyperspectral image as a list of spectral measurements with no spatial organization [40]. However, in remote sensing images, neighboring pixels are highly related or correlated since remote sensors acquire significant amount of energy from adjacent pixels and homogeneous structures in the image scene are generally larger than the size of a pixel. This is especially evident for the image of high spatial resolution. As an example, if a given pixel in an image represents the class "sea", its adjacent pixels belong to the same class with a high probability. As a result, spatial and contextual information of adjacent pixels can provide valuable information from the scene. Considering the spatial information can reduce the labeling uncertainty that exists when only spectral information is taken into account, and helps to overcome the salt and pepper appearance of the classification map. Furthermore, other relevant contextual information can be extracted when the spatial domain is considered. As an example, for a given pixel, it is possible to extract the size and the shape of the structure to which it belongs. Therefore, a joint spectral and spatial classifier is required in order to increase classification accuracies and quality of the final classification map. Figure 1.10. (the right image) shows an example regarding the quality improvement of conventional spectral classification map by considering spatial information into a classification framework. As a result, spectral-spatial classification methods (or context classifiers) must be developed, which assign each image pixel to one class based on: 1) its own spectral values (the spectral information) and 2) information extracted from its neighborhood (the spatial information) [41]. The use of spectral-spatial classification techniques is vitally important for processing of high resolution images with large spatial regions in the image scene. Broadly speaking, a spectral-spatial classification techniques consist of three main stages:

1. Extracting spectral information.

2. Extracting spatial information.
3. Combining the spectral information extracted from (1) and spatial information extracted from (2).

In the following, we describe the existing methods for the spectral-spatial classification of hyperspectral data.

In order to characterize the spatial information, two common strategies are available: Crisp neighborhood system and adaptive neighborhood system. While the first one mostly considers spatial and contextual dependencies in a predefined neighborhood system, the latter is more flexible and it is not confined to a given neighborhood system. In the following, existing methods based on each neighborhood system will be briefly discussed.

i. Crisp neighborhood system

One well-known way for extracting spatial information by using a crisp neighborhood system is the consideration of Markov Random Field (MRF) modeling. MRF is a family of probabilistic models and can be explained as a 2-D stochastic process over discrete pixels latices [42] and widely used to integrate spatial context into image classification problems. In MRFs, it is assumed that for a predefined pixel neighborhood of a given pixel, its closest neighbors belong with a high probability to the same object. Four- and eight-neighborhoods are the most frequently used in image analysis. By using this approach, the pixel in the center can be classified by taking into account the information from its neighbors according to one of those systems. MRFs are considered as a powerful tool for incorporating spatial and contextual information into the classification framework [43]. There is an extensive literature on the use of MRFs for increasing the accuracy of classification. In [44], Jackson and Landgrebe introduced spectral-spatial iterative statistical classifiers for hyperspectral data based on a MRF. Pixel-wise Maximum Likelihood classification was first performed and the classification map was regularized using the Maximum a Posteriori (MAP)-MRF framework. The spectral information was extracted by the Maximum Likelihood classification, while the spatial information was derived over the pixel neighborhood. In [45], the result of the Probabilistic SVM was regularized by a MRF. In [46], authors have further explored the MAP-MRF classification. They considered class-conditional PDFs estimated by the Mean Field-based SVM regression algorithm. Also, in [43, 47–50], MRFs were taken into consideration for modeling spatial and contextual information for improving the accuracy of the classification. Furthermore, a generalization of MRF, called conditional MRF, was investigated in [51] for the spectral and spatial classification of remote sensing images. In [52], the concept of Hidden Markov Model (HMM) was used for incorporating spectral and contextual information into a framework for performing unsupervised classification of remote sensing multispectral images. In [53, 54], *Ghamisi et. al* proposed to use a generalization of MRF named Hidden MRF (HMRF) for the spectral and spatial classification of hyperspectral data. In that work, spectral information was extracted by using a SVM and spatial information was extracted by using the HMRF and, finally, the spectral and spatial information were combined by using majority voting within each object [5]. In addition for the purpose of segmentation and anomaly detection, in [55] Gaussian MRF was employed.

Another common way to include spatial information into a classification technique is to consider texture measures. In [56, 57], authors have used texture measures derived from the Gray Level Co-occurrence Matrix (GLCM) for including the spatial information in order to classify hyperspectral data. In [56], texture images are produced using four measurements to describe the GLCM: Angular Second Moment, Contrast, Entropy and Homogeneity. Then, PCA is applied on the obtained texture images, and the PCs are selected as features for ML classification. In [57], authors have proposed to perform Non-negative Matrix Factorization feature extraction first, then extract spatial information using four measurements for the GLCM (Angular Second Moment, Entropy, Homogeneity and Dissimilarity), and apply a SVM classification on a stack of spatial and spectral features. The experimental results reported that, in most cases, this method did not demonstrate an improvement over the pixel-wise approaches. This may be explained by the fact that the hyperspectral remote sensing images only contains limited textural information [41].

However, the main disadvantages of considering a set of crisp neighbors are as follows:

- The crisp neighborhood system may not contain enough samples, which downgrades the effectiveness of the classifier (in particular, when the input data set is of high resolution and the neighboring pixels are highly correlated [58])

- A larger neighborhood system may lead to intractable computational problems [58]. Unfortunately, the closest fixed neighborhoods do not always accurately reflect information about spatial structures. For instance, they provoke assimilation of regions containing only few pixels with their larger neighboring structures and do not provide accurate spatial information at the border of regions.

- In general, the use of a crisp neighborhood system leads to acceptable results for big regions in the scene. Otherwise, it can disappear small structures in the scene and merge them with bigger surrounded objects.

[5] For performing majority voting within each object on the output of the segmentation and classification steps, first, the number of pixels with different class labels in each object is counted. Then, the set of pixels in each object is assigned to the most frequent class label.

In Chapter 2, the proposed HMRF, which is based on the crisp neighborhood system will be detailed in order to extract spatial and contextual information for the purpose of spectral-spatial classification.

ii. Adaptive neighborhood system

In order to address the shortcomings of using a set of crisp neighborhoods, an adaptive neighborhood system can be taken into account. One possible way for considering an adaptive neighborhood system is to take the advantage of different types of segmentation methods. Image segmentation is regarded as the process of partitioning a digital image into multiple regions or objects. In other words, in image segmentation a label is assigned to each pixel in the image such that pixels with the same label share certain visual characteristics [59]. These objects provide more information than individual pixels since the interpretation of images based on objects is more meaningful than based on individual pixels. Segmentation techniques extract large neighborhoods for large homogeneous regions, while not missing small regions consisting of one or a few pixels [58]. Image segmentation is considered as an important task in the analysis, interpretation and understanding of images and is also widely used for image processing purposes such as classification and object recognition [59, 60]. Image segmentation is a procedure which may lead to modify the accuracy of classification maps [61]. To make such an approach more effective, an accurate segmentation of the image is required [59]. There is an intensive literature on the use of segmentation techniques in order to extract the spatial information from remote sensing data (e.g., [62–64]).

In order to improve classification results, the integration of classification and segmentation steps has recently been taken into account [58, 65]. In such cases, the decision to assign a pixel to a specific class is simultaneously based on the feature vector of this pixel and some additional information derived from the segmentation step. A few methods for segmentation of multispectral and hyperspectral images have been introduced in the literature. Some of these methods are based on region merging techniques, in which neighboring image segments are merged with each other based on their homogeneity. For example, the multiresolution segmentation method in eCognition software uses this type of approach [66]. Tilton proposed a hierarchical segmentation algorithm [67], which alternately performs region growing and spectral clustering.

As mentioned in [63], image segmentation can be classified into four specific types including histogram thresholding based methods, texture analysis based methods, clustering based methods and region based split and merging methods. Thresholding is one of the most commonly used methods for the segmentation of images into two or more clusters. Many algorithms have been proposed in literature to address the issue of optimal thresholding (e.g., [68] and [69]). While several research papers address bi-level thresholding, others have considered the multilevel problem. Bi-level thresholding is reduced to an optimization problem to determine the threshold t that maximizes the σ_B^2 (between-class variance) and minimizes σ_W^2 (within-class variance). For two level thresholding, the problem is solved by finding the value T^* which results in max($\sigma_B^2(T^*)$) where $0 \leq T^* < L$ and L is the maximum intensity value. This problem could be extended to n-level thresholding through satisfying max $\sigma_B^2(T_1^*, T_2^*, ..., T_{n-1}^*)$ where $0 \leq T_1^* < T_2^* < ... < T_{n-1}^* < L$. One way for finding the optimal set of thresholds is by using exhaustive search. A commonly used exhaustive search is based on the Otsu criterion [70]. That approach is easy to implement, but it has the disadvantage that it is computationally expensive. Exhaustive search for $n - 1$ optimal thresholds involves evaluations of fitness of $n(L - n + 1)^{n-1}$combinations of thresholds [71]. Therefore, that method is not suitable from a computational cost point of view. The task of determining $n - 1$ optimal thresholds for n-level image thresholding could be formulated as a multidimensional optimization problem. To solve such a task, several biologically inspired algorithms have been explored in image segmentation (e.g. [71–73]). Bio-inspired algorithms have been used in situations where conventional optimization techniques cannot find a satisfactory solution or they take too much time to find it, e.g., when the function to be optimized is discontinuous, non-differentiable, and/or presents too many nonlinearly related parameters [73]. One of the best known bio-inspired algorithms is PSO [74]. The PSO consists of a number of particles that collectively move in the search space (e.g., pixels of the image) in search of the global optimum (e.g., maximizing the between-class variance of the distribution of intensity levels in the given image). However, a general problem with the PSO and similar optimization algorithms is that they may be trapped in local optimum points, and the algorithm may work in some problems but fail in others [59]. To overcome such a problem, Tillett *et al.* [75] presented the Darwinian PSO (DPSO). In the DPSO, multiple swarms of test solutions performing just like an ordinary PSO may exist at any time with rules governing the collection of swarms that are designed to simulate natural selection. In [61], DPSO was taken into account for the segmentation of multispectral remote sensing images. Results confirmed that DPSO outperforms the conventional PSO in terms of finding higher between class variance in less CPU processing time. More recently, in [76] and [63], for the purpose of image segmentation, *Ghamisi et. al* introduced further extension of the DPSO using fractional calculus to control the convergence rate of the algorithm [59] and evaluate the capability of that in order to segment hyperspectral images in [63] and a classification framework was proposed based on FODPSO based segmentation technique. The result of classification was promising and FODPSO based segmentation improved DPSO- and PSO-based segmentation techniques in terms of finding higher between class variance in less CPU processing time. In [77], in order to address the shortcomings of hard clustering approaches such as Kmeans, a new approach was proposed, i.e., fuzzy C-means which is optimized by FODPSO.

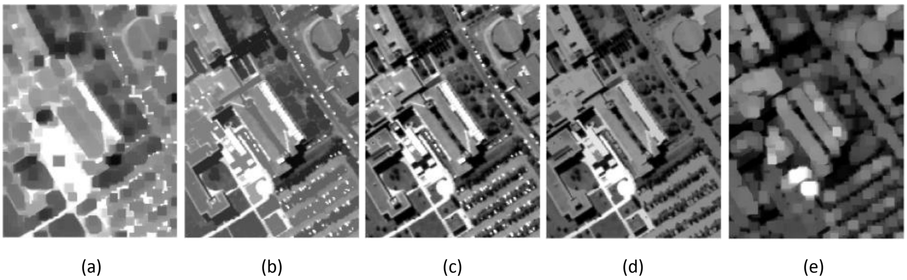


Figure 1.11: a) Morphological closing; b) Closing by reconstruction; c) Original VHR panchromatic image; d) Opening by reconstruction; e) Morphological opening. As can be seen, morphological opening and closing have influences on the shape of the structures and can introduce fake objects. However, opening and closing by reconstruction preserve the shape of different objects bigger than SE.

In Chapter 3, comprehensive information regarding the proposed thresholding-based segmentation techniques is given.

In [65, 78–80], Watershed, partitional clustering, and Hierarchical Segmentation (HSeg) have been considered in order to extract spatial information, and SVM has been considered in order to extract spectral information. Then, the spectral and spatial information have been integrated by using the majority voting [78]. The described approach leads to an improvement in terms of classification accuracies compared to spectral and spatial techniques using local neighborhoods for analyzing spatial information.

Another possible set of approaches which are able to extract spatial information by using an adaptive neighborhood system relies on morphological filters. Erosion and dilation are considered as the alphabet of mathematical morphology. These operators are carried out on an image with a set of known shape, called a Structuring Element (SE). Opening and closing are combinations of erosion and dilation. These operators simplify input data by removing structures with the size less than the SE. However, these operators have influences on the shape of the structures and can introduce fake objects in the image [9]. One possible way in order to handle this issue is to consider opening and closing by reconstruction [81]. Opening and closing by reconstructions are a family of connected operators which satisfies the following criterion: If the SE cannot fit the structure of the image, then it will be totally removed, otherwise it will be totally preserved. Reconstruction operators remove objects smaller than SE without altering the shape of those objects and reconstruct connected components from the preserved objects. For gray scale images, opening by reconstruction removes unconnected light objects and in dual, closing by reconstruction removes unconnected dark objects. Figure 1.11 illustrates an original

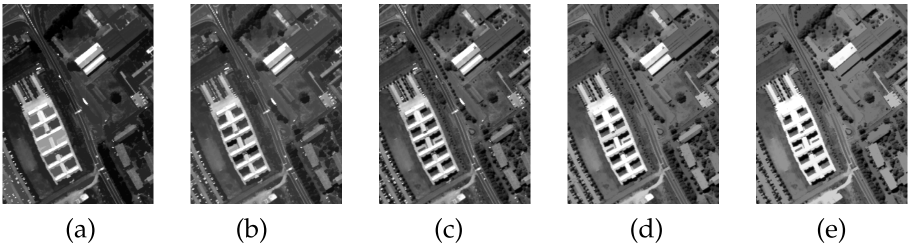


Figure 1.12: A simple example of MP consisting two sequential opening and closing by reconstruction.

VHR image along with its corresponding opening, opening by reconstruction, closing and closing by reconstruction.

In order to fully exploit the spatial information, filtering techniques should simultaneously attenuate the unimportant details and preserve the geometrical characteristics of the other regions. Pesaresi and Benediktsson [82] used morphological transformations to build a so-called Morphological Profile (MP). They carried out a multiscale analysis by computing an anti-granulometry and a granulometry, (i.e., a sequence of closings and openings with SE of increasing size), appended in a common data structure named MP. Figure 1.12 illustrates a simple example of MP consisting two sequential opening and closing by reconstruction.

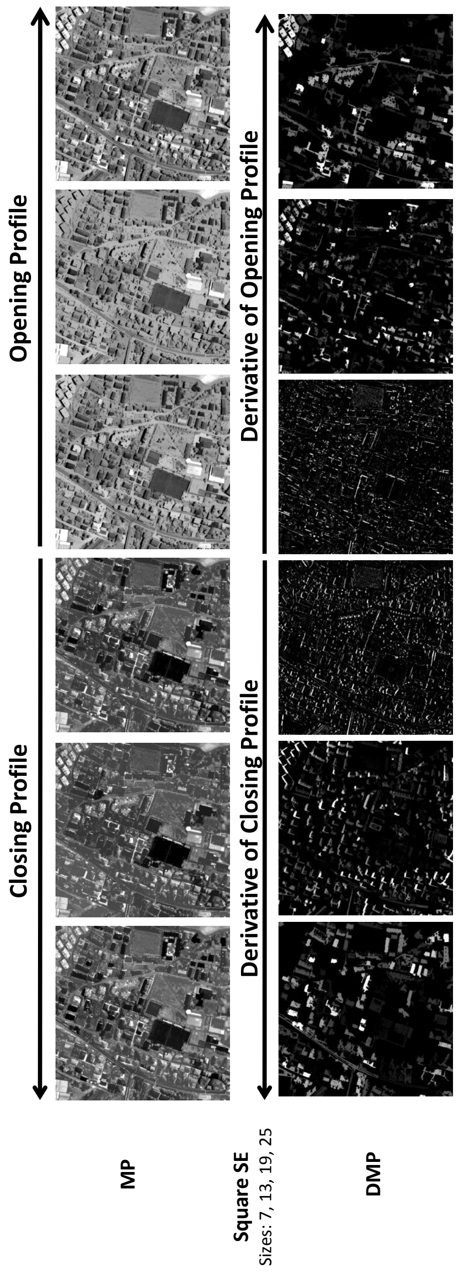


Figure 1.13: Examples of MP and DMP.

Another modification of using MP which was exploited for the classification of VHR panchromatic images is Differential Morphological Profile (DMP). DMP explains the residues of two successive filtering operations for two adjacent levels existing in the profile. The obtained map is generated by associating each pixel to the level where the maximum of the DMP (evaluated at the given pixel) occurs [83]. Figure 1.13 shows the examples of MP and DMP.

In [84], the MP generated by standard opening and closing was carried out on a Quickbird panchromatic image captured on Bam which was hit by the earthquake on 2003. In that work, the spatial features extracted by the MP were considered for assessing the damages caused by the earthquake. The standard opening and closing along with white and black top hat and opening and closing by reconstruction, were taken into account all together and classified by a SVM for the classification of a Quickbird panchromatic image, [85]. An automatic hierarchical segmentation technique based on the analysis of the DMP was proposed in [86]. The DMP was also analyzed in [87], by extracting a fuzzy measure of the characteristic scale and contrast of each structure in the image. The computed measures were compared with the possibility distribution predefined for each thematic class, generating a value of membership degree for each class used for classification. In [88], in order to reduce the dimensionality of data and address the curse of dimensionality, feature extraction techniques were taken into consideration for the DMP classified by a neural network classifier. In [89], the concept of MPs was successfully extended in order to handle hyperspectral images. For doing that, first the input hyperspectral data were transformed by using PCA and MPs have been performed on the PCs of the data (which were called Extended Morphological Profiles (EMPs)). Figure 1.14 shows a stacked vector consisting of the profiles based on the first and second PCs. Since the EMP do not fully exploit the spectral information and PCA discards class information, in [9], instead of PCA, different supervised feature extraction techniques were performed on the input data and the MP and extracted features are concatenated into a stacked vector and classified by an SVM.

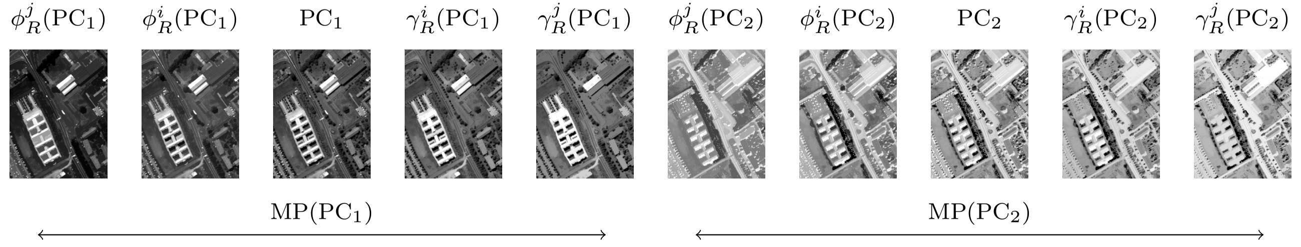

$\phi^j_R(\mathrm{PC}_1)$
$\phi^i_R(\mathrm{PC}_1)$
PC_1
$\gamma^i_R(\mathrm{PC}_1)$
$\gamma^j_R(\mathrm{PC}_1)$
$\phi^j_R(\mathrm{PC}_2)$
$\phi^i_R(\mathrm{PC}_2)$
PC_2
$\gamma^i_R(\mathrm{PC}_2)$
$\gamma^j_R(\mathrm{PC}_2)$
$\mathrm{MP}(\mathrm{PC}_1)$
$\mathrm{MP}(\mathrm{PC}_2)$

Figure 1.14: A simple example of MP consisting two sequential opening and closing by reconstruction

Some studies have been conducted in order to assess the capability of SEs with different shapes for the extraction of spatial information. For instance, MPs computed with a compact SE (e.g., square, disk, etc.) can be considered for modeling the size of the objects in the image (e.g., in [58] this information was exploited to discriminate small buildings from large ones). In [90], the computation of two MPs was introduced in order to model both the length and the width of the structures. In greater detail, one MP is built by disk-shaped SEs for extracting the smallest size of the structures, while the other employs linear SEs (which generate directional profiles [91]) for characterizing the objects maximum size (along with the orientation of the SE). This is appropriate for defining the minimal and maximal length but, as all the possible lengths and orientations cannot be practically investigated, such analysis is computationally intensive.

Based on the above-mentioned literature, it is easy to obtain that the computation of a multiscale processing (e.g., by MPs, DMPs, EMPs) has proven to be effective in extracting informative spatial features from the analyzed images. In order to characterize the shape or size of different structures present in an image, it is vitally important to consider a range of SEs with different sizes. MPs use successive opening/closing operations with an SE of an increasing size. The successive usage of opening/closing leads to a simplification of the input image and a better understanding of different available structures in the image. Although MP is a powerful technique for the extraction of spatial information, the concept of that is suffered by few limitations including:

1. The shape of SEs is fixed which is considered as a main limitation for the extraction of objects within a scene.

2. SEs are unable to describe information related to the gray-level characteristics of the regions such as spectral homogeneity, contrast and so on.

3. A final limitation associated with the concept of MPs is the computational complexity. The original image needs to be processed completely for each level of the profile, which requires two complete evaluations of the image; one performed by a closing transformation and the other by an opening transformation. Thus, the complexity increases linearly with the number of levels included in the profile [83].

A morphological Attribute Profile (AP) is considered as the generalization of the MP which provides a multilevel characterization of an image by using the sequential application of morphological Attribute Filters (AFs) [83]. Morphological attribute opening and thinning are AFs which were introduced in [92]. AFs are connected operators which process an image by considering only its connected components. For binary images, the connected components are simply the foreground and background regions present in the image. In order

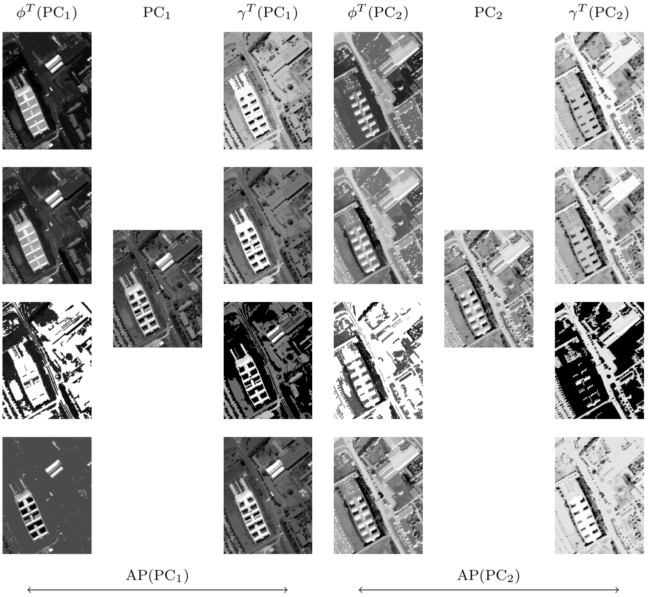

$\phi^T(\mathrm{PC}_1)$
PC_1
$\gamma^T(\mathrm{PC}_1)$
$\phi^T(\mathrm{PC}_2)$
PC_2
$\gamma^T(\mathrm{PC}_2)$
$\mathrm{AP}(\mathrm{PC}_1)$
$\mathrm{AP}(\mathrm{PC}_2)$

Figure 1.15: A simple example of EAP consisting four attributes on the first second PCs.

to deal with gray scale images, the set of connected components can be obtained by considering the image to be composed by a stack of binary images generated by thresholding the image at all its gray-level values [93]. Thus, they process the image without distorting or inserting new edges but only by merging existing flat regions [81]. AFs were employed for modeling the structural information of the scene in order to increase the effectiveness of a classification and building extraction in [83] and [94], respectively, where they proved to be efficient for the modeling of structural information in VHR images. AFs include in their definition, the morphological operators based on geodesic reconstruction [92]. Moreover, they are a flexible tool since they can perform a processing based on many different types of attributes. In fact, the attributes can be of any type. For example, they can be purely geometric, or related to the spectral values of the pixels, or on different characteristics. Furthermore, in [94], the problem of the tuning of the parameters of the filter was addressed by proposing an automatic selection procedure based on a genetic algorithm. Extended AP (EAP) is a stacked vector of different APs computed on the first C features extracted from the original data set. Figure 1.15 shows an example of an EAP for the first two PCs consisting of four attributes.

For the purpose of spectral-spatial classification of hyperspectral images, four attributes have been widely used in literature including: 1) Area of the region (related the size of the regions), 2) standard deviation (as an index for showing the homogeneity of the regions), 3) diagonal of the box bounding the regions, and 4) moment of inertia (as an index for measuring the elongation of the regions). When concatenation of different attributes, $\{a_1, a_2, ..., a_M\}$ are gathered into a stacked vector, the Extended Multi-AP (EMAP) is obtained [95]. In [96], the fusion of hyperspectral and LiDAR data is taken into account in order to develop a new classification framework for the accurate analysis of urban areas based on EMAPs.

A comprehensive information related to APs along with all its modifications and generalizations can be found in [97].

The application of the profiles for large volumes of data is computationally demanding and that is considered to be one of the main difficulties in using them. In order to solve this issue, the efficient implementation of attribute filters was proposed in [98]. Salembier *et al.* in [98], introduced a new data representation named Max-tree which has received much interest since it increases the efficiency of filtering by dividing the transformation process into three steps: 1) tree creation; 2) filtering; and 3) image restitution.

The main difficulties of using the EMAP are 1) to know which attributes lead to a better discrimination for different classes, and 2) which threshold values should be considered in order to initialize each AP. In this case, a few papers have tried to solve these issues and introduced automatic techniques in order to make the most of attribute profiles such as [99–101].

Chapter 4 provides a comprehensive survey regarding the use of AP in remote sensing. The rest of Chapter 4 is devoted to the proposed spectral-spatial classifiers, which consider AP for extracting spatial information.

1.3 Objectives

The **main objective** of this thesis is to develop spectral and spatial classification techniques which are efficient in terms of classification accuracies and CPU processing time. With respect to the above-mentioned description, the following objectives can be defined:

1. Besides the importance of classification accuracies, other critical issues for the purpose of hyperspectral image classification are simplicity and speed of the approaches. Therefore, in this thesis, special emphasis is given to proposing robust techniques that are accurate in classification as well as being fast. This is these types of techniques which can be used for handling real-time applications such as hazard monitoring and risk management. This point can be observed in a few papers developed in this PhD.

2. Most of the existing methods are not automatic and their performance is

very dependent on the initialization of the algorithms. Therefore, the existing techniques demand a significant effort by users to initialize the parameters in a trial and error way and different steps need to be done manually. This makes the existing techniques very time demanding which is undesirable for a wide variety of applications such as flood monitoring and bush fire management, which require a rapid response. In order to increase the efficiency of the existing techniques and reduce the laborious task of user interaction, a further development of automatic (or semi-automatic) techniques plays a key role in remote sensing data analysis. Consequently, this has been taken into consideration in most part of this thesis.

3. In order to reduce the laborious task observed for existing feature reduction techniques as well as increasing the efficiency of those, a proposition of powerful feature reduction approaches which are fast, robust against noise, and capable of handling high dimensional data with only a limited number of training samples would be of importance.

4. Conventional MRF techniques only consider spatial information in the label image without getting any feedback from the original observed image. In order to address this issue and increase the capability of conventional MRF techniques, further modifications and alternations are crucial.

5. In order to address the main disadvantage of considering crisp neighborhood systems as discussed in the introduction, a proposition of robust approaches based on adaptive neighborhood systems is highly demanded.

1.4 Main Contributions

The main contributions of this thesis are summarized in Figure 1.16, which depicts the proposed segmentation techniques, spectral–spatial classification approaches and feature reduction techniques. In order to achieve the objectives defined in the previous section, we have proposed and developed three general strategies for hyperspectral data classification. According to the figure, these three strategies cover three important concepts associated with the classification of high-dimensional data including: 1) Spectral-spatial classification using crisp neighborhood systems, 2) spectral-spatial classification using adaptive neighborhood systems as well as 3) feature reduction approaches. Below, the three strategies will be described in more detail.

1. *Strategy 1 (spectral-spatial classification using crisp neighborhood systems):* This strategy is detailed in Paper 1 (Chapter 2), which is related to the proposition of a novel spectral-spatial classification approach using closest fixed neighborhoods. This approach is based on the integration of the

SVM classifier and HMRF segmentation method via majority voting. In general, MRFs consider spatial information in the label image only, not in the original observed image. In order to address this issue, HMRF is defined with regard to a pair of random variable families (observation random field and hidden random field) while MRF is only defined with respect to hidden random field. To this extent, HMRF can provide more accurate segmentation maps. In HMRF, model-fitting procedure involves an initialization step and an iteration between two steps: maximum a posteriori estimation of the class labels and an expectation maximization algorithm for estimating the parameters of each Gaussian class. It should be noted that in remote sensing, the concept of HMRF was used for the first time in Paper 1 (Chapter 2).

2. *Strategy 2 (spectral-spatial classification using adaptive neighborhood systems):* This strategy is detailed in papers 2, 3, 4, 5, 6 and 7. Although the first strategy works well in terms of classification accuracies, it models the spatial dependencies of adjacent pixels based on a crisp neighborhood system. With respect to the general shortcomings of modeling spatial information based on a crisp neighborhood system and in order to model the spatial information in a more efficient way, in the second strategy, we propose to use adaptive neighborhood systems by considering different approaches based on image segmentation and APs. This strategy can be split into three subsections:

    (a) The first subsection is devoted to the proposition of a new thresholding based segmentation technique. The proposed segmentation method is based on a new optimization technique named FODPSO. In general, FODPSO is proposed to address the main shortcoming of PSO; the stagnation in local optimum by considering two main novelties: 1) Using many swarms of test solutions which may exist at any time, in which each swarm individually performs just like an ordinary (PSO) algorithm with a set of rules governing the collection of swarms that are designed to simulate natural selection; 2) using the concept of fractional derivative to control the convergence rate of particles. This work is mainly based on the segmentation of gray scale images and detailed in paper 2 (Chapter 3). It should be noted that FODPSO-based image segmentation was proposed in paper 2 (Chapter 3) for the first time in image processing and computer vision community.

    (b) The second subsection is on the generalization of the FODPSO-based segmentation technique from gray scale images to hyperspectral data sets. To this extent, first, in paper 3 (Chapter 3), the concept of the FODPSO is used for the segmentation of hyperspectral images and then, a novel spectral–spatial classification approach is introduced based on the integration of SVM and FODPSO. In paper 4 (Chapter 3), the integration of mean shift segmentation and FODPSO is proposed to use along with SVM for the spectral–spatial classification of hyperspectral images in order to address the shortcoming of using each of them individually.

(c) The third subsection is on the use of APs for the spectral–spatial classification of hyperspectral data. To do so, first, paper 5 (Chapter 4) is devoted to a comprehensive survey over all existing papers in terms of the use of APs for the extraction of spatial information from remote sensing data. The main contributions of this survey paper are to recall the concept of the AP and its all modifications and generalizations with special emphasis on remote sensing image classification and summarize the important aspects of its efficient utilization while also listing potential future works. Papers 6 and 7 (Chapter 4) propose new classification approaches by considering EMAP and RF. Those papers introduce efficient classifiers in terms of classification accuracies and CPU processing time. In addition, those techniques are also automatic.

3. *Strategy 3 (feature selection approaches):* Finally, the third strategy is on proposing novel feature selection approaches which are detailed in papers 8 and 9 (Chapter 5). Paper 8 is based on FODPSO and SVM, while Paper 9 is based on the new integration of GA, PSO and SVM. For both approaches, there is no need to set the number of output features manually, and the proposed approaches can automatically select the most informative features in terms of classification accuracies. In addition, since both approaches are based on evolutionary techniques, they are much faster than other well-known feature selection techniques that demand an exhaustive process to select the most informative bands. In this sense, the new approaches can work appropriately in a situation when other feature selection techniques are not applicable. Paper 8, also, solves the main problem of using EMAP for classification; 1) Which attributes should be used? and 2) what values should be used as threshold values?

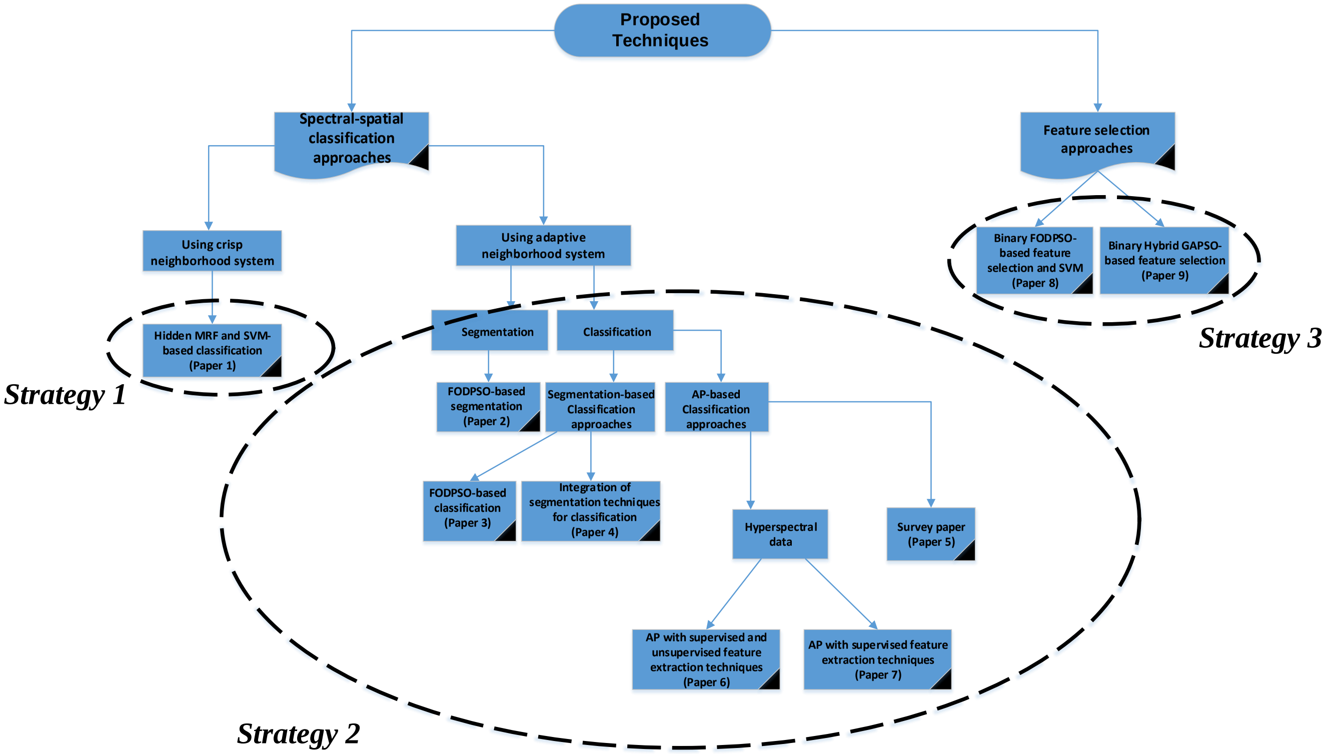

Proposed Techniques
Spectral-spatial classification approaches
Feature selection approaches
Using crisp neighborhood system
Using adaptive neighborhood system
Hidden MRF and SVM-based classification (Paper 1)
Segmentation
Classification
FODPSO-based segmentation (Paper 2)
Segmentation-based Classification approaches
AP-based Classification approaches
FODPSO-based classification (Paper 3)
Integration of segmentation techniques for classification (Paper 4)
Hyperspectral data
Survey paper (Paper 5)
AP with supervised and unsupervised feature extraction techniques (Paper 6)
AP with supervised feature extraction techniques (Paper 7)
Binary FODPSO-based feature selection and SVM (Paper 8)
Binary Hybrid GAPSO-based feature selection (Paper 9)
Strategy 1
Strategy 2
Strategy 3

Figure 1.16: The main contribution of this thesis

1.5 Thesis Outline

This thesis is composed as a collection of publications. To this extent, 10 journal papers written during the PhD present the main findings of this thesis. The papers are summarized bellow and detailed in the next chapters.

1.5.1 Paper 1

P. Ghamisi, J. A. Benediktsson and M. O. Ulfarsson, "Spectral-Spatial Classification of Hyperspectral Images Based on Hidden Markov Random Fields", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2565-2574, May 2014.

In this paper, a novel fully automatic spectral and spatial approach is introduced for the classification of hyperspectral images. This approach is based on the HMRF and SVM. In order to preserve the edges in the classification map, a gradient step based on the Sobel edge detector is taken into account. In the framework, SVM is used for the extraction of spectral information. In parallel, HMRF is used for the extraction of spatial information. In the final step, those results are combined by using majority voting. It should be noted that the concept of HMRF is used for the first time in the field of remote sensing in this paper, and the efficiency of that for the segmentation of hyperspectral images is demonstrated. it is shown in this paper that the method performs well in terms of classification accuracies. In addition, the proposed approach is fully automatic and user friendly in contrast to most of the methods.

1.5.2 Paper 2

P. Ghamisi, M. S. Couceiro, J. A. Benediktsson and N. M. F. Ferreira, "An Efficient Method for Segmentation of Images Based on Fractional Calculus and Natural Selection", *Expert Systems With Applications*, vol. 39, no. 16, pp. 12407-12417, Nov. 2012.

This paper presents two novel thresholding based segmentation methods based on the FODPSO and DPSO for determining the *n-1* optimal *n*-level threshold on a given image. The efficiency of the proposed methods is compared with other well-known thresholding based segmentation methods such as GA-, PSO- and Bacteria Foraging-based segmentation techniques. Results indicate that FODPSO is able to find the better thresholds with more stability in less CPU processing time.

1.5.3 Paper 3

P. Ghamisi, M. S. Couceiro, F. M. L. Martins and J. A. Benediktsson, "Multilevel Image Segmentation Based on Fractional-Order Darwinian Particle Swarm Optimization", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2382-2394, May 2014.

In this paper, a new method is proposed for the segmentation of multispectral and hyperspectral images, which is based on FODPSO. Segmentation methods were carried out on two different test cases. Experimental results indicate that the FODPSO is more robust than the two other methods (PSO and DPSO) and has a higher potential for finding the optimal set of thresholds with more between-class variance in less computational time, especially for higher segmentation levels and for images with a wide variety of intensities. In addition, to show the efficiency of the proposed segmentation method on the result of classification, a novel classification approach based on the new segmentation method and SVM is proposed. Results confirm that the new segmentation method improves on the SVM in terms of classification accuracies when compared to the standard SVM classification of the raw image data.

1.5.4 Paper 4

P. Ghamisi, M. S. Couceiro, M. Fauvel and J. A. Benediktsson, "Integration of Segmentation Techniques for Classification of Hyperspectral Images", *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 342-346, Jan. 2014.

In this letter, a new spectral–spatial classification approach is introduced for the accurate classification of hyperspectral images. The approach is based on the combination of FODPSO and MSS. In the proposed approach, the result of FODPSO is used as the input to MSS to develop a pre-processing method for the SVM classifier. Results indicate that the use of both segmentation methods can overcome the shortcomings of each other and the combination can improve the result of classification significantly.

1.5.5 Paper 5

P. Ghamisi, M. Dalla Mura and J. A. Benediktsson, "A Survey on Spectral–Spatial Classification Techniques Based on Attribute Profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2335-2353, May 2015.

The main objective of this survey paper is to recall the concept of the APs along with all its modifications and generalizations. This paper emphasizes on the use of APs for the classification of remote sensing data. Finally, this paper summarizes the important aspects of APs along with all its efficient utilization while also listing potential future works.

1.5.6 Paper 6

P. Ghamisi, J. A. Benediktsson and J. R. Sveinsson, "Automatic Spectral-Spatial Classification Framework Based on Attribute Profiles and Supervised Feature Extraction", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 9, pp. 5771-5782, Dec. 2014.

In this paper, a new approach is proposed for the spectral-spatial classification of hyperspectral images. The method can be implemented fully automatically. In order to use the spatial information, APs are taken into account. For reducing the redundancy of both the spatial information and the original spectral data in order to provide more accurate classification results, a few supervised feature extraction methods are considered. The obtained results confirm that considering spatial information by using APs in conjunction with spectral information can significantly improve the classification accuracies of the original data. In addition, by using supervised feature extractions, the classification accuracies can be increased further. Furthermore, in order to avoid the main difficulties of using APs, an automatic version of the proposed method is introduced which only considers area and standard deviation attributes. The automatic method obtained almost the same results as the manual method in terms of the classification accuracies and CPU processing time and solved the main difficulty of the manual method which is related to the initialization of the parameters in the EMAP. The proposed method worked well in terms of the classification accuracy and CPU processing time, which confirms the ability of the method to classify high-dimensional data sets.

1.5.7 Paper 7

P. Ghamisi, J. A. Benediktsson, G. Cavallaro and A. Plaza, "Automatic Framework for Spectral–Spatial Classification Based on Supervised Feature Extraction and Morphological Attribute Profiles", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2147 - 2160, Jun. 2014.

In this paper, we have developed a new automatic framework for the classification of hyperspectral images. Our framework uses both spectral and spatial information. In order to include the spatial information, morphological APs are taken into account. For reducing the redundancy of the extracted features and deal with the curse of dimensionality introduced by the Hughes effect, parametric supervised feature extraction methods (DAFE and DBFE) are considered. The new approach achieves well classification accuracies with acceptable CPU processing time in a fully automatic way.

1.5.8 Paper 8

P. Ghamisi, M. S. Couceiro and J. A. Benediktsson, "A Novel Feature Selection Approach Based on FODPSO and SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2935-2947, May 2015.

A novel feature selection approach is proposed based on a new binary optimization method inspired by FODPSO. A Support Vector Machine (SVM) classifier is used as a fitness function for cross validation samples and the overall accuracy of the classification is selected to evaluate the group of bands. In order to show the capability of the proposed method, two different applications are considered. In the first application, the proposed feature selection approach is directly carried out on the input hyperspectral data. The most informative bands selected from this step are classified by SVM. In the second application, the main shortcoming of using attribute profiles for spectral-spatial classification is addressed. In this case, a stacked vector of the input data and an attribute profile with all widely used attributes with wide ranges of thresholds is created. Then, the proposed feature selection approach automatically chooses the most informative features from the stacked vector. Experimental results successfully confirm that the proposed feature selection technique works better in terms of classification accuracies and CPU processing time than other studied methods.

1.5.9 Paper 9

P. Ghamisi and J. A. Benediktsson, "Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization", *IEEE Geoscience and Remote Sensing Letter*, vol. 12, no. 2, pp. 309-313, Jul. 2015.

A new feature selection approach which is based on the integration of a GA and PSO is proposed. A SVM classifier is used as the fitness function and its overall classification accuracy is considered as the fitness value. Results confirm that the new approach is able to automatically provide informative features in terms of classification accuracy within an acceptable CPU processing time. Furthermore, the usefulness of the proposed method is also tested for road detection. Results confirm that the proposed method is capable of discriminating between road and background pixels and performs better than the other approaches used for comparison in terms of performance metrics.

CHAPTER 2

Spectral-Spatial Classification based on a Crisp Neighborhood System: Hidden Markov Random Field

Spectral–Spatial Classification of Hyperspectral Images Based on Hidden Markov Random Fields

Pedram Ghamisi, *Student Member, IEEE*, Jón Atli Benediktsson, *Fellow, IEEE*, and Magnus Orn Ulfarsson, *Member, IEEE*

***Abstract*—Hyperspectral remote sensing technology allows one to acquire a sequence of possibly hundreds of contiguous spectral images from ultraviolet to infrared. Conventional spectral classifiers treat hyperspectral images as a list of spectral measurements and do not consider spatial dependences, which leads to a dramatic decrease in classification accuracies. In this paper, a new automatic framework for the classification of hyperspectral images is proposed. The new method is based on combining hidden Markov random field segmentation with support vector machine (SVM) classifier. In order to preserve edges in the final classification map, a gradient step is taken into account. Experiments confirm that the new spectral and spatial classification approach is able to improve results significantly in terms of classification accuracies compared to the standard SVM method and also outperforms other studied methods.**

***Index Terms*—Hidden Markov random field (HMRF), hyperspectral image analysis, image segmentation, support vector machine (SVM) classifier.**

I. Introduction

DUE TO recent advances in hyperspectral sensor technology, it is possible to capture hundreds of spectral channels for each image pixel from ultraviolet to infrared. By increasing the amount of spectral information, the accurate discrimination of different materials of interest is possible. In addition, the fine spatial resolution of the sensors enables the analysis of small spatial structures in the image. Furthermore, the high spectral resolution allows detailed physical analysis of the structures [1].

Classification plays a key role in the analysis of hyperspectral images. Examples of applications where it plays a key role are land-use and land-cover mapping, crop monitoring, forest applications, urban development, mapping, tracking, and risk management.

For hyperspectral images, several hundreds of spectral bands of the same scene are typically available, while for multispectral images, up to ten bands are usually available. By increasing the dimensionality of the images in the spectral domain, theoretical and practical problems arise. For instance, with a limited training set, beyond a certain limit, the classification accuracy actually decreases as the number of features increases [2]. For the purpose of classification, these problems are related to the curse of dimensionality.

Conventional spectral classifiers treat hyperspectral images as a list of spectral measurements [3]. For instance, support vector machine (SVM) classifiers have received significant attention lately because of their remarkable generalization capability for the classification of high dimensional data sets [4] and their considerable capability for handling big data sets with few number of training samples. The efficiency of SVM classifiers has been shown in terms of achieving very accurate results in a wide variety of applications [5], [6]. However, SVM classifiers do not consider spatial dependences and classify images only based on their spectral information. Therefore, this approach discards information associated with the correlations among distinct pixels in the image and is considered as the most vital limitation of SVM classifiers for the analysis of remote sensing images in which pixel neighborhoods provide important information [7].

To address the aforementioned problem, joint spectral and spatial classification techniques have recently received considerable attention. Consideration of spatial information helps us to overcome the salt and pepper appearance of the classification. More importantly, other relevant information can be extracted from the spatial domain: For a given pixel, it is possible to extract the size and the shape of the structure to which it belongs. Therefore, the combination of spectral information and spatial information can improve the result of the classification stage. The goal of considering spatial context in the classification step can partially be achieved by using methods such as morphological filters (e.g., [1]), morphological leveling (e.g., [8]), segmentation (e.g., [9]), and Markov random fields (MRFs) (e.g., [10]).

MRFs are a family of probabilistic models that can be described as 2-D stochastic processes over discrete pixel lattices [11]. They can be considered as a powerful tool for incorporating spatial and contextual information into the classification framework [12]. More recently [13], hidden MRF (HMRF) was introduced as a special case of the hidden Markov model (HMM). In HMRF, the underlying stochastic process is MRF, instead of Markov chains in HMM. Therefore, HMRF is not restricted to 1-D and can be used in order to extract spatial information from 2-D and 3-D images.

There is extensive literature on the use of MRFs for increasing the accuracy of classification. For instance, in [14], the

Manuscript received January 30, 2013; revised April 9, 2013 and May 4, 2013; accepted May 9, 2013. Date of publication June 14, 2013; date of current version February 27, 2014. This work was supported in part by the Icelandic Research Fund for Graduate Students.

The authors are with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland (e-mail: peg6@hi.is).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2263282

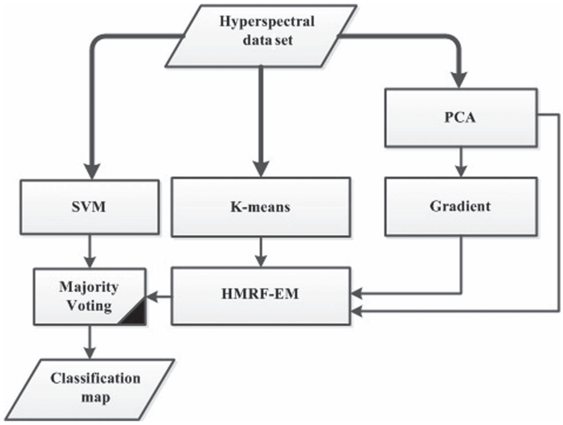


Fig. 1. Flowchart of the proposed method.

result of the probabilistic SVM was regularized by an MRF. In [10], the mean field-based SVM regression was used for image classification. Also, in [7], [12], and [15]–[17], MRFs were taken into consideration for modeling spatial and contextual information for improving the accuracy of the classification. Furthermore, a generalization of MRF, called conditional MRF, was investigated in [18] for the spectral and spatial classification of remote sensing images. In [19], the concept of HMM was used for incorporating spectral and contextual information into a framework for performing unsupervised classification of remote sensing multispectral images. In addition, Gaussian MRF was employed in [20] for the purpose of segmentation and anomaly detection.

Based on the previous discussion, the integration of SVM classifiers and MRFs for the accurate classification of remote sensing images by considering both spectral information and spatial information into the same framework is completely obvious. In this paper, a novel fully automatic spectral and spatial approach is introduced for the classification of hyperspectral images. The new approach is based on the HMRF and SVM. In order to preserve the edges in the classification map (CM), a gradient step based on the Sobel edge detector is taken into account. In addition, to our knowledge, this is the first time that HMRF is used in the field of remote sensing.

This paper is organized as follows. The proposed methodology is discussed in Section II. Then, Section III is devoted to experimental results. Finally, Section IV outlines the main conclusion.

II. Methodology

Fig. 1 illustrates the flowchart of the proposed method. In the following, specific parts of the proposed framework will be discussed in detail.

A. Notation

In the following, we let $\mathbf{y} = (y_1, \ldots, y_N)^T$ denote the first principal component map, where N is the number of pixels and $S = \{1, 2, \ldots, N\}$ is the set of pixel indices. Associated with pixel i is a class label x_i. A vector containing these labels is denoted by $\mathbf{x} = (x_1, \ldots, x_N)^T$.

B. HMRF-EM Segmentation by Preserving Edges

1) FGM: For better understanding of the concept of HMRF, we begin with the finite Gaussian mixture (FGM) model. For a pixel i, we have

$$\begin{aligned} q(l) &= q(x_i = l) \\ p(y_i|l) &= g(y_i; \theta_l) \end{aligned}$$

where $p(y_i|l)$ is a conditional probability of the intensity y_i given the class label l ($l \in L$, and L is regarded as the set of all possible labels). $q(l)$ is the probability mass function (pmf) of the class label, and $g(y_i; \theta_l)$ is a Gaussian probability density function (pdf) with parameter $\theta_l = (\mu_l, \sigma_l^2)$. The marginal distribution of $y = y_i$ dependent on the parameter set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_l, l \in L\}$ can be written as

$$p(y; \boldsymbol{\theta}) = \sum_{l \in L} g(y; \theta_l) q(l). \tag{1}$$

Although the FGM model is mathematically simple, it is not able to take the spatial information into consideration since all the data points are considered individually and are independent from the other neighborhood points. To overcome this limitation, the HMRF was proposed in [13].

2) HMRF Model: HMRF is a generalization of HMM. While HMM is based on 1-D Markov chains, HMRFs are based on MRFs. Due to its ability to handle a 2-D structure, HMRF is more suitable for image segmentation than HMM.

The Gaussian HMRF is given by

$$\begin{aligned} p(\mathbf{x}, \boldsymbol{y}; \boldsymbol{\theta}) &= f(\mathbf{x}) \prod_{i=1}^{N} p(y_i|x_i) \\ p(y_i|x_{\boldsymbol{N}i}; \boldsymbol{\theta}) &= \sum_{l \in L} g(y_i; \theta_l) q(l|x_{\boldsymbol{N}i}) \end{aligned} \tag{2}$$

where $f(\mathbf{x})$ is a pdf for x which follows the so-called Gibbs densities [21] and $q(l|x_{N_i})$ is a conditional pmf for the class label l given that x_{N_i} denotes a neighborhood for each pixel x_i. The difference between HMRF and FGM is the term $q(l|x_{N_i})$ in (2) and the term $q(l)$ in (1). If we do not consider the relationship between pixels in the neighboring system, HMRF and FGM are the same. In other words, spatial dependences can be modeled in HMRF, which are discarded in FGM. Therefore, the FGM model is a special case of HMRF. As a result, it can be concluded that HMRF is more flexible than FGM since it is able to model both the statistical and spatial properties of the image.

The model-fitting procedure [13] involves an initialization and an iteration between two steps: maximum a posteriori (MAP) estimation of the class labels and an expectation-maximization (EM) algorithm [22] for estimating $\boldsymbol{\theta}$. Now, we consider these three steps.

a) Initialization: The output of this step provides the initial label $\mathbf{x}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ for the MAP and EM algorithms, respectively. In this paper, K-means was used to provide the

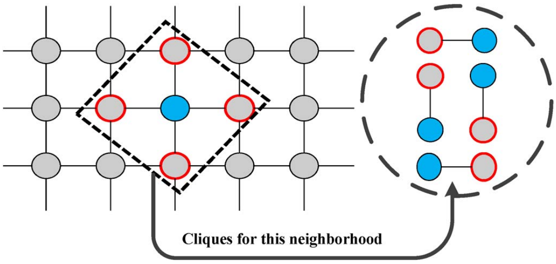


Fig. 2. All possible cliques for the predefined neighborhood system.

initial labels, and initial parameters $\boldsymbol{\theta}$ were computed for the initialization step. The initial parameters are obtained by estimating the mean and the standard deviation of the pixels within each cluster.

K-means [23] is one of the best-known clustering methods which was introduced by MacQueen. This method starts with a random initial partition of the pixel vectors into candidate clusters and then reassigns these vectors to clusters by reducing the squared error in each iteration until a convergence criterion is met.

b) MAP: From one point of view, image segmentation can be split into two categories: structural and statistical. The former is based on boundaries and regions. On the other hand, the latter is mostly based on the probability distribution function of image intensities and their associated class labels. Statistical approaches try to find the class label x, when only the intensity y for each pixel is given. MAP and maximum likelihood are widely used criteria for this kind of estimation. Using the MAP criterion, $\hat{\mathbf{x}}$ should be estimated based on

$$\hat{\mathbf{x}} = \arg\max_{x\in\chi} \{p(\boldsymbol{y}|\mathbf{x};\boldsymbol{\theta})f(\mathbf{x})\}\cdot \tag{3}$$

It is assumed that y_i and x_i are pairwise independent, so

$$p(\boldsymbol{y}|\mathbf{x},\boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i|x_i).$$

MRF can be completely explained by a Gibbs distribution using the Hammersley–Clifford theorem which describes the relation between MRF and Gibbs distribution [21]. Thus

$$f(\mathbf{x}) = \frac{1}{Z}\exp\left(-U(\mathbf{x})\right)$$

where Z is a normalizing constant and

$$U(\mathbf{x}) = \sum_{c\in C} V_c(x_i, x_j)$$

where $V_c(x_i, x_j)$ are the so-called clique potentials and C is the set of all possible cliques (see more details in [21]). A clique c is a subset of S where every pair of distinct sites is neighbors, except for single-site cliques. Fig. 2 depicts all possible cliques for the predefined neighborhood system. The general idea behind the HMRF model is that, if a pixel has a certain label, the pixels of its neighborhood system are also of that type. In this paper, it is assumed that each pixel has at most four neighbors in the image domain. Then, on pairs of neighboring pixels, the clique potentials are calculated by

$$V_c(x_i, x_j) = \frac{1}{2}(1 - I_{x_i,x_j})$$
$$I_{x_i,x_j} = \begin{cases} 0 & \text{if } x_i \neq x_j \\ 1 & \text{if } x_i = x_j. \end{cases} \tag{4}$$

MAP can be rewritten as a minimization problem

$$\hat{\mathbf{x}} = \arg\min_{x\in\chi} \{U(\boldsymbol{y}|\mathbf{x}) + U(\mathbf{x})\} \tag{5}$$

where $U(\boldsymbol{y}|\mathbf{x}) = \sum_i[((y_i - \mu_{x_i})^2/2\sigma_{x_i}^2) + (1/2)\log\sigma_{x_i}^2]$ measures the fit and $U(\mathbf{x})$ can be viewed as a penalty term that encourages spatial smoothness. The iterative MAP algorithm stops when the relative change in the cost function is below a prespecified threshold. There exist efficient algorithms for solving the MAP problem. Here, we use the same algorithms as in [13].

c) EM Algorithm: A statistical model is complete if and only if both its functional forms and parameters are determined. In HMRF, the parameter set $\boldsymbol{\theta} = \{\theta_l, l \in L\}$ should be estimated. If the Gaussian density function is assumed for the pixel intensity value y, the parameters of each Gaussian class are $\theta_l = (\mu_l, \sigma_l)$. Since both the class labels and parameters are unknown, the calculation of the parameters is not straightforward. One reliable way to solve this issue is the EM algorithm [22]. We use the EM algorithm to estimate the parameters $\boldsymbol{\theta}$. In the following discussion, the EM algorithm is briefly explained.

1) E-step: We compute the EM functional

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right) = E\left[\log p(\boldsymbol{y}, \mathbf{x}; \boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(k)}\right]. \tag{6}$$

2) M-step: For obtaining the next estimate, we maximize the EM functional

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right). \tag{7}$$

Then, let $\boldsymbol{\theta}^{(k)} \longrightarrow \boldsymbol{\theta}^{(k+1)}$ and return to the E-step.

The EM functional can be written as

$$Q = \sum_i \sum_l q^{(k)}(j|y_i)\left\{\ln q(l|x_{N_i}) - \frac{1}{2}\ln\sigma_l^2 - \frac{1}{2}\frac{(y_i - \mu_j)^2}{\sigma_l^2}\right\} \tag{8}$$

where the posterior $q^{(k)}(j|y_i)$ is obtained from the MAP step. The M-step yields the following updates:

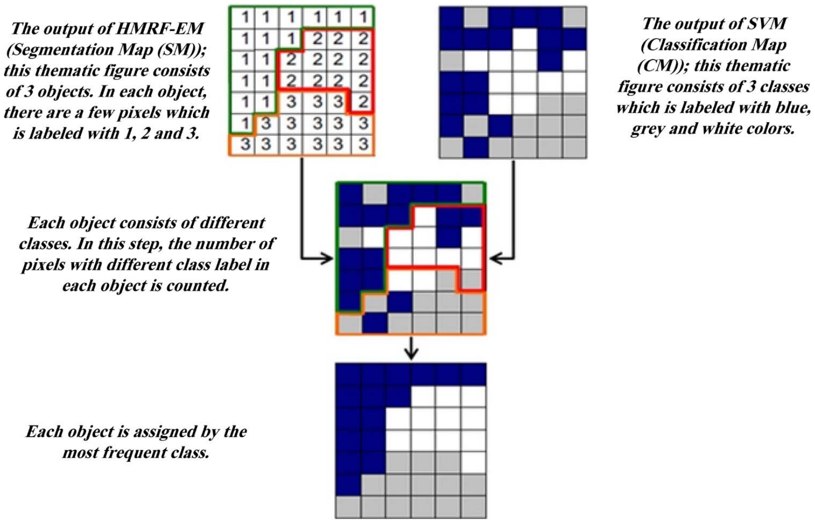
$$\mu_l^{(k+1)} = \frac{\sum_i q^{(k)}(j|y_i)y_i}{\sum_i q^{(k)}(j|y_i)} \tag{9}$$

$$\sigma_l^{2(k+1)} = \frac{\sum_i q^{(k)}(j|y_i)\left(y_i - \mu_l^{(k+1)}\right)}{\sum_i q^{(k)}(j|y_i)}. \tag{10}$$

The iterative algorithm will stop when the relative change in the cost function is less than a predefined threshold.

C. Gradient

Image segmentation provides a smoothing process. Provided that one image has strong discontinuities, MRFs may cause


The output of HMRF-EM (Segmentation Map (SM)); this thematic figure consists of 3 objects. In each object, there are a few pixels which is labeled with 1, 2 and 3.
The output of SVM (Classification Map (CM)); this thematic figure consists of 3 classes which is labeled with blue, grey and white colors.
Each object consists of different classes. In this step, the number of pixels with different class label in each object is counted.
Each object is assigned by the most frequent class.

1	1	1	1	1	1
1	1	1	2	2	2
1	1	2	2	2	2
1	1	2	2	2	2
1	1	3	3	3	2
1	3	3	3	3	3
3	3	3	3	3	3

Fig. 3. Procedure of MV for combining the spectral information and spatial information (based on [1]).

oversmoothing [24]. One way to address this issue is to combine the underlying label with an additional line process [24]. In order to preserve edges in the segmentation map (SM), the input image is first transformed by principal component analysis (PCA), and the PCs which have dominant variance (more than 99% of the total variation) are kept. Sobel edge detection is performed on each PC, and then, the output of Sobel edge-detected PCs are summed together. Finally, the output is transformed to a binary format. Let us assume that we have a binary edge map z; $z_i = 1$ if the ith pixel is edge, and $z_i = 0$ if not. In this case, (5) is modified to

$$\hat{\mathbf{x}} = \arg\min_{x \in \chi} \left\{ U(\boldsymbol{y}|\mathbf{x}) + \sum_{j \in N_i, z_i = 0} V_c\left(l, \mathbf{x}_N^{(k)}\right) \right\}. \tag{11}$$

This shows that the clique potentials are only estimated for the pixels which are not edge pixels.

D. SVM

The general idea behind SVM is to separate training samples belonging to different classes by tracing maximum margin hyperplanes in the space where the samples are mapped [25]. SVMs were originally introduced for solving linear classification problems. However, they can be generalized to nonlinear decision functions by considering the so-called kernel trick [26]. A kernel-based SVM is being used to project the pixel vectors into a higher dimensional space and to estimate the maximum margin hyperplanes in this new space in order to improve linear separability of data [26]. The sensitivity to the choice of the kernel and regularization parameters can be considered as the most important disadvantages of SVM. The latter is classically overcome by considering cross-validation techniques using training data [27]. The Gaussian radial basis function is widely used in remote sensing [26].

E. MV

In this paper, majority voting (MV) is used for combining the result of the segmentation and classification steps. Fig. 3 shows the general idea of MV. The output of the segmentation methods is a number of objects, where each object consists of several pixels with the same label. In other words, pixels in each object share the same characteristics. For performing MV on the output of the segmentation and classification steps, first, the number of pixels with different class labels in each object is counted. Then, the set of pixels in each object is assigned to the most frequent class label (coming from the classification step) in the object. Thus, each region from the SM is considered as an adaptive homogeneous neighborhood for all the pixels within this region. The described technique leads to a considerable improvement in terms of classification accuracies. In addition, MV provides more homogeneous CMs in comparison with classification methods which use local neighborhoods in order to take into account spatial information in a classifier [28]. For better understanding, the workflow of MV is given as follows.

1) The outputs of SVM (CM) and HMRF-EM (SM) are considered as the inputs for MV. SM consists of several objects (in Fig. 3, we have three different objects 1, 2, and 3), and CM consists of different classes (in Fig. 3, we have three different classes blue, gray, and white).
2) In each object, all of the pixels are assigned to the most frequent class within this object.

III. Experimental Results

Two hyperspectral data sets were used in the experiments. They are described in the following discussion.

A. Data Description

1) Indian Pines data: The first data set is the well-known AVIRIS data set captured on NW Indian Pines in 1992

TABLE I
INDIAN PINES: THE NUMBER OF TRAINING AND TEST SAMPLES; CLASSIFICATION ACCURACIES OF TEST SAMPLES IN PERCENTAGE FOR SVM, KMEANSSVM-16, HMRFSVM-NE-16, HMRFSVM-E-16, KMEANSSVM-20, HMRFSVM-NE-20, AND HMRFSVM-E-20

	Class	No. of Samples		SVM	KmeansSVM		HMRFSVM-NE		HMRFSVM-E	
No.	Name	Training	Test		16	20	16	20	16	20
1	Corn-notill	50	1384	79.1	63.7	73.0	**89.5**	88.3	85.5	89.4
2	Corn-mintill	50	784	83.4	89.4	93.3	96.1	96.0	**96.3**	95.9
3	Corn	50	184	92.9	97.2	**96.7**	92.9	94.0	96.2	94.5
4	Grass-pasture	50	447	**96.6**	95.7	95.9	95.5	93.9	96.2	95.1
5	Grass-trees	50	697	91.9	90.5	92.8	93.1	**93.8**	93.1	**93.8**
6	Hay-windrowed	50	439	96.8	**99.3**	98.4	98.4	98.6	98.4	97.7
7	Soybean-notill	50	918	84.6	**91.9**	91.2	72.6	89.4	66.9	90.8
8	Soybean-mintill	50	2418	69.1	**83.7**	83.5	81.5	80.6	82.3	82.4
9	Soybean-clean	50	564	87.2	83.5	85.6	89.1	88.6	**89.3**	88.3
10	Wheat	50	162	**99.3**	**99.3**	**99.3**	**99.3**	**99.3**	**99.3**	98.7
11	Woods	50	1244	88.5	96.1	**97.1**	96.4	96.3	96.5	96.3
12	Bldg-Grass-Tree-Drives	50	330	81.2	**93.9**	93.0	92.4	90.9	93.3	90.9
13	Stone-Steel-Towers	50	45	97.7	**100**	**100**	**100**	**100**	**100**	**100**
14	Alfalfa	15	39	89.7	76.9	92.3	**94.8**	**94.8**	89.7	**94.8**
15	Grass-pasture-mowed	15	11	90.9	**100**	90.9	90.9	90.9	90.9	90.9
16	Oats	15	5	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	Total Overall Accuracy	–	–	82.56	86.38	88.34	88.69	89.78	87.74	**90.50**
	Kappa Coefficient	–	–	0.8019	0.8446	0.8672	0.8709	0.8836	0.8601	**0.8917**

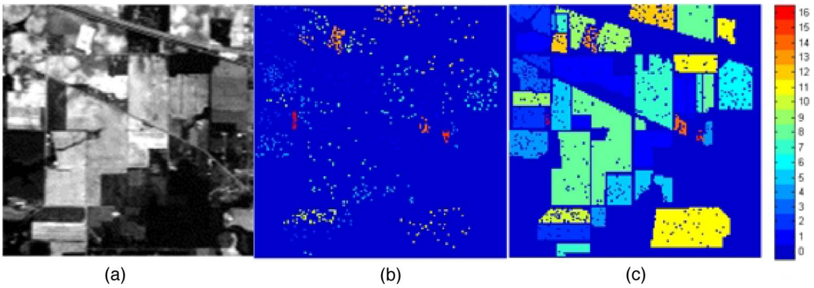


Fig. 4. Example of the Indian Pines test case. (a) Data channel 27. (b) Training samples. (c) Test samples. Each color represents a specific information class. The information classes are listed in Table I.

presenting 16 classes, mostly related to land covers. The data set consists of 145 by 145 pixels with a spatial resolution of 20 m/pixel. In this paper, we used 200 data channels, i.e., after the elimination of the bands affected by atmosphere absorption. The number of training and test samples is displayed in Table I. Fig. 4(a)–(c) illustrates one band of Indian Pines and its corresponding training and test sets, respectively.

2) Salinas data: This data set was captured by AVIRIS over Salinas Valley, CA, USA, and it is characterized by high spatial resolution (3.7-m pixels) consisting of 512 by 217 samples. The original data set consists of 224 data channels, but here, 20 water absorption bands are discarded. It includes vegetations, bare soils, and vineyard fields. The Salinas reference data contain 16 classes. Fig. 5(a) and (b) shows the Salinas data set and its corresponding reference map.

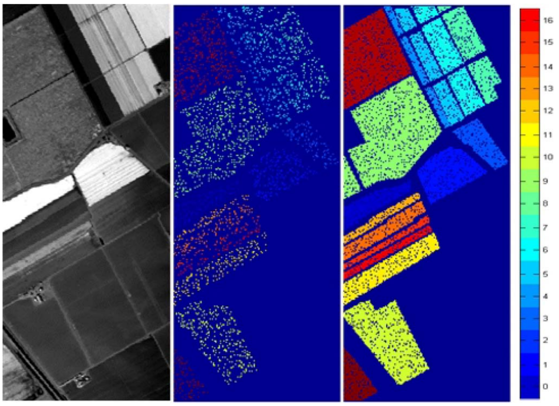


Fig. 5. Example of the Salinas test case. (a) Data channel 57. (b) Training samples. (c) Test samples. Each color represents a specific information class. The information classes are listed in Table III.

B. General Description

For the gradient step, the input image is transformed by PCA, and the first PCs with cumulative variance more than 99% are selected as the most effective components since they explain almost all of the variance in the data. Then, Sobel edge detection is performed on each component. Following that, the components are summed up, and the resulting image is transformed to binary format in order to create the gradient image.

Then, both data sets are classified by K-means, and 16 and 20 are selected as the number of clusters. Those numbers are selected in such a fashion that the former is equal to the number

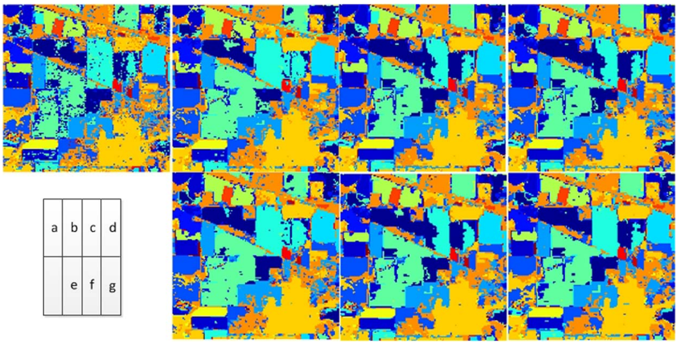


Fig. 6. CMs of different methods for Indian Pines. (a) SVM. (b) KmeansSVM-16. (c) HMRFSVM-NE-16. (d) HMRFSVM-E-16. (e) KmeansSVM-20. (f) HMRFSVM-NE-20. (g) HMRFSVM-E-20.

of classes in a reference map and the latter is superior to the minimum number in order to compare the efficiency of different methods in terms of different numbers of clusters in K-means. Ten iterations are chosen for this step, and the output of this step and the edge-detected image are regularized by HMRF-EM for providing the spatial information.

In parallel, for extracting spectral information, the data sets are classified by SVM with a Gaussian kernel. The hypertuning parameters are selected using fivefold cross-validation. To make the comparison as fair as possible, SVM is performed on each data set only once, and the CM of this step is directly used for other methods. In other words, the spectral part of all methods is the same, and only the spatial part is changed for each method.

In the final step, the results of the spectral and spatial steps are combined using the MV method, and the output of this step is the final CM.

In this paper, we use McNemar's test to assess our classification result. The aforementioned test is calculated as follows:

$$M = \frac{d_{12} - d_{21}}{\sqrt{d_{12} + d_{21}}} \tag{12}$$

where d_{12} is the number of pixels which are erroneously classified by the proposed method and not by the compared method and d_{21} has a dual meaning [29]. The differences between the proposed method and others are statistically significant at 5% significant level if $|M| > 1.96$.

In this paper, SVM denotes the traditional SVM, and HMRFSVM is the proposed method, which is the combination of HMRF and SVM. HMRFSVM-E and HMRFSVM-NE are HMRFSVM with and without including the gradient step, respectively, and 16 and 20 depict the number of predefined clusters for K-means clustering. KmeansSVM denotes a combination of K-means and SVM by using MV.

C. Results

1) Indian Pines: For the classification of Indian Pines, all of the available data channels are taken into consideration without performing feature reduction. It should be noted that all 16 classes were considered in order to evaluate the efficiency of different methods. The result of the classification for each class along with the overall accuracy and the kappa coefficient is given in Table I. Fig. 6 shows the CMs for SVM, KmeansSVM-16, HMRFSVM-NE-16, HMRFSVM-E-16, KmeansSVM-20, HMRFSVM-NE-20, and HMRFSVM-E-20.

TABLE II
INDIAN PINES: THE RESULT OF MCNEMAR'S TEST TO VALIDATE WHETHER THE DIFFERENCE BETWEEN CLASSIFICATION ACCURACIES OF THE PROPOSED METHOD WITH BOTH PREDEFINED 16 AND 20 CLUSTERS IS SIGNIFICANTLY DIFFERENT FROM OTHER METHODS

Indian Pines	M
HMRFSVM-E_16 vs. SVM	14.06
HMRFSVM-E_16 vs. KmeansSVM_16	4.24
HMRFSVM-E_16 vs. HMRFSVM-NE_16	4.41
HMRFSVM-E_20 vs. SVM	22.15
HMRFSVM-E_20 vs. KmeansSVM_20	7.11
HMRFSVM-E_20 vs. HMRFSVM-NE_20	4.71

The low spatial resolution of this data set adds more complexity since it leads to the presence of the highly mixed pixels. In this case, the unsupervised clustering (or/and clustering-based segmentation) might be degraded by spectrally mixed pixels in the image. In addition, the significant differences in the number of pixels in the reference data for different classes add more complexities on the data set and make the classification and the segmentation tasks more complicated [30].

As can be seen from Table I, the overall accuracy and kappa coefficient increase when the number of clusters increases from 16 to 20. For instance, the overall accuracies of KmeansSVM, HMRFSVM-NE, and HMRFSVM-E are improved by almost 1.9%, 1.1%, and 2.8%, respectively, when the number of clusters increases from 16 to 20. The main reason behind is undersegmentation which occurs when the number of predefined clusters is not sufficient. In this case, several regions are detected as one and merged together, which is not desired. This issue is easily solved by increasing the number of predefined clusters in the K-means.

Results confirm that the spectral and spatial classification approach using MV is able to improve the pixelwise classification accuracy considerably, particularly for the classification of large spatial structures in the data set. This fact helps in reducing the noisy behavior of the pixelwise classification significantly. However, for small structures, when the spatial information from adjacent neighbors is taken into account, the small structures are in danger of disappearing and merging with bigger structures. Accurate segmentation can improve the

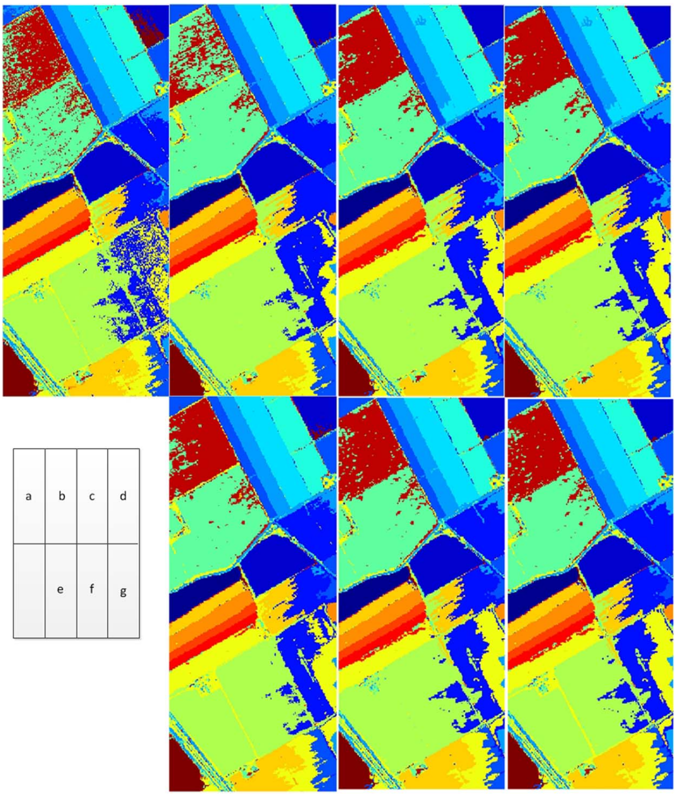


Fig. 7. Classification maps of different methods for Salinas: (a) SVM, (b) KmeansSVM-16, (c) HMRFSVM-NE-16, (d) HMRFSVM-E-16, (e) KmeansSVM-20, (f) HMRFSVM-NE-20, and (g) HMRFSVM-E-20.

spatial part of the spectral and spatial classification techniques and can help overcome the aforementioned problem.

Due to the fact that the data set contains large spatial structures and the reference data does not comprise region edges, the advantage of considering the gradient step for HMRFSVM-E compared to HMRFSVM-NE is not obvious. With reference to Table I, HMRFSVM-E-16 improves SVM and KmeansSVM by 5.1% and 1.3%, respectively. In the same way, HMRFSVM-E-20 increases the overall accuracy of the classification of SVM and KmeansSVM by 8.2% and 2.2%, respectively.

Table II shows the results from McNemar's test. As can be seen from the table, the differences in classification accuracy between the proposed method and others are statistically significant using 5% level of significance. In this case, HMRFSVM-20 is statistically different from SVM, KmeansSVM-20, and HMRFSVM-NE-20 by almost 22.15, 7.11, and 4.71 respectively.

2) Salinas: Fig. 7 shows the CMs for SVM, KmeansSVM-16, HMRFSVM-NE-16, HMRFSVM-E-16, KmeansSVM-20, HMRFSVM-NE-20, and HMRFSVM-E-20. Table III shows the classification accuracies for the approaches applied to the Salinas data. As can be seen from the table, HMRFSVM-E gives the best performance in terms of classification accuracies when compared with the other methods. For 16 clusters, HMRFSVM-E-16 improves the classification accuracies of KmeansSVM-16 and SVM by 5.7% and 2.7%, respectively. In the same way, when the number of clusters was selected as 20, the proposed method showed improvement over all studied methods. Results confirm that considering MV helps different methods to decrease the noisy behavior of the traditional SVM. The main assumption behind HMRF is that, in a predefined neighborhood structure, any given pixel is more likely to be allocated to a given cluster type if its neighboring pixels are also of that type. Therefore, it is easy to conclude that HMRF can be effective for images containing big structures.

KmeansSVM-16 shows the worst performance in terms of classification accuracies when compared to other methods. The main reasons for the bad performance of KmeansSVM-16 might be the following: 1) the spectral signature of Grapes-untrained and Vinyard-untrained are close to each other; in particular, considering only 16 clusters leads to merging of the clusters which have a close spectral response, and 2) KmeansSVM-16 does not consider spatial dependences of the image, and clustering is done by only considering the spectral information. In other words, since spatial dependences are not taken into account and the number of predefined clusters is not enough, MV is not able to determine the correct class within each segment.

As can be seen from Table IV, the differences between the proposed method which considers edges and others were

TABLE III
SALINAS: THE NUMBER OF TRAINING AND TEST SAMPLES; CLASSIFICATION ACCURACIES OF TEST SAMPLES IN PERCENTAGE FOR SVM, KMEANSSVM-16, HMRFSVM-NE-16, HMRFSVM-E-16, KMEANSSVM-20, HMRFSVM-NE-20, AND HMRFSVM-E-20

Class		No. of Samples		SVM	KmeansSVM		HMRFSVM-NE		HMRFSVM-E	
No.	Name	Training	Test		16	20	16	20	16	20
1	Brocoli_green_weeds_1	252	1757	99.5	**100**	**100**	**100**	**100**	**100**	**100**
2	Brocoli_green_weeds_2	474	3252	**100**	**100**	**100**	**100**	**100**	**100**	**100**
3	Fallow	239	1737	99.3	**99.7**	99.5	**99.7**	**99.7**	**99.7**	**99.7**
4	Fallow_rough_plow	169	1225	99.2	**99.8**	**99.8**	**99.8**	**99.8**	**99.8**	**99.8**
5	Fallow_smooth	342	2336	**99.4**	98.5	98.8	85.3	98.9	86.8	99.1
6	Stubble	516	3443	**99.9**	99.7	**99.9**	98.5	99.6	98.6	99.8
7	Celery	442	3137	99.4	99.4	99.4	99.4	99.4	99.4	**99.5**
8	Grapes_untrained	1395	9876	88.4	95.8	94.3	95.6	95.0	**96.2**	95.1
9	Soil_vinyard_develop	775	5428	**99.9**	**99.9**	**99.9**	**99.9**	99.7	**99.9**	99.6
10	Corn_senesced_green_weeds	407	2871	**97.2**	96.7	95.7	90.8	90.8	90.6	91.9
11	Lettuce_romaine_4wk	141	927	98.7	**99.0**	98.9	**99.0**	**99.0**	**99.0**	**99.0**
12	Lettuce_romaine_5wk	232	1695	99.8	99.8	99.8	**100**	99.8	**100**	**100**
13	Lettuce_romaine_6wk	124	792	**99.4**	98.9	98.8	98.6	98.8	98.8	99.0
14	Lettuce_romaine_7wk	121	949	95.4	95.4	95.8	**97.2**	96.0	97.1	96.3
15	Vinyard_untrained	906	6362	76.9	43.2	90.0	**92.9**	92.6	92.8	92.8
16	Vinyard_vertical_trellis	231	1576	98.9	98.9	98.9	**99.3**	99.0	99.1	99.2
	Total Overall Accuracy	–	–	94.02	91.01	96.93	96.57	97.10	96.76	**97.24**
	Kappa Coefficient	–	–	0.9334	0.8993	0.9658	0.9618	0.9677	0.9639	**0.9692**

TABLE IV
SALINAS: THE RESULT OF MCNEMAR'S TEST TO VALIDATE WHETHER THE DIFFERENCE BETWEEN CLASSIFICATION ACCURACIES OF THE PROPOSED METHOD WITH BOTH PREDEFINED 16 AND 20 CLUSTERS IS SIGNIFICANTLY DIFFERENT FROM OTHER METHODS

Salinas	M
HMRFSVM-E_16 vs. SVM	24.04
HMRFSVM-E_16 vs. KmeansSVM_16	41.74
HMRFSVM-E_16 vs. HMRFSVM-NE_16	5.41
HMRFSVM-E_20 vs. SVM	30.75
HMRFSVM-E_20 vs. KmeansSVM_20	3.83
HMRFSVM-E_20 vs. HMRFSVM-NE_20	3.66

TABLE V
INDIAN PINES: COMPARISON WITH THE STATE OF THE ART. THE METHODS WITH HIGHER ACCURACIES THAN THE PROPOSED APPROACH ARE SHOWN IN BOLDFACE

Method		Overall Accuracy	Kappa Coefficient
HMRFSVM-E		90.50	0.892
WH+MV [30]		89.63	0.848
EM+MV [27]		83.60	0.848
SVMMRF-E [13]		**91.83**	**0.907**
SVMMSF+MV [30]		**91.80**	**0.906**
MC-MSF [27]		86.66	0.848
MSSC-MSF [27]		**92.3**	**0.911**
M-HSEG^r [31]	SAM	77.53	0.744
S$_{wght}$ = 0.0	Inf	76.63	0.734
M-HSEG^p [31]	SAM	81.59	0.791
S$_{wght}$ = 0.0	Inf	81.16	0.786
M-HSEG^{op} [31]	SAM	89.23	0.877
S$_{wght}$ = 0.0	Inf	89.00	0.874
M-HSEG^{op} [31]	SAM	88.72	0.871
S$_{wght}$ = 0.2	Inf	89.01	0.874

significantly different when the Salinas data were clustered with 16 and 20 clusters.

D. Comparison of the Proposed Method With the State of the Art

In this section, the proposed method is compared with some recent approaches in terms of classification accuracy in order to provide a brief vision regarding the capability of HMRFSVM-E. Since Indian Pines is considered as one of best known data sets which many researchers have tested their algorithms on, that data set is used here for comparison. Table V reports the overall accuracy and kappa coefficient for the state of the art. In the following, we only analyze methods which have shown better classification accuracies than the proposed approach. The methods with better results than the proposed approach are shown in bold. For better understanding of the methods used for comparison, we refer readers to the references which can be found in front of each method in Table V.

As can be seen from Table V, the proposed method has an acceptable result in comparison with the other methods. In the following discussion, the proposed method is compared in more detail to SVMMRF-E [14], SVMMSF + MV [31], and MSSC-MSF [28].

1) HMRFSVM-E Versus SVMMRF-E: In SVMMRF-E, the input data set is at first classified by a probabilistic SVM and then regularized by MRF using a gradient step. The most important disadvantage of SVMMRF-E is that the parameter β must be carefully set, but that parameter controls the importance of the spatial energy terms versus the spectral energy term. With reference to [14], different values of β can considerably change the result of the classification, and that poses a problem for this approach. In contrast, the method proposed in this paper is fully automatic, i.e., there is no need to initialize the parameters in order to achieve good results.

2) HMRFSVM-E Versus SVMMSF + MV: SVMMSF + MV was proposed in [31]. In this method, the original data set is initially classified by using a probabilistic pixelwise classification technique. The output of this step provides both a CM and a probability map. The outputs of the first step helps one to select the most reliably classified pixels. For providing a map of *markers*, the classification and probability maps are considered to provide a connected component (CC) labeling of the CM. Then, for each CC, the region is compared to a threshold M in order to define whether the region is considered as being large or small. The M parameter is initialized by considering the resolution of the image along with typical sizes of the objects of interest. If the region is considered as small, the marker is the same with pixels of CC with probabilities more than S

percent. The S parameter is set by considering the probability of the presence of small structures in the image (which also depends on the image resolution and the classes of interests). If the region is considered as large, the marker is P (defining the percentage of pixels within the large region to be used as markers) percent of its pixels with the highest probabilities. The output of this step is a map of markers. Furthermore, the result of the previous step leads to the construction of a minimum spanning forest. Finally, MV within the CCs provides the final segmentation and CM. From the aforementioned description, it can be observed that the method is not automatic. In addition, in order to apply this method successfully, a comprehensive knowledge regarding the different structures of the input data is needed.

3) HMRFSVM-E Versus MSSC-MSF: The MSSC-MSF was introduced in [28]. In this method, the input image is at first classified by a pixelwise SVM. Second, The input image is segmented with watershed segmentation, and the result is combined with an SVM using MV. Third, the input data are segmented by EM and are combined with SVM through MV. Then, the input data set is segmented with recursive divide-and-conquer approximation of HSEG and is combined with SVM by using MV. Furthermore, the outputs of the three steps are used for marker selection. The output of this step is then used for the construction of a minimum spanning forest. Based on the aforementioned workflow, it is easy to see that MSSC-MSF is quite complicated and can become computationally very demanding without parallel processing.

IV. Conclusion

In this paper, a fully automated framework which takes into account both spectral information and spatial information has been introduced for classification of hyperspectral images. In the framework, SVM is used for the extraction of spectral information. In parallel, HMRF-EM is used for the extraction of spatial information. In the final step, those results are combined by using MV. The efficiency of the proposed method is tested in both situations with and without considering the gradient step. The proposed method is evaluated on two data sets (Indian Pines and Salinas). In both cases, the new approach outperforms other studied methods. The classification of the proposed method works better than SVM in terms of accuracies and improves the results of overall accuracy by almost 8% and 3.2% for Indian Pines and Salinas, respectively. It should be noted that the concept of HMRF is used for the first time in the field of remote sensing in this paper, and the efficiency of that for the segmentation of hyperspectral images is demonstrated. Finally, it is shown in this paper that the method performs well in terms of accuracies compared with the state of the art. In addition, the proposed approach is fully automatic and user-friendly in contrast to most of the methods.

Acknowledgment

The authors would like to thank Prof. Landgrebe from Purdue University for providing the Indian Pines data set.

References

[1] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[3] S. Tadjudin and D. Landgrebe, "Classification of high dimensional data with limited training samples," School Elect. Comput. Eng., Purdue Univ., West Lafayette, IN, USA, Tech. Rep., 1998.

[4] V. N. Vapnic, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.

[5] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. Hoboken, NJ, USA: Wiley, 2009.

[6] P. Mantero, G. Moser, and S. B. Serpico, "Partially supervised classification of remote sensing images using SVM-based probability density estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 559–570, Mar. 2005.

[7] G. Moser and S. B. Serpico, "Combining support vector machines and Markov random fields in an integrated framework for contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2734–2752, May 2013.

[8] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[9] P. Ghamisi, M. S. Couceiro, F. M. L. Martins, and J. A. Benediktsson, "Multilevel image segmentation based on fractional-order Darwinian particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2382–2394, May 2014.

[10] A. Farag, R. Mohamed, and A. El-Baz, "A unified framework for map estimation in remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1617–1634, Jul. 2005.

[11] H. Derin and P. A. Kelly, "Discrete-index Markov-type random processes," *Proc. IEEE*, vol. 77, no. 10, pp. 1485–1510, Oct. 1989.

[12] 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," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.

[13] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[14] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[15] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," in *Proc. PReMI*, 2005, pp. 260–265.

[16] D. Liu, M. Kelly, and P. Gong, "A spatial–temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery," *Remote Sens. Environ.*, vol. 101, no. 2, pp. 167–180, Mar. 2006.

[17] M. Khodadadzadeh, R. Rajabi, and H. Ghassemian, "Combination of region-based and pixel-based hyperspectral image classification using erosion technique and MRF model," in *Proc. 18th ICEE*, May 2010, pp. 294–299.

[18] G. Zhang and X. Jia, "Simplified conditional random fields with class boundary constraint for spectral–spatial based remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 5, pp. 856–860, Sep. 2012.

[19] B. Tso and R. C. Olsen, "Combining spectral and spatial information into hidden Markov models for unsupervised image classification," *Int. J. Remote Sens.*, vol. 26, no. 10, pp. 2113–2133, May 2005.

[20] G. Hazel, "Multivariate Gaussian MRF for multispectral scene segmentation and anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1199–1211, May 2000.

[21] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[22] A. P. Dempster, N. M. Laird, and D. B. Bubin, "Maximum likelihood from incomplete data via EM algorithm," *J. R. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[23] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, 1967, pp. 281–297.

[24] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE. Trans. Patt. Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.

[25] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[26] B. Scholkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[27] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for the classification of hyperspectral remote-sensing data over urban areas," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 11, Jan. 2009.

[28] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Multiple spectral–spatial classification approach for hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4122–4132, Nov. 2010.

[29] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.

[30] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.

[31] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 5, pp. 1267–1279, Oct. 2010.

[32] Y. Tarabalka, J. C. Tilton, J. A. Benediktsson, and J. Chanussot, "A marker-based approach for the automated selection of a single segmentation from a hierarchical set of image segmentations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 262–272, Feb. 2012.

Pedram Ghamisi (S'13) received the B.Sc. degree in civil (survey) engineering from Islamic Azad University, Tehran, Iran, and the M.Sc. degree in remote sensing from K. N. Toosi University of Technology, Tehran, in 2012. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Iceland, Reykjavik, Iceland.

His research interests are remote sensing and image analysis with the current focus on spectral and spatial techniques for hyperspectral image classification.

Mr. Ghamisi received the Best Researcher Award for M.Sc. students from K. N. Toosi University of Technology in 2010–2011. He serves as a reviewer for a number of journals including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Jón Atli Benediktsson (S'84–M'90–SM'99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984 and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is currently the Pro Rector for Academic Affairs and a Professor of electrical and computer engineering with the University of Iceland. He is the Cofounder of the biomedical start-up company Oxymap (www.oxymap.com). He is on the International Editorial Board of the *International Journal of Image and Data Fusion*. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in these fields.

Prof. Benediktsson was the 2011–2012 President of the IEEE Geoscience and Remote Sensing Society (GRSS), and he has been on the GRSS AdCom since 2000. He is a Fellow of SPIE. He is a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica, and Tau Beta Pi. He was the Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008, and he has served as an Associate Editor of TGRS since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and IEEE ACCESS since 2013. He was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2007–2010. He received the Stevan J. Kristof Award from Purdue University in 1991 as outstanding graduate student in remote sensing. He was the recipient of the Icelandic Research Council's Outstanding Young Researcher Award in 1997, he was granted the IEEE Third Millennium Medal in 2000, he was the corecipient of the University of Iceland's Technology Innovation Award in 2004, he received the yearly research award from the Engineering Research Institute of the University of Iceland in 2006, and he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society in 2007. He was the corecipient of the 2012 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Paper Award.

Magnus Orn Ulfarsson (S'00–M'02) received the B.S. and M.S. degrees from the University of Iceland, Reykjavik, Iceland, in 2002 and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA, in 2007.

He joined the University of Iceland in 2007, where he is currently an Associate Professor. His research interest includes statistical signal processing, image processing, remote sensing, genomics, and functional magnetic resonance imaging.

CHAPTER 3

Spectral-Spatial Classification based on an Adaptive Neighborhood System: Segmentation

Expert Systems with Applications 39 (2012) 12407–12417

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An efficient method for segmentation of images based on fractional calculus and natural selection

Pedram Ghamisi [a,*], Micael S. Couceiro [b,c], Jón Atli Benediktsson [d], Nuno M.F. Ferreira [c,e]

[a] Geodesy & Geomatics Engineering Faculty, K. N. Toosi University of Technology, Tehran, Iran
[b] Institute of Systems and Robotics, University of Coimbra, Pólo II, 3030-290 Coimbra, Portugal
[c] RoboCorp at the Electrical Engineering Department, Engineering Institute of Coimbra, Rua Pedro Nunes – Quinta da Nora, 3030-199 Coimbra, Portugal
[d] Faculty of Electrical and Computer Engineering, University of Iceland, Saemundargotu 2, 101 Reykjavik, Iceland
[e] GECAD – Knowledge Engineering and Decision Support Research Center Institute of Engineering, Polytechnic of Porto (ISEP/IPP) Porto, Portugal

ARTICLE INFO

Keywords:
Multilevel segmentation
Swarm Optimization
Image processing

ABSTRACT

Image segmentation has been widely used in document image analysis for extraction of printed characters, map processing in order to find lines, legends, and characters, topological features extraction for extraction of geographical information, and quality inspection of materials where defective parts must be delineated among many other applications. In image analysis, the efficient segmentation of images into meaningful objects is important for classification and object recognition. This paper presents two novel methods for segmentation of images based on the *Fractional-Order Darwinian Particle Swarm Optimization* (*FODPSO*) and *Darwinian Particle Swarm Optimization* (*DPSO*) for determining the *n*-1 optimal *n*-level threshold on a given image. The efficiency of the proposed methods is compared with other well-known thresholding segmentation methods. Experimental results show that the proposed methods perform better than other methods when considering a number of different measures.

1. Introduction

Image segmentation is the process of partitioning a digital image into multiple regions. In other words, image segmentation could assign a label to each pixel in the image such that pixels with the same label share certain visual characteristics. These objects contain more information than individual pixels since the interpretation of images based on objects is more meaningful than that based on individual pixels. Image segmentation is considered as an important basic task in the analysis and understanding of images, thus being widely used for further image processing purposes such as classification and object recognition (Sezgin & Sankur, 2004).

Image segmentation can be classified into four different types including texture analysis based methods, histogram thresholding based methods, clustering based methods and region based split and merging methods (Brink, 1995). One of the most common methods for the segmentation of images is the thresholding method, which is commonly used for segmentation of an image into two or more clusters (Kulkarni & Venayagamoorthy, 2010).

Thresholding techniques can be divided into two different types: optimal thresholding methods (Kapur, Sahoo, & Wong, 1985; Kittler & Illingworth, 1986; Otsu, 1979; Pun, 1980; Pun, 1981) and property-based thresholding methods (Lim & Lee, 1990; Tsai, 1995 & Yin & Chen, 1993). The former group search for the optimal thresholds which make the thresholded classes on the histogram reach the desired characteristics. Usually, it is made by optimizing an objective function. The latter group detects the thresholds by measuring some selected property of the histogram. Property-based thresholding methods are fast, which make them suitable for the case of multilevel thresholding. However the number of thresholds is hard to determine and needs to be specified in advance.

Several algorithms have been proposed in literature that addressed the issue of optimal thresholding (Brink, 1995; Cheng, Chen, & Li, 1998; Huang & Wang, 1995; Hu, Hou, & Nowinski, 2006; Kapur et al., 1985; Li, Zhao, & Cheng, 1995; Otsu, 1979; Pun, 1980; Saha & Udupa, 2001; Tobias & Seara, 2002; Yin & Chen, 1993). While many of them address the issue of bi-level thresholding, others have considered the multi-level problem. The problem of bi-level thresholding is reduced to an optimization problem to probe for the threshold t that maximizes the σ_B^2 and minimizes σ_W^2 (Kulkarni & Venayagamoorthy, 2010). For two level thresholding, the problem is solved by finding T^* which satisfies max

[*] Corresponding author. Address: No. 11, Fotouhi st., Farabi st., Ardakani st., Guyabadi st., Zafar st., Shariati st. Tehran, Iran Postal code: 1916759333. Tel.: +98 912 207 82 63.

E-mail addresses: p.ghamisi@gmail.com (P. Ghamisi), micaelcouceiro@isr.uc.pt, micael@isec.pt (M.S. Couceiro), benedikt@hi.is (J.A. Benediktsson), nunomig@isec.pt (N.M.F. Ferreira).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.078

($\sigma_B^2(T^*)$) where $0 \leqslant T^* < L$ and L is the maximum intensity value. This problem could be extended to n-level thresholding through satisfying max σ_B^2 $(T_1^*, T_2^*, \ldots, T_{n-1}^*)$ that $0 \leqslant T_1^* < T_2^* < \cdots < T_{n-1}^* < L$. One way for finding the optimal set of thresholds is the exhaustive search method. The exhaustive search method based on the Otsu criterion (Otsu, 1979) is simple, but it has a disadvantage that it is computationally expensive (Kulkarni & Venayagamoorthy, 2010). Exhaustive search for $n-1$ optimal thresholds involves evaluations of fitness of $n(L-n+1)^{n-1}$ combinations of thresholds (Kulkarni & Venayagamoorthy, 2010) so this method is not suitable from a computational cost point of view. The task of determining $n-1$ optimal thresholds for n-level image thresholding could be formulated as a multidimensional optimization problem.

Alternative to the Otsu method, several biologically inspired algorithms have been explored in image segmentation (Fogel, 2000; Kulkarni & Venayagamoorthy, 2010; Lai & Tseng, 2004; Yin, 1999). Bio-inspired algorithms have been used in situations where conventional optimization techniques cannot find a satisfactory solution, for example when the function to be optimized is discontinuous, non-differentiable, and/or presents too many nonlinearly related parameters (Floreano & Mattiussi, 2008). The Particle Swarm Optimization (PSO) is a machine-learning technique loosely inspired by birds flocking in search of food (Kennedy & Eberhart, 1995). It basically consists of a number of particles that collectively move in the search space (*e.g.*, pixels of the image) in search of the global optimum (*e.g.*, maximizing the between-class variance of the distribution of intensity levels in the given image). However, a general problem with the PSO and other optimization algorithms is that of becoming trapped in a local optimum, such that it may work in some problems but may fail on others (Couceiro, Ferreira, & Machado, 2011).

The Darwinian Particle Swarm Optimization (*DPSO*) was formulated by Tillett et al. in 2005 (Tillett, Rao, Sahin, Rao, & Brockport, 2005) in search of a better model of natural selection using the *PSO* algorithm. In this algorithm, multiple swarms of test solutions performing just like an ordinary *PSO* may exist at any time with rules governing the collection of swarms that are designed to simulate natural selection. More recently, an extension of the *DPSO* using fractional calculus to control the convergence rate of the algorithm was presented by Couceiro et al. in 2011 (Couceiro et al., 2011), being denoted as fractional-order *DPSO* (*FODPSO*). The novel algorithm was successfully compared with both the fractional-order *PSO* from Pires, Machado, Oliveira, Cunha, and Mendes (2010) and the traditional *DPSO*.

Significant progress has been made in the creative inspiration of bio-inspired computer algorithms applied to optimization, estimation, control and many others through the application of principles derived from the study of biology (Floreano & Mattiussi, 2008). Santana, Alves, Correia, and Barata (2010) presented a swarm-based model for trail detection in real-time. Experimental results on a large dataset revealed the ability of the model to produce a success rate of 91% using a 20 Hz camera with a resolution of 640 × 480 that was carried through a scenario at an approximate speed of 1 m s^{-1}. The authors in Kulkarni and Venayagamoorthy (2010) compared the *PSO* and Bacteria Foraging algorithm (*BF*) with the Otsu method to determine the optimal threshold level for the deployment of sensor nodes. It should be noted that all methods were run offline and the *PSO* presented a superior performance when compared to the Otsu and the *BF*. Omran (2004) presented the application of the *PSO* to the field of pattern recognition and image processing. He introduced a clustering algorithm based on *PSO*. Further, he developed a dynamic clustering algorithm that could find the "optimum" number of clusters in a dataset with minimum user interference. Sathya and Kayalvizhi (2010) proposed a multilevel thresholding method based on *PSO* and compared their method with *GA*-based thresholding method. Results showed that the *PSO*-based image segmentation executed faster and was more stable than *GA*.

This paper mainly focuses on using one of the best performing *PSO* main variants (*cf.*, (Couceiro, Luz, Figueiredo, Ferreira, & Dias, 2010; Couceiro et al., 2011), created by Tillett et al. (2005)), denoted as *DPSO*, and the recently fractional-order extension, denoted as *FODPSO* (Couceiro et al., 2011). This is the first work to verify and apply the *FODPSO* and *DPSO* to multilevel segmentation. Bearing this idea in mind, the problem formulation of image n-level thresholding is presented in the following sub-sections. Section 2 presents a brief review of particle swarm algorithms, focusing on the strengths and weaknesses of the traditional *PSO*, the *DPSO* and the *FODPSO*. In Section 3, several images used to compare the *PSO*-based image segmentation variants with other commonly used algorithm such as genetic algorithms (*GA*) and *BF*. Finally, Section 4 outlines the main conclusions.

1.1. Image thresholding

Multilevel segmentation techniques provide an efficient way to perform image analysis. However, the automatic selection of a robust optimum n-level threshold has remained a challenge in image segmentation. This section presents a more precise formulation of the problem, introducing some basic notation.

Let there be L intensity levels in each *RGB* (red-green–blue) component of a given image and these levels are in the range $\{0,1,2,\ldots,L-1\}$. Then one can define:

$$p_i^C = \frac{h_i^C}{N}, \quad \sum_{\substack{i=1 \\ C=\{R,G,B\}}}^{N} p_i^C = 1, \tag{1}$$

where i represents a specific intensity level, *i.e.*, $0 \leqslant i \leqslant L-1$,$C$ represents the component of the image, *i.e.*, $C=\{R,G,B\}$,N represents the total number of pixels in the image and h_i^C denotes the number of pixels for the corresponding intensity level i in the component C. In other words, h_i^C represents an image histogram for each component C, which can be normalized and regarded as the probability distribution p_i^C. The total mean (*i.e.*, combined mean) of each component of the image can be easily calculated as:

$$\mu_T^C = \sum_{\substack{i=1 \\ C=\{R,G,B\}}}^{L} i p_i^C. \tag{2}$$

The 2-level thresholding can be extended to generic n-level thresholding in which $n-1$ threshold levels $t_j^C, j = 1, \ldots, n-1$, are necessary and where the operation is performed as expressed below:

$$F^C(x,y) = \begin{cases} 0, & f^C(x,y) \leqslant t_1^C \\ \frac{1}{2}\left(t_1^C + t_2^C\right), & t_1^C < f^C(x,y) \leqslant t_2^C \\ & \vdots \\ \frac{1}{2}\left(t_{n-2}^C + t_{n-1}^C\right), & t_{n-2}^C < f^C(x,y) \leqslant t_{n-1}^C \\ L, & f^C(x,y) > t_{n-1}^C \end{cases} \tag{3}$$

where x and y are the width (W) and height (H) pixel of the image of size $H \times W$ denoted by $f^C(x,y)$ with L intensity levels in each *RGB* component. In this situation, the pixels of a given image will be divided into n classes $D_1^C, \ldots, D_n^C$, which may represent multiple objects or even specific features on such objects (*e.g.*, topological features).

The simplest and computationally most efficient method of obtaining the optimal threshold is the one that maximizes the between-class variance which can be generally defined by:

$$\sigma_B^{C^2} = \sum_{\substack{j=1 \\ C=\{R,G,B\}}}^{n} w_j^C \left(\mu_j^C - \mu_T^C\right)^2, \tag{4}$$

where j represents a specific class in such a way that w_j^C and μ_j^C are the probability of occurrence and mean of class j, respectively. The probabilities of occurrence w_j^C of classes $D_1^C, \ldots, D_n^C$ are given by:

$$w_j^C = \begin{cases} \sum\limits_{\substack{i=1 \\ C=\{R,G,B\}}}^{t_j^C} p_i^C, & j=1 \\ \sum\limits_{\substack{i=t_{j-1}^C+1 \\ C=\{R,G,B\}}}^{t_j^C} p_i^C, & 1<j<n \\ \sum\limits_{\substack{i=t_{j-1}^C+1 \\ C=\{R,G,B\}}}^{L} p_i^C, & j=n \end{cases}, \tag{5}$$

The mean of each class μ_j^C can then be calculated as:

$$\mu_j^C = \begin{cases} \sum\limits_{\substack{i=1 \\ C=\{R,G,B\}}}^{t_j^C} \frac{ip_i^C}{w_j^C}, & j=1 \\ \sum\limits_{\substack{i=t_{j-1}^C+1 \\ C=\{R,G,B\}}}^{t_j^C} \frac{ip_i^C}{w_j^C}, & 1<j<n \\ \sum\limits_{\substack{i=t_{j-1}^C+1 \\ C=\{R,G,B\}}}^{L} \frac{ip_i^C}{w_j^C}, & j=n \end{cases}, \tag{6}$$

In other words, the problem of n-level thresholding is reduced to an optimization problem to search for the thresholds t_j^C that maximizes the three objective functions (*i.e.*, fitness function) of each *RGB* component, generally defined as:

$$\varphi^C = \max_{\substack{1 < t_1^C < \ldots < t_{n-1}^C < L \\ C=\{R,G,B\}}} \sigma_B^{C^2}\left(t_j^C\right). \tag{7}$$

Computing this optimization problem involves a much larger computational effort as the number of threshold levels increase. This brings us to the question: which kind of method should be used to solve this optimization problem for real-time applications?

Many methods have been proposed in the literature (Sezgin & Sankur, 2004). However, more recently, biologically inspired methods have been used as computationally efficient alternatives to analytical methods to solve optimization problems (Couceiro et al., 2010; Couceiro et al., 2010).

1.2. Efficiency evaluation

The computational time is one of the most important indicators along with fitness value which determine the ability of the algorithm. Provided that the data is large, the efficiency of the method is restricted to a great extent (Fan, Han, & Wang, 2009). For instance, remote sensing (*RS*) data, in particular hyperspectral images, are considerably large most of the time so using a high speed and efficient algorithm is highly preferable. Moreover, in real-time applications, using a high-speed algorithm is the main objective (Kulkarni & Venayagamoorthy, 2010). As a result, the evaluation of the *CPU* process time and fitness value seems vitally important to show the efficiency of the new method. In addition, since all bio-inspired methods are random and stochastic, the results are not completely the same in each run. Consequently, the stability of different methods should be evaluated by an appropriate index such as standard deviation value which will be described in Section 3.

PSO-based segmentation algorithms have been one of the most used in recent years. In fact, the traditional PSO-based segmentation has already been compared with GA-based algorithms or even exhaustive ones and has been found to present better results and being faster than both. In Hammouche, Diaf, and Siarry (2010), PSO-based segmentation method was superior compared to other algorithms such as *GA*, Differential Evaluation (*DE*), Ant Colony Optimization (*ACO*), Simulated Annealing (*SA*) and Tabu Search (*TS*) in terms of precision, robustness of the results and runtime. In Sathya and Kayalvizhi (2010), the authors show that *PSO* outperforms *GA* in terms of *CPU* time and fitness value for Kapur's and Otsu's functions. In Jiang, Luo, and Yang (2007), results show that *PSO*-family methods act better than *GA* with a learning operator (*GA-L*) in different measures. As a result, it is easy to detect that *PSO*-based segmentation methods are considered an efficient way in terms of finding optimal thresholds in less *CPU* process time. In Kulkarni and Venayagamoorthy (2010), *PSO* has been shown to be significantly faster than *BF* and exhaustive methods. As a result, comparison of the new methods with *PSO* in terms of *CPU* process time can be completely satisfying.

In this paper, two novel methods for segmentation of images based on DPSO and *FODPSO* are proposed. Both *DPSO* and *FODPSO* were introduced by Tillett et al. (2005) and Couceiro et al. (2011), respectively, for optimization of some primary test functions. We herein extend the concept of the algorithms to image segmentation. Therefore, they will be used to solve the Otsu problem for delineating multilevel threshold values. In other words, proposing a new thresholding based segmentation method which is robust in terms of the results and runtime, makes up our main goal. Further, in order to show the advantages of the new methods, we compare both algorithms with other algorithms that have been commonly used in the literature to determine the $n-1$ optimal n-level threshold on given images.

2. A brief review of the algorithms

The original *PSO* algorithm was developed by Eberhart and Kennedy in 1995 (Kennedy & Eberhart, 1995). The *PSO* basically takes advantage of the swarm intelligence concept, which is the property of a system whereby the collective behaviors of unsophisticated agents that are interacting locally with their environment, create coherent global functional patterns (Del Valle, Venayagamoorthy, Mohagheghi, Hernandez, & Harley, 2008). Imagine a flock of birds where each bird cries at an intensity proportional to the amount of food that it finds at its current location. At the same time each bird can perceive the position of neighboring birds and can tell which of the neighboring birds emits the loudest cry. There is a good chance that the flock will find a spot with the highest concentration of food if each bird follows a trajectory that combines three rules: (i) keep flying in the same direction; (ii) return to the location where it found the highest concentration of insects so far; and (iii) move toward the neighboring bird that cries the loudest (Kulkarni & Venayagamoorthy, 2010).

2.1. Particle Swarm Optimization (PSO)

In the traditional *PSO*, the candidate solutions are called particles. These particles travel through the search space to find an optimal solution, by interacting and sharing information with neighbor particles, namely their individual best solution (local best) and computing the neighborhood best. Also, in each step of the procedure, the global best solution obtained in the entire swarm is updated. Using all of this information, particles realize the loca-

Table 1
The *PSO* algorithm.

```
Initialize swarm (Initialize $x_t^n$, $v_t^n$, $\breve{x}_t^n$, $\breve{n}_t^n$ and $\breve{g}_t^n$)
Loop:
  for all particles
    Evaluate the fitness $\varphi^C$ of each particle
    Update $\breve{x}_t^n$, $\breve{n}_t^n$ and $\breve{g}_t^n$
  Update $v_t^n$ and $x_t^n$
  end
until stopping criteria (convergence)
```

tions of the search space where success was obtained, and are guided by these successes.

In each step of the algorithm (Table 1), a fitness function is used to evaluate the particle success. To model the swarm, each particle n moves in a multidimensional space according to position (x_t^n) and velocity (v_t^n) values which are highly dependent on local best $\left(\breve{x}_t^n\right)$, neighborhood best $\left(\breve{n}_t^n\right)$ and global best $\left(\breve{g}_t^n\right)$ information:

$$v_{t+1}^n = wv_t^n + \rho_1 r_1\left(\breve{g}_t^n - x_t^n\right) + \rho_2 r_2\left(\breve{x}_t^n - x_t^n\right) + \rho_3 r_3\left(\breve{n}_t^n - x_t^n\right), \quad (8)$$

$$x_{t+1}^n = x_t^n + v_{t+1}^n. \quad (9)$$

The coefficients w, ρ_1, ρ_2 and ρ_3 assign weights to the inertial influence, the global best, the local best and the neighborhood best when determining the new velocity, respectively. Typically, the inertial influence is set to a value slightly less than 1. ρ_1, ρ_2 and ρ_3 are constant integer values, which represent "cognitive" and "social" components. However, different results can be obtained by assigning different influences for each component. For example, some methods do not consider the neighborhood best and ρ_3 is set to zero. Depending on the application and the characteristics of the problem, tuning these parameters properly will lead to better results. The parameters r_1, r_2 and r_3 are random vectors with each component generally a uniform random number between 0 and 1. The intent is to multiply a new random component per velocity dimension, rather than multiplying the same component with each particle's velocity dimension.

The particles in the *PSO* are evaluated for the fitness function, which is defined as the between-class variance σ_B^2 of the image-intensity distributions previously represented in (7).

In the beginning, the particles' velocities are set to zero and their position is randomly set within the boundaries of the search space. The search space will depend on the number of intensity levels L, *i.e.*, if the frames are 8-bit images then the particles will be deployed between 0 and 255.

The local, neighborhood and global bests are initialized with the worst possible values, taking into account the nature of the problem. The other parameters that need to be adjusted are population size and stopping criteria. The population size is very important to optimize to get an overall good solution in an acceptable time limit. Stopping criteria can be a predefined number of iterations without getting better results or other criteria, depending on the problem.

PSO reveals an effect of implicit communication between particles (similar to broadcasting) by updating neighborhood and global information, which affect the velocity and consequent position of particles. Also, there is a stochastic exploration effect due to the introduction of the random multipliers (r_1,r_2 and r_3).The *PSO* has been successfully used in many applications such as robotics (Couceiro, Luz, Figueiredo, & Ferreira, 2012; Pires, Oliveira, Machado, & Cunha, 2006; Tang, Zhu, & Sun, 2005), electric systems (Alrashidi & El-Hawary, 2006; Del Valle et al., 2008) and sports engineering (Couceiro et al., 2010).

2.2. Darwinian PSO

However, a general problem with the *PSO* and other optimization algorithms is that of becoming trapped in a local optimum such that it may work well on one problem but may fail on another problem. In order to overcome this problem many authors have suggested other adjustments to the parameters of the *PSO* algorithm combining fuzzy logic (*FAPSO*) where the inertia weight w is dynamically adjusted using fuzzy "*IF–THEN*" (Hammouche et al., 2010) rules or Gaussian approaches (*GPSO*) where the inertia constant w is no longer needed and the acceleration constants ρ_1, ρ_2 and ρ_3 are replaced by random numbers with Gaussian distributions (Jiang et al., 2007). More recently, Pires et al. used fractional calculus to control the convergence rate of the *PSO* (Sabatier et al., 2007). The authors rearrange the original velocity equation (8) in order to modify the order of the velocity derivative. Alternatively, many authors have considered incorporating selection, mutation and crossover, as well as the *DE*, into the *PSO* algorithm. The main goal is to increase the diversity of the population by either preventing the particles to move too close to each other and collide (Machado et al., 2010; Ortigueira & Tenreiro Machado, 2003) or to self-adapt parameters such as the constriction factor, acceleration constants (Podlubny, 1999), or inertia weight (Debnath, 2003).

The fusion between *GA* and the *PSO* originated the *GA-PSO* (Couceiro, Ferreira, & Machado, 2010) which combines the advantages of swarm intelligence and a natural selection mechanism, such as *GA*, in order to increase the number of highly evaluated agents, while decreasing the number of lowly evaluated agents at each iteration step. Similar to this last one, the *EPSO* is an evolutionary approach that incorporates a selection procedure to the original *PSO* algorithm, as well as self-adapting properties for its parameters. This algorithm adds a tournament selection method used in evolutionary programming (*EP*) (Pires et al., 2010). Based on the *EPSO*, a differential evolution operator has been proposed to improve the performance of the algorithm in two different ways. The first one (Yasuda, Iwasaki, Ueno, & Aiyoshi, 2008) applies the differential evolution operator to the particle's best position to eliminate the particles falling into local minima (*DEPSO*) while the second one (Venter & Sobieszczanski-Sobieski, 2002) applies it to find the optimal parameters (inertia and acceleration constants) for the canonical *PSO* (*C-PSO*).

In search of a better model of natural selection using the *PSO* algorithm, the Darwinian Particle Swarm Optimization (*DPSO*) was formulated (Tillett et al., 2005), in which many swarms of test solutions may exist at any time. Each swarm individually performs just like an ordinary *PSO* algorithm with rules governing the collection of swarms that are designed to simulate natural selection. Despite the similarities between the *PSO* and *GA*s, like randomly generated population, fitness function evaluation, population update, search for optimality with random techniques and not guaranteeing success; *PSO* does not use genetic operators like crossover and mutation, thus

Table 2
The *DPSO* algorithm.

Main program loop	Evolve swarm algorithm
For each swarm in the collection	For each particle in the swarm
Evolve the swarm (Evolve Swarm Algorithm: right)	Update Particles' Fitness
Allow the swarm to spawn	Update Particles' Best
Delete "failed" swarms	Move Particle
	If swarm gets better
	Reward swarm: spawn particle: extend swarm life
	If swarm has not improved
	Punish swarm: possibly delete particle: reduce swarm life

not being considered an evolutionary technique. On the other hand, the Darwinian Particle Swarm Optimization (*DPSO*) extends the *PSO* to determine if natural selection (Darwinian principle of survival of the fittest) can enhance the ability of the *PSO* algorithm to escape from local optima. The idea is to run many simultaneous parallel *PSO* algorithms, each one a different swarm, on the same test problem and a

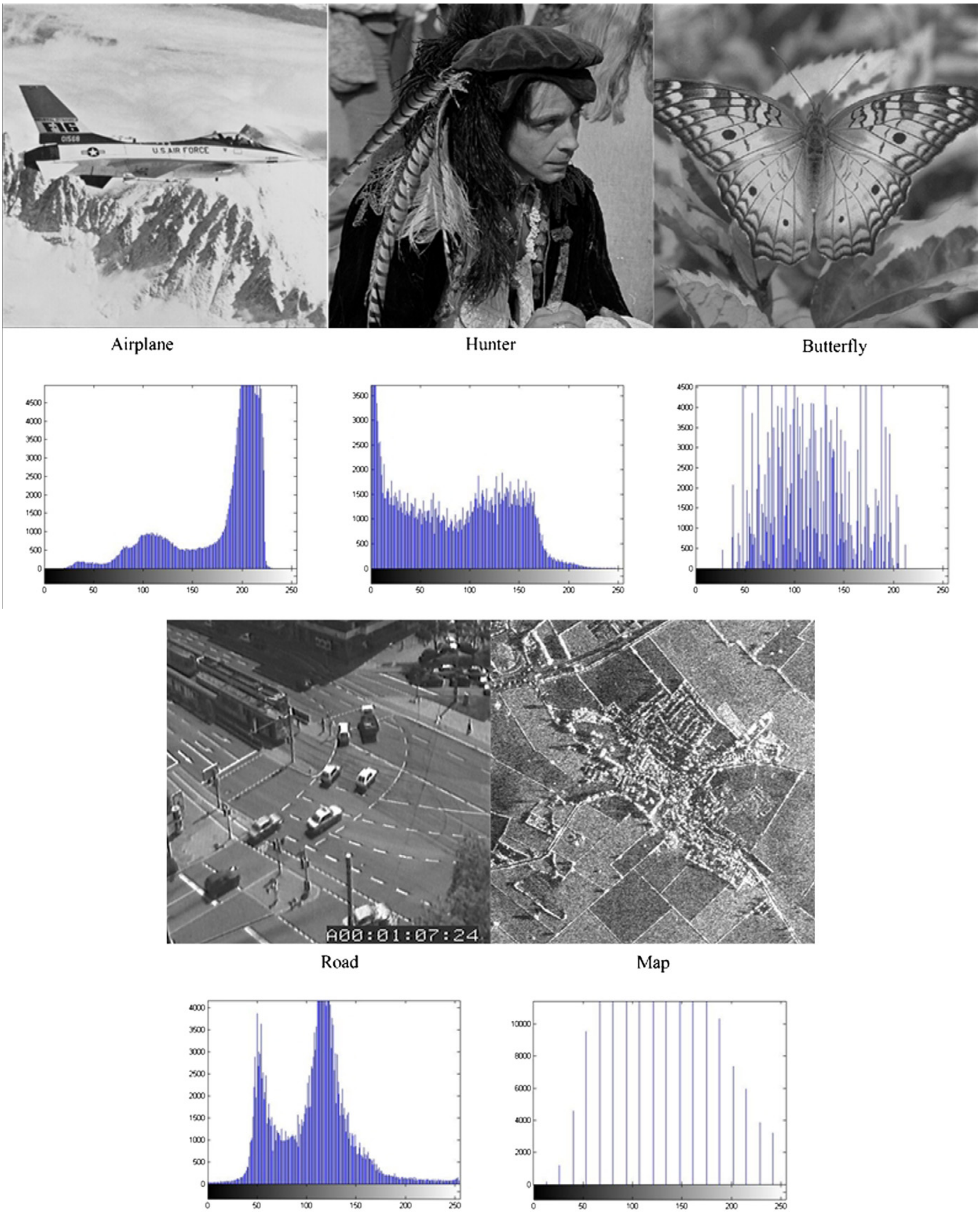


Fig. 1. Different test cases with their histograms.

Table 3
Initial parameters of the *PSO*, *DPSO* and *FODPSO*.

Parameter	PSO	DPSO	FODPSO
Num of Iterations	8	8	8
Population	200	30	30
ρ_1	1.5	1.5	1.5
ρ_2	1.5	1.5	1.5
W	1.2	1.2	1.2
Vmax	2	2	2
Vmin	−2	−2	−2
$X_{\max}$	255	255	255
$X_{\min}$	0	0	0
Min population	–	10	10
Max population	–	50	50
Num of swarms	–	4	4
Min swarms	–	2	2
Max swarms	–	6	6
Stagnancy	–	10	10
Fractional coefficient	–	–	0.75

simple selection mechanism is applied. When a search tends to a local optimum, the search in that area is simply discarded and another area is searched instead. In this approach, at each step, swarms that get better are rewarded (extend particle life or spawn a new descendent) and swarms which stagnate are punished (reduce swarm life or delete particles). To analyze the general state of each swarm, the fitness of all particles is evaluated and the neighborhood and individual best positions of each of the particles are updated. If a new global solution is found, a new particle is spawned. A particle is deleted if the swarm fails to find a fitter state in a defined number of steps (Table 2).

Some simple rules are followed to delete a swarm, delete particles, and spawn a new swarm and a new particle: (i) when the swarm population falls below a minimum bound, the swarm is deleted; and (ii) the worst performing particle in the swarm is deleted when a maximum threshold number of steps (search counter $SC_C^{\max}$) without improving the fitness function is reached. After the deletion of the particle, instead of being set to zero, the counter is reset to a value approaching the threshold number, according to:

$$SC_C(N_{kill}) = SC_C^{max}\left[1 - \frac{1}{N_{kill}+1}\right], \tag{10}$$

where N_{kill} is the number of particles deleted from the swarm over a period in which there was no improvement in fitness. To spawn a new swarm, a swarm must not have any particle ever deleted and the maximum number of swarms must not be exceeded. Still, the new swarm is only created with a probability of $p = f/NS$, with f a random number in [0,1] and *NS* the number of swarms. This factor avoids the creation of newer swarms when there are large numbers of swarms in existence. The parent swarm is unaffected and half of the parent's particles are selected at random for the child swarm and half of the particles of a random member of the swarm collection are also selected. If the swarm initial population number is not obtained, the rest of the particles are randomly initialized and added to the new swarm. A particle is spawned whenever a swarm achieves a new global best and the maximum defined population of a swarm has not been reached. Like the *PSO*, a few parameters also need to be adjusted to run the algorithm efficiently: (i) initial swarm population; (ii) maximum and minimum swarm population; (iii) initial number of swarms; (iv) maximum and minimum number of swarms; and (v) stagnancy threshold. In estimation problems previously studied (Del Valle et al., 2008) and robotic exploration strategies developed (Alrashidi & El-Hawary, 2006), the *DPSO* has been successfully compared with the *PSO* showing a superior performance.

2.3. Fractional-Order Darwinian PSO

The *FODPSO* presented in Couceiro et al. (2011) is an extension of the *DPSO* in which fractional calculus is used to control the convergence rate of the algorithm. Fractional calculus (*FC*) has attracted the attention of several researchers (Machado et al., 2010; Ortigueira & Tenreiro Machado, 2003; Sabatier et al., 2007), being applied in various scientific fields such as engineering, computational mathematics, fluid mechanics, among others (Couceiro et al., 2010; Debnath, 2003; Pires et al., 2010; Podlubny, 1999). The discrete time implementation of the *Grünwald–Letnikov* definition based on the concept of fractional differential with $\alpha \in \mathrm{C}$ of the signal $x(t)$, is given by:

$$D^{\alpha}[x(t)] = \frac{1}{T^{\alpha}}\sum_{k=0}^{r}\frac{(-1)^k\Gamma(\alpha+1)x(t-kT)}{\Gamma(k+1)\Gamma(\alpha-k+1)}, \tag{11}$$

Table 4
Average ± *STD* fitness values of different methods for different test cases.

Test image	Thresholds	FODPSO	DPSO	BF	PSO	GA
Airplane	2	1837.7974 ± 0.0140	1837.7787 ± 0.0298	1837.7517 ± 0.1021	1837.7222 ± 0.0796	1837.7144 ± 0.8355
	3	1911.6564 ± 0.0557	1911.5429 ± 0.1287	1910.7434 ± 1.5123	1905.7664 ± 1.2742	1844.5642 ± 2.0290
	4	1954.7374 ± 0.1624	1954.5612 ± 0.3566	1954.2480 ± 1.8991	1953.8872 ± 2.6057	1950.5919 ± 5.2334
	5	1979.6190 ± 0.2067	1978.7698 ± 0.5637	1978.4335 ± 2.1062	1977.9742 ± 3.1647	1973.0894 ± 7.4164
Hunter	2	3064.2066 ± 0.0292	3064.1684 ± 0.0470	3064.1188 ± 0.0322	3064.0688 ± 0.2534	3064.0156 ± 0.3781
	3	3213.2101 ± 0.1217	3212.9363 ± 0.1930	3213.4460 ± 0.9627	3212.0585 ± 1.5406	3211.7947 ± 2.0141
	4	3269.0574 ± 0.3526	3268.4573 ± 0.6478	3266.3504 ± 2.2936	3257.1767 ± 3.2342	3231.1313 ± 5.0298
	5	3307.5841 ± 0.7660	3305.6159 ± 1.6202	3291.1339 ± 3.6102	3276.3173 ± 4.1811	3244.7387 ± 9.7412
Butterfly	2	1553.0732 ± 0.0010	1553.0615 ± 0.0426	1553.0734 ± 0.0643	1553.0687 ± 0.0846	1552.4129 ± 1.5470
	3	1669.1929 ± 0.0706	1669.0419 ± 0.3586	1667.2801 ± 1.2113	1665.7589 ± 2.6268	1662.6963 ± 3.4022
	4	1710.6595 ± 0.4651	1709.9903 ± 0.6253	1707.0994 ± 2.2120	1702.9069 ± 3.7679	1696.6940 ± 5.3135
	5	1735.8941 ± 0.5968	1734.4957 ± 1.4495	1733.0317 ± 3.5217	1730.7879 ± 6.0747	1716.0428 ± 7.5842
Road	2	1321.3721 ± 0.0083	1321.3351 ± 0.0202	1321.3366 ± 0.0386	1321.1132 ± 0.0969	1320.7661 ± 0.8599
	3	1433.8901 ± 0.1036	1433.7674 ± 0.1643	1430.8712 ± 0.7123	1425.3853 ± 1.0941	1418.4695 ± 3.6425
	4	1490.1314 ± 0.2609	1489.5771 ± 0.9921	1488.2286 ± 1.9801	1483.1709 ± 2.8351	1476.7349 ± 5.8790
	5	1519.7184 ± 0.3494	1518.8896 ± 1.1087	1518.3493 ± 2.2354	1511.7474 ± 4.0074	1500.8104 ± 8.2580
Map	2	2340.3950 ± 0.0000	2340.3950 ± 0.0000	2340.3950 ± 0.0026	2340.3950 ± 0.0000	2252.3864 ± 1.2171
	3	2529.9384 ± 0.0000	2529.9384 ± 0.0000	2529.9348 ± 0.0548	2526.3034 ± 1.1180	2503.7932 ± 2.1368
	4	2621.1476 ± 0.0000	2621.1476 ± 0.0000	2621.1476 ± 0.3710	2618.4894 ± 2.9722	2617.9534 ± 3.7246
	5	2670.0640 ± 0.0000	2670.0640 ± 0.1969	2668.0699 ± 1.2189	2665.4116 ± 3.8519	2660.8599 ± 5.4901

Table 5
Average thresholds of different segmentation algorithm.

Image	Thresholds	FODPSO	DPSO	BF	PSO	GA
Airplane	2	116, 174	116, 174	117, 175	117, 174	116, 175
	3	93, 145, 190	95, 148, 193	91, 147, 190	99, 158, 193	86, 133, 204
	4	88, 132, 175, 203	88, 132, 174, 204	84, 127, 169, 202	84, 125, 168, 201	71, 119, 164, 200
	5	70, 106, 144, 180, 205	74, 109, 148, 181, 205	71, 110, 138, 175, 203	60, 101, 138, 177, 204	84, 124, 164, 188, 204
Hunter	2	51, 116	51, 116	51, 117	52, 116	51, 115
	3	35, 86, 134	35, 87, 134	36, 86, 135	39, 86, 135	36, 89, 133
	4	31, 72, 110, 146	26, 63, 102, 141	31, 80, 120, 152	36, 84, 130, 157	39, 93, 142, 163
	5	21, 53, 89, 122, 152	22, 52, 90, 118, 149	31, 73, 109, 141, 178	37, 85, 125, 154, 177	39, 94, 130, 169, 204
Butterfly	2	99, 151	98,151	99, 151	99, 150	100, 151
	3	82, 119, 159	81, 119, 160	78, 117, 162	79, 119, 164	74, 115, 155
	4	69, 98, 126, 162	71, 100, 130, 163	75, 105, 135, 165	80, 113, 145, 177	82, 119, 154, 184
	5	72, 98, 125, 150, 180	74, 103, 126, 153, 179	76, 104, 129, 154, 180	75, 106, 129, 157, 180	77, 107, 134, 171, 185
Road	2	90, 154	90, 154	90, 155	91, 155	90, 151
	3	86, 129, 186	86, 130, 187	89, 133,180	85, 127, 169	77, 121, 184
	4	72, 105, 135, 190	72, 106, 139, 193	78, 111, 139, 189	78, 114, 147, 205	74, 97, 139, 205
	5	69, 98, 124, 151, 202	73, 102, 124, 155, 205	70, 102, 128, 159, 211	71, 103, 134, 173, 225	79, 109, 142, 179, 204
Map	2	110, 186	113, 177	109, 176	113, 177	81, 173
	3	95, 148, 201	95, 140, 197	98, 146, 189	81, 145, 197	83, 132, 181
	4	81, 122, 169, 218	88, 122, 173, 224	88, 134, 173, 222	92, 133, 162, 206	90, 110, 158, 204
	5	80, 112, 140, 179, 218	76, 120, 145, 186, 221	80, 109, 135, 165, 224	79, 116, 139, 162, 204	68, 106, 138, 170, 214

where Γ is the gamma function, T is the sampling period and r is the truncation order.

An important property revealed by the *Grünwald–etnikov* is that while an integer-order derivative just implies a finite series, the fractional-order derivative requires an infinite number of terms. Therefore, integer derivatives are 'local' operators while fractional derivatives have, implicitly, a 'memory' of all past events. However, the influence of past events decreases over time.

The characteristics revealed by fractional calculus make this mathematical tool well suited to describe phenomena such as irreversibility and chaos because of its inherent memory property. In this line of thought, the dynamic phenomena of particle's trajectory configure a case where fractional calculus tools fit adequately.

Considering the inertial influence of (8) as $w = 1$, assuming $T = 1$ and similarly to Pires et al. (2010) work, the following expression can be defined:

$$D^{\alpha}\left[v^n_{t+1}\right] = \rho_1 r_1\left(\breve{g}^n_t - x^n_t\right) + \rho_2 r_2\left(\breve{x}^n_t - x^n_t\right) + \rho_3 r_3\left(\breve{n}^n_t - x^n_t\right). \quad (12)$$

Preliminary experimental tests in Couceiro et al. (2011) presented similar results for $r \geqslant 4$. Furthermore, the computational requirements increase linearly with r, *i.e.*, the *FODPSO* present a $\mathcal{O}(r)$ memory complexity. Hence, using only the first $r = 4$ terms of differential derivative given by (11) and (8) can be rewritten as (13):

$$v^n_{t+1} = \alpha v^n_t + \frac{1}{2}\alpha v^n_{t-1} + \frac{1}{6}\alpha(1-\alpha)v^n_{t-2} + \frac{1}{24}\alpha(1-\alpha)(2-\alpha)v^n_{t-3} \\ + \rho_1 r_1\left(\breve{g}^n_t - x^n_t\right) + \rho_2 r_2\left(\breve{x}^n_t - x^n_t\right) + \rho_3 r_3\left(\breve{n}^n_t - x^n_t\right). \quad (13)$$

The *DPSO* is then considered as being a particular case of the *FODPSO* when $\alpha = 1$ (without 'memory'). Hence, the value of α greatly affect the inertial particles. With a small α, particles will ignore their previous activities, thus ignoring the system dynamics and being susceptible to get stuck in local solutions (*i.e.*, exploitation behavior). On the other hand, with a large α, particles will present a more diversified behavior which allows exploring new solutions, thus improving the long-term performance (*i.e.*, exploration behavior). However, if the exploration level is too high, then the algorithm may take too much time to find the global solution. Based on the experimental results from Couceiro et al. (2011), it will be used a fractional coefficient of $\alpha = 0.6$, thus resulting in a balance between exploitation and exploration.

3. Experimental results

DPSO- and *FOPSO*-based image segmentation which proposed in this paper was programmed in *MATLAB* on a computer having *Intel Core 2 Duo T5800* processor (2.00 *GHz*) and 3*GB* of memory. The proposed methods are tested on a few common images including: *Airplane*, *Hunter*, *Butterfly*, *Road* and *Map*. Fig. 1 illustrates different test cases along with the histograms of the images. The efficiency of the proposed methods is evaluated by comparing their results with a few popular methods such as *GA*, *BF* and *PSO*.

The *PSO*, *DPSO* and *FODPSO* methods are parameterized algorithms. Therefore, one needs to be able to choose the parameter values that would result in faster convergence (Table 3). The cognitive, social and inertial weights were chosen taking into account several works focusing on the convergence analysis of the traditional *PSO* (*cf.*, Jiang et al. (2007) and Couceiro et al. (2011)).

Table 6
The average CPU process time of different segmentation methods.

Test image	No. of thresholds	FODPSO (*S*)	DPSO (*S*)	PSO (*S*)
Airplane	2	0.4221	0.4382	0.4127
	3	0.4623	0.4844	0.4936
	4	0.5438	0.5516	0.6057
	5	0.5890	0.6065	0.7020
Hunter	2	0.3871	0.3966	0.3927
	3	0.4544	0.4761	0.4941
	4	0.5479	0.5517	0.5973
	5	0.5913	0.6031	0.6956
Butterfly	2	0.3661	0.3805	0.3854
	3	0.4483	0.4689	0.4925
	4	0.5038	0.5235	0.6018
	5	0.5731	0.5811	0.7001
Road	2	0.3674	0.3875	0.3903
	3	0.4569	0.4726	0.4896
	4	0.5071	0.5241	0.5949
	5	0.5680	0.5696	0.6930
Map	2	0.3542	0.3463	0.3828
	3	0.4214	0.4366	0.4858
	4	0.4907	0.5047	0.5877
	5	0.5589	0.5621	0.6869

3.1. Fitness evaluation

Since all the optimization methods are stochastic and random population-based, each of them runs 20 times and the average and standard deviation fitness values are brought in Table 4. All fitness values are calculated for 2, 3, 4, 5 thresholds. It is noteworthy that, despite small differences, all algorithms seem to reach the vicinities of the optimal solution, *i.e.*, higher between-class variance. Nevertheless, those differences are more evident in most situations as the number of thresholds increase. The *FODPSO* leads with a slightly higher fitness value than the *DPSO*.

Note that they both use natural selection in order to avoid stagnation. However, the *FODPSO* has a fractional order mechanism that allows controlling the convergence rate of particles, thus presenting a more exploiting behavior when near the solution vicinities. In regards to the differences of *PSO* family and *GA*, the *PSO*

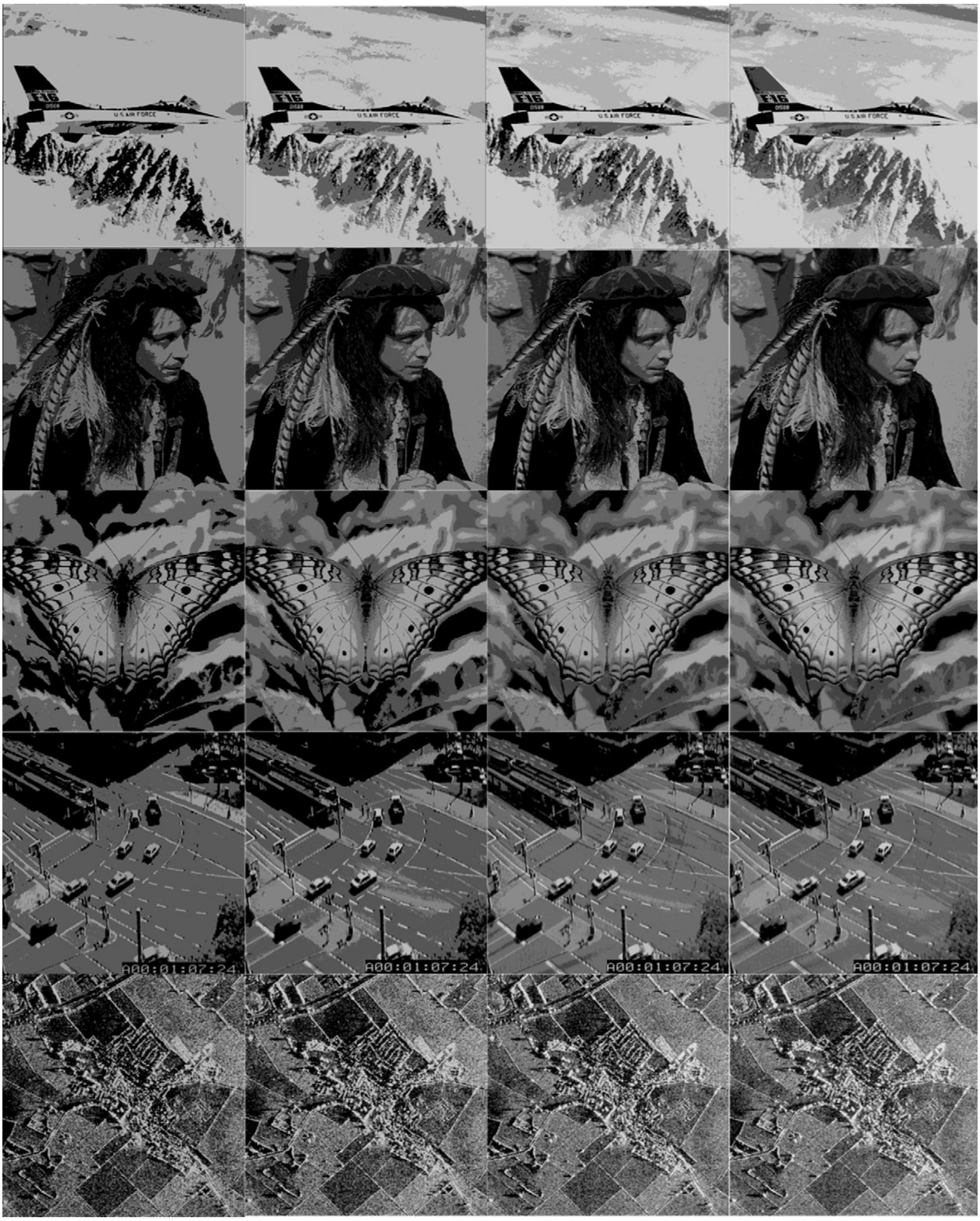


Fig. 2. The result of segmentation with 2, 3, 4, 5 thresholds, respectively (from left to right).

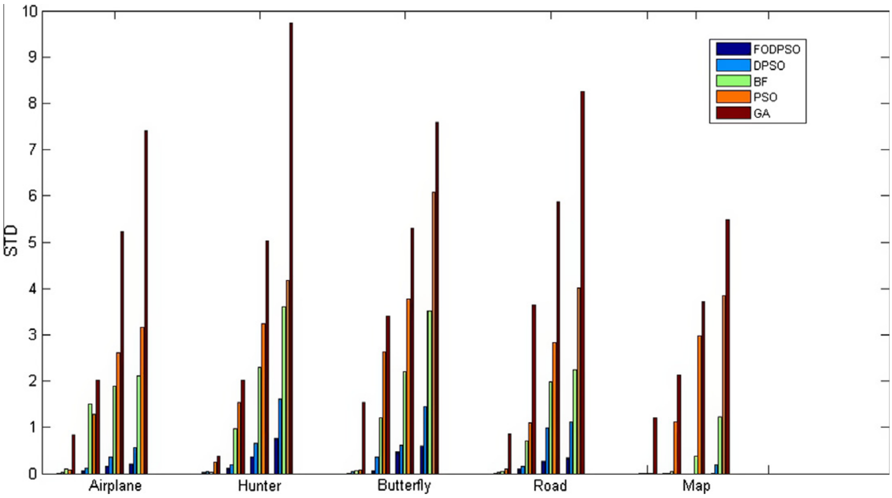


Fig. 3. Different *STD* values for different algorithms in face of different test cases (for each test case, 2, 3, 4, 5 thresholds, respectively (from left to right)).

family is an inherently continuous algorithm where as a *GA* is an inherently discrete algorithm (Venter & Sobieszczanski-Sobieski, 2002) and experiments conducted by Veeramachaneni, Peram, Mohan, and Osadciw (2003) showed that a *PSO* performed better than *GA*s when applied on some continuous optimization problems. In addition, *PSO* family was compared with a *GA* by Eberhart and Shi (1998) and Kennedy and Spears (1998). The results showed that *PSO* is generally faster and more robust to local solutions than *GA*s, especially when the dimension of a problem increases. As a result, when the number of dimension increases, a significant difference between the fitness values of the *PSO* family and *GA* happens and the *PSO* family shows better results than *GA* in higher dimensions. However, the tradition *PSO* suffers from premature convergence. Consequently, *BF* acts better than *PSO* in most cases. Table 5 demonstrates the optimal threshold values for the different methods.

3.2. CPU processing time

With regard to the *CPU* processing time, the *PSO* has been proven in the literature to require less *CPU* processing time for finding thresholds in comparison to *GA* and *BF* (*e.g.*, Sathya & Kayalvizhi, 2011). Therefore, we only compare the *CPU* time of *PSO*, *DPSO* and *FODPSO*. That brings us to Table 6 in which the *FODPSO* presents the best processing time, *i.e.*, it is able to reach its solution in less *CPU* time than *PSO* and *DPSO*. The *DPSO* still presents a lower *CPU* time than the *PSO* especially for higher threshold numbers. Nevertheless, it still needs more time to reach its solution than the *DPSO*. This is a small repercussion of having an exploitation activity when near the solution – a high level of exploitation allows a good short-term performance but slows down the convergence in order to reach a more feasible solution.

To visually compare the segmented results of different test cases by *FODPSO*, the segmented images with various threshold levels are given in Fig. 2. As can be seen from the figure, images with higher level of segmentation have more detail than the other images. In contrast, the 3 level segmented image is considered as the roughest image in different test cases. It is easy to conclude that by increasing the level of segmentation, the segmented image includes more detail. As a result, the 6-level segmented image in different test cases is smoother than the 3-level one.

3.3. Stability of different methods

Since almost all evolutionary methods are stochastic and random, the results are not completely the same in each run. Consequently, their results are affected by the nature and ability of the method. As a result, it seems necessary to evaluate the stability of the population-based algorithms. The comparison of the outputs gives us valuable information in terms of the stability of different algorithms, and which thresholding method is more suitable for segmentation of the images.

To evaluate the stability of the algorithm, the following index is used:

$$STD = \sqrt{\sum_{i=1}^{n} \frac{(\sigma_i - \mu)^2}{N}} \tag{14}$$

where *STD* is the standard deviation, σ_i is the best fitness value of the *i*th run of the algorithm, μ is the average value of σ and *N* is the repeated times of each algorithm ($N = 20$). It is easy to detect that the higher amount of *STD* represent more instability of the algorithm. The standard deviation of the different evolutionary algorithms for 20 runs, with 2, 3, 4, 5 thresholds are given in Table 4. From the Table 4, it can be seen that the *FODPSO* is the most stable evolutionary algorithms in comparison with others.

To improve the understanding of Table 4, Fig. 3 shows the standard deviation fitness values for the several algorithms in face of different test cases with different levels. As can be seen, in all experiments, *GA* shows the least stability among other bio-inspired method. According to the result, *FODPSO* is the most stable algorithm since illustrates the least *STD* values in comparison with other methods. *DPSO* and *BF* make up the second and third orders in terms of stability. In other words, the *FODPSO*-based segmentation is able to converge in approximately the same amount of time regardless on the image and the initial condition of particles.

Despite the observation that both *FODPSO* and *DPSO* present similar results, it is noteworthy that the fractional order algorithm is able to reach a slightly better fitness solution in less time. This should be highly appreciated as many applications require real-time segmentation methods such as the autonomous deployment of sensor nodes in a given environment or the detection of flaws in quality inspection of materials. In addition, *FODPSO* is slightly faster than *DPSO* since fractional calculus is used to control the convergence rate of the algorithm. As described in Yasuda et al. (2008), a swarm behavior can be divided into two activities: (i) *exploitation*; and (ii) *exploration*. The first one is related with the convergence of the algorithm, thus allowing a good short-term performance. However, if the exploitation level is too high, then the algorithm may be stuck on local solutions. The second one is related with the diversification of the algorithm which allows exploring new solutions, thus improving the long-term performance. However, if the exploration level is too high, then the algorithm may take too much time to find the global solution. In the *DPSO*, the trade-off between exploitation and exploration can only be handled by adjusting the inertia weight. While a large inertia weight improves exploration activity, the exploitation may be improved using a small inertia weight. Since the *FODPSO* presents a fractional calculus strategy to control the convergence of particles with memory effect, the coefficient α allows providing a higher level of exploration while ensuring the global solution of the algorithm.

4. Conclusion

In this paper, two new methods for segmentation of images were proposed which is based on *Fractional-Order Darwinian Particle Swarm Optimization* (*FODPSO*) and *Darwinian Particle Swarm Optimization* (*DPSO*). The new methods were used for solving the Otsu problem for delineating multilevel threshold values and to overcome the disadvantages of previous evolutionary methods in terms of trapping in local optimum and high *CPU* process time. In this paper, the fitness value, *STD* and *CPU* process time were selected as the measures for comparing the output of different methods. Results indicate that *FODPSO* is more efficient than other methods in particular, when the level of segmentation increases, thus being able to find the better thresholds with more stability in less *CPU* processing time. As future research direction, the *FODPSO* will be evaluated in remote sensing applications and further compared with exhaustive methods. Moreover, due to the low computational complexity of the algorithm, a future direction may be the use of the *FODPSO* method in image segmentation applications for the real-time autonomous deployment and distributed localization of sensor nodes from an unmanned aerial vehicle (*UAV*).

References

Alrashidi, M. R., & El-Hawary, M. E. (2006). A survey of particle swarm optimization applications in power system operations. *Electric Power Component Systems, v34 i12*, 1349–1357.

Brink, A. D. (1995). Minimum spatial entropy threshold selection. *IEE Proceedings on Vision Image and Signal Processing, 142*, 128–132.

Cheng, H. D., Chen, J., & Li, J. (1998). Threshold selection based on fuzzy c-partition entropy approach. *Pattern Recognition, 31*, 857–870.

Couceiro, M. S., Luz, J. M. A., Figueiredo, C. M., Ferreira, N. M. F., & Dias, G. (2010). Parameter estimation for a mathematical model of the golf putting. In *WACI'10 – Proceedings of workshop applications of computational intelligence ISEC-IPC, December 2, Coimbra, Portugal* (pp. 1-8).

Couceiro, M. S., Ferreira, N. M. F., & Machado, J. A. T. (2011). In *Fractional order Darwinian Particle Swarm Optimization, FSS'11 – Symposium on fractional signals and systems, November 4–5, Coimbra, Portugal.*

Couceiro, M. S., Ferreira, N. M. F., & Machado, J. A. T. (2010). Application of fractional algoritms in the control of a robotic bird. *Journal of Communications in Nonlinear Science and Numerical Simulation – Special Issue, 15*(4), 895–910.

Couceiro, M. S., Luz, J. M. A., Figueiredo, C. M., & Ferreira, N. M. F. (2012). Modeling and control of biologically inspired flying robots. *Robotica* (Vol. 30, pp. 107–121). Cambridge University Press. 1.

Debnath, L. (2003). Recent applications of fractional calculus to science and engineering. *International Journal of Mathematics and Mathematical Sciences, 54*, 3413–3442.

Del Valle, Y., Venayagamoorthy, G. K., Mohagheghi, S., Hernandez, J. C., & Harley, R. G. (2008). Particle swarm optimization: Basic concepts, variants and applications in power systems. *IEEE Transactions on Evolutionary Computation, 12*(2), 171–195.

Eberhart, R., & Shi, Y. (1998). Comparison between genetic algorithms and particle swarm optimization. In *Proceedings of the seventh annual conference on evolutionary programming* (pp. 611–619). Springer-Verlag.

Fan, J., Han, M., & Wang, J. (2009). Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation. *Pattern Recognition, 42*, 2527–2540.

Floreano, D., & Mattiussi, C. (2008). *Bio-inspired artificial intelligence: Theories, methods, and technologies*. Cambridge, MA: MIT Press.

Fogel, D. B. (2000). *Evolutionary computation: Toward a new philosophy of machine intelligence* (Second ed.). Piscataway, NJ: IEEE Press.

Hammouche, K., Diaf, M., & Siarry, P. (2010). A comparative study of various meta-heuristic techniques applied to the multilevel thresholding problem. *Engineering Applications of Artificial Intelligence, 23*, 676–688.

Huang, L. K., & Wang, M. J. (1995). Image thresholding by minimizing the measure of fuzziness. *Pattern Recognition, 28*, 41–51.

Hu, Q., Hou, Z., & Nowinski, W. (2006). Supervised range-constrained thresholding. *IEEE Transactions on Image Processing, 15*, 228–240.

Jiang, M., Luo, Y. P., & Yang, S. Y. (2007). Stochastic convergence analysis and parameter selection of the standard particle swarm optimization algorithm. *Information Processing Letters, 102*(1), 8–16.

Kapur, J. N., Sahoo, P. K., & Wong, A. K. C. (1985). A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision Graphics Image Processing, 2*, 273–285.

Kennedy, J., & Eberhart, R. (1995). A new optimizer using particle swarm theory. In *Proceedings of the IEEE sixth international symposium on micro machine and human science* (pp. 39–43).

Kennedy, J., & Spears, W. (1998). Matching algorithms to problems: An experimental test of the particle swarm and some genetic algorithms on the multimodal problem generator. In *IEEE international conference on evolutionary computation, Achorage, Alaska, USA*.

Kittler, J., & Illingworth, J. (1986). Minimum error thresholding. *Pattern Recognition, 19*, 41–47.

Kulkarni, R. V., & Venayagamoorthy, G. K. (2010). Bio-inspired algorithms for autonomous deployment and localization of sensor. *IEEE Transactions on Systems, 40*(6), 663–675.

Kulkarni, R. V., & Venayagamoorthy, G. K. (2010). Bio-inspired algorithms for autonomous deployment and localization of sensor nodes. *IEEE Transactions, SMC-C40*(6), 663–675.

Lai, C. C., & Tseng, D. C. (2004). A hybrid approach using Gaussian smoothing and genetic algorithm for multilevel thresholding. *International Journal of Hybrid Intelligent Systems, 1*(3), 143–152.

Lim, Y. K., & Lee, S. U. (1990). On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. *Pattern Recognition, 23*, 935–952.

Li, X., Zhao, Z., & Cheng, H. D. (1995). Fuzzy entropy threshold approach to breast cancer detection. *Information Sciences, 4*, 49–56.

Machado, J. A. T., Silva, M. F., Barbosa, R. S., Jesus, I. S., Reis, C. M., Marcos, M. G., et al. (2010). Some applications of fractional calculus in engineering. *Hindawi Publishing Corporation Mathematical Problems in Engineering*, 1–34.

Omran, M. G. H. (2004). Particle swarm optimization methods for pattern recognition and image processing. PhD Thesis, University of Pretoria, Pretoria.

Ortigueira, M. D., & Machado, J. A. T. (2003). Special Issue on fractional signal processing. *Signal Process, 83*, 2285–2480.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, Cybernetics, SMC-9*, 62–66.

Pires, E. J. S., Oliveira, P. B. M., Machado, J. A. T., & Cunha, J. B. (2006). Particle Swarm Optimization versus genetic algorithm in manipulator trajectory planning. In *7th Portuguese conference on automatic control, September 11–13*.

Pires, E. J. S., Machado, J. A. T., Oliveira, P. B. M., Cunha, J. B., & Mendes, L. (2010). Particle swarm optimization with fractional-order velocity. *Journal on Nonlinear Dynamics, 61*, 295–301.

Podlubny, I. (1999). *Fractional differential equations. Mathematics in Science and Engineering* (Vol. 198). San Diego, California: Academic Press.

Pun, T. (1980). A new method for grey-level picture thresholding using the entropy of the histogram. *Signal Processing, 2*, 223–237.

Pun, T. (1981). Entropy thresholding: A new approach. *Computer Vision Graphics Image Processing, 16*, 210–239.

Sabatier, J., Agrawal, O. P., & Tenreiro Machado, J. A. (Eds.). (2007). *Advances in Fractional Calculus - Theoretical Developments and Applications in Physics and Engineering*. Berlin: Springer. ISBN:978-1-4020-6041-0.

Saha, P. K., & Udupa, J. K. (2001). Optimum image thresholding via class uncertainty and region homogeneity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 689–706.

Santana, P., Alves, N., Correia, L., & Barata, J. (2010). Swarm-based visual saliency for trail detection. In *Proceedings of the IEEE/RSJ 2010 international conference on intelligent robots and systems (IROS 2010), Taiwan*.

Sathya, P. D., & Kayalvizhi, R. (2010). PSO based tsallistresholding selection procedure for image segmentation. *International Journal of Computer Applications, 5*(4), 39–46.

Sathya, P. D., & Kayalvizhi, R. (2011). Modified bacterial foraging algorithm based multilevel thresholding for image segmentation. *Journal Engineering Applications of Artificial Intelligence, 24*(4).

Sezgin, M., & Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging, 13*(1), 146–168.

Tang, J., Zhu, J., & Sun, Z. (2005). A novel path panning approach based on appart and particle swarm optimization. In *Proceedings of the 2nd international symposium on neural networks, LNCS* (Vol. 3498, pp. 253–258).

Tillett, J., Rao, T. M., Sahin, F., Rao, R., & Brockport, S. (2005). Darwinian Particle Swarm Optimization. In *Proceedings of the 2nd Indian international conference on artificial intelligence* (pp. 1474–1487).

Tobias, O. J., & Seara, R. (2002). Image segmentation by histogram thresholding using fuzzy sets. *IEEE Transactions on Image Processing, 11*, 1457–1465.

Tsai, D. M. (1995). A fast thresholding selection procedure for multimodal and unimodal histograms. *Pattern Recognition Letters, 16*, 653–666.

Veeramachaneni, K., Peram, T., Mohan, C., & Osadciw, L. (2003). Optimization using particle swarm with near neighbor interactions. *Lecture notes computer science* (Vol. 2723). Springer-Verlag.

Venter, G., & Sobieszczanski-Sobieski, J. (2002). Particle swarm optimization. In *The 43rd AIAA/ASME/ASCE/AHA/ASC structures, structural dynamics and materials conference, Denver, Colorado, USA*.

Yasuda, K., Iwasaki, N., Ueno, G., & Aiyoshi, E. (2008). Particle swarm optimization: A numerical stability analysis and parameter adjustment based on swarm activity. *IEEE Transactions on Electrical and Electronic Engineering, 3*, 642–659.

Yin, P. Y. (1999). A fast scheme for optimal thresholding using genetic algorithms. *Signal Processing, 72*, 85–95.

Yin, P. Y., & Chen, L. H. (1993). New method for multilevel thresholding using the symmetry and duality of the histogram. *Journal of Electronics and Imaging, 2*, 337–344.

Multilevel Image Segmentation Based on Fractional-Order Darwinian Particle Swarm Optimization

Pedram Ghamisi, *Student Member, IEEE*, Micael S. Couceiro, *Student Member, IEEE*,
Fernando M. L. Martins, and Jón Atli Benediktsson, *Fellow, IEEE*

***Abstract*—Hyperspectral remote sensing images contain hundreds of data channels. Due to the high dimensionality of the hyperspectral data, it is difficult to design accurate and efficient image segmentation algorithms for such imagery. In this paper, a new multilevel thresholding method is introduced for the segmentation of hyperspectral and multispectral images. The new method is based on fractional-order Darwinian particle swarm optimization (FODPSO) which exploits the many swarms of test solutions that may exist at any time. In addition, the concept of fractional derivative is used to control the convergence rate of particles. In this paper, the so-called Otsu problem is solved for each channel of the multispectral and hyperspectral data. Therefore, the problem of n-level thresholding is reduced to an optimization problem in order to search for the thresholds that maximize the between-class variance. Experimental results are favorable for the FODPSO when compared to other bioinspired methods for multilevel segmentation of multispectral and hyperspectral images. The FODPSO presents a statistically significant improvement in terms of both CPU time and fitness value, i.e., the approach is able to find the optimal set of thresholds with a larger between-class variance in less computational time than the other approaches. In addition, a new classification approach based on support vector machine (SVM) and FODPSO is introduced in this paper. Results confirm that the new segmentation method is able to improve upon results obtained with the standard SVM in terms of classification accuracies.**

***Index Terms*—Classification, image processing, multilevel segmentation, swarm optimization.**

Manuscript received December 27, 2012; revised April 4, 2013; accepted April 24, 2013. Date of publication June 5, 2013; date of current version February 27, 2014. This work was supported in part by the Icelandic Research Fund for Graduate Students, Ph.D. scholarship, (SFRH/BD /73382/2010) from the Portuguese Foundation for Science and Technology (FCT), the Institute of Systems and Robotics (ISR), the Institute of Telecommunications (IT-Covilhã), and RoboCorp.

P. Ghamisi and J. A. Benediktsson are with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavik, Iceland (e-mail: p.ghamisi@gmail.com; benedikt@hi.is).

M. S. Couceiro is with the Institute of Systems and Robotics-DEEC, Polo II, Pinhal de Marrocos, University of Coimbra, 3030-290 Coimbra, Portugal and also with the RoboCorp, Engineering Institute of Coimbra, Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal (e-mail: micaelcouceiro@isr.uc.pt).

F. M. L. Martins is with the Instituto de Telecomunicações (Covilhã), Coimbra College of Education, 6201-001 Coimbra, Portugal and also with the RoboCorp, Engineering Institute of Coimbra, Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal (e-mail: fmlmartins@esec.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2260552

I. Introduction

IMAGE segmentation is regarded as the process of partitioning a digital image into multiple regions or objects. In other words, in image segmentation, a label is assigned to each pixel in the image such that pixels with the same label share certain visual characteristics [1]. These objects provide more information than individual pixels since the interpretation of images based on objects is more meaningful than the interpretation based on individual pixels only. Image segmentation is considered as an important task in the analysis, interpretation, and understanding of images and is also widely used for image processing purposes such as classification and object recognition [1]–[3].

Image segmentation plays a key role in the field of remote sensing image analysis. For example, in order to improve classification results, the integration of classification and segmentation steps has recently been taken into account [4], [5]. In such cases, the decision to assign a pixel to a specific class is simultaneously based on the feature vector of this pixel and some additional information derived from the segmentation step. To make this approach effective, an accurate segmentation of the image is needed. A few methods for segmentation of multispectral and hyperspectral images have been introduced in the literature. Some of these methods are based on region-merging techniques, in which neighboring image segments are merged with each other based on their homogeneity. For example, the multiresolution segmentation method in eCognition software uses this type of approach [6]. Tilton proposed a hierarchical segmentation algorithm [7], which alternately performs region growing and spectral clustering. There are an extensive number of segmentation methods that have been proposed in the literature that exploit mathematical morphology approaches [8]–[16], for segmentation of multispectral and hyperspectral images.

Image segmentation can be classified into four specific types including histogram-thresholding-based methods, texture-analysis-based methods, clustering-based methods, and region-based split and merging methods [17]. *Thresholding* is one of the most commonly used methods for the segmentation of images into two or more clusters [18]. Thresholding techniques can be divided into two different types: optimal thresholding methods [19]–[23] and property-based thresholding methods [24]–[26]. Algorithms in the former group search for the optimal thresholds which make the thresholded classes on the histogram reach the desired characteristics. Usually, thresholds are selected by optimizing an objective function. The latter group detects the thresholds by measuring some selected property of the histogram. Property-based thresholding methods are fast, which makes them suitable for multilevel thresholding. However, the number of thresholds for these approaches is hard to determine and needs to be specified in advance.

Many algorithms have been proposed in literature to address the issue of optimal thresholding (e.g., [19], [21], [22], and [26]–[30]). While several research papers address bilevel thresholding, others have considered the multilevel problem. Bilevel thresholding is reduced to an optimization problem to determine the threshold t that maximizes σ_B^2 (between-class variance) and minimizes σ_W^2 (within-class variance) [18]. For two-level thresholding, the problem is solved by finding the value of T^* which satisfies $\max(\sigma_B^2(T^*))$, where $0 \leq T^* < L$ and L is the maximum intensity value. This problem could be extended to n-level thresholding by satisfying $\max \sigma_B^2(T_1^*, T_2^*, \ldots, T_{n-1}^*)$, where $0 \leq T_1^* < T_2^* < \cdots < T_{n-1}^* < L$. One way to find the optimal set of thresholds is by using exhaustive search. A commonly used exhaustive search is based on the Otsu criterion [21]. That approach is easy to implement, but it has the disadvantage that it is computationally expensive. Exhaustive search for $n-1$ optimal thresholds involves evaluations of fitness of $n(L-n+1)^{n-1}$ combinations of thresholds [18]. Therefore, that method is not suitable from a computational cost point of view. The task of determining $n-1$ optimal thresholds for n-level image thresholding could be formulated as a multidimensional optimization problem. To solve such a task, several biologically inspired algorithms have been explored in image segmentation [18], [31]–[34]. Bioinspired algorithms have been used in situations where conventional optimization techniques cannot find a satisfactory solution or they take too much time to find it, e.g., when the function to be optimized is discontinuous and nondifferentiable and/or presents too many nonlinearly related parameters [35].

One of the best known bioinspired algorithms is particle swarm optimization (PSO) [36]. The PSO consists of a number of particles that collectively move in the search space (e.g., pixels of the image) in search of the global optimum (e.g., maximizing the between-class variance of the distribution of intensity levels in the given image). However, a general problem with the PSO and similar optimization algorithms is that they may get trapped in local optimum points, and the algorithm may work in some problems but may fail in others [37]. To overcome such a problem, Tillett *et al.* [38] presented the Darwinian PSO (DPSO). In the DPSO, multiple swarms of test solutions performing just like an ordinary PSO may exist at any time, with rules governing the collection of swarms that are designed to simulate natural selection. More recently, Couceiro *et al.* [35] have further extended the DPSO using fractional calculus to control the convergence rate of the algorithm. In [37], fractional-order DPSO (FODPSO) was successfully compared to both the fractional-order PSO (FOPSO) from Pires *et al.* [39] and the traditional DPSO.

The main goal of this paper is to propose a computationally efficient bioinspired segmentation method, which is robust for partitioning remote sensing images into multiple regions. For this purpose, a new method for segmentation of multispectral and hyperspectral images based on the FODPSO is proposed. To demonstrate the performance of this new method, a methodical and statistical comparison with two other methods for thresholding segmentation of images, namely, the well-known PSO and the DPSO, is carried out. In summary, the main contributions of this paper are as follows:

1) formal presentation of the FODPSO algorithm for image segmentation;
2) evaluation of this novel algorithm using more complex data sets (i.e., multispectral/hyperspectral) and comparison with other thresholding-based segmentation methods;
3) proposition of a novel classification approach based on the concept of the new segmentation method to improve the classification accuracy of the traditional support vector machine (SVM) method.

It should be noted that this is the first time that the concept of FODPSO is used in remote sensing, thus showing the potential of its use in efficient image segmentation to determine broad groups of objects. The current paper partially builds on [36] with an important study on how FODPSO performs for remote sensing images while the segmentation level for such images is changed. Moreover, a deep statistical analysis is conducted so as to further sustain the proposed approach when compared to others. Many problems in remote sensing have been solved by considering optimization methods such as genetic algorithm (GA) and PSO. Therefore, this paper introduces a very powerful optimization method, both in terms of speed and optimal convergence, which can be considered for a wide variety of problems in remote sensing. Some optimization methods are fast but not efficient (for finding the global optimum) and vice versa. It has been recently proved in [37] for benchmarking optimization problems that the FODPSO is faster than the PSO (the most well-known optimization algorithm in terms of speed) and more efficient than the DPSO (in order to find the global optimum while avoiding local optima). Therefore, applying the FODPSO to the segmentation of images may allow achieving both vital important goals at once. More specifically, due to its convergence speed, this optimization method may present itself as a potential solution to a wide variety of complex problems in remote sensing such as hyperspectral image analysis—a problem that many researchers are struggling with since *hyperspectral* images in remote sensing are very volumetric.

Bearing these ideas in mind, the problem formulation of image n-level thresholding is presented in the following sections. Section II presents a brief review of PSO and DPSO algorithms and focuses on their strengths and weaknesses, thus paving the way for a detailed description on the method that is proposed in this paper. In Section III, two different remote sensing data sets are considered, and the performances of the three different methods are compared. In Section IV, the proposed segmentation approach is extended and applied for classification. Finally, the main conclusion is outlined in Section IV.

II. Methodology

Multilevel segmentation techniques provide an efficient way to perform image analysis. However, the automatic selection of a robust optimum n-level threshold remains a challenge in segmentation of remote sensing images. In the following discussion, a more precise formulation of the problem is introduced along with the basic notation used in this paper.

A. Problem Formulation

Let there be L intensity levels in each component, e.g., three color components for RGB images, of a given image, and these levels are in the range $\{0, 1, 2, \ldots, L-1\}$. Then, one can define

$$p_i^C = \frac{h_i^C}{N} \qquad \sum_{i=0}^{L-1} p_i^C = 1 \tag{1}$$

where i represents a specific intensity level, i.e., $0 \leq i \leq L-1$; C represents the component of the image, e.g., $C = \{R, G, B\}$ for RGB images; N represents the total number of pixels in the image; and h_i^c denotes the number of pixels for the corresponding intensity level i in component C. In other words, h_i^c represents an image histogram for each component C, which can be normalized and regarded as the probability distribution p_i^c. The total mean (i.e., combined mean) of each component of the image can be easily calculated as

$$\mu_T^C = \sum_{i=0}^{L-1} i p_i^C = 1. \tag{2}$$

The n-level thresholding presents $n-1$ threshold levels t_j^c, $j = 1, \ldots, n-1$, and the operation is performed as

$$F^C(x,y) = \begin{cases} 0, & f^C(x,y) \leq t_1^C \\ \frac{1}{2}\left(t_1^C + t_2^C\right), & t_1^C < f^C(x,y) \leq t_2^C \\ & \vdots \\ \frac{1}{2}\left(t_{n-2}^C + t_{n-1}^C\right), & t_{n-2}^C < f^C(x,y) \leq t_{n-1}^C \\ L-1, & f^C(x,y) > t_{n-1}^C \end{cases} \tag{3}$$

where x and y are the width (W) and height (H) of the pixel of the image of size $H \times W$ denoted by $f^c(x,y)$ with L intensity levels for each component. In this situation, the pixels of a given image will be divided into n classes $D_1^c, \ldots, D_n^c$, which may represent multiple objects or even specific features for such objects (e.g., topological features). The probabilities of occurrence w_j^c of classes $D_1^c, \ldots, D_n^c$ are given by

$$w_j^C = \begin{cases} \sum_{i=0}^{t_j^C} p_i^C, & j = 1 \\ \sum_{i=t_{j-1}^C+1}^{t_j^C} p_i^C, & 1 < j < n \\ \sum_{i=t_{j-1}^C+1}^{L-1} p_i^C, & j = n. \end{cases} \tag{4}$$

The mean of each class μ_j^c can then be calculated as

$$\mu_j^C = \begin{cases} \sum_{i=0}^{t_j^C} \frac{p_i^C}{w_j^C q}, & j = 1 \\ \sum_{i=t_{j-1}^C+1}^{t_j^C} \frac{p_i^C}{w_j^C}, & 1 < j < n \\ \sum_{i=t_{j-1}^C+1}^{L-1} \frac{p_i^C}{w_j^C}, & j = n. \end{cases} \tag{5}$$

The simplest and computationally most efficient method of obtaining the optimal threshold is the one that maximizes the between-class variance of each component which can be generally defined by

$$\sigma_B^{c^2} = \sum_{j=1}^{n} w_j^C \left(\mu_j^C - \mu_T^C\right)^2 \tag{6}$$

where j represents a specific class in such a way that w_j^c and μ_j^c are the probability of occurrence and the mean of class j, respectively.

In other words, the problem of n-level thresholding is reduced to an optimization problem to search for the thresholds t_j^c that maximize the objective functions (i.e., fitness function) of each image component C, generally defined as

$$\varphi^C = \max_{1 < t_1^C < \cdots < L-1} \sigma_B^{c^2}\left(t_j^C\right). \tag{7}$$

Computing the aforementioned optimization problem involves high computational complexity as the number of threshold levels and image components increases. Many optimization methods have been proposed in the literature [2]. However, more recently, biologically inspired methods, such as the well-known PSO, have been used as computationally efficient alternatives to analytical methods to solve optimization problems [18], [19].

B. General Approach

The original PSO[1] algorithm was developed by Eberhart and Kennedy in 1995 [36]. The PSO basically takes advantage of the swarm intelligence concept, which is the property of a system whereby the collective behaviors of unsophisticated agents that are interacting locally with their environment create coherent global functional patterns. More recently, based on the concepts inherent to the PSO, the DPSO [38], and the FOPSO [39], an extended version denoted as FODPSO has been presented in [37], in which several swarms compete using Darwin's survival-of-the-fittest principles and fractional calculus to control the convergence rate of the algorithm. Using those principles, the FODPSO enhances the ability of the PSO algorithm to escape from local optima by running several simultaneous parallel PSO algorithms, each being a different swarm, on the same test problem and applies a simple selection mechanism. When a search tends to a local optimum, the search in that area is simply discarded, and another area is searched instead. In this approach, at each step, swarms that show improvement are rewarded (extend particle life or spawn a new descendent), and swarms which stagnate are punished (reduce swarm life or delete particles). Moreover, the approximate *Grünwald–Letnikov FC* definition allows us to use the concept of fractional differential with α, $0 \leq \alpha \leq 1$, to control the convergence rate of particles.

Table I presents the FODPSO algorithm applied to image segmentation. Each particle a within each different swarm s moves in a multidimensional space according to position

[1] The software including PSO-, DPSO- and FODPSO-based segmentation methods are available on request by sensing an email to the authors.

TABLE I
FODPSO SEGMENTATION ALGORITHM

```
Initialize $\alpha, \rho_1, \rho_2$ // fractional coefficient, global and local weights
Initialize $N, N_{min}, N_{max}$ // initial, minimum and maximum number of particles within each swarm
Initialize $N^s, N^s_{min}, N^s_{max}$ // initial, minimum and maximum number of swarm
Initialize $\Delta v$ // maximum number of levels a particle can travel between iterations
Initialize $I_T, I_{kill}$ // total number of iterations and maximum stagnation of swarms
$p_i^C = \frac{h_i^C}{N},\ \sum_{i=0}^{L-1} p_i^C = 1$
$\mu_T^C = \sum_{i=0}^{L-1} i p_i^C$
Initialize $0 \le x_a^s[0] \le L-1$ // initial position of all particles from all swarms
Initialize $\breve{x}_a^s, \breve{g}_a^s$ based on $x_a^s[0]$ // initial local best of all particles and global best of all swarms
For each iteration $t$ until $I_T$ // main loop
    For each particle $a$ of swarm $s$
        $v_a^s[t+1] = \alpha v_a^s[t] + \frac{1}{2}\alpha v_a^s[t-1] + \frac{1}{6}\alpha(1-\alpha)v_a^s[t-2] + \frac{1}{24}\alpha(1-\alpha)(2-\alpha)v_a^s[t-3] + \rho_1 r_1(\breve{g}_a^s - x_a^s[t]) + \rho_2 r_2(\breve{x}_a^s - x_a^s[t]),\ |v_a^s[t+1]| \le \Delta v$
        $x_a^s[t+1] = x_a^s[t] + v_a^s[t+1],\ 0 \le x_a^s[t+1] \le L-1$
        Compute (4) and (5) based on the thresholds defined in $x_a^s[t+1]$
        ${\sigma^{c}_{B}}^2 = \sum_{j=1}^{n} w_j^C (\mu_j^C - \mu_T^C)^2$ // compute the solution of each particle $a$ of swarm $s$
        If ${\sigma^{c}_{a_B}}^2 > {\sigma^{c}_{abest_B}}^2$// particle $a$ has improved
            ${\sigma^{c}_{nbest_B}}^2 = {\sigma^{c}_{B}}^2$
            $\breve{x}_a^s = x_a^s[t+1]$
    For each swarm $s$
        If $\max {\sigma^{c}_{s_B}}^2 > \varphi^C$ // swarm $s$ has improved
            $\varphi^C = \max {\sigma^{c}_{s_B}}^2$
            $\breve{g}_a^s = x_a^s[t+1]$
            $I_k = 0$ // reset stagnancy counter
            If $N_s < N_{max}$ // the current number of particles within swarm $s$ is inferior to the maximum number of allowed particles
                $N_s = N_s + 1$
                Randomly spawns a new particle in swarm $s$
                If $N^s < N^s_{max}$ and $rand()\frac{N_s}{N_{max}} > rand()$ // small probability of creating a new swarm
                    $N^s = N^s + 1$
                    Randomly spawns a new swarm with an initial number of $N$ particles
        Else // swarm $s$ has not improved
            $I_k = I_k + 1$
            If $I_k = I_{kill}$ // swarm $s$ has improved for too long
                If $N_s > N_{min}$ // swarm $s$ has currently more than the minimum number of allowed particles to form a swarm
                    Delete worse particle from swarm $s$, i.e., lower local solution
                Else // swarm $s$ does not currently have the minimum number of allowed particles to form a swarm
                    Delete whole swarm $s$, i.e., all particles from swarm $s$
End
```

$(x_a[t])$, $0 \le x_a[t] \le L-1$, and velocity $(v_a[t])$. The position and velocity values are highly dependent on the local best $(\breve{x}_a[t])$ and global best $(\breve{g}_a[t])$ information. The coefficients w, ρ_1, and ρ_2 are assigned weights, which control the inertial influence, i.e., according to "the globally best" and "the locally best," respectively, when the new velocity is determined. Typically, the inertial influence is set to a value slightly less than 1. ρ_1 and ρ_2 are constant integer values, which represent "cognitive" and "social" components. However, different results can be obtained by assigning different influences for each component. Depending on the application and the characteristics of the problem, tuning these parameters properly will lead to better results. The parameters r_1 and r_2 are random vectors, with each component generally a uniform random number between 0 and 1. The intent is to multiply a new random component per velocity dimension, rather than multiplying the same component with the velocity dimension of each particle.

It is noteworthy that the α value greatly affects the inertial particles. With a small α, particles ignore their previous activities, thus ignoring the system dynamics and being susceptible to get stuck in local solutions (i.e., exploitation behavior). On the other hand, with a large α, particles will present a more diversified behavior, which allows exploration of new solutions and improves the long-term performance (i.e., exploration behavior). However, if the exploration level is too high, then the algorithm may take too much time to find the global solution. Based on the experimental results from [37], a fractional coefficient of $\alpha = 0.6$ will be used, thus resulting in a balance between exploitation and exploration.

TABLE II
COMPUTATIONAL AND MEMORY COMPLEXITIES OF PSO, DPSO, AND FODPSO

Complexity	*PSO*	*DPSO*	*FODPSO*
Memory	$O(C)$	$O(C)$	$O(4C)$
Computational	$O(nN^P)$	$O(n\sum_{\forall s} N^S)$	$O(n\sum_{\forall s} N^S)$

One may summarize both computational and memory complexities as Table II depicts.

Note that the memory complexity of the FODPSO is larger than the alternatives since it intrinsically has memory properties related to the fractional extension. In other words, due to the truncation order of the approximate fractional derivative, it needs to track the last four steps of each particle's velocity that depends on the number of components C (i.e., bands) of the image. The computational complexity of the algorithms was considered, excluding the initial computation of (1) and (2). Note that this may be accomplished since the three algorithms require the same initial computation that depends on the size of the image. After that initial setup, the three algorithms may be adjusted in such a way to ensure a similar computational complexity. Likewise, the computational complexity of the three algorithms will increase with the number of desired thresholds n. Nevertheless, while the PSO depends on the number of particles N^P within the population, the DPSO and FODPSO depend on the accumulated number of particles within each swarm, i.e., $\sum_{\forall s} N^S$. In other words, one may ensure that the computational complexity of both DPSO and FODPSO will be inferior to the PSO by defining the maximum number of particles within each swarm as $N_{\max} \leq N^P/N^s_{\max}$, wherein $N^s_{\max}$ represents the maximum number of allowed swarms. It is, however, noteworthy that one may avoid holding this assumption since the evolutionary features of both DPSO and FODPSO are stochastic and depend on uniformly distributed variables. In other words, by setting an adequate combination between the minimum and maximum acceptable numbers of particles to form a swarm $N_{\min}$ and $M_{\max}$ and the minimum and maximum numbers of swarms within the population $N_{\min}$ and $N_{\max}$, one may ensure that $\mathcal{O}(n \sum_{\forall s} N^S) \leq \mathcal{O}(nN^P)$ for a steady-state regime. Such a condition may be achieved by adhering to the following condition: $N^P \geq (N^s_{\min} N_{\min} + N^s_{\max} N_{\max}/2)$.

C. Algorithm Evaluation

The computational time is one of the most important indicators, along with the fitness value, to determine the performance of the algorithm. Provided that the data are large, the efficiency of the method is restricted to a great extent [46]. For instance, hyperspectral images are, in general, large, so using a high-speed and efficient algorithm is highly preferable. Moreover, in real-time applications, using a high-speed algorithm is the main objective [18]. As a result, the evaluation of the CPU processing time and the fitness value seems vitally important to show the efficiency of the new method. In addition, since all bioinspired methods are random and stochastic, the results are not completely the same in each run. Consequently, the stability of different methods should be evaluated by an appropriate index such as standard deviation value.

PSO-based segmentation algorithms have been widely used in recent years. In fact, the ability of the traditional PSO-based segmentation has already been compared with other thresholding-based methods such as GA-based algorithms and exhaustive ones. Results confirm that the PSO-based method presents better results in terms of fitness value and CPU processing time. In [47], authors illustrated that the PSO-based segmentation method acted better than other methods such as GAs, differential evaluation, ant colony optimization, simulated annealing, and tabu search in terms of precision, robustness of the results, and runtime. In [28], PSO outperforms GA in terms of the CPU time and the fitness value for Kapur's and Otsu's functions. In [48], results indicate that PSO family methods act better than GA with a learning operator (GA-L) from different points of view. Consequently, it is easy to detect that PSO-based segmentation methods are considered an efficient way to find optimal thresholds in short CPU processing time.

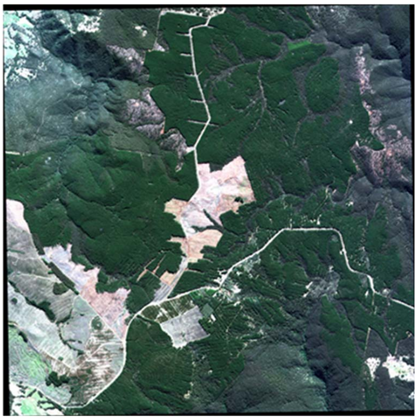


Fig. 1. Our test case study site (data channels 5, 3, and 2 are mapped to the R, G, and B channels).

III. Experimental Results

To compare the performance of the proposed FODPSO method with the PSO and DPSO approaches, all methods are tested on two different types of images, i.e., multispectral and hyperspectral images. In all cases, the image segmentation approaches were programmed in MATLAB on a computer having *Intel Core 2 Duo T5800* processor (2.00 GHz) and 3 GB of memory.

A. Description of Data Sets

1) First Test Case—Multispectral Worldview Image: The first data set is an 8 × 8 km multispectral Worldview satellite image consisting of eight bands captured at Tamworth, Northern New South Wales, Australia (Fig. 1). The pixel size was 2.4 × 2.4 m. The image covered large sections of pine plantations, interspersed with native vegetation, grasslands, logged areas, barren soil, and roads. This introduced a high level of natural variability to the segmentation problem. Unlike artificial objects, natural vegetation has multiple levels of variation. For example, within the pine plantation class, there are age differences and differences in reflectance due to slope, aspect, sun position, soil types, etc., and all of these cause added complexities in the segmentation scheme. Fig. 1 shows an image of the data where data channels 5, 3, and 2 are used in showing R, G, and B components, respectively, while Fig. 2 depicts the histogram for all eight data channels.

Table III gives the initial parameters of the PSO, the DPSO, and the proposed FODPSO-based methods for the first test case. The PSO, DPSO, and FODPSO methods are parameterized algorithms. Therefore, one needs to be able to choose the parameter values that would result in faster convergence. The cognitive, social, and inertial weights were chosen by taking into account several works focusing on the convergence analysis of the traditional PSO (cf., [1], [37], and [48]). For instance, to guarantee the convergence of the process,

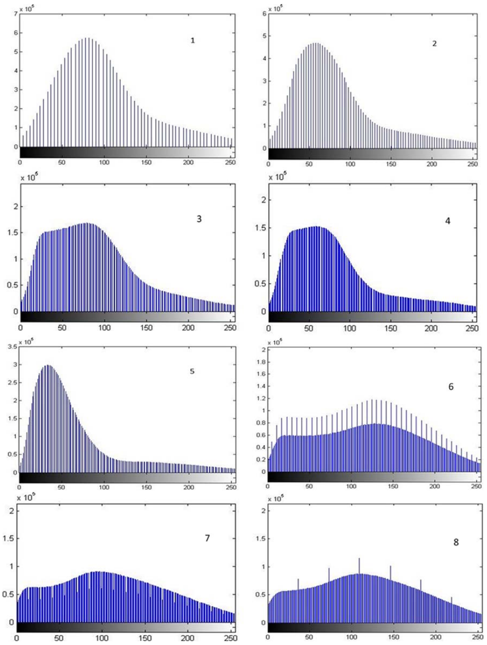


Fig. 2. Histograms of different data channels (data channel number inserted in each figure). Gray values on the x-axis and value count on y-axis.

TABLE III
INITIAL PARAMETERS OF THE PSO, DPSO, AND FODPSO FOR THE FIRST DATA SET

Parameter	***PSO***	***DPSO***	***FODPSO***
I_T	100	100	100
N	150	20	20
ρ_1	1.2	1.2	1.2
ρ_2	0.8	0.8	0.8
w	0.8	-	-
Δv	2	2	2
N_{min}	-	10	10
N_{max}	-	30	30
N^s	-	4	4
N^s_{min}	-	2	2
N^s_{max}	-	6	6
N_{kill}	-	10	10
α	-	-	0.6

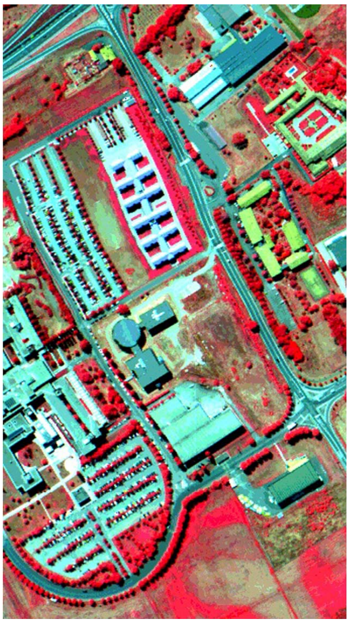


Fig. 3. Image of the second test case.

TABLE IV
INITIAL PARAMETERS OF THE PSO, DPSO, AND FODPSO FOR THE SECOND DATA SET

Parameter	***PSO***	***DPSO***	***FODPSO***
I_T	100	100	100
N	150	15	15
ρ_1	1.2	1.2	1.2
ρ_2	0.8	0.8	0.8
w	0.8	-	-
Δv	5	5	5
N_{min}	-	10	10
N_{max}	-	50	50
N^s	-	4	4
N^s_{min}	-	2	2
N^s_{max}	-	6	6
N_{kill}	-	10	10
α	-	-	0.6

Jiang *et al.* [48] presented a set of attraction domains that altogether present a relation between ρ_1, ρ_2, and w, wherein $0 \leq w < 1$ and $\rho_1 + \rho_2 > 0$. Based on the attraction domain in [4], if one would choose an inertial coefficient $w = 0.8$, the sum between the cognitive and social components would need to be less than 7, i.e., $\rho_1 + \rho_2 < 7$. The parameters in Table II were selected by considering that many works present a larger cognitive coefficient (cf., [48]). Note that the threshold velocities of particles and the maximum number of particles within each swarm in the DPSO are smaller than the PSO algorithm. This was experimentally adjusted to provide swarms of 20 particles with the same level of diversity (i.e., exploration and exploitation) than swarms of 150 particles.

2) Second Data Set—Hyperspectral ROSIS Image: The second test case is a hyperspectral data set which was captured on the city of Pavia, Italy, by airborne data from the Reflective Optics System Imaging Spectrometer (ROSIS-03). The ROSIS-03 sensor has 115 data channels with a spectral coverage ranging from 0.43 to 0.86 μm. In our experiments, we eliminate 12 noisy data channels and use 103 data channels for processing. The spatial resolution is 1.3 m per pixel. The original data set is 610 by 340 pixels. This data set is captured on the Engineering School, University of Pavia, Pavia, consisting of different classes including trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow, and soil. Fig. 3 shows an image of the second test case.

The proposed multilevel thresholding techniques based on PSO, DPSO, and FODPSO were implemented with the specific parameters shown in Table IV for the second test case. Table III presents the initial parameters of the PSO- and DPSO-based methods for the second test case. The main differences here in comparison to Table II are that the maximum velocity of particles and the capacity of each swarm, in the case of the

TABLE V
AVERAGE AND STD CPU PROCESSING TIMES (IN SECONDS) OF EACH ALGORITHM FOR DIFFERENT LEVELS

Level	FODPSO	DPSO	PSO	Difference (%) between FODPSO and DPSO	Difference (%) between FODPSO and PSO
6	41.05 ± 0.63	43.04 ± 0.78	46.69 ± 0.54	4.8	13.73
8	54.15 ± 1.21	56.21 ± 1.27	60.03 ± 0.65	3.8	10.85
10	65.46 ± 0.90	67.12 ± 1.39	73.60 ± 2.22	2.5	12.43

TABLE VI
AVERAGE FITNESS VALUES FOR ALL BANDS AT EACH LEVEL FOR THE FIRST TEST CASE

Level	FODPSO	DPSO	PSO
6	3812.23 ± 0.25	3812.31 ± 0.04	3811.45 ± 0.59
8	3877.56 ± 0.23	3877.21 ± 0.07	3876.04 ± 0.60
10	3907.52 ± 0.24	3907.15 ± 0.15	3905.95 ± 0.61

DPSO and FODPSO algorithms, need to be increased in order to overcome the increased complexity of using data.

B. Results and Discussion

1) First Test Case— Multispectral Image: The CPU average processing times of the PSO, DPSO, and FODPSO for six-, eight-, and ten-level thresholding are presented in Table V, and they were calculated over 40 different runs. PSO is referred to as a fast optimization algorithm. However, as can be seen from Table V, the computation time for PSO-based segmentation was significantly higher than that for both the FODPSO and DPSO methods. The main reason for this is that the PSO has a fixed population of 150 particles, which, in other words, means that 150 different solutions are needed to be evaluated within the same swarm. The FODPSO and DPSO, on the other hand, are composed of multiple smaller swarms (between 2 and 6 swarms of 10 and 30 particles each), being faster than the PSO even with an equal or larger number of particles in the whole DPSO and FODPSO. The dynamical clustering of particles inherent to both FODPSO and DPSO allows releasing most of the processing effort necessary to compute the local and global solutions. In other words, the CPU processing time decreases as the number of particles within the same swarm decreases. The difference percentages of CPU processing time between FODPSO and DPSO remain almost the same regardless of the segmentation level in the range of 2%–5%. Although the difference may be considered small to justify the choice between the FODPSO and the DPSO, it still represents an improvement that can be highly pondered depending upon the fitness value of the algorithms. Moreover, the stability of the traditional PSO highly deteriorates for a segmentation level of 10, contrary to both FODPSO and DPSO.

The fitness and optimal threshold values were calculated for all different data channels separately. The average and standard deviation fitness values of all data channels were calculated for each level of segmentation, and the obtained results are presented in Table VI. FODPSO generally performed slightly better than other methods in terms of fitness value. An exception may be observed for a segmentation level of 6 in which the DPSO presented a slightly better result than the FODPSO. It is noteworthy that such behavior should be expected for specific situations since the DPSO is a particular case of the FODPSO. In general, both DPSO and FODPSO give a better fitness because the PSO may get stuck in the vicinities of the global solution, while both FODPSO and DPSO use natural selection in order to avoid stagnation (cf., [37] and [38]). Hence, it can be concluded that both Darwinian algorithms are able to find better thresholds in less CPU time than the traditional PSO. Fig. 4 shows ten-level segmented image based on FODPSO and their histograms.

Despite the minor differences in fitness values between the FODPSO and the DPSO with respect to the between-class variance, one should note that the FODPSO-based thresholding is able to achieve segmentation of the image faster than both DPSO and PSO. Consequently, the proposed FODPSO method is very attractive for image segmentation, especially for more complex images and/or high segmentation levels.

Fig. 5 shows a subset of the main image and six-level and ten-level FODPSO-based segmented images zoomed by 200%. As can be seen from the figure, the main image [Fig. 5(c)] has more details than the other images. In contrast, the six-level segmented image [Fig. 5(a)] is the roughest image. It is easy to conclude that, by increasing the level of segmentation, the segmented image includes more details. As a result, the ten-level segmented image [Fig. 5(b)] is smoother than the six-level one. Our segmented image is also less pixelated compared to the original image.

To further improve the comparison between the three algorithms, the significance of the segmentation method and the segmentation level (independent variables) on the fitness value and the CPU processing time (dependent variables) was analyzed using the two-way MANOVA technique after checking the assumptions of multivariate normality and homogeneity of variance/covariance [50], [51]. The assumption of normality for each of the univariate dependent variables was examined using univariate tests of *Kolmogorov–Smirnov* (p-value < 0.05). The univariate normality of each dependent variable has not been verified. However, since $n \geq 30$, the multivariate normality was assumed based on the central limit theorem [50]–[52]. Furthermore, the assumption of multivariate normality was validated [50], [51]. The assumption about the homogeneity of variance/covariance matrix in each group was examined with the *Box's M test* ($M = 605.13$, $F(24; 376576.64) = 24.693$; p-value $= 0.001$). Although the homogeneity of variance/covariance matrices has not been verified (i.e., p-value $= 0.001$), the MANOVA technique is robust to this violation because all the samples have the same size [50], [51]. When the MANOVA detected significant statistical differences, we proceeded to the commonly-used ANOVA for each dependent variable followed by the Tukey's HSD post hoc. The classification of the size effect (i.e., measure of the proportion of the total variation in the dependent variable explained by the independent variable) was done according to Maroco [50] and Pallant [51]. This analysis was performed using the *IBM SPSS Statistics* software with a significance level of 5%.

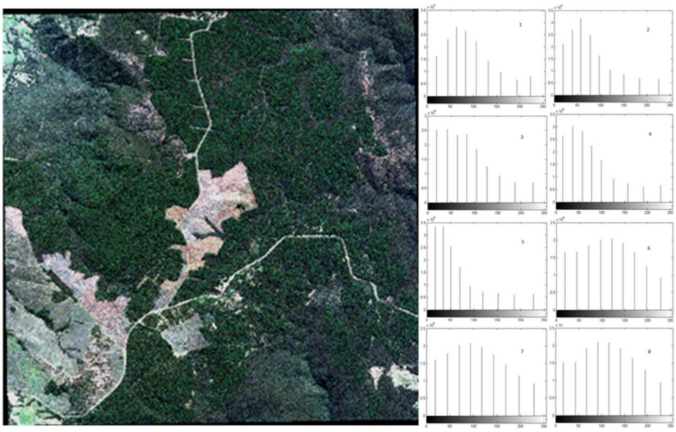


Fig. 4. Ten-level segmented image (data channels 5, 3, and 2 are mapped to the R, G, and B channels of the display) based on FODPSO and the histograms of all data channels.

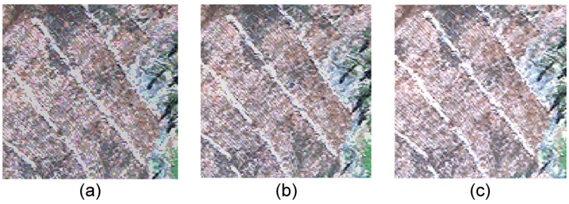


Fig. 5. Subset of (a) six-level, (b) ten-level, and (c) input images zoomed by 200%.

TABLE VII
TUKEY'S HSD POST HOC TEST TO THE MAXIMUM COMMUNICATION DISTANCE

Algorithm	Fitness Value	CPU Time
PSO *vs* DPSO	-1.06*	4.61*
PSO *vs* FODPSO	-1.25*	6.34*
DPSO *vs* FODPSO	-0.18*	1.73*

* The corresponding mean difference is significant at the 0.05 level

All *p*-values corresponding to the mean differences are equal to 0.001

A two-way MANOVA analysis was carried out to assess whether the algorithms used on this study have statistically significant differences with respect to the segmentation process. The MANOVA analysis revealed that the type of algorithm had a large and significant effect on the multivariate composite ($Pillai's\ Trace = 0.973$; $F(4;702) = 166.19$; p-value $=$ 0.001; $Partial\ Eta\ Squared\ \eta_p^2 = 0.486$; $Power = 1.0$). The segmentation level had a very large and significant effect on the multivariate composite ($Pillai's\ Trace = 1.847$; $F(4;702) = 2116.515$; p-value $= 0.001$; $\eta_p^2 = 0.923$; $Power = 1.0$). Finally, the interaction between the two independent variables had a moderate and significant effect on the multivariate composite ($Pillai's\ Trace = 0.469$; $F(8;702) = 26.901$; p-value $=$ 0.001; $\eta_p^2 = 0.235$; $Power = 1.0$).

After observing the multivariate significance for different algorithm types and segmentation levels, a univariate ANOVA for each dependent variable followed by the Tukey's HSD test was carried out. For the type of algorithm, the dependent variable fitness value presents statistically significant differences ($F(2,$ $351) = 469.97$; p-value$= 0.001$; $\eta_p^2 = 0.728$; $Power = 1.0$), as well as the dependent variable CPU processing time ($F(2,$ $351) = 2138.04$; p-value$= 0.001$; $\eta_p^2 = 0.92$; $Power = 1.0$). For the segmentation level, the dependent variable fitness value also demonstrates statistically significant differences ($F(2,351) =$ 2445064.03; p-value$= 0.001$; $\eta_p^2 = 1$; $Power = 1.0$), as well as the dependent variable CPU processing time ($F(2,351) =$ 1864.22; p-value$= 0.001$; $\eta_p^2 = 0.99$, $Power = 1.0$).

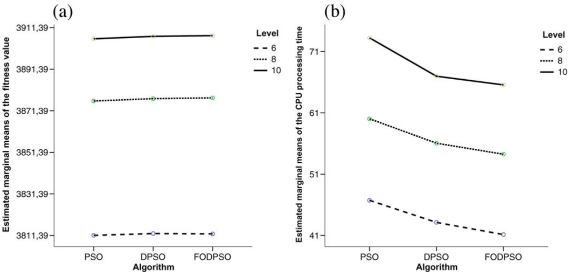


Fig. 6. Estimated marginal means of the (a) fitness value and (b) CPU processing time. (Dashed line) Leve 6. (Dotted line) Level 8. (Solid line) Level 10.

Using the Tukey's HSD post hoc, it is possible to verify the differences between the algorithms. Analyzing the fitness value and the CPU processing time, there are statistically significant differences between the obtained experimental results using the PSO, DPSO, and FODPSO segmentation algorithms.

It is noteworthy that the FODPSO produces better solutions than both the PSO and the DPSO. As expected, the FODPSO algorithm produces better solutions than the DPSO, and on the other hand, this last one produces better solutions than the PSO. In fact, using the PSO segmentation algorithm proves to be the "worse" segmentation method.

As shown in Table VII, which is based on Tukey's HSD post hoc test, the FODPSO is able to reach a slightly better fitness solution in less time. Nevertheless, the differences between the

TABLE VIII
AVERAGE AND STD CPU PROCESSING TIMES FOR EACH ALGORITHM AND DIFFERENT LEVELS

Level	FODPSO	DPSO	PSO	Percentage Difference between FODPSO and DPSO	Percentage Difference between FODPSO and PSO
10	689.83 ± 66.22	740.62 ± 49.73	1138.81 ± 103.02	7.4	65.1
12	691.96 ± 7.6	800.84 ± 8.4	1387.84 ± 123.43	15.7	100.1
14	753.59 ± 37.51	991.02 ± 82.60	1654.89 ± 141.82	31.5	119.6

TABLE IX
AVERAGE AND STD FITNESS VALUES AT EACH LEVEL

Level	FODPSO	DPSO	PSO
10	2971.69 ± 1.13	2971.22 ± 0.40	2970.09 ± 0.12
12	3002.73 ± 5.14	2991.78 ± 0.65	2984.92 ± 0.90
14	3090.13 ± 14,06	3035.97 ± 10.65	2997.53 ± 3.04

algorithms are not clearly seen in Fig. 6. Although it is possible to observe significant differences in the global CPU processing time between the FODPSO and the other algorithms, the improvement of the solution is not perceptible. Hence, in the next section, the same analysis will be performed on a hyperspectral image.

2) Second Data Set— Hyperspectral Image: As for the first data set, the CPU processing times in the second test case for each algorithm for 10-, 12-, and 14-level thresholding were calculated as the average value of 40 different runs, and the results are being presented in Table VIII. According to Table VIII, the FODPSO-based method has the least CPU processing time in comparison with other studied methods as was observed for the first data. On the contrary, PSO is the worst method among others in terms of CPU processing time. As can be seen from Table VIII, FODPSO significantly outperforms the PSO-based method, in particular, when the level of segmentation increases. FODPSO improves the result of the PSO-based segmentation method by 119.6% and 65.1% in the best and worst cases, respectively. In the same way, the CPU processing time of the FODPSO is considerably less than that for the DPSO and shows an improvement by 7.4% and 31.5% for the best and worst cases, respectively.

Table IX gives information regarding the average fitness values of 103 data channels in 40 different iterations. As in the case of the first multispectral data set, in the hyperspectral test case, FODPSO finds optimal threshold values which are better than that for the other methods. This shows that FODPSO is able to find optimal thresholds with better fitness values in less CPU processing time compared to the other studied methods. The fitness value of the FODPSO-based method is followed by DPSO, which is more efficient than the conventional PSO. As can be seen from the table, by increasing the level of segmentation, the fitness of FODPSO increases more than the fitness of the other methods. PSO gives almost the same fitness for 10, 12, and 14 levels of segmentation since it is not endowed with any kind of mechanism to improve the convergence of particles when in the vicinities of the optimal solution.

Fig. 7 shows 10-level and 14-level FODPSO-based segmented images using a 200% zoom. As can be seen from the figure, the 14-level-based segmented image [Fig. 7(b)] provides more details than the 10-level segmentation.

Similar to the first data set, the assumption of normality for each of the univariate dependent variables was examined using univariate tests of *Kolmogorov–Smirnov* (p-value $<$ 0.05) [50]–[52]. The assumption about the homogeneity of variance/covariance matrix in each group was examined with the *Box's M test* ($M = 1239.38$, $F(24; 376576.64) = 50.58$; p-value $= 0.001$). When the MANOVA detected significant statistical differences, we proceeded to the commonly-used ANOVA for each dependent variable followed by the Tukey's HSD post hoc.

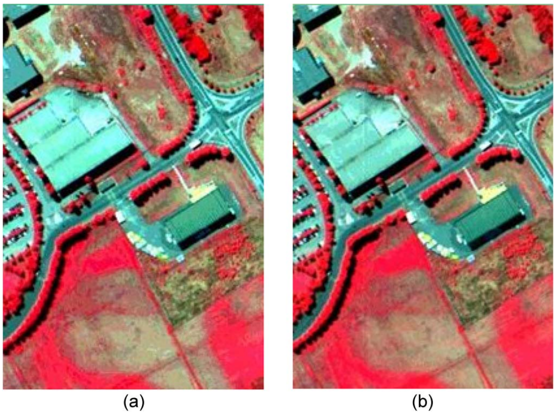


Fig. 7. Subset of (a) 10-level and (b) 14-level FODPSO-based segmented images zoomed by 200%.

TABLE X
TUKEY'S HSD POST HOC TEST TO THE MAXIMUM COMMUNICATION DISTANCE

Algorithm	Fitness Value	CPU Time
PSO *vs* DPSO	-15.47*	549.69*
PSO *vs* FODPSO	-37.33*	682.05*
DPSO *vs* FODPSO	-21.86*	132.37*

* The corresponding mean difference is significant at the 0.05 level

All p-values corresponding to the mean differences are equal to 0.001

The MANOVA analysis revealed that the algorithm type had a very large and significant effect on the multivariate composite ($Pillai's\ Trace = 1.40$; $F(4; 702) = 405.97$; p-value $=$ 0.001; $Partial\ Eta\ Squared\ \eta_p^2 = 0.698$; $Power = 1.0$). The segmentation level also had a large and significant effect on the multivariate composite ($Pillai's\ Trace = 0.97$; $F(4; 702) =$ 165.03; p-value $= 0.001$; $\eta_p^2 = 0.49$; $Power = 1.0$). Finally, the interaction between the two independent variables had a very large and significant effect on the multivariate composite ($Pillai's\ Trace = 1.02$; $F(8; 702) = 91.82$; p-value $= 0.001$; $\eta_p^2 = 0.51$; $Power = 1.0$).

After observing the multivariate significance in the type of algorithm and the segmentation level, a univariate ANOVA for

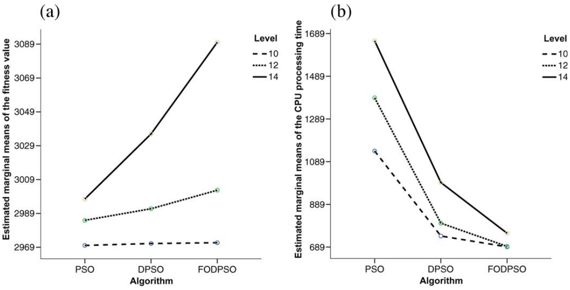


Fig. 8. Estimated marginal means of the (a) fitness value and (b) CPU processing time. (Dashed line) Level 10. (Dotted line) Level 12. (Solid line) Level 14.

each dependent variable followed by the Tukey's HSD test was carried out. For the type of algorithm, the dependent variable fitness value presents statistically significant differences ($F(2, 351) = 1066.64$; p-value $= 0.001$; $\eta_p^2 = 0.86$; $Power = 1.0$), as well as the dependent variable CPU processing time ($F(2, 351) = 2309.24$; p-value $= 0.001$; $\eta_p^2 = 0.93$; $Power = 1.0$). For the segmentation level, the dependent variable fitness value also presents statistically significant differences ($F(2, 351) = 3907.10$; p-value $= 0.001$; $\eta_p^2 = 0.96$; $Power = 1.0$), as well as the dependent variable CPU processing time ($F(2, 351) = 77.58$; p-value $= 0.001$; $\eta_p^2 = 0.66$, $Power = 1.0$).

Using the Tukey's HSD post hoc, one can observe that there are statistically significant differences between experiments using the PSO, DPSO, and FODPSO segmentation algorithms, for both CPU processing time and fitness function.

Once again, the FODPSO produces better solutions than both the PSO and the DPSO in terms of fitness value. Furthermore, as expected, the DPSO produces better solutions than the PSO. As shown in Table X (also shown in Fig. 8), based on Tukey's HSD post hoc test, the fractional-order algorithm is able to once again reach a better fitness solution in less time. Moreover, the differences between the FODPSO and the other algorithms are more evident as the segmentation level increases. This should be highly appreciated as many applications require real-time multisegmentation methods (e.g., autonomous deployment of sensor nodes in a given environment).

In summary, it is possible to observe that the FODPSO is faster than the DPSO since fractional calculus is used to control the convergence rate of the algorithm. As described in [49], a swarm behavior can be divided into *exploitation* and *exploration*. The exploitation behavior is related with the convergence of the algorithm, allowing a good short-term performance. However, if the exploitation level is too high, then the algorithm may be stuck on local solutions. On the other hand, the exploration behavior is related with the diversification of the algorithm which allows exploring new solutions, thus improving the long-term performance. However, if the exploration level is too high, then the algorithm may take too much time to find the global solution. In the DPSO, the tradeoff between exploitation and exploration can only be handled by adjusting the inertia weight w. While a large inertia weight improves exploration activity, the exploitation may be improved using a small inertia weight. Since the FODPSO presents a fractional calculus strategy to control the convergence of particles with memory effect, the coefficient α allows providing a higher level of exploration while ensuring the global solution of the algorithm (cf., [38]).

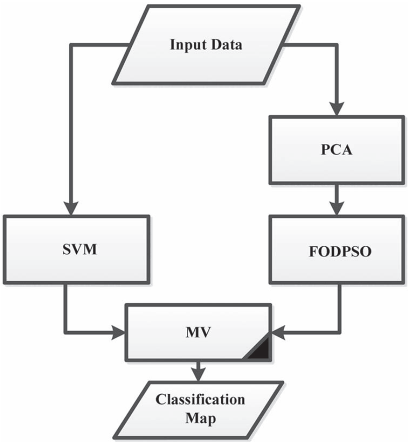


Fig. 9. Illustrative flowchart of the new classification approach.

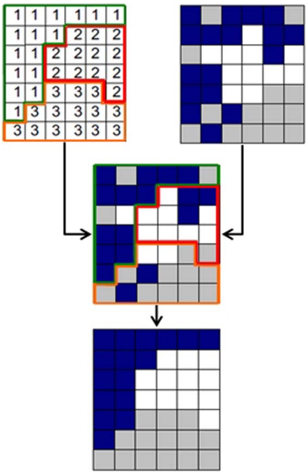


Fig. 10. Integration of the classification and segmentation steps using MV [4].

IV. Classification

Although the main idea behind this paper is to introduce a thresholding-based segmentation technique, it is of interest to see the effectiveness of the new segmentation method on classification. In this way, this section presents a novel framework to prove the efficiency of the proposed method for classification. The proposed classification method is based on the FODPSO and the SVM classifier. Since we do not have reference samples for the first data set, the classification is only performed on the second data set. Fig. 9 shows the general idea of the proposed classification approach. As can be seen, the data have been first classified with SVM and a Gaussian kernel. The hyperparameters have been selected using five-fold cross validation. Each variable has been scaled between -1 and 1.

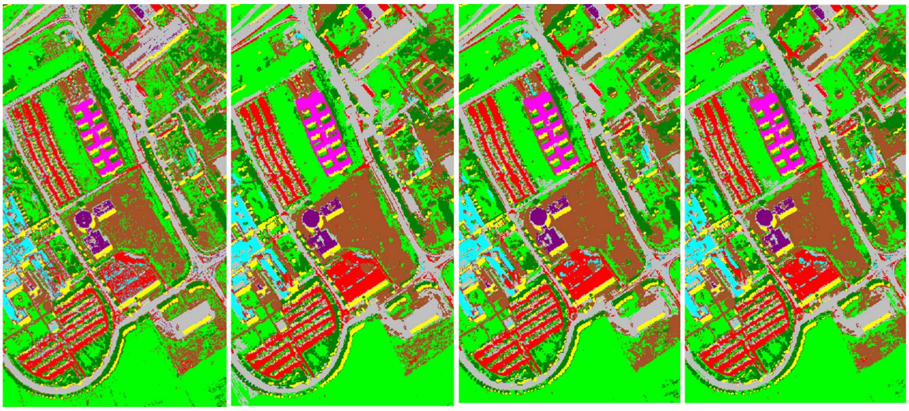


Fig. 11. Classification map of the standard SVM and the proposed classification method with 10-, 12-, and 14-level segmentation by FODPSO.

To carry out a fair evaluation, the input is classified only once, while the output of this step is used for all different levels. By doing that, the accuracy of the classification for different methods is only dependent on the effect of the segmentation method. In parallel, the input data are transformed using the principal component analysis (PCA), and the first principal component (PC) is kept since most of the variance is provided by that. The output of this step is segmented by the proposed FODPSO method. In the final step, the results of the SVM and the FODPSO are combined by using majority voting (MV).

Fig. 10 depicts the general idea of the proposed approach with MV. The output of the segmentation methods is a few number of objects, and each object consists of several pixels with the same label. In other words, pixels in each object share the same characteristics. To perform the MV on the output of the segmentation and classification steps, counting the number of pixels with different class labels in each object is first carried out. Subsequently, all pixels in each object are assigned to the most frequent class label for the object. In the case where two classes have the same (most frequent) proportions in one object, the object is not assigned to any of those classes, and the result of the traditional SVM is considered for each pixel in the object directly.

The procedure of the new classification approach is described step by step as follows:

1) The input data are classified by SVM.
2) The input data are transformed by PCA, and the first PC is kept.
3) The output of step 2 is segmented by FODPSO.
4) The results of steps 1 and 2 are combined using MV.

Fig. 11 illustrates the classification map of the standard SVM and the proposed classification method with 10-, 12-, and 14-level segmentation by FODPSO. The output of the SVM presents a lot of noisy pixels which decrease the accuracy of the classification. The results of the overall accuracy and kappa coefficient for the SVM and the new method with 10, 12, and 14 levels are shown in Table XI. For a better understanding, the classification accuracy for each class is also included in the table. All three segmentation levels improve the result of the SVM classification. The accuracy increases when the number of levels increases from 10 to 14. The main reason behind that phenomenon is denoted as *under segmentation* in which several objects are merged into a single one. This problem can be easily solved by increasing the number of levels. SVM + FODPSO with 10, 12, and 14 levels improves the overall accuracy of SVM by almost 0.7, 1.7, and 2 percent.

V. Conclusion

In this paper, a novel multilevel thresholding segmentation method has been proposed for grouping the pixels of multispectral and hyperspectral images into different homogenous regions. The new method is based on FODPSO which is used in finding the optimal set of threshold values and uses many swarms of test solutions which may exist at any time. In the FODPSO, each swarm individually performs just like an ordinary (PSO) algorithm with a set of rules governing the collection of swarms that are designed to simulate natural selection. Moreover, the concept of fractional derivative is used to control the convergence rate of particles. Experimental results compare the FODPSO with the classical PSO and DPSO within multilevel segmentation problems on remote sensing images from different points of view such as CPU time and corresponding fitness value. Segmentation methods were carried out on two different test cases. The first test case was a multispectral image related to native vegetation, grasslands, logged areas, and barren soil. The second test case was a hyperspectral image which is from an urban area, showing a wide variety of human artifacts. Experimental results indicate that the FODPSO is more robust than the two other methods and has a higher potential for finding the optimal set of thresholds with more between-class variance in less computational time, especially for higher segmentation levels and for images with a wide variety of intensities. In addition, to show the efficiency of the proposed segmentation method on the result of classification, a novel classification approach based on the new segmentation method and SVM is proposed. Results confirm that the new segmentation method improves the SVM in terms of classification accuracies when compared to the standard SVM classification of the raw image data. It should be noted that this is the first time that the concept of FODPSO is used in remote sensing, thus showing the potential of its use in efficient image segmentation

TABLE XI
RESULTS OF THE STANDARD SVM AND THE PROPOSED CLASSIFICATION METHOD WITH 10, 12, AND 14 LEVELS OF SEGMENTATION BY FODPSO. CLASSIFICATION ACCURACIES ARE GIVEN IN PERCENTAGE

Class		**Samples**		**SVM**	**SVM+ FODPSO(10)**	**SVM+ FODPSO(12)**	**SVM+ FODPSO(14)**
No	**Name**	**Training**	**Test**				
1	**Asphalt**	840	5791	94.4	87.1	95.4	96.1
2	**Meadow**	2317	16332	98.1	96.9	97.6	96.9
3	**Gravel**	253	1846	77.9	98.8	98.2	98.5
4	**Trees**	373	2691	93.0	99.7	98.6	99.0
5	**Metal sheets**	149	1196	99.2	100	99.9	100
6	**Bare soil**	619	4410	89.4	96.0	93.7	97.4
7	**Bitumen**	181	1149	85.8	97.3	97.9	97.8
8	**Bricks**	480	3202	92.0	91.0	87.4	86.5
9	**Shadow**	135	812	99.4	99.1	99.9	99.9
Overall Accuracy				*94.3*	95.0	*96.0*	96.2
Kappa Coefficient				*.924*	.934	*.947*	.950

to determine broad groups of objects. As future work, due to the low computational complexity of the algorithm, the FODPSO will be evaluated in image segmentation applications for the real-time autonomous deployment and distributed localization of sensor nodes. The objective is to deploy the nodes only in the terrains of interest, which are identified by segmenting the images captured by a camera onboard an unmanned aerial vehicle using the FODPSO algorithm. Such a deployment has importance for emergency applications, such as disaster monitoring and battlefield surveillance. In addition, finding a way for the estimation of the number of thresholds (parameter n) and joint multichannel segmentation instead of segmenting data set band by band would be of interest.

ACKNOWLEDGMENT

The Worldview and ROSIS data and corresponding reference information were obtained from Dr. L. Kumar of the University of New England, Madgwick, Australia, and Prof. P. Gamba of the University of Pavia, Pavia, Italy, respectively.

REFERENCES

[1] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, “An efficient method for segmentation of images based on fractional calculus and natural selection,” *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12 407–12 417, Nov. 2012.

[2] M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” *J. Electron. Imag.*, vol. 13, no. 1, pp. 146–168, Jan. 2004.

[3] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, “Use of Darwinian particle swarm optimization technique for the segmentation of remote sensing images,” in *Proc. IEEE IGARSS*, Jul. 22–27, 2012, pp. 4295–4298.

[4] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson, “Segmentation and classification of hyperspectral images using watershed transformation,” *Pattern Recognit.*, vol. 43, no. 7, pp. 2367–2379, Jul. 2010.

[5] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, “Advances in spectral-spatial classification of hyperspectral images,” *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[6] A. Darwish, K. Leukert, and W. Reinhardt, “Image segmentation for the purpose of object-based classification,” in *Proc. IGARSS*, 2003, vol. 3, pp. 2039–2041.

[7] J. Tilton, “Analysis of hierarchically related image segmentations,” in *Proc. IEEE Workshop Adv. Tech. Anal. Remotely Sensed Data*, 2003, pp. 60–69.

[8] M. Pesaresi and J. A. Benediktsson, “A new approach for the morphological segmentation of high-resolution satellite imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[9] G. Noyel, J. Angulo, and D. Jeulin, “Morphological segmentation of hyperspectral images,” *Image Anal. Stereol.*, vol. 26, pp. 101–109, 2007.

[10] H. G. Akcay and S. Aksoy, “Automatic detection of geospatial objects using multiple hierarchical segmentations,” *IEEE Trans. Geosci. Remote Sens*, vol. 46, no. 7, pp. 2097–2111, Jul. 2008.

[11] J. Chanussot and P. Lambert, “Bit mixing paradigm for multivalued morphological filters,” in *Proc. IEEIPA*, 1997, pp. 804–808.

[12] J. Chanussot and P. Lambert, “Total ordering based on space filling curves for multivalued morphology,” in *Proc. ISMM*, 1998, pp. 51–58.

[13] P. Lambert and J. Chanussot, “Extending mathematical morphology to color image processing,” in *Proc. CGIP*, 2000, pp. 158–163.

[14] A. G. Hanbury and J. Serra, “Morphological operators on the unit circle,” *IEEE Trans. Image Process.*, vol. 10, no. 12, pp. 1842–1850, Dec. 2001.

[15] J. Angulo and J. Serra, “Morphological coding of color images by vector connected filters,” in *Proc. ISSPA*, 2003, vol. 1, pp. 69–72.

[16] E. Aptoula and S. Lefevre, “A comparative study on multivariate mathematical morphology,” *Pattern Recognit.*, vol. 40, no. 11, pp. 2914–2929, Nov. 2007.

[17] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, “Extending the fractional order Darwinian particle swarm optimization to segmentation of hyperspectral images,” in *Proc. SPIE 8537—Image Signal Process. Remote Sens. XVIII*, Nov. 8, 2012, 85730F.

[18] R. V. Kulkarni and G. K. Venayagamoorthy, “Bio-inspired algorithms for autonomous deployment and localization of sensor nodes,” *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 663–675, Nov. 2010.

[19] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, “A new method for gray-level picture thresholding using the entropy of the histogram,” *Comput. Vis. Graph. Image Process.*, vol. 29, no. 3, pp. 273–285, Mar. 1985.

[20] J. Kittler and J. Illingworth, “Minimum error thresholding,” *Pattern Recognit.*, vol. 19, no. 1, pp. 41–47, 1986.

[21] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[22] T. Pun, “A new method for grey-level picture thresholding using the entropy of the histogram,” *Signal Process.*, vol. 2, no. 3, pp. 223–237, Jul. 1980.

[23] T. Pun, “Entropy thresholding: A new approach,” *Comput. Vis. Graph. Image Process.*, vol. 16, no. 3, pp. 210–239, Jul. 1981.

[24] Y. K. Lim and S. U. Lee, “On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques,” *Pattern Recognit.*, vol. 23, no. 9, pp. 935–952, 1990.

[25] D. M. Tsai, “A fast thresholding selection procedure for multimodal and unimodal histograms,” *Pattern Recognit. Lett.*, vol. 16, no. 6, pp. 653–666, Jun. 1995.

[26] P. Y. Yin and L. H. Chen, “New method for multilevel thresholding using the symmetry and duality of the histogram,” *J. Electron. Imag.*, vol. 2, no. 4, pp. 337–344, Oct. 1993.

[27] A. D. Brink, “Minimum spatial entropy threshold selection,” *Proc. Inst. Elect. Eng.—Vis. Image Signal Process.*, vol. 142, no. 3, pp. 128–132, Jun. 1995.

[28] Q. Hu, Z. Hou, and W. Nowinski, “Supervised range-constrained thresholding,” *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 228–240, Jan. 2006.

[29] P. K. Saha and J. K. Udupa, “Optimum image thresholding via class uncertainty and region homogeneity,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 689–706, Jul. 2001.

[30] O. J. Tobias and R. Seara, “Image segmentation by histogram thresholding using fuzzy sets,” *IEEE Trans. Image Process.*, vol. 11, no. 12, pp. 1457–1465, Dec. 2002.

[31] D. B. Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, 2nd ed. Piscataway, NJ, USA: IEEE Press, 2000.

[32] C. C. Lai and D. C. Tseng, “A hybrid approach using Gaussian smoothing and genetic algorithm for multilevel thresholding,” *Int. J. Hybrid Intell. Syst.*, vol. 1, no. 3, pp. 143–152, Dec. 2004.

[33] P. Y. Yin, “A fast scheme for optimal thresholding using genetic algorithms,” *Signal Process.*, vol. 72, no. 2, pp. 85–95, Jan. 1999.

[34] Y. Kao, E. Zahara, and I. Kao, “A hybridized approach to data clustering,” *Expert Syst. Appl.*, vol. 34, no. 3, pp. 1754–1762, Apr. 2008.

[35] D. Floreano and C. Mattiussi, *Bio-Inspired Artificial Intelligence: Theories, Methods, Technologies*. Cambridge, MA, USA: MIT Press, 2008.

[36] J. Kennedy and R. Eberhart, “A new optimizer using particle swarm theory,” in *Proc. IEEE 6th Int. Symp. Micro Mach. Human Sci.*, 1995, pp. 39–43.

[37] M. S. Couceiro, N. M. F. Ferreira, and J. A. T. Machado, “Fractional order Darwinian particle swarm optimization,” in *Proc. Symp. FSS*, Coimbra, Portugal, Nov. 4–5, 2011.

[38] J. Tillett, T. M. Rao, F. Sahin, R. Rao, and S. Brockport, “Darwinian particle swarm optimization,” in *Proc. 2nd Indian Int. Conf. Artif. Intell.*, 2005, pp. 1474–1487.

[39] E. J. S. Pires, J. A. T. Machado, P. B. M. Oliveira, J. B. Cunha, and L. Mendes, “Particle swarm optimization with fractional-order velocity,” *J. Nonl. Dyn.*, vol. 61, no. 1/2, pp. 295–301, Jul. 2010.

[40] J. Sabatier, O. P. Agrawal, and J. A. T. Machado, Eds., *Advances in Fractional Calculus—Theoretical Developments and Applications in Physics and Engineering*. New York, NY, USA: Springer-Verlag, 2007.

[41] M. D. Ortigueira and J. A. T. Machado, “Special issue on fractional signal processing,” *Signal Process.*, vol. 83, no. 11, pp. 2285–2286, Nov. 2003.

[42] J. A. T. Machado, M. F. Silva, R. S. Barbosa, I. S. Jesus, C. M. Reis, M. G. Marcos, and A. F. Galhano, *Some Applications of Fractional Calculus in Engineering*. New York, NY, USA: Hindawi Publ. Corp. Math. Problems Eng., 2010, pp. 1–34.

[43] I. Podlubny, “Fractional differential equations,” in *Mathematics in Science and Engineering*, vol. 198. San Diego, CA, USA: Academic, 1999.

[44] L. Debnath, “Recent applications of fractional calculus to science and engineering,” *Int. J. Math. Sci.*, vol. 2003, no. 54, pp. 3413–3442, 2003.

[45] M. S. Couceiro, N. M. F. Ferreira, and J. A. T. Machado, “Application of fractional algorithms in the control of a robotic bird,” *J. Commun. Nonl. Sci. Numer. Simul.-Special Issue*, vol. 15, no. 4, pp. 895–910, Apr. 2010.

[46] J. Fan, M. Han, and J. Wang, “Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation,” *Pattern Recognit.*, vol. 42, no. 11, pp. 2527–2540, Nov. 2009.

[47] K. Hammouche, M. Diaf, and P. Siarry, “A comparative study of various meta-heuristic techniques applied to the multilevel thresholding problem,” *Eng. Appl. Artif. Intell.*, vol. 23, no. 5, pp. 676–688, Aug. 2010.

[48] M. Jiang, Y. P. Luo, and S. Y. Yang, “Stochastic convergence analysis and parameter selection of the standard particle swarm optimization algorithm,” *Inf. Process. Lett.*, vol. 102, no. 1, pp. 8–16, Apr. 2007.

[49] K. Yasuda, N. Iwasaki, G. Ueno, and E. Aiyoshi, “Particle swarm optimization: A numerical stability analysis and parameter adjustment based on swarm activity,” *IEEJ Trans. Elect. Electron. Eng.*, vol. 3, no. 6, pp. 642–659, Nov. 2008.

[50] J. Maroco, *Análise Estatística com Utilização do SPSS*. Lisboa, Portugal: Edições Silabo, 2010.

[51] J. Pallant, *SPSS Survival Manual*, 4th ed. Berkshire, U.K.: Open Univ. Press, 2011, Kindle Edition.

[52] A. C. Pedrosa and S. M. A. Gama, *Introdução Computacional à Probabilidade e Estatística*. Porto, Portugal: Porto Editora, 2004.

Pedram Ghamisi (S’11) received the B.Sc. degree in civil (survey) engineering from Islamic Azad University, Tehran, Iran, and the M.Sc. degree in remote sensing from K. N. Toosi University of Technology, Tehran, in 2012. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Iceland, Reykjavík, Iceland.

His research interests are remote sensing and image analysis with the current focus on spectral and spatial techniques for hyperspectral image classification. He serves as a reviewer for a number of journals including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Mr. Ghamisi received the Best Researcher Award from K. N. Toosi University of Technology in 2010–2011.

Micael S. Couceiro (S’11) received the M.Sc. degree in automation and communications from the Engineering Institute of Coimbra, Coimbra, Portugal, in 2010. He is currently working toward the Ph.D. degree in electrical and computer engineering in the Faculty of Sciences and Technology, University of Coimbra, Coimbra.

He conducts research on multirobot systems and swarm robotics at the Mobile Robotics Laboratory, Institute of Systems and Robotics, and at RoboCorp, Polytechnic Institute of Coimbra, Coimbra. He has published papers on mobile robotics, biomimetics, fractional-order control, sports engineering, biomechanics, and mathematical methods.

He has been a Regular Reviewer for IEEE top conferences such as IEEE/RSJ International Conference on Intelligent Robots and Systems and IEEE International Conference on Robotics and Automation.

Fernando M. L. Martins graduated with a Mathematics Teaching Licensure and received the M.Sc. and Ph.D. degrees in applied mathematics from the University of Beira Interior (UBI), Covilha, Portugal, in 2001, 2003, and 2007, respectively.

Since 2009, he has been with the Coimbra College of Education (ESEC), Coimbra, Portugal, where he is currently a Professor with the Department of Education (Mathematics and Mathematics Education). He conducts research at the Institute of Telecommunications (IT-Covilhã) on Applied Mathematics. He has published papers on mathematics education, applied mathematics, sports engineering, and robotics.

Jón Atli Benediktsson (S’84–M’90–SM’99–F’04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984 and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is currently the Pro Rector for Academic Affairs and a Professor of electrical and computer engineering with the University of Iceland. He is the Cofounder of the biomedical start up company Oxymap. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in these fields.

Prof. Benediktsson was the 2011–2012 President of the IEEE Geoscience and Remote Sensing Society (GRSS), and he has been on the GRSS AdCom since 2000. He is a Fellow of SPIE and a member of Societas Scinetiarum Islandica and Tau Beta Pi. He was the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008, and he has served as Associate Editor of TGRS since 1999 and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003. He was the Chairman of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2007–2010. He received the Stevan J. Kristof Award from Purdue University in 1991 as outstanding graduate student in remote sensing. He was the recipient of the Icelandic Research Council’s Outstanding Young Researcher Award in 1997, he was granted the IEEE Third Millennium Medal in 2000, he was a corecipient of the University of Iceland’s Technology Innovation Award in 2004, he received the yearly research award from the Engineering Research Institute of the University of Iceland in 2006, and he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society in 2007. He was the corecipient of the 2012 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Paper Award.

Integration of Segmentation Techniques for Classification of Hyperspectral Images

Pedram Ghamisi, *Student Member, IEEE,* Micael S. Couceiro, *Student Member, IEEE,* Mathieu Fauvel, and Jon Atli Benediktsson, *Fellow, IEEE*

***Abstract*—A new spectral–spatial method for classification of hyperspectral images is introduced. The proposed approach is based on two segmentation methods, fractional-order Darwinian particle swarm optimization and mean shift segmentation. The output of these two methods is classified by support vector machines. Experimental results indicate that the integration of the two segmentation methods can overcome the drawbacks of each other and increase the overall accuracy in classification.**

***Index Terms*—Hyperspectral image analysis, mean shift segmentation, multilevel segmentation.**

I. Introduction

ACCURATE classification of remote sensing images plays a key role in many applications, including crop monitoring, forest applications, urban development, mapping and tracking, and risk management. One way for achieving this goal would be to use the spectral and the spatial information sequentially [15]. The goal of considering spatial context in the classification step can be partially achieved by using some specific methods, such as morphological filters [15] and Markov random fields [4]. The above-mentioned methods significantly increase the accuracy of the classification by incorporating spatial and spectral information. Another way for considering the spatial structures would be to perform image segmentation.

Image segmentation is a procedure that can be used to modify the accuracy of classification maps. To make such an approach effective, an accurate segmentation of the image is needed. A few methods for segmentation of multispectral and hyperspectral images have been introduced in the literature. Some of these methods are based on region merging methods, where neighboring segmented regions are merged with each other according to their homogeneity criterion, for instance multiresolution segmentation method in the eCognition software is used this type of approach [5]. In [6], hierarchical segmentation algorithm is proposed, which performs region growing and spectral clustering alternately.

One of the best known methods for image segmentation is thresholding. Different types of optimal thresholding methods have been proposed in the literature (e.g., [16]). One strategy to find the optimal set of thresholds is to take into account an exhaustive search. A commonly used exhaustive search method is based on the Otsu criterion [1]. However, exhaustive search to find $n - 1$ optimal thresholds involves evaluation of the fitness for $n(L - n + 1)^{n-1}$ combinations of thresholds and L is the intensity level in each component [9]. Therefore, this method is not desirable from a computational point of view. Alternatively, the issue of determining $n - 1$ optimal thresholds for n-level image thresholding can be formulated as a multidimensional optimization problem. To solve the aforementioned issue, several biologically inspired algorithms have been explored in image segmentation [9].

One of the most commonly used methods based on split and merging segmentation is mean shift segmentation (MSS) that is widely used in image processing. MSS is a nonparametric clustering technique, which does not need embedded assumptions on the shape of the distribution and the number of clusters compared with the classic K-means clustering. MSS is a powerful method for segmentation of images with high redundancy [10], such as remote sensing images.

Fractional-order Darwinian particle swarm optimization (FODPSO) segmentation (as all thresholding-based methods in general) suffers from the following disadvantages: 1) It cannot handle inhomogeneity; 2) it fails when the intensity of object of interest does not appear as a peak in the histogram; and 3) the traditional FODPSO-based segmentation takes into account only the between-class variance, thus disregarding any feedback from the within-class variance. In the MSS method, a kernel size needs to be tuned by the user. The tuning may be a difficult task and the final results may be affected by that dramatically.

In this letter, a new spectral–spatial classification approach is introduced for accurate classification of hyperspectral images. First, an input image will be segmented by FODPSO. Then, the output of this step will be segmented again by MSS. At the end, the segmented image will be classified by support vector machine (SVM). The letter is organized as follows: Methodology is discussed in Section II. Then, Section III is

Manuscript received December 3, 2012; revised January 24, 2013 and March 14, 2013; accepted March 21, 2013. Date of publication June 21, 2013; date of current version November 8, 2013. This work was supported in part by the Icelandic Research Fund for Graduate Students and the Ph.D. scholarship SFRH/BD/73382/2010 granted by the Portuguese Foundation for Science and Technology (FCT).

P. Ghamisi and J. A. Benediktsson are with the Faculty of Electrical and Computer Engineering, University of Iceland, Iceland (e-mail: peg6@hi.is; benedikt@hi.is).

M. S. Couceiro is with the Engineering Institute of Coimbra, RoboCorp, 3030–199 Coimbra, Portugal (e-mail: micael@isec.pt).

M. Fauvel is with the DYNAFOR Lab, INRA and the University of Toulouse, France (e-mail: mathieu.fauvel@ensat.fr).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2013.2257675

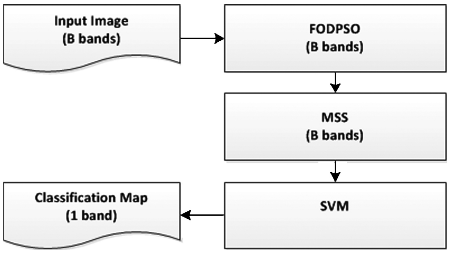


Fig. 1. Flowchart of the proposed methodology.

devoted to experimental results. Finally, in Section IV the main conclusions are outlined.

II. Methodology

The flowchart of the proposed method is illustrated in Fig. 1. The segmentation part consists of two different approaches: 1) multilevel thresholding method based on FODPSO; and 2) MSS. Then, the output of the segmentation methods will be classified by SVM. The following sections present a brief description of both segmentation methods.

A. *Multilevel Thresholding Method Based on Fractional Order Particle Swarm Optimization (FOPSO)*

Multilevel segmentation techniques provide an efficient way to carry out image analysis. However, the automatic selection of a robust optimum n-level threshold has remained a challenge in remote sensing image segmentation.

Let L represents the intensity levels in each component of a given image, where a component is defined in the range $\{0, 1, 2, \ldots, L-1\}$. Then, one can calculate the probability distribution p_i^C as

$$p_i^C = \frac{h_i^C}{N}, \qquad \sum_{i=1}^{N} p_i^C = 1 \tag{1}$$

where i represents a specific intensity level, i.e., $0 \leq i \leq L-1$, C represents the component of the pixel, e.g., $C = \{R, G, B\}$ for *RGB* images, N represents the total number of pixels in the image, and h_i^C denotes the number of pixels for the corresponding intensity level i in the component C. In other words, h_i^C represents an image histogram for each component C, which can be normalized and regarded as the probability distribution p_i^C.

Hence, the n-level thresholding presents $n-1$ threshold levels t_j^C, $j = 1, \ldots, n-1$, and the operation is performed as

$$F^C(a,b) = \begin{cases} 0, & f^C(a,b) \leq t_1^C \\ \frac{1}{2}\left(t_1^C + t_2^C\right), & t_1^C \leq f^C(a,b) \leq t_2^C \\ \vdots & \vdots \\ \frac{1}{2}\left(t_{n-2}^C + t_{n-1}^C\right), & t_{n-2}^C < f^C(a,b) \leq t_{n-1}^C \\ L, & f^C(a,b) > t_{n-1}^C \end{cases} \tag{2}$$

where a and b are the width (W) and height (H) pixel of the image of size $H \times W$ represented by $f^C(a, b)$. The pixels of a given image will be divided into n classes $D_1^C, \ldots, D_n^C$, which may represent multiple objects or even specific features on such objects (e.g., topological features).

The simplest method of obtaining the optimal threshold is the one that maximizes the between-class variance of each component, which can be generally defined by

$$\sigma_B^{c^2} = \sum_{j=1}^{n} w_j^C (\mu_j^C - \mu_T^C)^2 \tag{3}$$

where j represents a specific class in such a way that w_j^C and μ_j^C are the probability of occurrence and mean of class j, respectively. The total mean value of a component is represented by μ_T^C.

For classes $D_1^C, \ldots, D_n^C$, the probabilities of occurrence w_j^C and the means μ_j^C can be defined by (4) and (5), respectively

$$w_j^C = \begin{cases} \displaystyle\sum_{i=1}^{t_j^C} p_i^C, & j = 1 \\ \displaystyle\sum_{i=t_{j-1}^C}^{t_j^C} p_i^C, & 1 < j < n, \\ \displaystyle\sum_{i=t_{j-1}^C}^{L} p_i^C, & j = n \end{cases} \tag{4}$$

$$\mu_j^C = \begin{cases} \displaystyle\sum_{\substack{i=1 \\ C}}^{t_j^C} \frac{ip_i^C}{w_j^C}, & j = 1 \\ \displaystyle\sum_{\substack{i=t_{j-1}^C+1 \\ C}}^{t_j^C} \frac{ip_i^C}{w_j^C}, & 1 < j < n \\ \displaystyle\sum_{\substack{i=t_{j-1}^C+1 \\ C}}^{L} \frac{ip_i^C}{w_j^C}, & j = n \end{cases} . \tag{5}$$

The problem of n-level thresholding is reduced to an optimization problem to search for the thresholds t_j^C that maximize the objective functions of each image component C, generally defined as

$$\varphi^C = \max_{\substack{1<t_1^C<\cdots<t_{n-1}^C<L \\ C}} \sigma_B^{C^2}(t_j^C). \tag{6}$$

Computing this optimization problem involves a huge computational effort because the number of threshold levels and image components increases. Recently, biologically inspired methods, such as the well-known particle swarm optimization (PSO), have been used as computationally efficient alternatives to analytical methods to solve optimization problems [13].

An example of such methods is the FODPSO recently presented in [16]. This method is a natural extension of the Darwinian particle swarm optimization (DPSO) presented by Tillett *et al.* [14] using fractional calculus to control the convergence rate and was extended for the classification of remote sensing images in [8].

As in the classical PSO, particles within the FODPSO travel through the search space to find an optimal solution by interacting and sharing information with other particles. In each step of the algorithm t, a fitness function is used to evaluate the success for a particle. To model the swarm s, each particle n, moves in a multidimensional space according to a position $x_n^s[t]$, $0 \leq x_n^s[t] \leq L-1$, and velocity $v_n^s[t]$. The position and velocity values are highly dependent on the individually best $\breve{x}_n^s[t]$ and the globally best $\breve{g}_n^s[t]$, information

$$\begin{aligned} v_n^s[t+1] = {} & \alpha v_n^s[t] + \frac{1}{2}\alpha v_n^s[t-1] + \frac{1}{6}\alpha(1-\alpha)v_n^s[t-2] \\ & + \frac{1}{24}\alpha(1-\alpha)(2-\alpha)v_n^s[t-3] + \rho_1 r_1(\breve{g}_n^s - x_n^s[t]) \\ & + \rho_2 r_2(\breve{x}_n^s - x_n^s[t]) \end{aligned} \tag{7}$$

$$x_n^s[t+1] = x_n^s[t] + v_n^s[t+1]. \tag{8}$$

The coefficients ρ_1 and ρ_2 are weights, which control the global and individual performance, respectively. Within the FODPSO algorithm, the inertial influence of particles depends on the fractional coefficient. The parameters r_1 and r_2 are random vectors with each component is generally a uniform random number between 0 and 1. The parameter α, commonly known as the fractional coefficient, will weigh the influence of past events in determining a new velocity, $0 < \alpha < 1$.

When applying the FODPSO to multilevel thresholding of images, the particles' velocities are initially set to zero and their position is randomly set within the boundaries of the search space, i.e., $v_n^s[0] = 0$ and $0 \leq x_n^s[0] \leq L-1$. In other words, the search space depends on the number of intensity levels L, i.e., if one wishes to perform a segmentation of a 8-bit image, then particles will be deployed between 0 and 255. Hence, associated to each particle, a possible solution φ^c will be found and compared between all particles of the same swarm. The particle that has found the higher between-class variance φ^c so far will be the best performing particle (i.e., $\breve{g}_n^s$), thus luring other particles toward it. It is also noteworthy that when a particle improves, i.e., when a particle is able to find a higher between-class variance from one step to another, the fractional extension of the algorithm outputs a higher exploitation behavior. This allows achieving an improved collective convergence of the algorithm, thus allowing a good short-term performance.

FODPSO is a promising method to specify a predefined number of clusters with a higher between-class variance. In [9], the authors demonstrated that the FODPSO-based segmentation method performs considerably better in terms of accuracies than genetic algorithm, bacterial algorithm, PSO, and DPSO, thus finding different number of clusters with a higher between-class variance and more stability in less computational processing time. For further information on the FODPSO algorithm, please refer to [9] and [16].[1]

B. Mean Shift Segmentation

MSS is a nonparametric clustering technique, which requires neither embedded assumptions on the shape of the distribution nor the number of clusters in comparison to the classic K-means clustering approach. Mean shift was firstly introduced in [11]. This approach has been more recently developed for different purposes of low-level vision problems, including adaptive smoothing and segmentation [10].

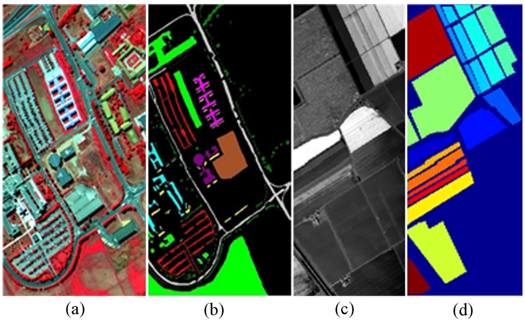


Fig. 2. Example of our test cases. (a) False color composition of Pavia dataset. (b) Reference map where each color represents a specific class. (c) Salinas dataset. (d) Reference.

TABLE I
INITIAL PARAMETERS OF THE FODPSO FOR OUR DATASETS

I_T	N	ρ_1, ρ_2	Δ_v	N_{min}	N_{max}	N^s	$N^s_{\min}$	$N^s_{\max}$	N_{kill}	α
100	30	0.8	5	10	50	4	2	6	10	0.6

The most important limitation of the standard MSS is that the value of the kernel size is unspecified. More information regarding the MSS can be found in [10].

III. EXPERIMENTAL RESULTS

A. Description of Datasets

1) *Pavia Data:* The first test case is a hyperspectral dataset captured on the city of Pavia, Italy, by airborne data from the *ROSIS-03*. In our experiments, 12 noisy bands were eliminated and 103 bands were processed. The spatial resolution is 1.3 m per pixel. The original dataset is 610 × 340 pixels. This dataset consisted of different classes, including trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow, and soil. Fig. 2(a) and (b) depicts Pavia dataset and its reference map.

2) *Salinas Data:* This scene was captured by AVIRIS sensor over Salinas Valley, CA, USA, and is characterized by high-spatial resolution (3.7-m pixels) consisting 512 lines and 217 samples. It includes vegetables, bare soils, and vineyard fields. The Salinas reference data contains 16 classes. Fig. 2(c) and (d) shows the Salinas dataset and its corresponding reference map.

The datasets have been classified with SVM and a Gaussian kernel. The hyper parameters have been selected using fivefold cross validation. The training set was randomly composed of 12.5% of the referenced set, the experiments have been repeated 20 times, and the mean accuracy and the standard deviation have been reported in Table II and IV.

The proposed multilevel thresholding techniques based on FODPSO were implemented with the specific parameters shown in Table I for our test cases. These parameters are chosen based on some studies from [3], [16].

Since the MSS approach is very dependent on the kernel size, two different kernel sizes were selected (5 and 20) in

[1] MATLAB code is available upon request from the authors.

TABLE II

THE κ COEFFICIENT AND OVERALL TEST ACCURACY OF DIFFERENT METHODS FOR THE PAVIA DATASET

Method	Kernel size	Mean(OA)	Mean (κ coefficient)
FODPSO + SVM		90.8 ± 0.192	0.887 ± 0.003
MSS ($R = 5$) + SVM	$R = 5$	98.84 ± 0.079	0.985 ± 0.001
FODPSO + MSS ($R = 5$) + SVM	$R = 5$	**98.92** ± 0.106	**0.986** ± 0.001
MSS ($R = 20$) + SVM	$R = 20$	97.72 ± 0.120	0.970 ± 0.002
FODPSO + MSS ($R = 20$) + SVM	$R = 20$	**98.04** ± 0.128	**0.974** ± 0.002
SVM		94.32 ± 0.174	0.925 ± 0.002

TABLE III

MANOVA RESULTS FOR THE PAVIA DATASET

Method	κ coefficient	OA
FODPSO + MSS ($R = 20$) + SVM *vs* MSS ($R = 20$) + SVM	0.004*	0.003*

All p-values corresponding to the mean differences are equal to 0.001.
*The corresponding mean difference is significant at the 0.05 level.

TABLE IV

THE κ COEFFICIENT AND OVERALL TEST ACCURACY OF DIFFERENT METHODS FOR THE SALINAS DATASET

Method	Kernel size	Mean (OA)	Mean (κ coefficient)
FODPSO + SVM		91.97 ± 0.17	0.91 ± 0.0020
MSS ($R = 5$) + SVM	$R = 5$	99.14 ± 0.05	0.99 ± 0.0005
FODPSO + MSS ($R = 5$) + SVM	$R = 5$	**99.13 ± 0.03**	**0.99 ± 0.0004**
MSS ($R = 20$) + SVM	$R = 20$	94.76 ± 0.19	0.94 ± 0.0021
FODPSO + MSS ($R = 20$) + SVM	$R = 20$	**96.27 ± 0.11**	**0.958 ± 0.0013**
SVM		94.06 ± 0.13	0.93 ± 0.0014

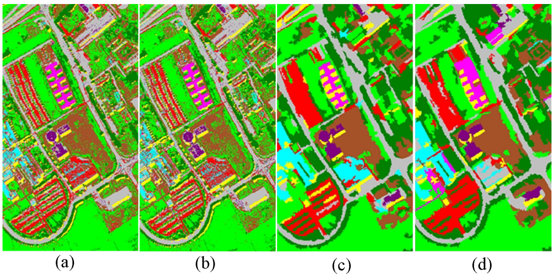


Fig. 3. Pavia classification result for (a) original image, (b) FODPSO, (c) MSS, and (d) FODPSO + MSS.

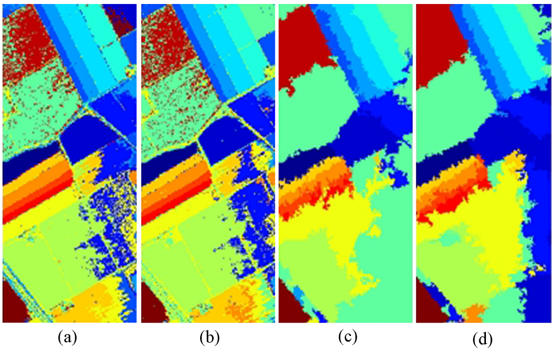


Fig. 4. Salinas classification result for (a) original image, (b) FODPSO, (c) MSS, and (d) FODPSO + MSS.

the experiments. The experimental evaluation will demonstrate whether the proposed method is highly dependent on the size of kernel or not.

B. Results and Discussion

1) *Pavia Dataset:* Table II illustrates the κ coefficient and overall accuracy (OA) for different methods for the Pavia dataset. As can be observed from Table II, FODPSO + SVM gave comparatively the worst performance in terms of accuracies. In histogram-based methods, the spatial information of data such as size and shape are not taken into consideration, and the final result is spatially independent and can be determined by considering only the histogram of the data. On the contrary, the MSS + SVM outperforms the FODPSO + SVM in terms of accuracies because it does not suffers from the above-mentioned disadvantages and can handle images with more complexity such as remote sensing images in a significant way. As can be seen in the table, FODPSO + MSS + SVM gave comparatively the best accuracies. Fig. 3 shows the output of classification for different methods.

To further improve the comparison between MSS + SVM and FODPSO + MSS + SVM, the significance of the method on the OA and the κ coefficient (dependent variables) was analyzed using the multivariate analysis of variance (MANOVA) technique after checking the assumptions of multivariate normality and homogeneity of variance/covariance. This is a statistical test procedure that allows comparing multivariate means of several groups. In other words, it allows comparing different methods (as it is the case) with more than one dependent variable (i.e., OA and the κ coefficient). In other words, the MANOVA merges the multiple dependent variables, thus creating a single dependent variable. For more information regarding MANOVA, it is referred to [2].

The assumption of normality for each of the univariate dependent variables was examined using univariate tests of Kolmogorov–Smirnov ($p < 0.05$). When the MANOVA detected significant statistical differences, we proceeded to the commonly used ANOVA for each dependent variable followed by the Tukey's HSD *post hoc*. The classification of the size effect (i.e., measure of the proportion of the total variation in the dependent variable explained by the dependent variable) was done according to Maroco [7] and Pallant [2]. This analysis was carried out using IBM SPSS Statistics for a significance level of 5%.

A two-way MANOVA analysis was carried out to assess whether the algorithms used in this letter have statistically significant differences with respect to the classification process. The MANOVA analysis revealed that the dependent variable κ coefficient presents statistically significant differences with large effect [$F_{(1,38)} = 66.656$; $p = 0.001$; $\eta_p^2 = 0.637$; power = 1.0], as well as the dependent variable OA [$F(1,38) = 66.491$; $p = 0.001$; $\eta_p^2 = 0.636$; power = 1.0] (see Table III).

2) *Salinas Dataset:* For the Salinas dataset, FODPSO + SVM gave the worst accuracies (Table IV).

TABLE V
MANOVA RESULTS FOR THE SALINAS DATASET

Method	κ	OA
FODPSO + MSS ($R=5$) + SVM *vs* MSS ($R=5$) + SVM	−9.258E-5	−8.023E–5
FODPSO + MSS ($R=20$) + SVM *vs* MSS ($R=20$) + SVM	0.017*	0.015*

All p-values corresponding to the mean differences are equal to 0.001.
*The corresponding mean difference is significant at the 0.05 level.

Furthermore, the overall classification accuracy by MSS + SVM dropped from 99.14% to 94.76% when the kernel size was increased from 5 to 20. This dramatic decrease in accuracy shows that the result of the classification by using MSS is highly dependent on the kernel size. The kernel size must still be tuned by user who might find the task difficult since the size can dramatically influence the final result. A larger or smaller kernel may influence the result of the segmentation and considerably reduce the efficiency of the MSS method. Mode candidates with a distance that is less than the kernel size are merged and may cause to lose information on an image. In contrast, a small kernel size may cause a high increase on the CPU processing time. As the FODPSO is able to find modes with maximum between-class distance, the influence of tuning the size of the kernel size is significantly reduced. In other words, the two methods can solve each other's problems and, thus, complement each other. Fig. 4 illustrates the result of the classification for the different methods. Compared with the MSS + SVM, the FODPSO + MSS + SVM increased the accuracy from 94.76% to 96.27% with the kernel size of 20. Considering both kernel sizes ($R=5$ and 20), it can be stated that the FODPSO + MSS + SVM shows more stability. FODPSO + MSS + SVM improved the result of the traditional SVM by almost 5% and 2% by considering kernels with the size of 5 and 20, respectively (Table IV).

For the Salinas dataset, the MANOVA was separately carried out on MSS ($R=5$) + SVM, FODPSO + MSS ($R=5$) + SVM and MSS ($R=20$) + SVM, FODPSO + MSS ($R=20$) + SVM (Table V). The MANOVA analysis revealed that the κ coefficient does not present statistically significant differences ($F_{(1,38)}=0.415$; $p=0.523$; $\eta_p^2 = 0.011$; power = 0.96). A similar result was observed for the OA ($F(1,38)=0.386$; $p=0.538$; $\eta_p^2 = 0.010$; power = 0.93).

For $R=20$, the MANOVA analysis depicted that the dependent variable κ coefficient presents statistically significant differences with large effect ($F_{(1,38)}=1111.827$; $p=1.00$; $\eta_p^2 = 0.967$; power = 1.00), as well as the dependent variable OA ($F_{(1,38)}=1112.876$; $p=0.001$; $\eta_p^2 = 0.967$; power = 1.00). Results show that by increasing the size of the kernel, the proposed method works better than others in terms of accuracies.

IV. CONCLUSION

In this letter, a new spectral–spatial classification approach is introduced for accurate classification of hyperspectral images. The approach is based on the combination of FODPSO and MSS. FODPSO is a very powerful approach for finding the predefined number of clusters with the highest between-class value. In the proposed approach, the result of FODPSO is used as the input to MSS to develop a pre-processing method for classification. Tuning the size of the kernel can be considered as the main difficulty of MSS and the obtained result may considerably be affected by the kernel size. The SVM is used for classification on the outcome of these two segmentation methods. Results indicate that the use of both segmentation methods can overcome the shortcomings of each other and the combination can improve the result of classification significantly.

ACKNOWLEDGMENT

The ROSIS Pavia data were obtained from Prof. Paolo Gamba of the University of Pavia, Italy.

REFERENCES

[1] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[2] J. Pallant, *SPSS Survival Manual*, 4th ed. London, U.K.: Open Univ. Press, 2011.

[3] M. S. Couceiro, F. M. L. Martins, R. P. Rocha, and N. M. F. Ferreira, "Analysis and parameter adjustment of the RDPSO—Towards an understanding of robotic network dynamic partitioning based on Darwin's theory," *Int. Math. Forum*, Hikari, Ltd., vol. 7, no. 32, pp. 1587–1601, 2012.

[4] A. Farag, R. Mohamed, and A. El-Baz, "A unified framework for map estimation in remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.* vol. 43, no. 7, pp. 1617–1634, Jul. 2005.

[5] A. Darwish, K. Leukert, and W. Reinhardt, "Image segmentation for the purpose of object-based classification," *Proc. IGARSS'03*, vol. 3, pp. 2039–2041, 2003.

[6] J. Tilton, "Analysis of hierarchically related image segmentations," in *Proc. IEEE Workshop Advances in Techniques for Analysis of Remotely Sensed Data*, 2003, pp. 60–69.

[7] J. Maroco, *Análise Estatística com utilização do SPSS*. Lisboa: Edições Silabo, 2010.

[8] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, "Use of Darwinian particle swarm optimization technique for the segmentation of remote sensing images," in *Proc. IGARSS*, 2012, pp. 4295–4298.

[9] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, "An efficient method for segmentation of images based on fractional calculus and natural selection," *Expert Syst. Appl.*, vol. 39, pp. 12407–12417, 2012.

[10] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 603–619, Aug. 2002.

[11] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 1, pp. 32–40, Jan. 1975

[12] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003

[13] M. S. Couceiro, J. M. A. Luz, C. Figueiredo, and N. M. F. Ferreira, "Modeling and control of biologically inspired flying robots," *Robotica*, vol. 30, no. 1, pp. 107–121, 2012.

[14] J. Tillett, T. M. Rao, F. Sahin, R. Rao, and S. Brockport, "Darwinian particle swarm optimization," in *Proc. 2nd Indian Int. Conf. Artificial Intelligence*, 2005, pp. 1474–1487.

[15] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[16] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, "Extending the fractional order Darwinian particle swarm optimization to segmentation of hyperspectral images," *Proc. SPIE , Image and Signal Processing for Remote Sensing XVIII, 85370F,* Edinburgh, U.K., Nov. 2012.

CHAPTER 4

Spectral-Spatial Classification based on an Adaptive Neighborhood System: Attribute Profiles

A Survey on Spectral–Spatial Classification Techniques Based on Attribute Profiles

Pedram Ghamisi, *Student Member, IEEE*, Mauro Dalla Mura, *Member, IEEE*, and Jon Atli Benediktsson, *Fellow, IEEE*

***Abstract*—Just over a decade has passed since the concept of morphological profile was defined for the analysis of remote sensing images. Since then, the morphological profile has largely proved to be a powerful tool able to model spatial information (e.g., contextual relations) of the image. However, due to the shortcomings of using the morphological profiles, many variants, extensions, and refinements of its definition have appeared stating that the morphological profile is still under continuous development. In this case, recently introduced theoretically sound attribute profiles (APs) can be considered as a generalization of the morphological profile, which is a powerful tool to model spatial information existing in the scene. Although the concept of the AP has been introduced in remote sensing only recently, an extensive literature on its use in different applications and on different types of data has appeared. To that end, the great amount of contributions in the literature that address the application of the AP to many tasks (e.g., classification, object detection, segmentation, change detection, etc.) and to different types of images (e.g., panchromatic, multispectral, and hyperspectral) proves how the AP is an effective and modern tool. The main objective of this survey paper is to recall the concept of the APs along with all its modifications and generalizations with special emphasis on remote sensing image classification and summarize the important aspects of its efficient utilization while also listing potential future works.**

***Index Terms*—Attribute profile (AP), hyperspectral image analysis, morphological attribute filters (AFs), spatial features, spectral–spatial classification.**

I. Introduction

SUPERVISED classification is an important process in remote sensing image analysis. A wide range of applications such as crop monitoring, forest applications, urban development, mapping and tracking, and risk management can be handled by using appropriate data and efficient classifiers. A large amount of data with different spectral, spatial, and temporal resolutions are currently being made available for different applications. Hyperspectral imaging sensors are able to capture hundreds of narrow spectral channels with a very fine spectral resolution, which is helpful for detailed physical analysis of structures in the captured image [1]. In addition, due to recent advances in remote sensing technologies, spatial resolution of sensors is also improving [1], which has led to a better identification of relatively small structures.

Conventional spectral classifiers consider the image as an ensemble of spectral measurements without exploiting their spatial arrangement. In other words, the spatial organization of distinct pixels is not considered in spectral classification [2]. In order to make use of the spatial organization, a joint spectral and spatial classifier is required to reduce the labeling uncertainty that exists when only spectral information is taken into account. Furthermore, more spatially homogeneous classification maps are produced. Moreover, spatial information provides additional discriminant information related to the shape and size of different structures, which, if properly exploited, leads to more accurate classification maps.

In order to model the spatial information of a scene, two common strategies are available: a crisp neighborhood system and an adaptive neighborhood system [1]. While the first one mostly relies on considering spatial and contextual dependence relations in a predefined neighborhood system, the latter shows more flexibility and is not confined to a given neighborhood system.

One well-known way for extracting spatial information by using a crisp neighborhood system is the use of Markov Random Field (MRF) modeling[1] (see the list of abbreviations shown in Table I). MRF is a family of probabilistic models and can be explained as a 2-D stochastic process over discrete pixel lattices [3]. MRF is considered to be a powerful tool for incorporating spatial and contextual information into the classification framework [4]. There is a considerable literature on the use of MRFs in classification. For example, in [5], the result of the probabilistic support vector machine (SVM) was regularized by an MRF. In [6], a fully automated framework for the spectral and spatial classification of hyperspectral (multispectral) data was proposed, which was based on the integration of a modification of MRF (hidden MRF) and SVM. However, the main disadvantages of considering a set of crisp neighbors are as follows.

1) The standard neighborhood system may not contain enough samples to characterize the object of interest, and this downgrades the effectiveness of the classifier (in

Manuscript received May 28, 2014; revised August 2, 2014; accepted September 15, 2014. This work was supported in part by the Icelandic Research Fund for Graduate Students and the program J. Verne 2014 under Project 31936TD.

P. Ghamisi and J. A. Benediktsson are with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavík, Iceland (e-mail: peg6@hi.is; benedikt@hi.is).

M. Dalla Mura is with the Speech Signals and Automatics Laboratory (GIPSA-Lab), Grenoble Institute of Technology (Grenoble INP), 38031 Grenoble Cedex 1, France (e-mail: mauro.dalla-mura@gipsa-lab.grenoble-inp.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2014.2358934

[1] Please note that, here, we are discussing the most well-known MRF model, which models the spatial information of adjacent pixels by considering a crisp neighborhood system. However, MRFs based on an adaptive neighborhood system can be found in the literature as well.

TABLE I
LIST OF ABBREVIATIONS

VHR	Very High Resolution
MRF	Markov Random Field
SE	Structuring Element
MP	Morphological Profile
EMP	Extended Morphological Profile
DMP	Differential Morphological Profile
AF	Attribute Filter
AP	Attribute Profile
EAP	Extended Attribute Profile
EMAP	Extended Multi-Attribute Profile
EEMAP	Entire Extended Multi-Attribute Profile
SDAP	Self-Dual Attribute Profile
SVM	Support Vector Machine
RF	Random Forest
MLR	Multinomial Logistic Regression
SRC	Sparse Representation Classification
RBF	Radial Basis Function
FS	Feature Selection
FE	Feature Extraction
PC	Principal Component
PCA	Principal Component Analysis
KPCA	Kernel Principal Component Analysis
ICA	Independent Component Analysis
DAFE	Discriminant Analysis Feature Extraction
DBFE	Decision Boundary Feature Extraction
NWFE	Nonparametric Weighted Feature Extraction
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
BFODPSO	Binary Fractional Order Darwinian Particle Swarm Optimization
GLCM	Gray-Level Co-occurrence Matrix

particular, when the input data set is of high resolution and the neighboring pixels are highly correlated [1]).

2) A larger neighborhood system leads to intractable computational problems [1].

In order to address the aforementioned issues, an adaptive neighborhood system can be considered. One possible way for considering the adaptive neighborhood system is to utilize different types of segmentation methods. Segmentation of images into spatially homogenous regions may improve the accuracy of classification maps [7]. To make such an approach more effective, an accurate segmentation of the image is required [8]. There is an extensive literature on the use of segmentation techniques in order to extract the spatial information (e.g., [9]–[12]).

Another possible set of approaches that are able to extract spatial information by using an adaptive neighborhood system relies on the concept of the morphological profile (MP). An MP is constructed based on the repeated use of openings and closings by reconstruction with an structuring element (SE) of increasing size, applied to a scalar image. MPs simultaneously attenuate some spatial details and preserve the geometrical characteristics of the other regions. Pesaresi and Benediktsson [13] used morphological transformations to build the so-called MP. In [14], the MP generated by morphological opening and closing operations was used for classifying a Quickbird panchromatic image captured over Bam, Iran, which was hit by an earthquake in 2003. To do so, the spatial features extracted by the MP were considered for assessing the damages caused by the earthquake. The standard opening and closing along with white and black top hat [15] and opening and closing by reconstruction were computed, and the resulting features were classified using an SVM classifier for the classification of a Quickbird panchromatic image in [16]. An automatic hierarchical segmentation technique based on the analysis of the differential morphological profile (DMP) (the derivative of the MP) was proposed in [17]. The DMP was also analyzed in [18], by extracting a fuzzy measure of the characteristic scale and contrast of each structure in the image. The computed measures were compared with the possibility distribution predefined for each thematic class, generating a value of membership degree for each class used for classification. In [19], in order to reduce the dimensionality of data and address the so-called curse of dimensionality [20], feature extraction (FE) techniques were taken into consideration for the DMP classified by a neural network classifier. In [21], the concept of MPs was successfully extended to handle hyperspectral images, resulting in the extended morphological profile (EMP). The EMP is obtained by first reducing the dimensionality of the hyperspectral image with a principal component analysis (PCA) and by computing an MP on each of its first few components.

Some studies have been conducted in order to assess the capability of SEs with different shapes for the extraction of spatial information. For instance, MPs computed with a compact SE (e.g., square, disk, etc.) can be considered for modeling the size of the objects in the image (e.g., in [1], this information was exploited to discriminate small buildings from large ones). In [22], the computation of two MPs was introduced in order to model both the length and the width of the structures. In more detail, one MP is built using disk-shaped SEs for extracting the smallest size of the structures, whereas the other employs linear SEs (which generate directional profiles [23]) for characterizing the object's maximum size (along with the orientation of the SE). This is appropriate for defining the minimal and maximal length. However, such analysis is computationally intensive as all the possible lengths and orientations cannot be practically investigated. Moreover, in [22], Bellens *et al.* proposed the use of operators based on "partial reconstruction" instead of the conventional geodesic reconstruction in order to reduce the "leakage effect." In [24], a new binary optimization method inspired by the fractional-order Darwinian particle swarm optimization (PSO) [8] is introduced in order to select the most informative features extracted by MP.

Based on the aforementioned literature, it is easy to infer that multiscale processing based on morphological filters (e.g., by MPs, DMPs, and EMPs) has proven to be effective in extracting informative spatial features from the images to be analyzed. Although MP is a powerful technique for the extraction of spatial information, the concept has a few limitations: 1) The shape of SEs is fixed; and 2) SEs are unable to characterize information related to the gray-level characteristics of the regions. To overcome this, the morphological AP has been proposed as the generalization of the MP, which provides a multilevel characterization of an image by using the sequential application of a morphological attribute filter (AF) [25]. AFs are connected operators that process an image by considering only its connected components. For binary images, the connected components are simply the foreground and background regions present in the image. In order to deal with grayscale images, the set of connected components can be obtained by considering the image to be composed by a stack of binary images generated by thresholding the image at all its gray-level values [26]. Thus, they process the image without distorting or inserting new edges but only by merging existing flat regions [15]. AFs were employed for modeling the structural information of the scene in order to increase the effectiveness of a classification and building extraction in [25] and [27], respectively, where they proved to be efficient for the modeling of structural information in very high resolution (VHR) images. In [28], AFs were used in a scheme for building height retrieval by considering VHR images acquired on the same area with different acquisition angles. That is, the filters were used for reducing the complexity of the images in order to isolate regions that correspond to the rooftops of buildings. A neural network was used to find correspondences between the extracted regions in the different images by considering moment invariant descriptors as features. After finding correspondences and since the building considered had a flat roof, the horizontal displacement of matching regions was converted in height by trigonometry.

AFs include in their definition the morphological operators based on geodesic reconstruction [29]. Moreover, an AP is a flexible tool since images can be processed based on many different types of attributes. In fact, the attributes can be of any type. For example, they can be purely geometric, related to the spectral values of the pixels, or on different characteristics such as spatial relations to other connected components. Furthermore, in [27], the problem of tuning the parameters of the filters was addressed by proposing an automatic feature selection (FS) procedure based on a genetic algorithm (GA). In [30], it was proved that the automatic method with considering only two attributes (area and standard deviation) is comparative with a manual technique with four attributes in terms of classification accuracy and CPU processing time. In [31], a topographic map of the image was used, which is autodual (it is invariant to contrast inversion) and does not require any SE in order to extract the profiles. In addition, the concept of MPs was extended to the profiles of other features (e.g., perimeters, scales, and total variations).

This work presents a survey over the existing papers related to AP with special emphasis on multispectral and hyperspectral image classification, while still providing a general framework for other types of data. The rest of this survey paper is organized as follows: First, a few primary concepts related to morphological profiles will be disscused in Section II. Then, the concept of AP and its extension for hyperspectral data will be explained in Section III. Then, Section IV is devoted to spectral and spatial classification of remote sensing data by considering AP. Section V is on the use of AP for other types of applications such as change detection and other types of data such as LiDAR. In Section VI, the main obtained points, the advantages and disadvantages of different techniques, and the survey of existing experiments with respect to classification results will be briefly discussed. Finally, Section VII outlines the main conclusions and possible future works.

II. Morphological Profile

Here, first, we recall a few primary concepts such as *connected components*, *basic morphological operators*, and *morphological profile* and their modifications.

A. Connected Components

A connected component is regarded as a group of isolevel pixels that are connected according to a predefined connectivity rule. The most well-known connectivity rules are 4- and 8-connected, where a pixel is considered as adjacent to four or eight of its neighboring pixels, respectively.

B. Basic Morphological Operators

Erosion and dilation are considered as the basic building blocks of mathematical morphology. These operations are carried out on an image with a set of known shape, which is called an SE. Opening and closing are combinations of erosion and dilation. These operators simplify the input data by removing structures with size less than that of the SE. However, they can influence the shape of the structures and can introduce fake objects in the image [32]. One way to handle this issue is to consider opening and closing by reconstruction [15].

Opening and closing by reconstruction filters are a family of connected operators that satisfy the following criterion: If the SE cannot fit in an object, then it will be totally removed; otherwise, it will be totally preserved. Reconstruction operators remove objects smaller than the SE without altering the shape of those objects and reconstruct connected components from the preserved objects. For gray-scale images, opening by reconstruction removes unconnected light objects, and in dual, closing by reconstruction removes unconnected dark objects. Fig. 1 illustrates the different results that are obtained when considering operators with or without geodesic reconstruction.

C. Morphological Profile and Its Modifications

In order to characterize the scale of different structures present in an image, it is very important to consider a range of SEs with different sizes. MPs use successive opening/closing operations with an SE of increasing size. The successive application of opening/closing leads to a simplification of the input image and a better understanding of different available structures in the image. An MP consists of an opening profile and a closing profile. In order to fully exploit the spatial information, filtering techniques should simultaneously attenuate the unimportant details and preserve the geometrical characteristics of the other regions. In [13], morphological transformations were used to build an MP. They carried out a multiscale analysis by computing an antigranulometry and a granulometry (i.e., a sequence of closing and opening with an SE of increasing size), appended in a common data structure named MP.

Another modification of using MP, which was exploited for the classification of VHR panchromatic images, is DMP. DMP is composed of the residues of two subsequent filtering operations for two adjacent levels existing in the profile. Since

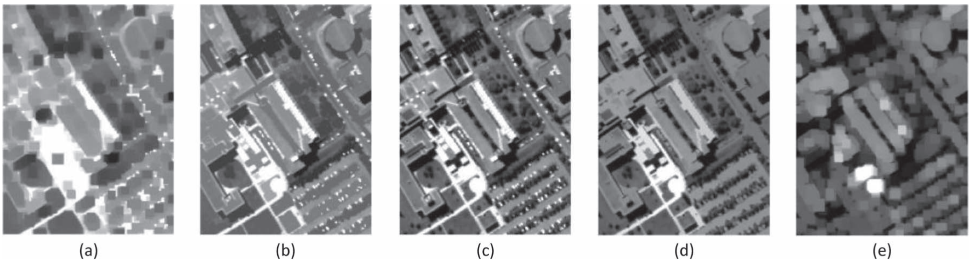


Fig. 1. (a) Morphological closing. (b) Closing by reconstruction. (c) Original VHR panchromatic image. (d) Opening by reconstruction. (e) Morphological opening. As can be seen, morphological opening and closing have influences on the shape of the structures and can introduce fake objects. However, opening and closing by reconstruction preserve the shape of different objects bigger than the SE. The disk-shaped SE with a radius size of three pixels is taken into account.

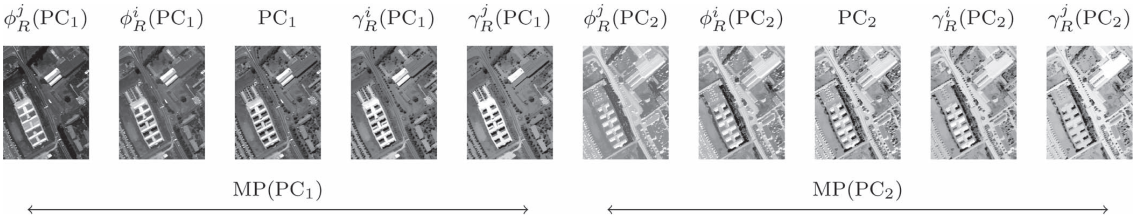

$\phi_R^j(\mathrm{PC}_1)$
$\phi_R^i(\mathrm{PC}_1)$
PC_1
$\gamma_R^i(\mathrm{PC}_1)$
$\gamma_R^j(\mathrm{PC}_1)$
$\phi_R^j(\mathrm{PC}_2)$
$\phi_R^i(\mathrm{PC}_2)$
PC_2
$\gamma_R^i(\mathrm{PC}_2)$
$\gamma_R^j(\mathrm{PC}_2)$
$\mathrm{MP}(\mathrm{PC}_1)$
$\mathrm{MP}(\mathrm{PC}_2)$

Fig. 2. Example of a simple EMP consisting of two PCs.

the DMP is the derivative of the MP, it has a number of levels, which is one less than the number of levels in the MP.

In [21], the concept of MPs was successfully extended to handle hyperspectral images by using PCA for reducing the dimensionality of the hyperspectral data. The dimensionality is reduced by only considering the few first components of the transformation, which retain most of the information (here expressed by the variance) of the original data. MPs were generated for the selected PCs of the data and stacked into a single profile named EMP. Fig. 2 shows a stacked vector consisting of the profiles based on the first and second PCs. Since the EMP does not fully exploit the spectral information and PCA does not consider class information, in [32], different supervised FE techniques are used instead of PCA, and the EMP stacked along with other extracted features is classified using an SVM.

Although MP is a powerful technique for the extraction of spatial information, the concept suffers from a few limitations as follows.

1) The shape of SEs is fixed, which is considered as a main limitation for the extraction of objects within a scene.
2) SEs are unable to describe information related to the gray-level characteristics of the regions such as spectral homogeneity, contrast, and so on.
3) A final limitation associated with the concept of MPs is the computational complexity. The original image needs to be completely processed for each level of the profile, which requires two complete evaluations of the image: one performed by a closing transformation and the other by an opening transformation. Thus, the complexity linearly increases with the number of levels included in the profile [25].

To address the aforementioned issues, the concept of attribute profile (AP) was proposed in [25].

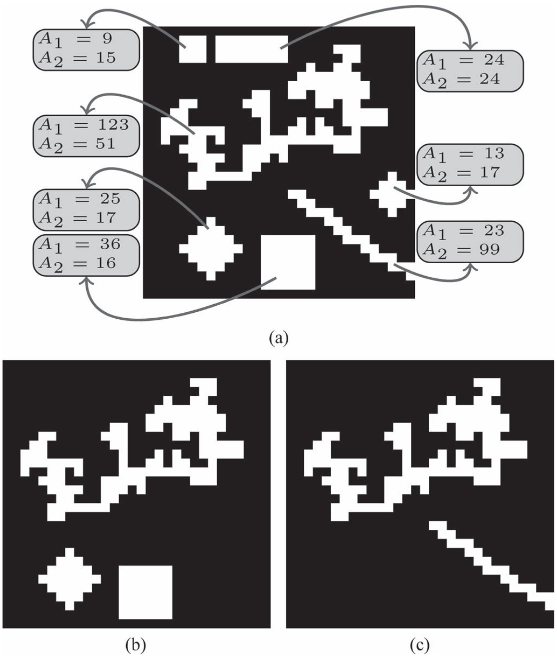
III. AP

A. Morphological AFs

Morphological AFs are *connected filters* [33]; hence, they process an image by only merging its connected components. We will now detail how such set of connected components can be derived from an image.

Let us consider a discrete 2-D image $f : E \rightarrow T$, with E as the discrete image domain ($E \subseteq \mathbf{Z}^2$) and $T \subseteq \mathbf{Z}$ as the set of possible scalar values associated to the elements (i.e., pixels) of E. It is well known that it is possible to decompose a scalar image into a set of binary images by the so-called *threshold decomposition principle* (i.e., $f = \sum_t f_t$ with $f_t : f \geq t$ and $t \in T$) [34], [35]. Using the analogy of an image with a topographic surface in which the elevation of the map corresponds to the intensity of the gray level, the image can be seen as a superposition of all the isolevel maps (i.e., slices of the 3-D map at all the possible height levels). Each binary image is composed of connected components, and in this representation, by varying the threshold's value (i.e., the height of the plane), connected components can merge, enlarge, shrink, split, appear, or disappear according to the spatial organization and the intensity values of the pixels in the image. Since elements of T are ordered, f can be equivalently decomposed into an *upper* or a *lower level set*, which are defined as the sets of binary maps obtained by considering the upper (i.e., $\geq t$) or the lower (i.e., $< t$) threshold for all the possible values of the pixels [36].

AFs operate through a transformation based on a predicate P (i.e., $P : S \rightarrow \{false, true\}$, with S as a generic set of values),



(a)

(b) (c)

Fig. 3. Illustrative example for attribute filtering. (a) Binary image in which two different attributes were computed for each connected component of the image upper level set. Only connected components of the foreground are considered in this example. Attribute A_1 is a scale-dependent measure (i.e., the area in number of pixels for each region), and A_2 is a shape index invariant to scale and rotation (i.e., the moment of inertia multiplied by a factor 10^2 was considered as measure). (b) Result of a thinning with predicate $A_1 > 25$. (c) Result of a thinning with predicate $A_2 > 30$.

which is evaluated on each connected component (obtained by the image decomposition). Different filtering effects are obtained by considering either the components of the upper or the lower level set. The predicates implement a comparison between the value of a generic attribute $\mathcal{A}$ computed on a component C and a predefined threshold value λ, e.g., $P(C) = \mathcal{A}(C) \geq \lambda$. Any measure that can be computed for an image region can act as an attribute [33]. Moreover, even multiple attributes can be considered in the same transformation if they are evaluated in a single joint predicate. The filtering operates on each connected component according to the output of the predicate: If P is fulfilled, the component is preserved; otherwise, it is merged to one of its adjacent components (i.e., setting its gray level to the gray level of the component to whom it will be merged to). An important property of P is *increasingness*. A criterion is said to be increasing when it is verified for a connected component, then it will be also true for all the components in which the component is nested. This property leads to have, for example, $P(C_j) = true$ when also $P(C_i) = true$ for any $C_j \subseteq C_i$. Examples of increasing criteria involve increasing attributes (e.g., area, volume, size of the bounding box, etc.) and an inequality relation (e.g., $\geq$). In contrast, nonincreasing attributes, such as scale-invariant measures (e.g., gray-level homogeneity, shape descriptors, region orientation, etc.), lead to nonincreasing criteria.

When considering the components of the set, the result of the filtering is a *thinning*, which is denoted by γ, since the transformation obtained in this case is idempotent and antiextensive.[2] If the predicate is increasing, the filter also increases, leading to an opening. Analogous considerations can be done for the dual transformation by considering the lower level set. The transformation is a *thickening*, which is denoted by ϕ, and if the criterion is increasing, it becomes a *closing*. Fig. 3 shows an example of attribute filtering on a binary image. Two attributes, namely, A_1 based on the size and A_2 based on the shape of the regions, were computed on the connected components of the image [see Fig. 3(a)]. The results of two thinning operators, i.e., one based on A_1 and another on A_2 (with an arbitrary predicate), are shown in Fig. 3(b) and (c). It is worth noting that it is not possible to achieve the result in Fig. 3(c) (i.e., removing all the foreground objects with a compact shape) with a single filtering based on A_1. In the same way, the removal of objects based on their scale cannot be obtained considering A_2. Furthermore, when considering connected operators based on SEs (such as opening and closing by reconstruction), the result in Fig. 3(b) can be equivalently obtained with any SE that is not contained in the foreground objects of smaller scale (but contained in the three largest objects). However, as in Fig. 3(c), the result cannot be straightforwardly achieved with a single filtering based on an SE due to the different scale of the structures meant to be preserved.

There is an inclusion relations between the connected components in the image (obtained from the upper or the lower set), which means that any two components are either nested or disjoint. Due to this, the set of connected components can be represented by a tree, its nodes being the components and the links between nodes being the inclusion relations between components. The tree derived by the components in the upper (resp., lower) level set is called *max-tree* (resp., *min-tree*) [37].[3] Another representation of a gray-scale image as a hierarchy of regions is the *inclusion tree*. Here, the set of connected components is obtained by progressively “filling” regions internal to others (i.e., considered as “holes”), and the connections in the tree are determined by the considered region-filling operator [39].

Such hierarchical representations of an image can be effectively exploited for the computation of morphological AFs [37], [40]. Thinnings and thickenings will be obtained from a max-tree and a min-tree, respectively. The image transformation done by the filter on the image is equivalent to a pruning of the tree, i.e., the removal of single nodes or branches.

Different pruning strategies exist depending on whether the predicate evaluated by the AF is increasing or not [40]. The increasingness of a predicate leads to the removal of entire branches (i.e., a node with all its descendants up to their leaves) from the tree. Conversely, for nonincreasing predicates, intermediate nodes in a branch can either fulfill the predicate or

[2] We recall that for a generic transformation ψ on an image f (and g), idempotence means $\psi(\psi(f)) = \psi(f)$, increasingness $f \leq g \Leftrightarrow \psi(f) \leq \psi(g)\ \forall\, f, g$, and antiextensivity (resp., extensivity) refers to $f \geq \psi(f)$ (resp., $f \leq \psi(f)$).

[3] In [38], min- and max-trees are called *component trees*.

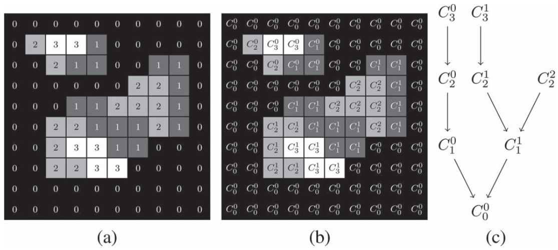

0 0 0 0 0 0 0 0 0 0
0 2 3 3 1 0 0 0 0 0
0 0 2 1 1 0 0 1 1 0
0 0 0 0 0 0 2 2 1 0
0 0 0 1 1 2 2 2 1 0
0 0 2 2 1 1 1 2 1 0
0 0 2 3 3 1 1 0 0 0
0 0 2 2 3 3 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
(a)
C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0
C_0^0 C_2^0 C_3^0 C_3^0 C_1^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0
C_0^0 C_0^0 C_2^0 C_1^0 C_1^0 C_0^0 C_0^0 C_1^1 C_1^1 C_0^0
C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_2^2 C_2^2 C_1^1 C_0^0
C_0^0 C_0^0 C_0^0 C_1^1 C_1^1 C_2^2 C_2^2 C_2^2 C_1^1 C_0^0
C_0^0 C_0^0 C_2^1 C_2^1 C_1^1 C_1^1 C_1^1 C_2^2 C_1^1 C_0^0
C_0^0 C_0^0 C_2^1 C_3^1 C_3^1 C_1^1 C_1^1 C_0^0 C_0^0 C_0^0
C_0^0 C_0^0 C_2^1 C_2^1 C_3^1 C_3^1 C_0^0 C_0^0 C_0^0 C_0^0
C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0
C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0 C_0^0
(b)
C_3^0
C_3^1
C_2^0
C_2^1
C_2^2
C_1^0
C_1^1
C_0^0
(c)

Fig. 4. Example of max-tree. (a) Gray-scale image with intensities ranging from 0 to 3. (b) Image in (a) with its connected components labeled. (c) Max-tree of (a). This shows the relations between the nodes associated to the connected components in (b) [42].

not. For that case, different filtering rules have been defined in the literature [37], [41].

An example of max-tree is shown in Fig. 4. As one can see in Fig. 4(b), the image is composed by connected components of isointensity pixels. The max-tree maps each of all the connected components of the image to a node organized in a hierarchical tree structure [see Fig. 4(c)]. The root node of the tree represents the whole image at his lowest gray level. The tree grows by connecting the nodes of the progressively nested connected components in the image up to the leaves of the tree that correspond to the regional maxima in the image.

B. AP and Its Extension to Vectorial Images

Although this section should be considered self-sufficient for understanding the concept of AP, extended AP (EAP), and extended multiattribute profile (EMAP), for more information regarding the aforementioned concepts, please refer to [25] and [42]. More useful references can be found throughout the paper.

APs are obtained by the outputs of a sequence of thinning and thickening transformations applied on a scalar image [25]. APs can be seen as a generalization of MPs since both opening and closing by reconstruction can be implemented as AFs [33]. The motivation for using AFs is to overcome the limitation of the conventional operators based on geodesic reconstruction in defining a decomposition of the image based on characteristics different from the scale [43]. In fact, MPs naturally perform a multiscale decomposition of an image, since structuring elements of a fixed shape and an increasing size are employed. However, trying to compute an MP based on the shape cannot be easily achieved since scale-invariant characteristics are poorly modeled by SEs (it would require a very large set of SEs, since the analysis should be scale invariant and many different shapes of the SE should be considered). Instead, by using AFs, the image decomposition can be based on the scale (as for MP), shape, texture, etc., according to the type of attribute considered.

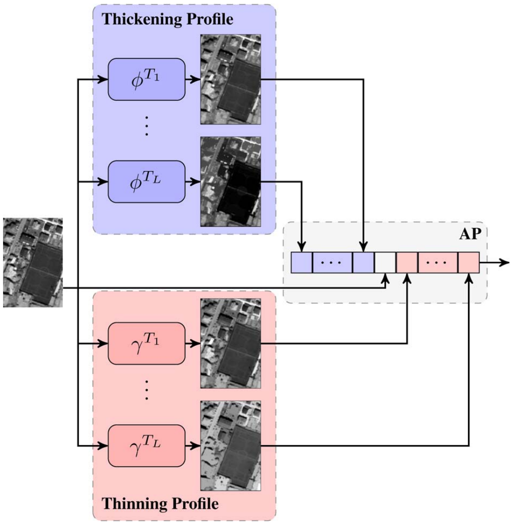


Fig. 5. Example of the general architecture of AP.

More formally, an AP is defined as in (1), shown at the bottom of the page [25], with $P_\lambda : \{P_{\lambda_i}\}$ $(i = 1, \ldots, L)$ as a set of L ordered predicates (i.e., $P_{\lambda_i} \subseteq P_{\lambda_k}$, $i \leq k$). The sequence of criteria considered for constructing the profile has to be ordered for guaranteeing the fulfillment of the absorption property, which might not be verified for nonincreasing predicates. Thus, the AP can be seen as a stack of thickening and thinning profiles. The thickening profile is considered in reversed order, such that the coarser image appears first and the original image last. The original image f also appears in the profile since it can be considered as the level zero of both the thickening and thinning profiles (i.e., $\phi^{P_{\lambda_0}}(f) = \gamma^{P_{\lambda_0}}(f) = f$, where P_{λ_0} is a predicate that is fulfilled by all the components in the image, leading to no filtering). According to the attribute and criterion considered, different information can be extracted from the structures in the scene leading to different multilevel characterizations of the image. [25]. We refer the reader to [25] for further details. Fig. 5 shows an example for the general architecture of AP.

We recall also that in [45], an inclusion tree was used for computing the AP instead of the min-tree and max-tree data representation, leading to a self-dual AP. In that work, the application of self-dual connected operators led to an image simplification characterized by more homogeneous regions with respect to the results obtained by extensive or antiextensive connected operators.

When dealing with vectorial images ($\mathbf{f} : E \to T$, with $E \subseteq \mathbf{Z}^2$ and $T \subseteq \mathbf{Z}^n$, $n > 1$ and $\mathbf{f} = \{f_1, f_2, \ldots, f_n\}$) such as multispectral and hyperspectral images (where n is the number

$$AP(f) = \left\{ \underbrace{\phi^{P_{\lambda_L}}(f), \phi^{P_{\lambda_{L-1}}}(f), \ldots, \phi^{P_{\lambda_1}}(f)}_{\text{thickening profile}}, f, \underbrace{\gamma^{P_{\lambda_1}}(f), \ldots, \gamma^{P_{\lambda_{L-1}}}(f), \gamma^{P_{\lambda_L}}(f)}_{\text{thinning profile}} \right\} \tag{1}$$

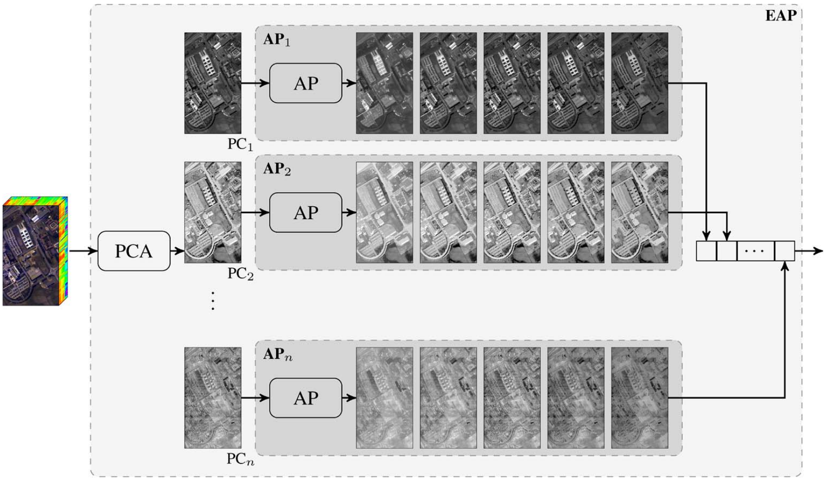


Fig. 6. General architecture of EAP.

of spectral bands), the application of morphological filters (here specifically APs) is not straightforward since there is no unique approach for extending them to vectorial images [46–50, Ch. 11]. A possible way of applying the concept of the profile to vectorial images was proposed in [21] and [42] and recalled in Section II-B. The proposed approach is based on the reduction of the dimensionality of the image values from T to $T' \subseteq \mathbf{Z}^m$ $(m \leq n)$ with a generic transformation $\Psi : T \rightarrow T'$ applied to an input image $\mathbf{f}$ (i.e., $\mathbf{g} = \Psi(\mathbf{f})$) and then on the application of the AP to each g_i $(i = 1, \ldots, m)$ of the transformed image. This can be formalized as

$$EAP(\mathbf{g}) = \{AP(g_1), AP(g_2), \ldots, AP(g_m)\}\,. \tag{2}$$

Fig. 6 shows the general architecture of EAP.

It can be convenient to compute multiple EAPs considering different attributes in order to derive a more complete descriptor of an image. This is the underlyining idea of the EMAP [42] (see Fig. 7), which is consequently defined considering k different attributes as

$$EMAP(\mathbf{g}) = \left\{EAP_{\mathcal{A}_1}(\mathbf{g}), EAP'_{\mathcal{A}_2}(\mathbf{g}), \ldots, EAP'_{\mathcal{A}_k}(\mathbf{g})\right\} \tag{3}$$

where $EAP_{\mathcal{A}_i}$ is an EAP built with a set of predicates evaluating the attribute $\mathcal{A}_i$ and $EAP' = EAP \setminus \{g_i\}_{i=1,\ldots,m}$ in order to avoid redundancy since the original components $\{g_i\}$ are present in each EAP. Fig. 7 shows the general architecture of EMAP. The following attributes have been widely used in literature in order to produce EMAP:

1) area of the region (related to the size of the regions);
2) standard deviation (as an index for showing the homogeneity of the regions);
3) diagonal of the box bounding the regions;
4) moment of inertia (as an index for measuring the elongation of the regions).

Fig. 8 shows an example of different APs (area, moment of inertia, and standard deviation) with different threshold values.

IV. Spectral–Spatial Classification Based on the AP

Although this section should be considered self-sufficient for understanding the concept of spectral–spatial classification based on the AP, for more information regarding the aforementioned concepts, please refer to [30], [51], and [52]. More references can be found throughout the paper.

This section aims at reviewing the main steps composing the techniques based on APs for land cover classification. We will focus on the classification of hyperspectral images (thus considering EAPs/EMAPs) since most of those techniques were proposed for this imagery. It is underlined that this choice is done without loss of generality since all the classification architectures proposed for other types of data (e.g., panchromatic images in [53]) can be reconducted to the general scheme presented in this section. A general workflow of the spectral–spatial classification with EMAP is shown in Fig. 9. First, FE/FS is performed on remote sensing data, and the resulting features are used as bases to build the EMAP. It should be noted that FE/FS are mostly taken into account for hyperspectral images in order to reduce the redundancy of the data and address the so-called curse of dimensionality. For other types of data, this step can be discarded. In [54], it has been noted that a prior spectral decomposition based on kernel FE before building the APs can lead to better classification results. The FE/FS can either be supervised or unsupervised. A further FE/FS operation applied to the EMAP can both reduce the effect of the Hughes phenomenon [20] and the redundancy in the profiles for classification. The classification is usually performed using nonlinear classifiers due to the fact that the resulting EMAP is characterized by highly nonlinear class distributions. In the following, the main components of the flowchart in Fig. 9 (FS/FE and classification) will be discussed

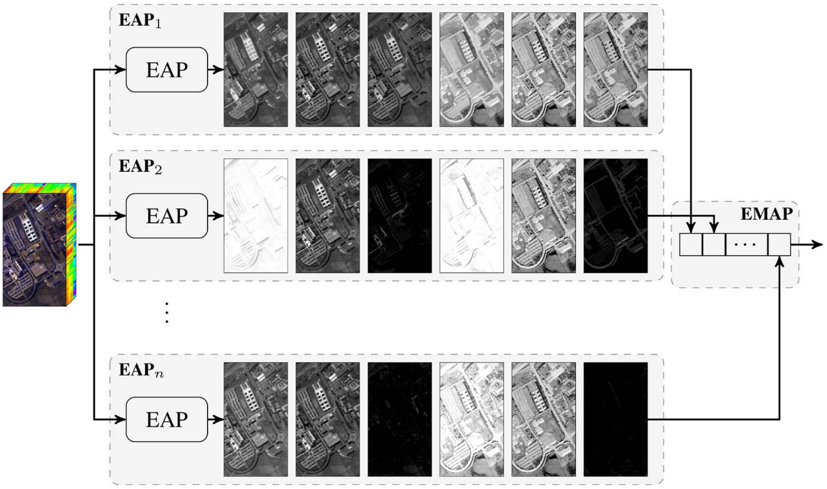


Fig. 7. General architecture of EMAP.

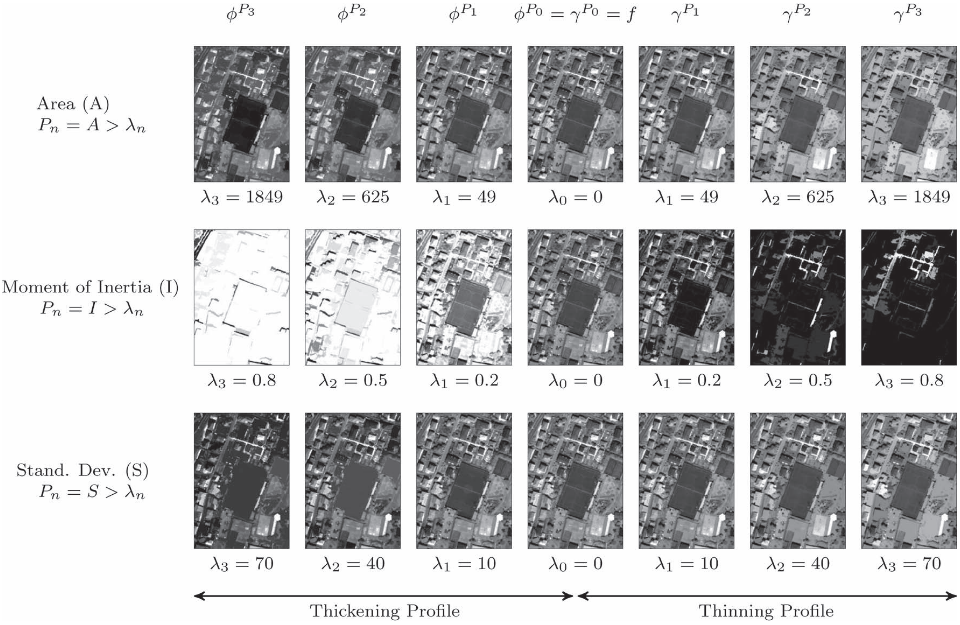

ϕ^{P_3}
ϕ^{P_2}
ϕ^{P_1}
$\phi^{P_0} = \gamma^{P_0} = f$
γ^{P_1}
γ^{P_2}
γ^{P_3}
Area (A)
$P_n = A > \lambda_n$
$\lambda_3 = 1849$ $\lambda_2 = 625$ $\lambda_1 = 49$ $\lambda_0 = 0$ $\lambda_1 = 49$ $\lambda_2 = 625$ $\lambda_3 = 1849$
Moment of Inertia (I)
$P_n = I > \lambda_n$
$\lambda_3 = 0.8$ $\lambda_2 = 0.5$ $\lambda_1 = 0.2$ $\lambda_0 = 0$ $\lambda_1 = 0.2$ $\lambda_2 = 0.5$ $\lambda_3 = 0.8$
Stand. Dev. (S)
$P_n = S > \lambda_n$
$\lambda_3 = 70$ $\lambda_2 = 40$ $\lambda_1 = 10$ $\lambda_0 = 0$ $\lambda_1 = 10$ $\lambda_2 = 40$ $\lambda_3 = 70$
Thickening Profile
Thinning Profile

Fig. 8. Example of different APs (area, moment of inertia, and standard deviation) with different threshold values.

in detail. Furthermore, at the end of this section, the automatic generation of EMAP for the accurate classification of remote sensing data will be discussed.

A. FE and FS

In the spectral domain, each spectral channel is considered as one dimension. By increasing the features in the spectral domain, theoretical and practical problems may arise. For instance, while keeping the number of training samples constant, the classification accuracy actually decreases when the number of features becomes large [20]. For the purpose of classification, these problems are related to the curse of dimensionality. In [55], it was shown that too many spectral bands can be undesirable from the standpoint of expected classification accuracy because the accuracy of the statistics estimation decreases (Hughes phenomenon). The aforementioned issue demonstrates that there is an optimal number of bands for classification accuracy, and more features do not necessarily lead to better results. Therefore, the use of feature reduction techniques may lead to better classification accuracy. The Hughes phenomenon highly influences parametric classifiers where the higher set of

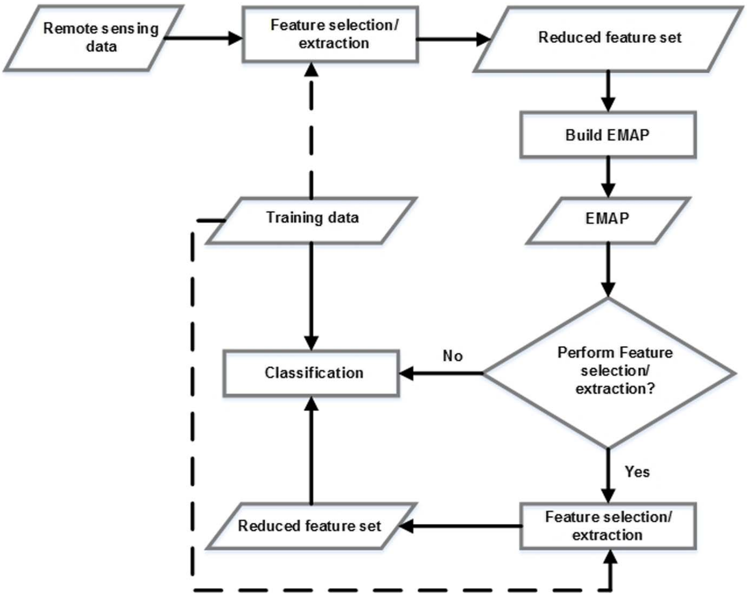


Fig. 9. General workflow of spectral–spatial classification with EMAP. The dotted lines indicate the possibility of switching between supervised and unsupervised feature reduction. An optional feature reduction step can be used to reduce the dimensionality of EMAP before classification [51].

statistical estimations needs to be estimated, and their classification accuracy values are dramatically downgraded by that effect. However, this issue has less influence on nonparametric classifiers such as SVM [56] and random forest (RF) [57].

In order to fully exploit spatial information from different structures in the scene, different attributes with a considerable range of threshold values should be considered. Nevertheless, considering many attributes with many threshold values can result in hyperdimensional profiles and, thus, hyperdimensional feature vectors that can lead to the Hughes phenomenon [20] (i.e., the curse of dimensionality) and high redundancy since filters with slightly different parameters may produce similar results. The issue of the high dimensionality of the profile can be addressed by considering feature reduction techniques. However, the selection of appropriate filter parameters is an essential step in order to guarantee a good tradeoff between the descriptive power of the profile and its redundancy [58]. In this case, FS and FE techniques have been gaining significant considerations in order to select the most effective features of the APs.

FS methods choose features from the original data set based on a criterion that is used to filter out unimportant or redundant features. FE can be explained as finding a set of vectors that represents an observation while reducing the dimensionality by transforming data to another domain. FE/FS can be split into two categories, namely, unsupervised and supervised FE/FS, where the former is used for the purpose of data representation, and the latter is considered for solving the Hughes phenomena [20] and reducing the redundancy of data in order to improve classification accuracy values by getting feedback from a set of available training samples. Although a reduction in dimensionality is of importance, the error arising from the reduction in dimension has to be without sacrificing the discriminative power of classifiers [32].

As shown in Fig. 9, FE/FS can be performed on the input data (in order to reduce the redundancy of the input data and select informative features as basis for producing APs) or on the obtained APs (in order to reduce the redundancy of obtained features by APs and increase the classification accuracy values). It should be noted that FE/FS can as well be used for both steps in one classification framework.

1) FE: PCA, KPCA (Kernel PCA), and independent component analysis (ICA) are the most commonly used unsupervised FE, which are used along with the concept of EMAP (e.g., [30], [51], and [59]). Moreover, Discriminant analysis feature extraction (DAFE), decision boundary FE (DBFE), and nonparametric weighted feature extraction (NWFE) are considered as the best known supervised FE, which are taken into account along with the concept of EMAP (e.g., [51]). It should be noted that the unsupervised FEs are mostly applied in order to extract informative features as the basis for producing APs. However, the supervised FEs can be performed on either the input data or the features obtained by APs.

The choice of the FE method has also been found to greatly influence the classification results using EMAPs. In [51], various supervised and unsupervised FE methods were compared when EMAPs were built using corresponding features and classified using RF and SVM classifiers. It has been concluded that kernel FE methods (in this case, kernel PCA) provides more consistent performance even if supervised FE (e.g., DBFE, NWFE, etc.) produces more accurate maps when a sufficient number of training samples are available. Furthermore, it has been noted in [54] that a prior FE from multispectral data using kernel methods to build the EMAP produces significantly improved classification maps.

In order to classify informative features produced by supervised FE, the first features with cumulative eigenvalues above 99% are retained. In the case of DAFE and NWFE, the criterion is related to the size of the eigenvalues of the scatter matrices. In the case of DBFE, the criterion is related to the size of the eigenvalues of the decision boundary feature matrix. For PCA, the first PCs with a cumulative variance of more than 99% are kept, since they contain almost all the variance in the data. However, different percentages can be used for different data.

Very recently in [60], it has been proposed to compute multiple profiles composed of APs built on different base images obtained by linear, nonlinear, manifold learning-based, and multilinear transformations of the original hyperspectral image. The individual APs computed on the extracted features (obtained by the different strategies) are either considered separately or jointly in a stacked vector. In order to deal with the high dimensionality of the profile, it was proposed to use a decision fusion approach or a sparse-based classifier.

2) FS: In [58], an automatic method was introduced for the classification of hyperspectral data, which is based on the FS step. In order to reduce the number of features by only keeping those that are important, GAs based on a measure of the relevance of the features were used. The main idea here is to construct a large profile from input hyperspectral data, which is called the EEMAP, that covers all the reasonable range of values for the filter parameters in order to provide a complete and detailed characterization of the spatial information of the scene. Then, for reducing the number of features by only keeping those that are important, GAs based on a measure of the relevance of the features are taken into account.

In [61], a new FS technique was proposed, which is based on the integration of GA and PSO. Then, the FS technique was applied on several features produced by EMAP for selecting the most informative features in order to detect road networks.

In [62], a new FS technique is introduced, which is based on a new binary optimization method named binary fractional order Darwinian particle swarm optimization (BFODPSO). In that method, SVM is used as the fitness function, and its corresponding classification overall accuracy is chosen as the fitness value in order to evaluate the efficiency of different group of features. In that paper, first, an AP feature bank is built consisting of different attributes with a wide range of threshold values. Then, BFODPSO-based FS is performed on the feature bank. In this case, SVM is chosen as the fitness function. The fitness of each particle is evaluated by the overall accuracy of SVM over the validation samples. After a few iterations, the BFODPSO-based FS approach finds the most informative features (resulted by EMAP) with respect to the overall accuracy of SVM over the validation samples.

In [63], a strategy for the selection of spatial features (among those APs were considered) relevant for classification was proposed. The relevance of the features is determined with respect to their capability in maximizing the SVM margin in the separation of classes. A research procedure based on the random generation of spatial filter banks and use of an active set criterion to rank the candidate features according to their benefits to margin maximization is proposed. This way, it is possible to explore the virtually infinite feature space (constituted by all the possible spatial features that could be computed) in order to retain the relevant ones for guaranteeing a final classification scheme, which is compact (uses as few features as possible), discriminative (enhances class separation), and robust (works well in small-sample situations).

B. Classification Using Different Methods

As discussed above, APs have been successfully exploited as efficient tools for spectral–spatial classification of remote sensing data. APs are inherently characterized by large dimensionality and high redundancy. This poses a great challenge for classification particularly to counter the Hughes phenomenon [20]. Due to highly nonlinear characteristics of the class distributions in the APs, the classification should be performed using nonlinear classifiers. A majority of the studies on classification of APs employed SVM [56] and RF [57] classifiers (e.g., [51], [58], and [64]).

SVM is a well-known classifier that separates training samples of different classes by tracing maximum margin hyperplanes in the space where the samples are mapped [65]. SVMs were originally introduced to solve linear classification problems. However, they were generalized for solving nonlinear decision functions by using the so-called kernel trick [66]. A kernel-based SVM is used to project the pixel vectors into higher dimensional space and estimate maximum margin hyperplanes in the new space, for improving linear separability of data [66]. The two main critical aspects of SVMs are sensitivity to the choice of the kernel and selection of the regularization parameters. The second issue can be classically overcome by considering cross-validation techniques using training data [67]. However, that can be computationally expensive [51]. The Gaussian radial basis function (RBF) is the most widely used kernel in remote sensing [66]. The cross validation explores an appropriate bandwidth parameter that provides the minimum error when the kernel-based SVM classifier is applied on the training data set. The main shortcomings of the cross validation are that 1) the bandwidth parameter needs to be discretized between a minimum and a maximum value, and the SVM classifier has to be trained and tested in a fivefold way for each of the discrete values of the bandwidth parameter. By increasing the number of discrete levels, the probability of finding the best parameter increases, which leads to higher computational time. On the contrary, by decreasing the number of levels, a suboptimal bandwidth parameter might be selected [68]. 2) In most of data sets, the cross-validation procedure does not consider a convex error curve over the selected discrete bandwidth parameter values, which makes the selection of discrete bandwidth parameter values a difficult task [68]. In order to tackle the aforementioned problems, in [69], a gradient-based method was proposed to minimize the upper bound of the leave-one-out generalization error of SVM over the set of full-diagonal bandwidth parameters. In [68] and [70], the upper bound was estimated based on the radius margin bound.

RF is an ensemble method for classification and regression. Ensemble classifiers get their name from the fact that several classifiers, i.e., an ensemble of classifiers, are trained, and their individual results are then combined to provide a final classification. For the purpose of classification of an object from an input vector, the input vector is run down each tree in the forest. Each tree provides a single vote for a particular class, and the forest chooses the classification label having the most votes (based on studies in [71]).

RF is not computationally intensive but demands a considerable amount of memory. RF can provide a good classification result in terms of accuracy values and does not assume any underlying probability distribution for input data. Another advantage of the RF classifier is that it is insensitive to noise in the training labels [71]. In addition, RF provides an unbiased estimate of the test set error as trees are added to the ensemble, and finally, it is less prone to overfit.

Apart from using SVM and RF classifiers, a composite kernel framework for spectral–spatial classification using APs has been recently investigated. In [72], a linearly weighted composite kernel framework with SVMs has been used for spectral–spatial classification using APs. A linearly weighted composite kernel is a weighted combination of different kernels computed using the available features [73]. For classification using APs, probabilistic SVMs were employed to classify the spectral information to obtain different rule images. The kernels are computed using the obtained rule images and are combined using the weighting factor. The choice of the weighting factor can be subjectively given or estimated using cross validation. However, classification using composite kernels and SVMs require a convex combination of kernels and a time-consuming optimization process. To overcome these limitations, a generalized composite kernel framework for spectral–spatial classification using APs has been proposed in [72]. MLR ([74]–[76]) has been employed instead of an SVM classifier and a set of generalized composite kernels, which can be linearly combined without any constraint of convexity, were proposed.

Very recently, sparse representation classification (SRC) techniques have been proposed for the classification of EMAPs [77]. SRC relies on the concept that an unknown sample can be represented as a linear combination of a set of labeled ones (i.e., the training set), called the dictionary. The representation of the samples is cast as an optimization problem in which the weights of each sample of the dictionary should be estimated with a constraint enforcing sparsity on the weights (i.e., limiting the contribution in the representation to only few samples). After representation, the sample is assigned to the class that shows the minimum reconstruction error when considering only the samples of the dictionary belonging to that class.

The importance of sparse-based classification methods has been further confirmed in [74] where a sparse-based MLR efficiently proved to effectively handle the very high dimensionality of the AP-based features used as input to the classifier.

In [78], a new technique was introduced for the combined classification of a high spatial resolution color image and a lower spatial resolution hyperspectral image of the same scene. To this extent, 1) contextual information is extracted from the high spatial resolution color image by transforming the image into CIE-Lab space. In the new space, instead of working on the "R," "G," and "B" bands separately, APs are carried out on the "L" band, which corresponds to the luminance, whereas the "a" and "b" bands (which contain the color information) are kept intact. Finally, the resulting images are transformed back into RGB space. 2) In parallel, the spectral information is extracted from the low spatial resolution hyperspectral data. 3) Finally, a composite decision fusion technique was investigated for combining the result of spectral and spatial information.

C. Automatic Scheme for EMAP

Although this section should be considered self-sufficient for understanding the concept of automatic EMAP, for more information regarding the aforementioned concepts, please refer to [30], [52], and [58]. More references can be found throughout the paper.

In order to tackle the main difficulties of using the EMAP, namely, 1) which attributes lead to a better discrimination for different classes and 2) which threshold values should be considered in order to initialize each AP, automatic schemes of using EMAP have been investigated. While the APs can be constructed by using different attributes, in the automatic scheme, generally the area and standard deviation attributes are only used, since these attributes can be adjusted in an automatic way and are well related to the object hierarchy in the images [30], [52], [58].

The standard deviation is adjusted with respect to the mean of the individual features, since the standard deviation shows dispersion from the mean [64]. Therefore, λ_s is initialized in a way to cover a reasonable amount of deviation in the individual feature and can be mathematically given by

$$\lambda_s(PC_i) = \frac{\mu_i}{100}\{\sigma_{\min}, \sigma_{\min} + \delta_s, \sigma_{\min} + 2\delta_s, \ldots, \sigma_{\max}\} \tag{4}$$

where μ_i is the mean of the ith feature, and $\sigma_{\min}$, $\sigma_{\max}$, and δ_s are the inner bound, the upper bound, and the step size for the standard deviation attribute, respectively.

With regard to adjusting λ_a for the area attribute, the resolution of the image should be taken into account in order to construct an EAP [58]. The automatic scheme of the attribute area is given as

$$\lambda_a(PC_i) = \frac{1000}{\upsilon}\{a_{\min}, a_{\min} + \delta_a, a_{\min} + 2\delta_a, \ldots, a_{\max}\} \tag{5}$$

where $a_{\min}$ and $a_{\max}$ are considered as the inner and upper bounds, respectively, with a step size increase δ_a, and υ shows the spatial resolution of the input data.

Here, "automatic" means that the framework only needs to establish a range of parameter values in order to automatically obtain a classification result with high accuracy for different data sets, instead of adjusting different thresholds with crisp values. More information regarding appropriate values for inner bound, upper bound, and step sizes can be found in [30], [52], and [58].

In [30], a spectral–spatial classification approach was introduced [please see the general idea of the model in Fig. 10(a); here, this method is called MANUAL, since the threshold values for EMAP are manually adjusted]. Then, in that paper,

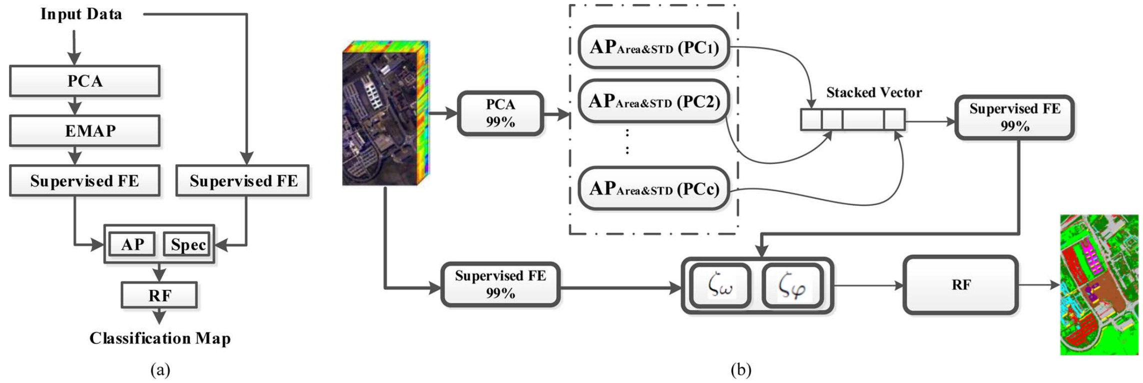


Fig. 10. (a) Flowchart of the method introduced in [30] for the MANUAL classification of hyperspectral images using AP and FE techniques. (b) General idea of the AUTOMATIC scheme of the method introduced in [30]. The main idea here is that since PCA cannot consider the class specific information for producing EMAP, supervised FE is carried out on both the input data and features obtained by EMAP. The main difference of this figure with Fig. 9 is that in this method, features obtained by performing a supervised FE on both the EMAP and the input data are concatenated into a stacked vector and classified by RF. In this case, first, PCA is performed on the input data, and first PCs with a cumulative variance of more than 99% are kept. Then, $\mathrm{AP}_{\mathrm{area,STD}}$ including area and standard deviation attributes with respect to (4) and (5) are built for each PC. Furthermore, the $\mathrm{AP}_{\mathrm{area,STD}}$ of different PCs are concatenated into a stacked vector. The output of this step provides spectral information. After that, the supervised FE is carried out on the stacked vector, and the first features corresponding to the top few eigenvalues, which account for 99% of the total sum of the eigenvalues, are kept. The output of this part provides the spatial information of the method. In parallel, the supervised FE is performed on the input data, and the first features corresponding to the top few eigenvalues, which account for 99% of the total sum of the eigenvalues, are selected. The output of this step is considered as spectral information. As the last stage, the spectral and spatial information is concatenated in a stacked vector, and the stacked vector is classified by RF [30].

the automatic scheme of that method is developed [please see the general idea of the model in Fig. 10(b); here, this method is called AUTOMATIC, since the threshold values for EMAP are automatically adjusted with respect to (4) and (5)]. Results reported in [30] for both schemes (MANUAL and AUTOMATIC) are very close in terms of classification accuracy values. The little difference obtained in the classification accuracy values between the MANUAL and AUTOMATIC schemes can show that the use of only two attributes, i.e., area and standard deviation, can model the spatial information on the used data sets considerably, and other attributes (diagonal of the box bounding the region and the moment of inertia) cannot add significant improvement to classification accuracy values although they carry information on the shape of regions. It is generally accepted that the use of different attributes will lead to the extraction of complementary (and redundant) information from the scene leading to increased accuracy values when used in classification (provided the Hughes effect is efficiently solved by only keeping those features that are most informative). In summary, it can be inferred that the AUTOMATIC can provide classification maps comparable with the MANUAL in terms of both classification accuracy values and CPU processing time when only two attributes (area and standard deviation) are used instead of four (area, standard deviation, moment of inertia, and diagonal of the box). However, the whole procedure in AUTOMATIC, as the name indicates, is automatic, and there is no need for any parameters to be set.

In [64], the automatic generation of standard deviation attributes was introduced. As it was mentioned there, features commonly follow different statistics, and also, the individual classes have different statistics for different features. Therefore, different thresholds are needed to build the standard deviation profiles from different features. This way, the thresholds for the standard deviation attribute are estimated based on the statistics of the classes of interest. The general idea behind that paper was that the standard deviation of the training samples of different classes of interest is related to the maximum standard deviation of the pixel values within individual segments of the corresponding classes of interest. The obtained results infer that the automatic method with only one attribute (standard deviation) along with supervised feature reduction can provide good results in terms of classification accuracy values.

V. Use of the AP for Different Types of Data and Applications

Although the concept of AP was introduced in order to extract spatial information from optical data (e.g., multispectral and hyperspectral data), this concept recently has been successfully exploited for characterizing spatial information of different types of data.

In [79] and [80], the effectiveness of using EMAP for the classification of the joint use of optical and LiDAR data has been investigated. In [79], first, the hyperspectral data were transformed by using PCA, and the first effective PCs were used as the base images for building EMAP. In parallel, different the intensity and the first return of LiDAR data were considered as inputs for building $\mathrm{AP}_{\mathrm{area,STD}}$. Then, all obtained profiles are concatenated into a stacked vector and classified by using either SVM or RF. Good classification accuracy values obtained in that paper confirmed that APs can be effectively applied to LiDAR data, since they provide a simplification of the image reducing the noise caused by the irregular spatial sampling of the LiDAR pulse and the interpolation phase. To show the effectiveness of using AP, a comparison considering texture features computed on gray-level co-occurrence matrix (GLCM) has been taken into account. The results obtained by considering FE techniques along with the introduced technique have outperformed those achieved with GLCM features. Finally, in all the experiments, the application of EMAP on both optical and LiDAR data led to the best classification accuracy values. Based on the results obtained in [80], the use of automatic EMAP can extract valuable information from LiDAR data by simultaneously filtering the unnecessary details and preserve the geometrical characteristics of the other regions.

In [53], FE was carried out on polarimetric synthetic aperture radar images based on the decomposition of the covariance matrix, and the corresponding features are used for building the APs. In the paper, the standard deviation is used as the attribute to build the EAP. The classification results show that there is an improvement when the EAP is used, and a smoother classification result is obtained for visual examination of results.

The concept of the AP has been also used for different applications such as change detection. Although the main objective of this paper is to investigate the usefulness of the AP for remote sensing image classification, other applications of the AP are briefly discussed below.

In [81], APs were employed for detecting changes occurred on the ground by analyzing two images acquired over the same areas before and after the occurred event. APs computed on each of the two images were compared in order to detect differences in the geometrical and morphological characteristics of the underlining structures present in the two images in corresponding zones. The reason motivating the use of APs for the detection of changes relies on the fact that if an abrupt change has occurred in the scene (e.g., a man-made change or a natural disaster), it will be likely that the spatial characteristics of the affected areas will have changed too. Thus, detecting differences in the behaviors of the APs in corresponding positions in the image can be useful for spotting a modification in the spatial arrangement of the pixel values between the two images. The change detection technique was based on three main steps: 1) application of the APs to each image; 2) region extraction and reliable level selection by analyzing the DAPs; and 3) comparison of the APs and generation of the change-detection map.

VI. Discussion

As shown in Fig. 9, the use of appropriate FE/FS techniques and efficient classifiers can significantly influence the obtained classification accuracy values and the quality of the classification map. Therefore, here, leveraging the outcomes presented in the literature, the capability of different FE/FS techniques as well as different classifiers will be investigated.

A. Influence of Different FE/FS Along With AP on the Classification Accuracy Values

Although this section should be considered self-sufficient for understanding the influence of different FE/FS along with AP on the classification accuracy values, for more information regarding the aforementioned concepts, please refer to [30], [51], and [52]. More references can be found throughout the paper.

1) When only spectral information derived by NWFE, DAFE, and DBFE is used, the result of the classification is almost the same. However, when the corresponding EMAP based on DAFE, DBFE, and NWFE is made, the accuracy values are quite different, which shows that the classification with EMAPs does not necessarily follow the trend of classification with spectral information only [51], [52].
2) When the number of training samples is limited, a supervised FE leads to less accurate results in terms of classification accuracy compared with unsupervised techniques. In [54], it was mentioned that the combination of KPCA and EMAPs can be a simple and even powerful strategy to perform spectral–spatial classification of data sets with limited spectral resolution (RGB and multispectral images). With reference to [51], in general, EMAP based on KPCA can be found more consistent even though it sometimes produces slightly inferior accuracy values in comparison with the supervised FE techniques. However, it is difficult to anticipate which supervised FE technique is appropriate for a problem at hand, and the performance of that is highly dependent on the number of available training samples.
3) In a case when the number of training samples is sufficient, according to the experiments shown in [51], DBFE seems to be able to provide better results in terms of classification accuracy values than DAFE and NWFE in order to produce EMAP.
4) Based on the results reported in [30], the CPU processing time for both schemes (MANUAL and AUTOMATIC) is almost the same. For AUTOMATIC, there is no need to adjust the initial parameters for the APs, which is considered as the main shortcoming of the usage of AP.
5) The results obtained in [63] show that the selection strategy is able to retrieve for each class its optimal discriminant features. Remarkably, that technique effectively handled different types of spatial features (e.g., textural features and APs). In addition, it was shown that the models trained on the features discovered reached, at worst, the same performances as considering predefined filter banks (i.e., manual selection of the filter parameters requiring prior knowledge).
6) In [52], a spectral–spatial classification framework was developed, which is specifically related to the use of parametric supervised FE techniques (DAFE and DBFE) and EMAP. Results show that when different parametric supervised FE techniques are used for the first and second steps (e.g., DBFE is applied on the input data, and DAFE is carried out on the features extracted by EMAP, or vice versa) and the first features corresponding to the top few eigenvalues of both steps are concatenated into a stacked vector, the result of the classification is good, and RF can classify the stacked vector of features accurately.

B. Comparison of Different Classifiers Used With EMAP

Although this section should be considered self-sufficient for understanding the influence of different classifiers on the classification accuracy values for the features produced by AP, for more information regarding the aforementioned concept,

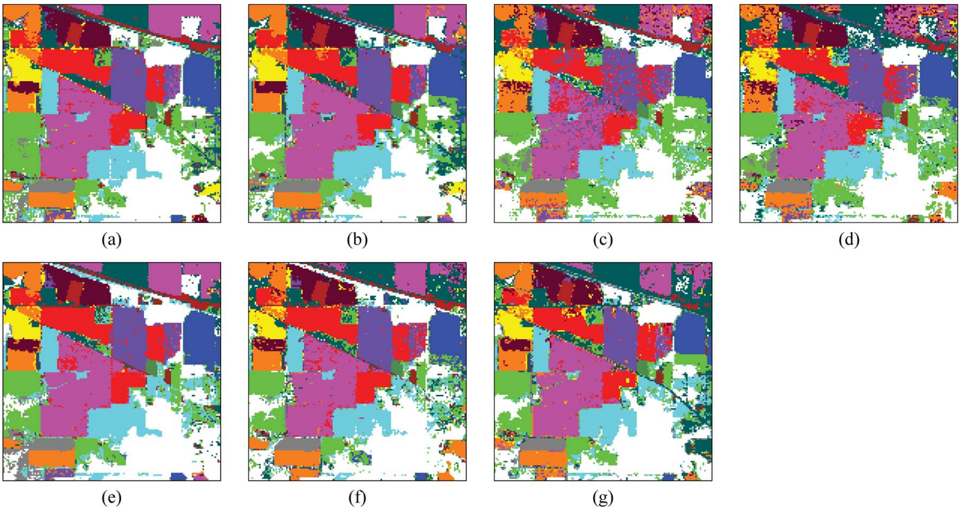


Fig. 11. Classification maps for Indian Pines data with RF classifier (with 200 trees) using EMAPs of (a) PCA (OA = 92.83%), (b) KPCA (OA = 94.76%), (c) DAFE (OA = 84.33%), (d) DBFE (OA = 87.23%), and (e) NWFE (OA = 95.06%) and feature reduction applied on EMAP using (f) NW-NW (OA = 91.03%) and (g) KP-NW (OA = 90.36%) [51].

please refer to [30], [51], and [52]. More references can be found throughout the paper.

Below, the main points regarding the applicability of SVM and RF are listed.

1) While both SVM and RF classification methods are shown to be effective classifiers for nonlinear classification problems, SVM requires a computationally demanding parameter tuning process (cross validation) in order to tune hyperplane parameters and consequently achieve optimal results, whereas RF does not require such tuning process. In this sense, RF is much faster than SVM, and for volumetric data, using RF instead of SVM is favorable.
2) The effect of the Hughes phenomenon is more evident when the number of dimensionality is high and the data are classified by SVM. Better accuracy values are achieved with SVM only after further feature reduction [51].
3) RF is more stable when limited training samples are available. Even when a sufficient number of training samples are available, the SVM classifier required further feature reduction of the profile to achieve acceptable classification accuracy [64].
4) Based on the results reported in [51], [52], and [64], RF provides higher classification accuracy values compared with SVM when it is directly performed on EMAP; however, SVM performs better in terms of classification accuracy values when further FE is performed on EMAP. This shows the capability of RF in order to handle higher dimensional space as an input to the classifier. On the contrary, the second FE on EMAP downgrades the classification accuracy values of the RF classifier. The reason might be that the RF classifier is based on a collection of weak classifiers, which can statistically take advantage of a large set of redundant features. In contrast, the SVM classifier seems to be more effective in designing a discriminant function when a subset of nonredundant features defines a highly nonlinear problem.
5) The experiments in [74] showed significant improvement in classification accuracy values when the generalized composite kernel framework is used compared with the regular composite kernel framework.
6) From the result obtained with SRC coupled with EMAPs [72], it can be stated that SRC outperforms other classifiers such as SVM and SVM with composite kernels particularly when the number of training samples is small.

C. Comparison of Different Classification Maps Obtained by EMAP

This section is based on the comparison of different classification maps obtained by EMAP and different FEs. In order to precisely evaluate different classification maps, two different scenarios are taken into account. The first scenario is devoted to a situation when the number of training samples is not sufficient. For this purpose, a frequently used *Indian Pines* data set is used. For more information related to the number of training and test samples, please see [51]. The second scenario is devoted to a situation when the number of training samples is sufficient. For this purpose, a frequently used *Pavia University* data set is used. For more information related to the number of training and test samples, please see [30].

Scenario (1) When the number of training samples is limited: By visually comparing the classification maps shown in Figs. 11 and 12, the following points can be obtained.

1) When the number of training samples is limited, features obtained by NWFE can be considered as good bases in order to produce EMAP. In this case, NWFE may outperform other FE techniques such as PCA, KPCA, DAFE, and DBFE in terms of classification accuracy values, when it is considered for building EMAP.

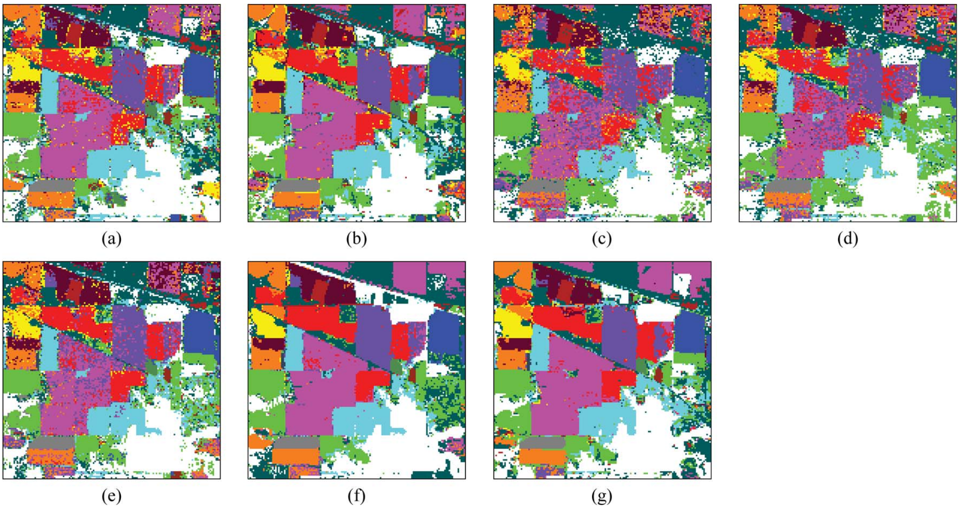


Fig. 12. Classification maps for Indian Pines data with SVM (with fivefold cross validation and RBF kernel) classifier using EMAPs of (a) PCA $(\text{OA} = 86.53\%)$, (b) KPCA $(\text{OA} = 90.20\%)$, (c) DAFE $(\text{OA} = 70.69\%)$, (d) DBFE $(\text{OA} = 78.39\%)$, and (e) NWFE $(\text{OA} = 81.85\%)$ and feature reduction applied on EMAP using (f) NW-NW $(\text{OA} = 94.17\%)$ and (g) KP-NW $(\text{OA} = 93.75\%)$ [51].

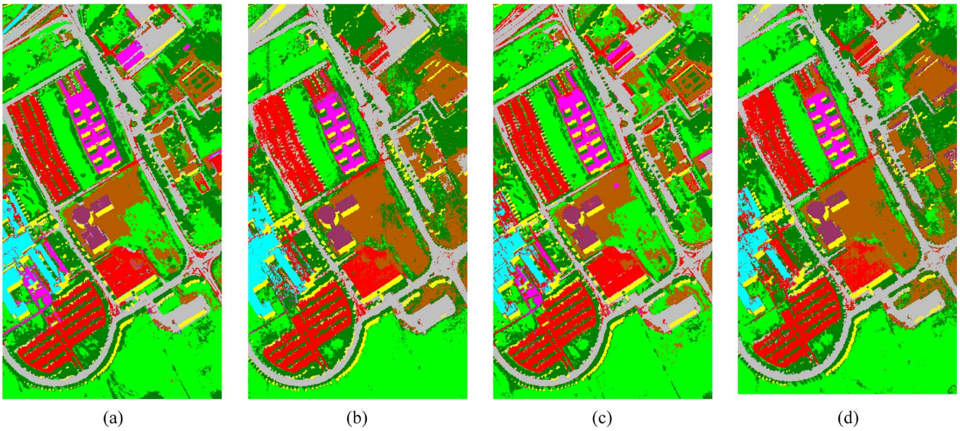


Fig. 13. Classification results of Pavia University with RF classifier (with 200 trees) using EMAPS of (a) KPCA $(\text{OA} = 92.37\%)$, (b) DBFE $(\text{OA} = 95.83\%)$, and (c) NWFE $(\text{OA} = 92.19\%)$ and feature reduction applied on EMAP using (d) DB-DB $(\text{OA} = 96.81\%)$ [51].

2) In order to produce EMAP, when the number of training samples is too small, supervised FE techniques lead to salt-and-pepper effects, and the object cannot be properly exploited after performing a classification. In other words, the shape of different objects may not be properly preserved when a supervised FE method is taken into account even while APs are used. In this case, the use of unsupervised feature reduction (in particular, KPCA), can extract the shape of the object in a better way.
3) As it was mentioned before, RF shows more stability when limited training samples are available.
4) RF is able to provide higher classification accuracy values compared with SVM when it is directly performed on EMAP. However, SVM performs better in terms of classification accuracy values when further FE is performed on EMAP.
5) The overall accuracy of Indian Pines when it is classified by RF (with 200 trees) and SVM (with fivefold cross validation) is 65.6% and 69.70%, respectively. Based on classification accuracy values reported in Figs. 11 and 12, one can easily obtain that the use of AP can significantly improve the classification results.

Scenario (2) When an adequate number of training samples is available:

By visually comparing the classification maps shown in Figs. 13–15, the following can be concluded.

1) When an adequate number of training samples is available, DBFE seems to be able to provide better results in terms of overall classification accuracy.
2) Based on the experiments reported in [30], when the number of training samples is adequate, the use of DAFE

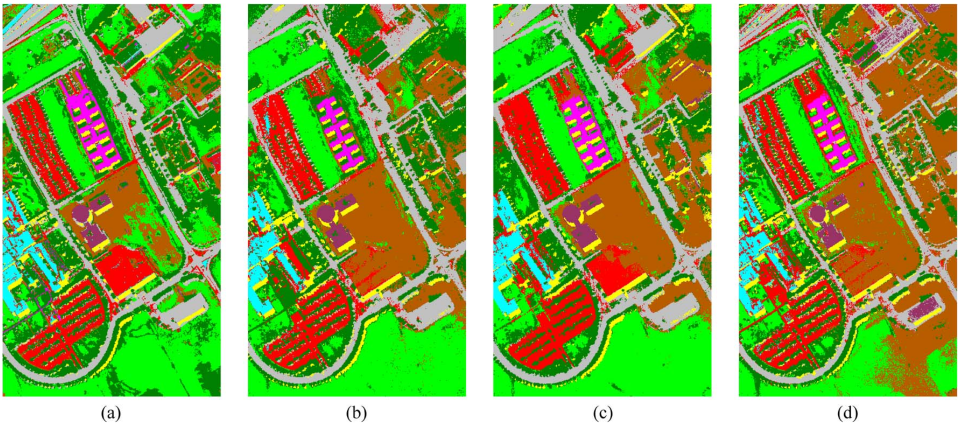


Fig. 14. Classification maps of Pavia University with SVM classifier (with fivefold cross validation and RBF kernel) using EMAPS of (a) KPCA (OA $=$ 91.52%), (b) DBFE (OA $=$ 91.64%), and (c) NWFE (OA $=$ 89.27%) and feature reduction applied on EMAP using (d) DB-DB (OA $=$ 97.89%) [51].

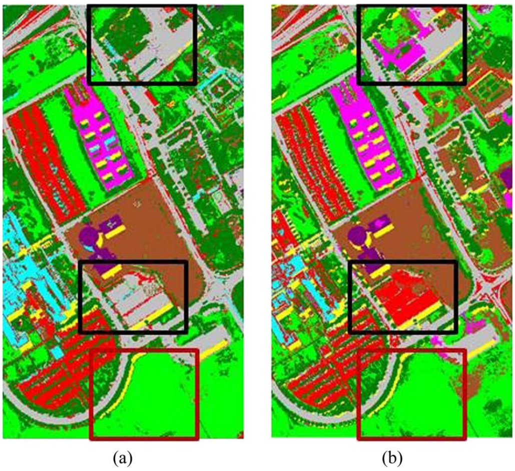


Fig. 15. Comparison between classification maps obtained by (a) DAFE (OA $=$ 97.00%) and (b) NWFE (OA $=$ 97.58%) by using RF classifier (with 200 trees) based on Fig. 10(a) [30].

may lead to better classification accuracy values by using the frameworks developed in [30]. In this case, the use of DAFE improves the overall accuracy of NWFE by almost 2.5%. Fig. 15 shows that not only the number of training samples is important on the efficiency of DAFE and NWFE, but also the distribution of training samples on the whole data set is of importance. As an example, the black boxes in Fig. 15 shows two parts of the input data, which do not contain training samples. In this case, although the overall accuracy of DAFE (97.00%) is significantly higher than the overall accuracy of NWFE (94.58%), some objects are missing in the classification map obtained by DAFE because the data do not have training samples in that regions. On the other hand, for the region where there is an adequate number of training samples (the red box), DAFE leads to a comparatively smoother classification map.

3) The overall accuracy of Pavia University when it is classified by RF (with 200 trees) and SVM (with fivefold cross validation) is 71.57% and 81.44%, respectively. Based on classification accuracy values reported in Figs. 13–15, one can easily obtain that the use of AP can significantly improve the classification results.

VII. Conclusion and Future Works

In this paper, a survey of recent works dealing with APs has been presented. From the various contributions in the literature dealing with APs, the effectiveness of using APs for modeling the spatial information of an image can be assessed.

Indeed, the AP, being based on an attribute that models some regional characteristics (e.g., scale, shape, and contrast), provides a multilevel decomposition of an image. As shown by many works referred to in this paper, the sequence of filtered images composing the AP can be employed for classification in a simple yet effective architecture by considering it as a set of features feeding a classifier (as complement of the original spectral data).

We have focused this survey on the classification of remotely sensed images with particular attention to the hyperspectral data for which extensions of the AP have been proposed (i.e., EAP and EMAP).

From the analysis, it emerged that APs can extract spatial features useful for classification but present some aspects that might be critical. As clearly seen by the achievements reported in many works, AP provides features that can greatly improve the discriminability of the samples in classification. Despite their usefulness, we recall that the AP, being composed of attribute thinnings and thickenings, relies on a representation of the image as a min-tree and a max-tree. Thus, these filters can only process image extrema. This can be a limitation when the image structures do not correspond to regional extrema such as for noisy or highly textured images. Further research should be addressed to a deeper investigation of the extension of APs for the analysis of images different from the optical ones (e.g., SAR data).

Many works in the literature address the classification of hyperspectral images. In this context, since AFs cannot be uniquely extended to multivariate images (e.g., multi- or hyperspectral images), different strategies can be considered. The mostly used architecture relies on a reduction of the dimensionality of the data, followed by the application of an AP to each component (leading to an EAP) after the reduction. Although this strategy has the great advantage of dealing with only few components, the resulting profile heavily relies on the transformation employed. Several supervised and unsupervised FE techniques have been proposed in the literature showing the criticality of this step, which still presents margins of improvement.

The selection of attributes and their related thresholds is also another aspect of utmost importance. Strategies for the selection of the attribute thresholds have been proposed in order to automatically perform this task, proving to achieve results that are comparable to those obtained by manual tuning. However, the proposed techniques are specific to some attributes (i.e., area and standard deviation) and might not be applicable to others, thus opening the need for developing more generic selection strategies for the filter parameters.

Although informative, APs are typically a set of highly dimensional and redundant features. Consequently, these aspects should be properly handled in order to a make full exploitation of the informative content of the profiles. Thus, the selection of the classifier is another key aspect to consider. Nonparametric classifiers such as SVMs and RF have largely proven to deal well with the high dimensionality of the profiles. More recently, SVM with composite kernels and sparse representation classification have been proposed, leading to accurate and robust results even in cases of a reduced number of training samples.

In order to reduce the redundancy of the APs, particularly when considered in their extended architecture (i.e., the EMAP), it has been proposed to use dimensionality reduction techniques. Conventional FE techniques (e.g., DAFE, DBFE, and NWFE) have been considered proving their usefulness. Alternatively, FS techniques (e.g., based on evolutionary algorithms such as GAs and PSO) have also been proposed to address this task.

In conclusion, although the concept of AP and its extensions EAP and EMAP have proven to be effective in the analysis of remote sensing images particularly for classification, many lines of research remain open.

ACKNOWLEDGMENT

The authors would like to thank Prof. P. Gamba from the University of Pavia, Italy, for providing the ROSIS data and corresponding reference information and Dr. P. Marpu for his contributions.

REFERENCES

[1] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, “Advances in spectral–spatial classification of hyperspectral images,” *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[2] S. Tadjudin and D. Landgrebe, “Classification of high dimensional data with limited training samples,” School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA, Tech. Rep., 1998.

[3] H. Derin and P. A. Kelly, “Discrete-index Markov-type random processes,” *Proc. IEEE*, vol. 77, no. 10, pp. 1485–1510, Oct. 1989.

[4] 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,” *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.

[5] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, “SVM- and MRF-based method for accurate classification of hyperspectral images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[6] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, “Spectral–spatial classification of hyperspectral images based on hidden Markov random fields,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2565–2574, May 2014.

[7] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, “Use of Darwinian particle swarm optimization technique for the segmentation of remote sensing images,” in *Proc. IEEE IGARSS*, 2012, pp. 4295–4298.

[8] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, “An efficient method for segmentation of images based on fractional calculus and natural selection,” *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12407–12417, Nov. 2012.

[9] P. Marpu, M. Neubert, H. Herold, and I. Niemeyer, “Enhanced evaluation of image segmentation results,” *J. Spatial Sci.*, vol. 55, no. 1, pp. 55–68, 2010.

[10] P. Ghamisi, M. S. Couceiro, F. M. Martins, and J. A. Benediktsson, “Multilevel image segmentation approach for remote sensing images based on fractional-order Darwinian particle swarm optimization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2382–2394, May 2014.

[11] P. Ghamisi, M. Couceiro, M. Fauvel, and J. A. Benediktsson, “Integration of segmentation techniques for classification of hyperspectral images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 342–346, Jan. 2014.

[12] J. Li, H. Zhang, and L. Zhang, “Supervised segmentation of very high resolution images by the use of extended morphological attribute profiles and a sparse transform,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1409–1413, Aug. 2014.

[13] M. Pesaresi and J. A. Benediktsson, “A new approach for the morphological segmentation of high-resolution satellite imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[14] M. Chini, N. Pierdicca, and W. Emery, “Exploiting SAR and VHR optical images to quantify damage caused by the 2003 Bam earthquake,” *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 145–152, Jan. 2009.

[15] P. Soille, *Morphological Image Analysis, Principles and Applications*, 2nd ed. New York, NY, USA: Springer-Verlag, 2003.

[16] M. K. D. Tuia, F. Pacifici, and W. Emery, “Classification of very high spatial resolution imagery using mathematical morphology and support vector machines,” *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3866–3879, Nov. 2009.

[17] H. Akcay and S. Aksoy, “Automatic detection of geospatial objects using multiple hierarchical segmentations,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2097–2111, Jul. 2008.

[18] J. B. J. Chanussot and M. Fauvel, “Classification of remote sensing images from urban areas using a fuzzy possibilistic model,” *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 40–44, Jan. 2006.

[19] J. A. Benediktsson, M. Pesaresi, and K. Arnason, “Classification and feature extraction for remote sensing images from urban areas based on morphological transformations,” *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1940–1949, Sep. 2003.

[20] G. Hughes, “On the mean accuracy of statistical pattern recognizers,” *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[21] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, “Classification of hyperspectral data from urban areas based on extended morphological profiles,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[22] R. Bellens *et al.*, “Improved classification of VHR images of urban areas using directional morphological profiles,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 2803–2813, Oct. 2008.

[23] P. Soille and M. Pesaresi, “Advances in mathematical morphology applied to geoscience and remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 9, pp. 2042–2055, Sep. 2002.

[24] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, “Classification of hyperspectral images with binary fractional order Darwinian PSO and random forests,” in *Proc. SPIE*, 2013, vol. 8892, pp. 88920S-1–88920S-8.

[25] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, “Morphological attribute profiles for the analysis of very high resolution

images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[26] N. Bouaynaya and D. Schonfeld, "Theoretical foundations of spatially-variant mathematical morphology Part II: Gray-level images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 837–850, May 2008.

[27] M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, "Modeling structural information for building extraction with morphological attribute filters," in *Proc. SPIE Eur. Remote Sens.*, Berlin, Germany, Aug. 2009, pp. 747 703-1–747 703-9.

[28] G. A. Licciardi *et al.*, "Retrieval of the height of buildings from worldview-2 multi-angular imagery using attribute filters and geometric invariant moments," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 71–79, Feb. 2012.

[29] E. J. Breen and R. Jones, "Attribute openings, thinnings and granulometries," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2486–2494, Nov. 2013.

[30] P. Ghamisi, J. A. Benediktsson, and J. R. Sveinsson, "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5771–5782, Sep. 2014.

[31] B. Luo and L. Zhang, "Robust autodual morphological profiles for the classification of high-resolution satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1451–1462, Feb. 2014.

[32] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[33] E. J. Breen and R. Jones, "Attribute openings, thinnings, granulometries," *Comput. Vis. Image Understanding*, vol. 64, no. 3, pp. 377–389, Nov. 1996.

[34] P. Maragos and R. Ziff, "Threshold superposition in morphological image analysis systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 498–504, May 1990.

[35] P. Salembier and J. Serra, "Flat zones filtering, connected operators, filters by reconstruction," *IEEE Trans. Image Process.*, vol. 4, no. 8, pp. 1153–1160, Aug. 1995.

[36] V. Caselles and P. Monasse, *Geometric Description of Images as Topographic Maps*. New York, NY, USA: Springer-Verlag, 2010.

[37] P. Salembier, A. Oliveras, and L. Garrido, "Antiextensive connected operators for image and sequence processing," *IEEE Trans. Image Process.*, vol. 7, no. 4, pp. 555–570, Apr. 1998.

[38] L. Najman and M. Couprie, "Building the component tree in quasi-linear time," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3531–3539, Nov. 2006.

[39] P. Monasse and F. Guichard, "Fast computation of a contrast-invariant image representation," *IEEE Trans. Image Process.*, vol. 9, no. 5, pp. 860–872, May 2000.

[40] P. Salembier and M. Wilkinson, "Connected operators," *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 136–157, Nov. 2009.

[41] E. R. Urbach, J. B. T. M. Roerdink, and M. H. F. Wilkinson, "Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 272–285, Feb. 2007.

[42] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, Nov. 2010.

[43] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute filters for the analysis of very high resolution remote sensing images," in *Proc. IEEE IGARSS*, Jul. 2009, vol. 3, pp. III-97–III-100.

[44] M. H. F. Wilkinson, H. Gao, W. H. Hesselink, J.-E. Jonker, and A. Meijster, "Concurrent computation of attribute filters on shared memory parallel machines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1800–1813, Oct. 2008.

[45] M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, *Self-Dual Attribute Profiles for the Analysis of Remote Sensing Images*. Berlin, Germany: Springer-Verlag, 2011.

[46] L. Najman and H. Talbot, *Mathematical Morphology*. Hoboken, NJ, USA: Wiley-ISTE, Aug. 2010.

[47] J. Chanussot and P. Lambert, "Total ordering based on space filling curves for multivalued morphology," in *Proc. 4th Int. Symp. Math. Morphology Appl.*, 6 1998, pp. 51–58.

[48] L. Garrido, P. Salembier, and D. Garcia, "Extensive operators in partition lattices for image sequence analysis," *Signal Process.*, vol. 66, no. 2, pp. 157–180, Apr. 1998.

[49] P. Lambert and J. Chanussot, "Extending mathematical morphology to color image processing," in *Proc. 1st Int. Conf. CGIP*, Saint-Etienne, France, 2000, pp. 158–163.

[50] E. Aptoula and S. Lefevre, "A comparative study on multivariate mathematical morphology," *Pattern Recog.*, vol. 40, no. 11, pp. 2914–2929, Nov. 2007.

[51] P. R. Marpu *et al.*, "Classification of hyperspectral data using extended attribute profiles based on supervised and unsupervised feature extraction techniques," *Int. J. Image Data Fusion*, vol. 3, no. 3, pp. 269–298, 2012.

[52] P. Ghamisi, J. A. Benediktsson, G. Cavallaro, and A. Plaza, "Automatic framework for spectral–spatial classification based on supervised feature extraction and morphological attribute profiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2147–2160, Jun. 2014.

[53] P. R. Marpu, K.-S. Chen, and J. A. Benediktsson, "Spectral–spatial classification of polarimetric SAR data using morphological attribute profiles," in *Proc. SPIE*, 2011, vol. 8180, pp. 81 800K-1–81 800K-6.

[54] S. Bernabé, P. R. Marpu, and A. Plaza, "Spectral–spatial classification of multispectral images using kernel feature space representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 288–292, Jan. 2014.

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

[56] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, Beyond*. Cambridge, MA, USA: MIT press, 2002.

[57] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[58] M. Pedergnana, P. R. Marpu, M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, "A novel technique for optimal feature selection in attribute profiles based on genetic algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 6, pp. 3514–3528, Jun. 2013.

[59] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

[60] X. Huang *et al.*, "Multiple morphological profiles from multicomponent base images for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. PP, no. 99, pp. 1–17, 2014.

[61] P. Ghamisi and J. A. Benediktsson, "Feature selection based on hybridization of genetic algorithm and particle swarm optimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 309–313, Feb. 2015.

[62] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, "FODSPO based feature selection for hyperspectral remote sensing data," in *Proc. WHISPERS*, 2014.

[63] D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, and R. Flamary, "Automatic feature learning for spatio-spectral image classification with sparse SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6062–6074, Oct. 2014.

[64] P. 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," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 293–297, Mar. 2013.

[65] V. N. Vapnic, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.

[66] B. Scholkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[67] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for the classification of hyperspectral remote-sensing data over urban areas," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 11, Jan. 2009.

[68] P. Gurram and H. Kwon, "Sparse kernel-based ensemble learning with fully optimized kernel parameters for hyperspectral classification problems," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 787–802, Feb. 2013.

[69] A. Villa, M. Fauvel, J. Chanussot, P. Gamba, and J. Benediktsson, "Gradient optimization for multiple kernel's parameters in support vector machines classification," in *Proc. IEEE IGARSS*, Jul. 2008, vol. 4, pp. IV-224–IV-227.

[70] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[71] L. Breiman, "RF tools a class of two eyed algorithms," in *Proc. SIAM Workshop*, 2003, pp. 1–56.

[72] J. Li, P. Marpu, A. Plaza, J. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

[73] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[74] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.

[75] J. Li, J. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.

[76] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.

[77] B. Song *et al.*, "Remotely sensed image classification using sparse representations of morphological attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5122–5136, Aug. 2014.

[78] G. Thoonen, Z. Mahmood, S. Peeters, and P. Scheunders, "Multisource classification of color and hyperspectral images using color attribute profiles and composite decision fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 510–521, Apr. 2012.

[79] M. Pedergnana, P. R. Marpu, M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, "Classification of remote sensing optical and lidar data using extended attribute profiles," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 7, pp. 856–865, Nov. 2012.

[80] P. Ghamisi1, J. A. Benediktsson1, and S. Phinn, "Fusion of hyperspectral and LiDar data in classification of urban areas," in *Proc. IEEE IGARSS*, Quebec City, Canada, Jul. 13–18, 2014, pp. 181–184.

[81] N. Falco, M. Dalla Mura, F. Bovolo, J. A. Benediktsson, and L. Bruzzone, "Change detection in VHR images based on morphological attribute profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 636–640, May 2013.

Pedram Ghamisi (S'12) received the B.Sc. degree in civil (survey) engineering from the Islamic Azad University South Tehran Branch, Tehran, Iran, and the M.Sc. degree in remote sensing from K.N.Toosi University of Technology, Tehran, in 2012. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Iceland, Reykjavík, Iceland.

His research interests include remote sensing and image analysis with the current focus on spectral and spatial techniques for hyperspectral image classification and the integration of LiDAR and hyperspectral data for land cover assessment.

Mr. Ghamisi received the Best Researcher Award for M.Sc. students from K.N.Toosi University of Technology for the academic year 2010–2011. He was a recipient of the IEEE Mikio Takagi Prize, which was awarded for the first placer in the Student Paper Competition at the 2013 IEEE International Geoscience and Remote Sensing Symposium in Melbourne, Australia. He serves as a Reviewer for a number of journals, including the IEEE Transactions on Image Processing, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters, and *Pattern Recognition Letters*.

Mauro Dalla Mura (S'08–M'11) received the Laurea (B.E.) and Laurea Specialistica (M.E.) degrees in telecommunication engineering from the University of Trento, Trento, Italy, in 2005 and 2007, respectively, and the joint Ph.D. degree in information and communication technologies (telecommunications area) from the University of Trento and in electrical and computer engineering from the University of Iceland, Reykjavík, Iceland, in 2011.

In 2011, he was a Research Fellow with the Fondazione Bruno Kessler, Trento, conducting research on computer vision. He is currently an Assistant Professor with the Grenoble Institute of Technology (Grenoble INP), Grenoble Cedex 1, France. He is conducting his research in the Grenoble Images Speech Signals and Automatics Laboratory (GIPSA-Lab). His main research activities are in the fields of remote sensing, image processing, and pattern recognition. In particular, his interests include mathematical morphology, classification, and multivariate data analysis.

Dr. Dalla Mura was a recipient of the IEEE Geoscience and Remote Sensing Society (GRSS) Second Prize in the Student Paper Competition of the 2011 IEEE International Geoscience and Remote Sensing Symposium in Vancouver, Canada. He is a Reviewer of the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters, the IEEE Journal of Selected Topics in Earth Observations and Remote Sensing, the IEEE Journal of Selected Topics in Signal Processing, *Pattern Recognition Letters*, the *ISPRS Journal of Photogrammetry and Remote Sensing*, and *Photogrammetric Engineering and Remote Sensing*. He is a member of the GRSS and the IEEE GRSS Data Fusion Technical Committee and the Secretary of the IEEE GRSS French Chapter (2013–2016). He was a Lecturer at the RSSS12—Remote Sensing Summer School 2012 (organized by the IEEE GRSS) in Munich, Germany.

Jon Atli Benediktsson (S'84–M'90–SM'99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavík, Iceland, in 1984 and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is currently the Pro Rector for Academic Affairs and a Professor of electrical and computer engineering with the University of Iceland. He has extensively published in his fields of interest. His research interests include remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing.

Dr. Benediktsson was President of the IEEE Geoscience and Remote Sensing Society (GRSS) during 2011–2012 and has been on the GRSS Administrative Committee since 2000. He was Editor-in-Chief of the IEEE Transactions on Geoscience and Remote Sensing (TGRS) from 2003 to 2008 and has served as an Associate Editor for the IEEE TGRS in 1999, the IEEE Geoscience and Remote Sensing Letters in 2003, and the IEEE Access in 2013. He is on the Editorial Board of the Proceedings of the IEEE and on the International Editorial Board of the *International Journal of Image and Data Fusion* and was the Chairman of the Steering Committee of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing during 2007–2010. He is a cofounder of the biomedical startup company Oxymap. He is a Fellow of the International Society for Optics and Photonics (SPIE). He is a member of the 2014 IEEE Fellow Committee. He received the Stevan J. Kristof Award from Purdue University in 1991 as outstanding graduate student in remote sensing. He was a recipient of the Icelandic Research Council's Outstanding Young Researcher Award in 1997 and he was granted the IEEE Third Millennium Medal in 2000. He was a corecipient of the University of Iceland's Technology Innovation Award in 2004. In 2006, he received the yearly research award from the Engineering Research Institute of the University of Iceland, and in 2007 he received the Outstanding Service Award from the IEEE GRSS. He was a corecipient of the 2012 IEEE Transactions on Geoscience and Remote Sensing Paper Award, and in 2013, he was a corecipient of the IEEE GRSS Highest Impact Paper Award. In 2013, he received the IEEE/VFI Electrical Engineer of the Year Award. He is a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica, and Tau Beta Pi.

Automatic Spectral–Spatial Classification Framework Based on Attribute Profiles and Supervised Feature Extraction

Pedram Ghamisi, *Student Member, IEEE*, Jón Atli Benediktsson, *Fellow, IEEE*, and Johannes R. Sveinsson, *Senior Member, IEEE*

***Abstract*—A robust framework for the classification of hyperspectral images which takes into account both spectral and spatial information is proposed. The extended multivariate attribute profile (EMAP) is used for extracting spatial information. Moreover, for solving the so-called curse of dimensionality, supervised feature extraction is carried out on both the original hyperspectral data and the output of the EMAP. After performing the dimensionality reduction, two output vectors of the original data and attributes are concatenated into one stacked vector. The final classification map is achieved by using a random-forest classifier. The main difficulties of using an EMAP is to initialize the attribute parameters. Therefore, a fully automatic scheme of the proposed method is introduced to overcome the shortcomings of using EMAP. The proposed method is tested on two widely known data sets. Experimental results confirm that the proposed method provides an accurate classification map in an acceptable CPU processing time.**

***Index Terms*—Attribute profile (AP), automatic classification, feature extraction (FE), hyperspectral image analysis, random forest (RF) classifier, spectral–spatial classification.**

I. Introduction

SUPERVISED classification plays a key role in remote sensing image processing and is important in many applications, including crop monitoring, forest applications, urban development, mapping and tracking, and risk management. Owing to recent advances in remote sensing technologies, the spatial resolution of the sensed images has increased. This has led to a better identification of relatively small structures such as roads and houses. Hyperspectral sensors capture hundreds of spectral bands from ultraviolet to infrared for each image pixel, which is helpful for the detailed physical analysis of structures in the captured image [1].

In the spectral domain, each spectral channel is considered as one dimension. By increasing the features in the spectral domain, theoretical and practical problems may arise. For instance, while keeping the number of training samples constant, the classification accuracy actually decreases when the number of features becomes large [2]. For the purpose of classification, these problems are related to the curse of dimensionality. In [3], Landgrebe shows that too many spectral bands can be undesirable from the standpoint of expected classification accuracy because the accuracy of the statistics estimation decreases (Hughes phenomenon). The aforementioned issue demonstrates that there is an optimal number of bands for classification accuracy, and more features do not necessarily lead to better results. Therefore, the use of feature reduction techniques may lead to a better classification accuracy [4].

Conventional spectral classifiers consider the hyperspectral image as a list of spectral measurements with no spatial organization [5]. A joint spectral and spatial classifier is required in order to reduce the labeling uncertainty that exits when only spectral information is taken into account and helps to overcome the salt-and-pepper appearance of the classification map. Furthermore, other relevant contextual information can be extracted when the spatial domain is considered. As an example, for a given pixel, it is possible to extract the size and the shape of the structure to which it belongs. Therefore, using a combination of spectral and spatial information can improve the accuracy of the classification.

In order to extract the spatial information, two neighborhood systems are available: crisp neighborhood system and adaptive neighborhood system. One way for extracting spatial information by using crisp neighbors is to consider the Markov random field (MRF) modeling. MRF is a family of probabilistic models that can be described as 2-D stochastic processes over discrete pixel latices [6]. They can be considered as a powerful tool for incorporating spatial and contextual information into the classification framework. There is an extensive literature on the use of MRFs in classification such as [7]–[12]. Texture analysis is another way of considering a crisp neighbor system to extract spatial information. That approach is widely used in remote sensing (e.g., in [13]). However, the main shortcoming of considering a set of crisp neighbors is the following: 1) The standard neighborhood system may not contain enough samples, which decreases the effectivity of the classifier, particularly when the input data set is of high resolution and the neighboring pixels are highly correlated [1], and 2) a larger neighborhood system leads to intractable computational problems [1].

To solve the aforementioned problem, an adaptive neighborhood system can be taken into account. One way for

Manuscript received June 25, 2013; revised September 13, 2013 and November 18, 2013; accepted November 19, 2013. This work was supported in part by the Icelandic Research Fund for Graduate Students.

The authors are with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland (e-mail: peg6@hi.is).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2292544

considering the adaptive neighborhood system is to use different types of segmentation methods. Image segmentation is a procedure which can be used to modify the accuracy of classification maps [14]. To make such an approach effective, an accurate segmentation of the image is needed [15]. An extensive literature is available on the extraction of spatial information using segmentation techniques (e.g., [16]–[18]).

Another set of methods which can extract spatial information by using an adaptive neighbor system is based on morphological filters. Pesaresi and Benediktsson [19] used morphological transformations to build a morphological profile (MP). In [20], the concept of MP was extended in order to handle hyperspectral images, and the extension was named the extended MP (EMP). In [21], EMP was used along with the support vector machine (SVM) classifier in order to perform spectral and spatial classification on hyperspectral images. The attribute profile (AP) is another extension of MP and provides a multilevel characterization of an image by using the sequential application of morphological attribute filters (AFs) which can be considered for modeling different specifications of the structural information [22]. AP is a powerful tool to increase the discrimination of different classes [22], [23].

In this paper, a new automatic approach is proposed for the accurate classification of remote sensing images. Although the method can be used for the classification of multispectral images with a coarse spectral resolution, it is used here for the spectral and spatial classification of hyperspectral images. The spatial part of the method consists of both the unsupervised and supervised feature extraction (FE) and the extended multivariate AP (EMAP). Supervised FE is applied on the spectral part. Furthermore, the results of the spatial and spectral features are gathered into a stacked vector. In the field of hyperspectral image analysis, the most widely used classifiers are random forest (RF) and SVM classifiers. These two methods are comparable in the sense of classification accuracies. However, while both methods are shown to be effective classifiers for nonlinear classification problems, SVM requires an exhaustive (computationally intensive) parameter tuning (e.g., model selection performed on a grid) for optimal results, whereas RF does not require any such tuning. In addition, the overall accuracy (OA) is not the only critical issue for the purpose of hyperspectral image classification. Another critical index which can evaluate the efficiency of a classifier is the CPU processing time. In this sense, RF is faster than SVM, and therefore, we prefer using RF instead of SVM in the final step, where the produced stacked vector is classified. The proposed method is performed in experiments on two well-known data sets. Results confirm that the new method is able to effectively classify hyperspectral images both in terms of classification accuracies and CPU processing time. To make the new approach as efficient as possible, an automatic scheme for the new method is introduced in this paper in order to solve the main difficulties in using the EMAP.

This paper is organized as follows. The proposed methodology is discussed in Section II. Then, Section III is devoted to experimental results. Finally, Section IV outlines the main conclusions.

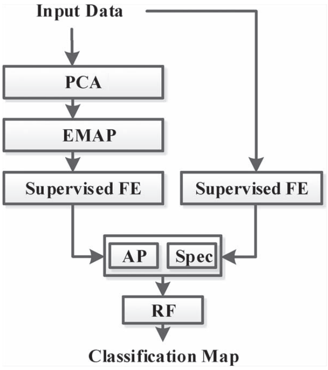


Fig. 1. Flowchart of the proposed method.

II. Methodology

In the proposed method, supervised FE is performed initially on the input data. In parallel, the input data are transformed by principal component (PC) analysis (PCA), and the most important PCs are used as base images for the EMAP. Then, the supervised FE is performed on the output of the EMAP. Furthermore, the most important features with cumulative eigenvalues of more than 99% are selected as the output of the supervised FE. Finally, the spatial and spectral features are gathered into a stacked vector and classified by the RF classifier, and a final classification map is achieved. Fig. 1 illustrates the flowchart of the new method. In the following, specific parts of the proposed framework will be discussed in detail.

A. FE

FE can be explained as finding a set of vectors that represents an observation while reducing the dimensionality. From one point of view, FE can be classified into two categories: unsupervised and supervised FE, where the former is used for the purpose of data presentation and latter is considered for solving the so-called Hughes phenomena [2] and reducing the redundancy of data in order to improve classification accuracies. In pattern recognition, it is desirable to extract features which are focused on the discrimination between classes of interest. Although a reduction in dimensionality is of importance, the error rising from the reduction in dimension has to be without sacrificing the discriminative power of classifiers [21]. In this paper, PCA is used for the purpose of unsupervised FE, and discriminant analysis FE (DAFE), decision boundary FE (DBFE), and nonparametric weighted FE (NWFE) are taken into consideration for the purpose of supervised FE. Below, PCA, DAFE, DBFE, and NWFE are described in more detail.

1) PCA: The general aim of PCA is to transform the data into a lower dimensional subspace which is optimal in terms of sum of squared error [24]. PCA reduces the dimensionality of a data set with interrelated variables while retaining as much as possible the variation in the data set. The dimensionality reduction is obtained by a linear transformation of the data into a new set of variables, the PCs. The PCs are orthogonal to each other and are ordered in such a fashion that the first

PC corresponds to the greatest variance, the second component corresponds to the second greatest variance, and so on.

2) DAFE: This approach is widely used for dimension reduction in classification problems [25]. Since DAFE uses the mean vector and the covariance matrix of each class, it is considered as supervised FE. In DAFE, within-class, between-class, and mixture scatter matrices are usually considered as the criteria of class separability. DAFE is fast and works well when the distribution of data is normal. Otherwise, the performance of DAFE may not be satisfactory. Another problem associated with this method is that, if the difference in the class mean vectors is small, the feature chosen will not be reliable. In the same way, if one class mean vector is very different from others, its class will eclipse the others in the computation of the between-class covariance matrix [26]. As a consequence, the FE process will be ineffective. In addition, DAFE performs the computations at full dimensionality, which requires a large number of training samples in order to accurately estimate parameters. The main shortcoming of DAFE is that DAFE is not full rank, but its rank at maximum is equal to $L - 1$, where L is the number of classes. Assuming that the rank of the within-class scatter matrix is u, then only $\min(L - 1, u)$ features are selected by using DAFE. Since, in real situations, the data distribution is complicated, using only $L - 1$ features usually is not sufficient.

3) DBFE: This method was proposed in [27] where it was shown that both discriminantly informative features and redundant features can be extracted from the decision boundary between two classes. The features are extracted from the decision boundary feature matrix (DBFM). In order to obtain the same classification accuracy as in the original space, keeping the eigenvectors of the DBFM corresponding to nonzero eigenvalues is crucial. The performance of this method does not deteriorate even when there is no difference in the mean vectors or the covariance matrices, and the approach does not rely on the number of classes in the same way as the DAFE. The efficiency of DBFE is highly dependent on training samples, which is not desirable. Another shortcoming of DBFE is that it can be computationally intensive.

4) NWFE: In order to overcome the limitations of DAFE and DBFE, NWFE was introduced in [28]. NWFE is developed based on DAFE by focusing on samples near the eventual decision boundary. The main ideas behind NWFE are to put different weights on different samples in order to compute "weighted means" and define new nonparametric within-class and between-class scatter matrices.

B. EMAP

1) Connected Components: A connected component is regarded as a group of *iso*-level pixels which are connected according to a predefined connectivity rule. Two pixels are connected based on a connectivity rule. The most well-known connectivity rules are 4- and 8-connected, where a pixel is considered as adjacent to four or eight of its neighboring pixels, respectively.

2) Basic Morphology Operators: Erosion and dilation are considered as the alphabets of mathematical morphology. These operators are performed on an image with a set of known shape, called a structuring element (SE). Opening and closing are combinations of erosion and dilation. These operators simplify the input image by removing structures with size less than the SE. However, these operators make changes on the shape of the structures which are still present in the image after the opening/closing. Therefore, they can introduce fake objects in the image [21]. One way to solve this issue is to consider opening and closing by reconstruction. Opening and closing by reconstructions are connected operators that satisfy the following criterion: If the SE cannot fit the structure of the image, then it is totally removed; otherwise, it is totally preserved. Reconstruction operators remove objects smaller than the SE without altering the shape of those objects and reconstruct connected components from the preserved objects. For grayscale images, opening by reconstruction removes unconnected light objects, and in dual, closing by reconstruction removes unconnected dark objects.

C. MP

To determine the shape or size of all structures present in an image, it is crucial to use a range of different SE sizes for the better analysis of structures in the image. MPs are defined using successive opening/closing operations with an SE of an increasing size. The successive usage of opening/closing leads to a simplification of the input image and a better understanding of different structures in the image. An MP is composed of the opening profile and the closing profile. Although MP is a powerful tool for the extraction of spatial information, it suffers by few limitations such as the following.

1) The shape of SEs is fixed which is considered as a main limitation for the extraction of objects within a scene.
2) SEs are unable to describe information related to the gray-level characteristics of regions such as spectral homogeneity, contrast, and so on.
3) A final limitation associated with the concept of MPs is the computational complexity. The original image needs to be completely processed for each level of the profile, which demands two complete processings of the image: one performed by a closing transformation and the other performed by an opening transformation. Thus, the complexity increases linearly with the number of levels included in the profile [22].

D. AP

A morphological AP is considered as the generalization of the MP which provides a multilevel characterization of an image by using the sequential application of morphological AFs [22]. Morphological attribute opening and thinning are morphological AFs which were introduced in [29]. AFs are connected operators which process an image by considering only its connected components. For binary images, the connected components are simply the foreground and background regions present in the image. In order to deal with grayscale images, the set of connected components can be obtained by considering the image to be composed by a stack of binary images generated by thresholding the image at all its gray-level values [30].

AFs process an image based on a given criterion. AFs keep or merge the connected component CC_i based on a logical predicate T if a given attribute is greater/lower than an arbitrary reference, such as $T_\kappa^a(CC_i) = a(CC_i) > \kappa$, where a is an attribute and κ is an arbitrary reference value [23]. The criterion is evaluated on all the connected components of the image, and if the criterion is not met, the region is merged to the adjacent region with a closer gray-level value. If the regions with lower (greater) gray-level values are taken into account in the merging process, then the transformation is considered as antiextensive (extensive) [22]. The transformation is idempotent if the result of the transformation is not dependent on the number of times that a transformation with the same parameter is performed. A transformation with the aforementioned specifications is called thinning (thickening).

An AP is obtained by the sequence of attribute thinning and thickening transformations defined with a sequence of progressively stricter criteria [22]. Let ϕ^κ and γ^κ be the attribute thickening and attribute thinning, respectively. The AP of the image f with the set of N criteria is shown by AP(f). Mathematically, AP(f) can be expressed as

$$\text{AP}(f) = \{\phi^{\kappa_N}(f), \phi^{\kappa_{N-1}}(f), \ldots, \phi^{\kappa_1}(f), f, \gamma^{\kappa_1}(f), \ldots, \gamma^{\kappa_{N-1}}(f), \gamma^{\kappa_N}(f)\}. \quad (1)$$

To handle hyperspectral images, the extension of AP was proposed in [31]. The extended AP (EAP) is a stacked vector of different APs computed on the first C features extracted from the original data set (I^D with D dimensions), and fe shows a feature. EAP is given by

$$\text{EAP}(I^D) = \left\{\text{AP}\left(fe_1(I^D)\right), \text{AP}\left(fe_2(I^D)\right), \ldots, \text{AP}\left(fe_C(I^D)\right)\right\}. \quad (2)$$

During the concatenation of different attributes, $a_1, a_2, \ldots, a_M$ are gathered into a stacked vector, and the EMAP is obtained [31] and is given mathematically by

$$\text{EMAP}(I^D) = \left\{\acute{\text{EAP}}_{a_1}(I^D), \acute{\text{EAP}}_{a_2}(I^D), \ldots, \acute{\text{EAP}}_{a_M}(I^D)\right\} \quad (3)$$

where $\acute{\text{EAP}} = \text{EAP}\{\text{PC}_1, \ldots, \text{PC}_c\}$ and a_i is a generic attribute.

The application of the profiles for large volumes of data is computationally demanding, and that is considered to be one of the main difficulties in using them. In order to solve this issue, the efficient implementation of AFs was proposed in [32]. Salembier *et al.* in [32] introduced a new data representation named Max-tree which has received much interest since it increases the efficiency of filtering by dividing the transformation process into three steps: 1) tree creation; 2) filtering; and 3) image restitution [22].

E. Fusion of Extracted Features via Vector Stacking

As can be seen from Fig. 1, the original data are transformed by a supervised FE approach in order to provide a few effective features that contain the spectral information of the input data. Let ζ_φ be the features associated with the spectral bands.

With reference to Fig. 1, the input data are transformed by PCA, and the first effective PCs are used in order to reduce the redundancy in the data but keeping most of the variation. Then, EMAP is computed by using only the first effective PCs that correspond to 99% of the cumulative variance. Afterward, each AP is composed of n thickening and n thinning transformations of the corresponding PC for each attribute. In order to produce the MAP for each PC, depending on the number of attributes (e.g., m different attributes), we come up with $m(2n) + 1$ number of features in each MAP. Finally, the number of features in the EMAP by considering P PCs is equal to $P(m(2n) + 1)$. Let ζ_ω be the features associated to the EMAP. Finally, the obtained stack vector is $\zeta = [\zeta_\varphi, \zeta_\omega]^T$.

F. RF

RF was first introduced in [33] and is an ensemble method for classification and regression. Ensemble classifiers get their name from the fact that several classifiers, i.e., an ensemble of classifiers, are trained and their individual results are then combined through a voting process. For the purpose of classification of an object from an input vector, the input vector is run down each tree in the forest. Each tree provides a unit vote for a particular class, and the forest chooses the classification having the most votes. Based on studies in [34], the computational complexity of the RF algorithm is $cT\sqrt{MN}\log(N)$, where c is a constant, T denotes the number of trees in the forest, M is regarded as the number of variables, and N is the number of samples in the data set. It is easy to detect that RF is not computationally intensive but demands a considerable amount of memory since it needs to store an N-by-T matrix while running. RF can provide a good classification result in terms of accuracies and does not assume any underlying probability distribution for input data. Another advantage of the RF classifier is that it is insensitive to noise in the training labels. In addition, RF provides an unbiased estimate of the test set error as trees are added to the ensemble, and finally, it does not overfit.

G. Automatic Scheme for EMAP

The main difficulties of using the EMAP are as follows: 1) to know which attributes lead to a better discrimination for different classes and 2) to know which threshold values should be considered in order to initialize each AP. In this section, an automatic scheme of the proposed method is introduced in order to solve the latter problem. While the APs can be constructed by using a wide variety of attributes, in the automatic scheme, the area and standard deviation attributes are only used since the aforementioned attributes can be adjusted in an automatic way and are well related to the object hierarchy in the images. The standard deviation is adjusted with respect to the mean of the individual features since the standard deviation shows dispersion from the mean [35]. Therefore, λ_s is initialized in a fashion to cover a reasonable amount of deviation in the individual feature, which is mathematically given by

$$\lambda_s(\text{PC}_i) = \frac{\mu_i}{100}\{\sigma_{\min}, \sigma_{\min} + \delta_s, \sigma_{\min} + 2\delta_s, \ldots, \sigma_{\max}\} \quad (4)$$

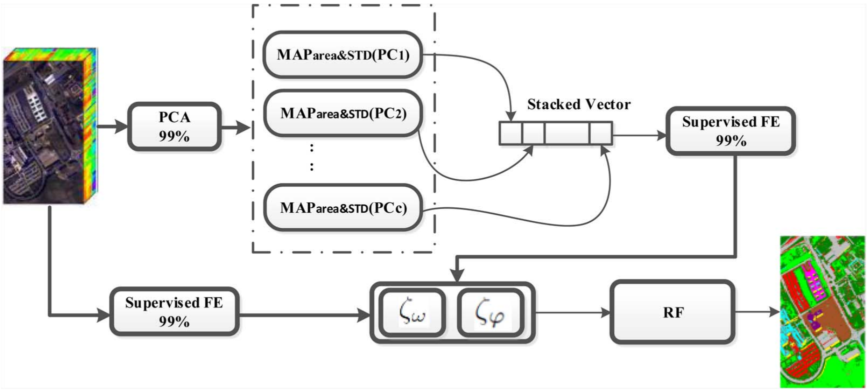


Fig. 2. General idea of the AUTOMATIC scheme of the proposed method. First, PCA is performed on the input data, and first PCs with a cumulative variance of more than 99% are kept. Then, MAP including area and standard deviation attributes with respect to (4) and (5) is built for each PC. Furthermore, the MAPs of different PCs are concatenated into a stacked vector. After that, the supervised FE is carried out on the stacked vector, and first features with cumulative eigenvalues of more than 99% are kept. The output of this part provides the spatial information of the method. In parallel, the supervised FE is performed on the input data, and first features with cumulative eigenvalues of more than 99% are selected. The output of this step is considered as spectral information. As the last stage, the spectral and spatial information are concatenated in a stacked vector, and the stacked vector is classified by RF.

where μ_i is the mean of the ith feature and $\sigma_{\min}$, $\sigma_{\max}$, and δ_s are 2.5%, 27.5%, and 2.5%, respectively, which leads to 11 thinning and 11 thickening operations.

With regard to adjusting λ_a for the area attribute, the resolution of the image should be taken into account in order to construct EAP [23]. The automatic scheme of the attribute area is given as follows:

$$\lambda_a(\mathrm{PC}_i) = \frac{1000}{\upsilon}\{a_{\min}, a_{\min}+\delta_a, a_{\min}+2\delta_a, \ldots, a_{\max}\} \tag{5}$$

where $a_{\min}$ and $a_{\max}$ are initialized by 1 and 14, respectively, with a step increase δ_a equal to 1 and υ shows the spatial resolution of the input data. The EAP for the area attribute includes 14 thinning and 14 thickening operations for each feature. Each level is provided in square meters by considering the resolution of the image υ in meters. As an example, for an image with a spatial resolution of 1 m per pixel, each profile covers structures in the range of 1000–14 000 m^2, which might be a reasonable range of sizes for different structures in both urban and rural cases in remote sensing images [23]. However, different ranges can be considered for different applications. It should be noted that the aforementioned parameters have been tested on other well-known data sets such as the Indian Pines in [23], and results show that these parameters are data set distribution independent and can provide excellent results in terms of classification accuracies. In the introduced framework, one only needs to establish a range of parameter values in order to automatically obtain a classification result with high accuracy for different data sets.

Fig. 2 shows the general idea of the automatic scheme of the proposed method. First, the input data are transformed via PCA, and the first PCs with a cumulative variance of more than 99% are kept since they provide most of the data variation. Then, MAP including area and standard deviation attributes with respect to (4) and (5) is built for each PC. Furthermore, the MAPs of different PCs are concatenated into a stacked vector. Finally, in order to extract spatial information, the stacked vector is transformed by a supervised FE, the first features with cumulative eigenvalues more than 99% are kept, and ζ_ω is the output of this step. In parallel, in order to provide the spectral information, a supervised FE is performed on the input data set, and ζ_φ is the output of this step. The final classification map is provided by performing RF on the output of the stack vector, $\zeta = [\zeta_\varphi, \zeta_\omega]^T$.

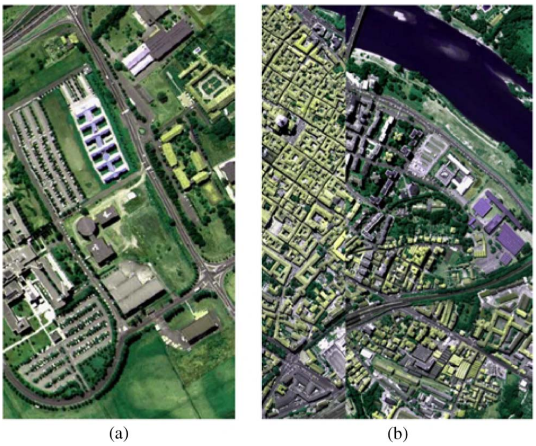


Fig. 3. ROSIS data. (a) University. (b) Center. Data specifications are detailed in Tables I and IV.

III. Experimental Results

A. Data Description

The test cases are hyperspectral data sets which were captured of the city of Pavia, Italy, by airborne data from the ROSIS-03 (Reflective Optics System Imaging Spectrometer). The ROSIS-03 sensor has 115 data channels with a spectral coverage ranging from 0.43 to 0.86 μm. The data have been

TABLE I
PAVIA UNIVERSITY: NUMBER OF TRAINING AND TEST SAMPLES. CLASSIFICATION ACCURACIES OF TEST SAMPLES IN PERCENTAGE. THE NUMBER OF FEATURES IS GIVEN IN BRACKETS. THE BEST ACCURACY IN EACH ROW IS SHOWN IN BOLD

Class		No. of Samples		Spectral	Attribute Profile	Spectral + AP	DAFE	DBFE	NWFE
No.	Name	Training	Test	(103)	(99)	(103 + 99)	(6 + 8)	(29 + 25)	(11 + 8)
1	Asphalt	548	6631	80.8	**96.4**	95.9	93.7	96.0	93.4
2	Meadows	540	18649	56.1	92.5	94.2	**97.3**	95.3	95.6
3	Gravel	392	2099	53.5	68.2	67.6	**96.2**	75.2	57.6
4	Trees	524	3064	98.7	98.4	**99.8**	97.1	96.8	99.2
5	Metal Sheets	265	1345	99.1	99.5	**99.6**	**99.6**	99.7	99.5
6	Soil	532	5029	78.1	68.4	68.4	**98.3**	88.6	97.7
7	Bitumen	375	1330	84.3	**99.9**	**99.9**	**99.9**	99.9	99.5
8	Bricks	514	3682	91.0	99.5	99.4	**99.5**	99.5	98.8
9	Shadows	231	947	98.3	99.7	99.7	88.4	99.3	**99.6**
AA		–	–	82.25	91.43	91.66	**96.72**	94.55	93.51
OA		–	–	71.64	90.74	90.90	**97.00**	94.55	94.58
Kappa		–	–	0.6511	0.8773	0.8794	**0.9604**	0.9280	0.9287

atmospherically corrected but not geometrically corrected. The spatial resolution is 1.3 m per pixel.

1) Pavia University: The first data set is of the Engineering School at the University of Pavia and consists of different classes including the following: trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow, and soil. In the experiments, 12 noisy data channels are eliminated, and 103 data channels are used for processing. The original data set is 610 by 340 pixels. Fig. 3(a) shows a false color composite of the Pavia University scene. The available test and training samples are listed in Table I.

2) Pavia Center: The second data set is captured of the center of Pavia. The original data set is 1096 by 1096 pixels. A 381-pixel-wide black stripe in the left part of the data set was removed, leading to 1096 by 715 pixels. Thirteen data channels were removed due to the noise, and 102 bands were processed. This data set consists of nine classes, i.e., water, trees, meadows, bricks, soil, asphalt bitumen, tiles, and shadows. Fig. 3(b) depicts a false color composite of Pavia Center. The available test and training samples are listed in Table IV.

B. General Information

The input image is transformed by PCA, and the first PCs with a cumulative variation of more than 99% are kept since they contain almost all of the variance in the data sets. For Pavia University, three PCs are necessary to retain 99% of the variance criterion (the cumulative sum of eigenvalues in percentage is 99.55%). In the same way, for Pavia Center, three PCs are necessary to retain 99% of the variance criterion (the cumulative sum of eigenvalues in percentage is 99.47%). In the same fashion, for all supervised FEs (DAFE, DBFE, and NWFE), the first features with cumulative eigenvalues of more than 99% are selected. The number of trees for the RF classifier is equal to 200.

In this paper, Spectral refers to when only spectral information is classified by RF. In the same way, when only spatial information is taken into account, it is called AP. AP+Spectral is referred to the classification of the stacked vector including both spectral and spatial information without performing feature reduction. For simplification, the proposed approaches using DAFE, DBFE, and NWFE are called DAFE, DBFE, and NWFE, respectively.

For the MANUAL scheme of the proposed method, four attributes are considered:

1) (a) area of the region (related the size of the regions);
2) (s) standard deviation (as an index for showing the homogeneity of the regions);
3) (d) diagonal of the box bounding the regions;
4) (i) moment of inertia (as an index for measuring the elongation of the regions).

The values of each attribute are adjusted based on studies in [31] and given as follows:

$\lambda_a = \{100, 500, 1000, 5000\}$;
$\lambda_s = \{20, 30, 40, 50\}$;
$\lambda_d = \{10, 25, 50, 100\}$;
$\lambda_i = \{0.2, 0.3, 0.4, 0.5\}$.

As mentioned before, in the AUTOMATIC scheme of the proposed method, the area (a) and standard deviation (s) attributes are only used in constructing our APs since these attributes can be adjusted automatically and are well related to the object hierarchy in the images. It should be noted that, in order to provide a comparative evaluation of the results, the results of the MANUAL scheme and AUTOMATIC scheme are evaluated separately.

The following measures are used in order to evaluate the performance of different classification methods.

1) Average Accuracy (AA): This index shows the average value of the class classification accuracy.

2) OA: This index represents the number of samples which is classified correctly divided by the number of test samples.

3) Kappa Coefficient: This index provides information regarding the amount of agreement corrected by the level of agreement that could be expected due to chance alone.

4) McNemar's Test: This test is used to assess classification results and is calculated by

$$M = \frac{d_{21} - d_{12}}{\sqrt{d_{12} + d_{21}}} \tag{6}$$

where d_{12} is the number of samples that are incorrectly classified by a first classifier but not the second one and d_{21} has a dual meaning [36]. The difference between the proposed method and others is statistically significant at 5% significant level if $|M| > 1.96$. It should be noted that, in each comparison, the method which provides a better OA has been considered

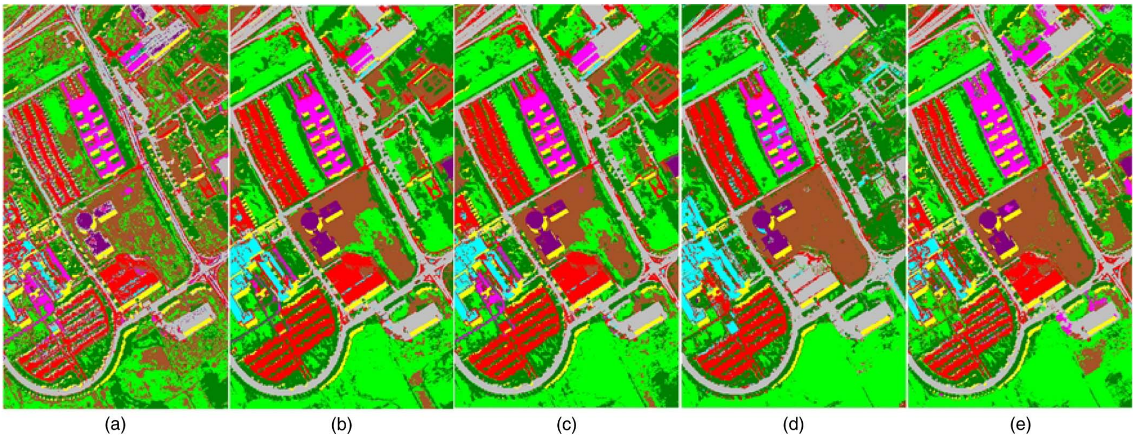


Fig. 4. Classification maps of different methods for Pavia University. (a) Spectral. (b) AP. (c) Spectral+AP. (d) DAFE. (e) NWFE.

as classifier 1 and the other method has been considered as classifier 2.

5) CPU Processing Time: This measure shows the speed of different algorithms. It should be noted that, since, in all algorithms (except Spectral), EMAP is carried out, the CPU processing time of this step is discarded from all methods. Hence, the CPU processing time is only provided for AP, AP+Spectral, DAFE, and NWFE. All methods used were programmed in MATLAB on a computer having an Intel Pentium 4 3.20-GHz CPU and 4 GB of memory.

C. Results

1) MANUAL Scheme:

a) Pavia University: As can been seen in Table I, DAFE gives the highest classification accuracy compared to other methods used and improves the OA of Spectral, AP, Spectral+AP, DBFE, and NWFE by almost 25%, 6%, 6%, 2.5%, and 2.5%, respectively. This shows that when a sufficient set of training samples is available, DAFE leads to more discriminant features in comparison with those achieved by DBFE and NWFE. The main reason for that might be that the number of selected features used by DBFE and NWFE is not sufficient. As a result, more features need to be considered in order to provide more promising results in the case of NWFE and DBFE.

AP shows a better performance than Spectral in terms of accuracies and improves the OA by almost 19%. Spectral has better class accuracies for classes Trees and Soils where spectral information can lead to better discrimination of those classes than the spatial information. According to Fig. 4, by considering the spatial dependences using AP, the noisy behavior of classified pixels by RF has been decreased significantly.

As can be seen from the table, by considering both Spectral and AP in the same stacked vector (Spectral+AP), the OA of the classification is improved by only 0.2% (see Table I), and the CPU processing time is increased by 63 s (see Table II). This infers that having both the spectral and AP features in the same vector and using more features (202 features instead of 99 features) do not necessarily lead to better classification results. Spectral+AP improves the class accuracies of Meadows, Trees, and Metal Sheets in comparison with the cases when Spectral and AP have been classified separately and degrades the class accuracies of Asphalt, Gravel, and Bricks compared with AP and the class accuracy of Soil compared with Spectral. In other words, the consideration of the full features obtained by AP along with the input data (Spectral) sometimes can lead to a better discrimination of different classes and sometimes downgrades class accuracies in comparison with the individual use of either AP or Spectral.

Table II shows that DAFE has the least CPU processing time in comparison with the other methods used. DAFE is a very fast FE method and is able to find more effective features in less CPU processing time than NWFE.

In Table III, the difference in classification accuracy between the DAFE and others is statistically significant using the 5% level of significance. In the same way, in comparison with

TABLE II
PAVIA UNIVERSITY: CPU PROCESSING TIME IN SECONDS FOR DIFFERENT METHODS OF THE GENERAL METHOD. SINCE AP IS USED FOR ALL METHODS, THE CPU PROCESSING TIME FOR MAKING AP IS DISREGARDED. FOR DAFE AND NWFE, THE CPU PROCESSING TIME IS THE SUMMATION OF THE FE PART AND THE CLASSIFICATION STEP

Spectral	Attribute Profile	Spectral + AP	DAFE	DBFE	NWFE
62	39	102	12	65	50

TABLE III
PAVIA UNIVERSITY: RESULT OF MCNEMAR'S TEST TO VALIDATE WHETHER THE DIFFERENCE BETWEEN CLASSIFICATION ACCURACIES OF THE PROPOSED METHOD IS SIGNIFICANTLY DIFFERENT FROM OTHER METHODS

Pavia University	M
DAFE vs. Spec	99.01
DAFE vs. AP	41.34
DAFE vs. ALL	41.28
DAFE vs. NWFE	20.16
DAFE vs. DBFE	22.70
NWFE vs. ALL	27.61
NWFE vs. AP	27.63
NWFE vs. Spectral	94.55
NWFE vs. DBFE	0.19
AP vs. Spectral	74.75

TABLE IV
PAVIA CENTER: NUMBER OF TRAINING AND TEST SAMPLES. CLASSIFICATION ACCURACIES OF TEST SAMPLES IN PERCENTAGE. THE NUMBER OF FEATURES IS GIVEN IN BRACKETS. THE BEST ACCURACY IN EACH ROW IS SHOWN IN BOLD

Class		No. of Samples		Spectral	Attribute Profile	Spectral + AP	DAFE	DBFE	NWFE
No.	Name	Training	Test	(102)	(99)	(102 + 99)	(7 + 8)	(30 + 25)	(13 + 33)
1	Water	824	65147	98.8	98.6	98.2	98.9	96.9	**99.3**
2	Trees	820	6778	88.1	91.0	88.7	88.3	91.2	**92.7**
3	Meadows	824	2266	95.9	95.6	**98.1**	96.3	95.9	95.8
4	Bricks	808	1891	66.4	99.0	99.1	**99.6**	98.8	**99.6**
5	Soil	820	5764	89.9	**99.7**	99.5	98.5	98.4	98.8
6	Asphalt	816	8432	93.7	**99.4**	99.3	99.2	98.6	99.0
7	Bitumen	808	6479	93.3	98.1	98.1	99.4	**99.1**	97.9
8	Tiles	1260	41566	97.6	99.6	99.7	99.7	99.7	**99.8**
9	Shadows	476	2387	99.6	98.5	99.1	63.6	**100**	**100**
AA		–	–	91.51	97.77	97.81	93.71	97.66	**98.14**
OA		–	–	96.56	98.58	98.39	98.05	97.83	**99.02**
Kappa		–	–	0.9503	0.9795	0.9767	0.9717	0.9688	**0.9858**

other methods used, NWFE is statistically significant when considering the 5% level of significance. As can be seen from Table III, the difference between NWFE and DBFE is not statistically significant at the 5% significance level.

b) Pavia Center: In Table IV, it can be seen that NWFE works better than DAFE in terms of the classification accuracies. The main reason for that might be the following: 1) The between-scatter matrix in DAFE is not full rank, its rank is equal to $L - 1$, and only $L - 1$ features are selected when using DAFE; as it was mentioned before, since, in real situations, the data distribution is complicated, using only $L - 1$ features is not enough; and 2) DAFE works well when the distribution of data is normal. Otherwise, the performance of DAFE is not satisfactory. As also can be seen, NWFE improves the result of DAFE by 1% in terms of OA. NWFE improves the OA of the classification obtained by DBFE. That infers that, for Pavia Center, NWFE which is a nonparametric FE can discriminate different classes of interest in comparison with DAFE and DBFE which are parametric and based on a normal distribution.

AP works better than Spectral+AP in terms of the OA and Kappa coefficient. The number of features in Spectral+AP is 201, which increases the possibility of the problems with the curse of dimensionality. In other words, by increasing the dimensionality of the data set, although class separability increases, the accuracy of the class statistics estimation decreases, which means a higher dimensional set of statistics must be estimated with a fixed number of samples.

The classification of AP also works better than the classification of Spectral in terms of the classification accuracies. One reason for that would be that, since this data set contains a very dense urban area, AP can provide discriminative information which might be more useful than spectral information.

The CPU processing time for different methods is given in Table V. DAFE has the lowest CPU processing time. The main reasons can be the following: 1) DAFE is a very fast FE approach, and 2) only 15 features are classified by RF. Spectral+AP has the worst performance in terms of CPU processing time since 201 features need to be classified by RF.

Fig. 5 shows the classification map for different classifiers. As can be seen, spectral–spatial methods improve the noisy behavior of using RF on the original data since those methods consider spatial dependences as well.

TABLE V
PAVIA CENTER: CPU PROCESSING TIME IN SECONDS FOR DIFFERENT METHODS OF THE GENERAL METHOD. SINCE AP IS USED FOR ALL METHODS, THE CPU PROCESSING TIME FOR MAKING AP IS DISREGARDED. FOR DAFE AND NWFE, THE CPU PROCESSING TIME IS THE SUMMATION OF THE FE PART AND THE CLASSIFICATION STEP

Spectral	Attribute Profile	Spectral + AP	DAFE	DBFE	NWFE
447	313	1096	32	124	156

As can be seen from Table VI, the difference in classification accuracy between the proposed method using NWFE and the others is statistically significant using the 5% level of significance.

Based on the results given in this section, it is obvious that the importance of including the spatial information leads to an increase in classification accuracies when compared to the classification of the original data set. Moreover, considering EMAP allows us to obtain a representation of the image based on complementary characteristics, which helps to a great extent in improving the result of the classification in terms of accuracies. Furthermore, the redundancy of the AP and the original data set can be easily solved by using supervised FE.

2) AUTOMATIC Scheme: Table VII shows the result of the classification in terms of the accuracies for the automatic schemes of the proposed method for both test cases. As can be seen from Table VII, the achieved accuracies are almost the same as the accuracies reported in Tables I and IV. However, for the automatic scheme, there is no need to adjust the initial parameters for the APs, which is considered as the main shortcoming of the usage of AP. The little difference obtained in the classification accuracies between the MANUAL and AUTOMATIC settings of the proposed method can show that the use of only two attributes—area and standard deviation—can model the spatial information on the used data sets considerably and other attributes (diagonal of the box bounding the region and the moment of inertia) do not add significant improvement to classification accuracies, although they carry information on the shape of regions. It is generally accepted that the use of different attributes will lead to the extraction of complementary (and redundant) information from the scene, leading to increased accuracies when used in classification (provided that the Hughes effect is efficiently solved by only keeping those features which are most informative).

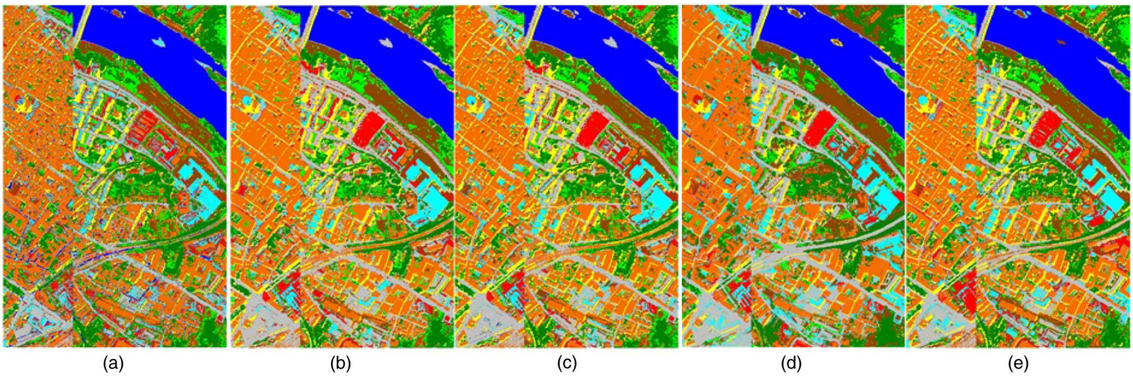


Fig. 5. Classification maps of different methods for Pavia Center. (a) Spectral. (b) AP. (c) Spectral+AP. (d) DAFE. (e) NWFE.

TABLE VI
PAVIA CENTER: RESULT OF MCNEMAR'S TEST TO VALIDATE WHETHER THE DIFFERENCE BETWEEN CLASSIFICATION ACCURACIES OF THE PROPOSED METHOD IS SIGNIFICANTLY DIFFERENT FROM OTHER METHODS

Pavia Center	M
NWFE vs. Spec	54.36
NWFE vs. AP	15.48
NWFE vs. ALL	23.63
NWFE vs. DAFE	28.20
NWFE vs. DBFE	34.34
DAFE vs. ALL	-9.28
DAFE vs. AP	-14.90
DAFE vs. Spectral	28.89
DAFE vs. DBFE	26.83
AP vs. Spectral	44.27

TABLE VII
RESULT OF THE CLASSIFICATION IN TERMS OF THE ACCURACIES FOR THE AUTOMATIC SCHEMES FOR THE PROPOSED METHOD FOR BOTH TEST CASES. THE BEST ACCURACY IN EACH ROW IS SHOWN IN BOLD

Class	Pavia Center			Pavia Uni		
No.	DAFE (7+6)	DBFE (30+14)	NWFE (13+24)	DAFE (6+8)	DBFE (29+16)	NWFE (11+27)
1	99.3	96.5	**100**	**97.2**	96.9	96.5
2	89.2	86.0	**89.9**	**94.8**	91.4	94.7
3	96.5	96.4	**96.7**	**84.2**	71.6	70.5
4	**99.6**	**99.6**	99.4	99.1	99.8	**99.9**
5	97.9	**99.5**	98.3	99.6	**99.9**	99.6
6	98.1	97.4	**98.7**	**99.1**	98.6	85.0
7	**98.8**	96.9	97.8	99.9	**100**	99.7
8	**99.8**	**99.5**	99.8	**99.8**	99.5	99.5
9	99.6	**100**	**100**	96.4	**99.7**	**99.7**
AA	97.69	96.89	**97.87**	**96.72**	95.30	93.96
OA	98.83	97.19	**99.16**	**96.30**	94.20	93.92
Kappa	0.9830	0.9596	**0.9878**	**0.9514**	0.9244	0.9197

TABLE VIII
PAVIA CENTER AND PAVIA UNIVERSITY: CPU PROCESSING TIME OF THE AUTOMATIC SCHEME OF THE PROPOSED METHOD IN SECONDS FOR DIFFERENT METHODS PER SECOND. SINCE AP IS USED FOR ALL METHODS, THE CPU PROCESSING TIME FOR MAKING AP IS DISREGARDED. FOR DAFE AND NWFE, THE CPU PROCESSING TIME IS THE SUMMATION OF THE FE PART AND THE CLASSIFICATION STEP

Pavia Center			Pavia Uni		
DAFE	DBFE	NWFE	DAFE	DBFE	NWFE
31	127	149	13	81	63

As can be seen in Table VII, for Pavia University, DAFE works better than the other methods used in the experiment in terms of accuracies and improves the OA of DBFE and NWFE by almost 2% and 2.5%, respectively. It should be noted that DAFE provides very good accuracies by considering only 14 features. DBFE works slightly better than NWFE in terms of accuracies but with a higher number of features.

For Pavia Center, NWFE works better than others in terms of accuracies. Furthermore, DAFE is more accurate than DBFE with a less number of features. It should be noted that both data sets follow the same trend as both the AUTOMATIC and MANUAL schemes.

Table VIII shows the CPU processing time of the AUTOMATIC scheme of the proposed method in seconds for the University of Pavia and Pavia Center data sets. As expected for both data sets, DAFE has the least CPU processing time. Generally, the AUTOMATIC scheme follows the same trend of the MANUAL scheme; for Pavia University, NWFE works better than DBFE; and for Pavia Center, DBFE has a shorter processing time than NWFE.

In summary, it can be concluded that the AUTOMATIC version of the proposed method can provide classification maps comparable with the MANUAL version of the proposed method in terms of both classification accuracies and CPU processing time when only two attributes (area and standard deviation) are used instead of four (area, standard deviation, moment of inertia, and diagonal of the box). However, the whole procedure in the AUTOMATIC version of the proposed method as the name indicates is automatic, and there is no need for any parameters to be set.

Based on our literature review, the proposed method improves all methods in the literature in terms of classification accuracies. For example, the proposed method improves the classification accuracy of the classification technique proposed in [21] for Pavia University by almost 10 percentage points [the best OA of 95% for the Pavia University data set reported in [21] is achieved by DBFE (see [21, Table V]) which is equal to 87.97% (same size of train and test sets)]. The best OA for the Pavia Center data set in [21] is achieved by NWFE and is, on the other hand, equal to 98.87% (same size of train and test sets but slightly larger test set). Thus, the best improvement in OA

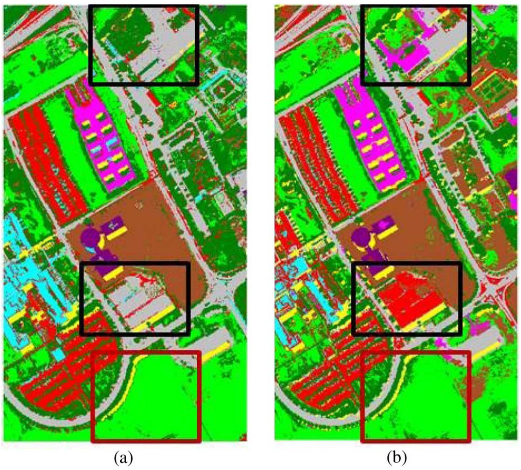


Fig. 6. Comparison between classification maps obtained by (a) DAFE and (b) NWFE.

is from 87.97% to 97.00% (almost by 10 percentage points with the MANUAL version) for the Pavia University data set. Based on the results reported in [37], the performance of the proposed method is as follows.

1) Against the independent component analysis (ICA) results in [37], the proposed approach gives improved overall accuracies of 2.5% (with the MANUAL version) and 1.83% (with the AUTOMATIC version) for the Pavia University data set.
2) Against the PCA results in [37], the proposed approach gives improved overall accuracies of 19.19% (with the MANUAL version) and 18.49% (with the AUTOMATIC version) for the Pavia University data set. It should be noted that the test samples of Pavia Center used in [37] are different from the test samples of Pavia Center used in this paper. Therefore, the results reported in [37] and in this paper for Pavia Center are not fully comparable.

It should be again noted that, in the proposed AUTOMATIC method, there is no need to adjust any parameters, which increases the desirability of using the method.

The following results have been figured out in this paper.

1) When the number of training samples is adequate, the use of DAFE may lead to better classification accuracies. With reference to Table I, DAFE improves the OA of NWFE by almost 2.5%.
2) Fig. 6 shows that not only the number of training samples is important for the efficiency of DAFE and NWFE but also the distribution of training samples on the whole data set is of importance. As an example, the black boxes in Fig. 6 show two parts of the input data which do not contain training samples. In this case, although the OA of DAFE (97.00%) is significantly higher than the OA of NWFE (94.58%), some objects are missing in the classification map obtained by DAFE since the data do not have training samples in those regions. On the contrary, for the region where there is an adequate number of training samples (the red box), the use of DAFE leads to a smoother classification map.
3) The classification of AP works much better than the classification of Spectral in terms of the classification accuracies. In data sets which contain a very dense urban area, AP can provide discriminative information which can be more useful than spectral information.
4) The AUTOMATIC version of the proposed method can provide classification maps that are comparable with the MANUAL version of the proposed method in terms of both classification accuracies and CPU processing time when only two attributes (area and standard deviation) are used instead of four (area, standard deviation, moment of inertia, and diagonal of the box). However, the advantage of the AUTOMATIC version is that there is no need for any parameters to be set.

IV. Conclusion

In this paper, a new approach is proposed for the classification of hyperspectral images which uses both spectral and spatial information. The method can be implemented fully automatically. In order to use the spatial information, APs are taken into account. For reducing the redundancy of both the spatial information and the original spectral data in order to provide more accurate classification results, a few supervised FE methods are considered. The new method was tested on two data sets, and the obtained results confirm that considering spatial information by using APs in conjunction with spectral information can significantly improve the classification accuracies of the original data. In addition, by using supervised FE, the classification accuracies can be increased further. Furthermore, in order to avoid the main difficulties of using APs, an automatic version of the proposed method is introduced which only considers area and standard deviation attributes. The AUTOMATIC method obtained almost the same results as the MANUAL method in terms of the classification accuracies and CPU processing time and solved the main difficulty of the MANUAL method which is related to the initialization of the parameters in the EMAP. The proposed method was tested on two widely used ROSIS data sets named Pavia Center and Pavia University. The proposed method worked well in terms of the classification accuracy and CPU processing time, which confirms the ability of the method to classify high-dimensional data sets.

In experiments, the proposed approach can be thought of as a general framework, and some of its steps can be replaced by other techniques, possibly to improve the CPU processing time and classification accuracies for the proposed approach. For example, other types of feature reduction techniques such as kernel PCA, ICA, and supervised feature extraction and supervised feature selection can be used instead of PCA in the proposed approach.

Acknowledgment

The Reflective Optics System Imaging Spectrometer data and corresponding reference information were obtained from Prof. P. Gamba of the University of Pavia, Italy.

REFERENCES

[1] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

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

[4] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, "Classification of hyperspectral images with binary fractional order Darwinian PSO and random forests," in *Proc. SPIE, Image Signal Process. Remote Sens. XIX*, 2013, pp. 88920S–88920S-8.

[5] S. Tadjudin and D. Landgrebe, "Classification of high dimensional data with limited training samples," School Elect. Comput. Eng., Purdue Univ., West Lafayette, IN, USA, Tech. Rep., 1998.

[6] H. Derin and P. A. Kelly, "Discrete-index Markov-type random processes," *Proc. IEEE*, vol. 77, no. 10, pp. 1485–1510, Oct. 1989.

[7] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," in *Proc. PReMI*, 2005, pp. 260–265.

[8] D. Liu, M. Kelly, and P. Gong, "A spatial-temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery," *Remote Sens. Environ.*, vol. 101, no. 10, pp. 167–180, Mar. 2006.

[9] 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," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.

[10] G. Moser and S. B. Serpico, "Combining support vector machines and Markov random fields in an integrated framework for contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2734–2752, May 2013.

[11] M. Khodadadzadeh, R. Rajabi, and H. Ghassemian, "Combination of region-based and pixel-based hyperspectral image classification using erosion technique and MRF model," in *Proc. 18th ICEE*, May 2010, pp. 294–299.

[12] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral-spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, May 2014, to be published.

[13] G. Zhang, X. Jia, and N. M. Kwok, "Spectral-spatial based super pixel remote sensing image classification," in *Proc. 4th Int. Congr. Image Signal Process.*, 2011, no. 3, pp. 1680–1684.

[14] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, "Use of Darwinian particle swarm optimization technique for the segmentation of remote sensing images," in *Proc. IEEE IGARSS*, 2012, pp. 4295–4298.

[15] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, "An efficient method for segmentation of images based on fractional calculus and natural selection," *Exp. Syst. Appl.*, vol. 39, no. 16, pp. 12 407–12 417, Nov. 2012.

[16] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 2973–2987, Aug. 2009.

[17] P. Ghamisi, M. S. Couceiro, F. M. Martins, and J. A. Benediktsson, "Multilevel image segmentation approach for remote sensing images based on fractional-order Darwinian particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, May 2014, to be published.

[18] P. Ghamisi, M. Couceiro, M. Fauvel, and J. A. Benediktsson, "Integration of segmentation techniques for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 342–346, Jan. 2014.

[19] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[20] J. A. Palmason, J. A. Benediktsson, J. R. Sveinsson, and J. Chanussot, "Classification of hyperspectral data from urban areas using morphological preprocessing and independent component analysis," in *Proc. IEEE IGARSS*, 2005, no. 3, pp. 176–179.

[21] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[22] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[23] M. Pedergnana, P. R. Marpu, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "A novel technique for optimal feature selection in attribute profiles based on genetic algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 6, pp. 3514–3528, Jun. 2013.

[24] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 2002, ser. Springer Series in Statistics.

[25] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, NY, USA: Academic, 1974.

[26] L. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, Nov. 1999.

[27] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 388–400, Apr. 1993.

[28] B. C. Kuo and D. A. Landgrebe, "A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2486–2494, Nov. 2002.

[29] E. J. Breen and R. Jones, "Attribute openings, thinnings and granulometries," *Comput. Vis. Image Understand.*, vol. 64, no. 3, pp. 377–389, Nov. 1996.

[30] N. Bouaynaya and D. Schonfeld, "Theoretical foundations of spatially variant mathematical morphology part II: Gray-level images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 837–850, May 2008.

[31] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, Jul. 2010.

[32] P. Salembier, A. Oliveras, and L. Garrido, "Anti-extensive connected operators for image and sequence processing," *IEEE Trans. Image Process.*, vol. 7, no. 4, pp. 555–570, Apr. 1998.

[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 532–539, Oct. 2001.

[34] L. Breiman, "RF tools a class of two eyed algorithms," in *SIAM Workshop*, San Francisco, CA, USA, 2003, pp. 1–56. [Online]. Available: http://www.stat.berkeley.edu/~breiman/siamtalk2003.pdf

[35] P. Marpu, M. Pedergnana, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "Automatic generation of standard deviation attribute profiles for spectral-spatial classification of remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 293–297, Mar. 2013.

[36] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.

[37] M. D. Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

Pedram Ghamisi (S'13) received the B.Sc. degree in civil (survey) engineering from the Tehran South Campus of Azad University, Tehran, Iran, and the M.Sc. degree in remote sensing from the K. N. Toosi University of Technology, Tehran, in 2012. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Iceland, Reykjavik, Iceland.

His research interests are in remote sensing and image analysis with the current focus on spectral and spatial techniques for hyperspectral image classification.

Mr. Ghamisi serves as a reviewer for a number of journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the recipient of the Best Researcher Award for M.Sc. students in the K. N. Toosi University of Technology in the academic year 2010–2011. He was the recipient of the IEEE Mikio Takagi Prize which was awarded for the first place in the Student Paper Competition at the 2013 IEEE International Geoscience and Remote Sensing Symposium, Melbourne, July 2013.

Jón Atli Benediktsson (S'84–M'90–S'M99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984 and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is currently the Prorector for Academic Affairs and a Professor of Electrical and Computer Engineering at the University of Iceland. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in those fields. He is a cofounder of the biomedical start up company Oxymap.

Prof. Benediktsson was the 2011–2012 President of the IEEE Geoscience and Remote Sensing Society (GRSS) and has been on the GRSS Administrative Committee since 2000. He was an Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008 and has served as an Associate Editor of TGRS since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and the IEEE Access since 2013. He is on the International Editorial Board of the *International Journal of Image and Data Fusion* and was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2007–2010. He was the recipient of the Stevan J. Kristof Award from Purdue University in 1991 as outstanding graduate student in remote sensing. In 1997, he was the recipient of the Icelandic Research Council's Outstanding Young Researcher Award; in 2000, he was granted the IEEE Third Millennium Medal; in 2004, he was a corecipient of the University of Iceland's Technology Innovation Award; in 2006, he received the yearly research award from the Engineering Research Institute of the University of Iceland; and in 2007, he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society. He is a corecipient of the 2012 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Paper Award and of the 2013 IEEE Geoscience and Remote Sensing Society Highest Impact Paper Award. He is a Fellow of SPIE and a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica, and Tau Beta Pi.

Johannes R. Sveinsson (S'86–M'90–SM'02) received the B.S. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, and the M.S. and Ph.D. degrees in electrical engineering from Queen's University, Kingston, ON, Canada.

He is currently the Head and a Professor with the Department of Electrical and Computer Engineering, University of Iceland, where he was with the Laboratory of Information Technology and Signal Processing from 1981 to 1982 and the Engineering Research Institute and the Department of Electrical and Computer Engineering as a Senior Member of the research staff and a Lecturer from November 1991 to 1998, respectively. He was a Visiting Research Student with the Imperial College of Science and Technology, London, U.K., from 1985 to 1986. At Queen's University, he held teaching and research assistantships. His current research interests are in systems and signal theory.

Dr. Sveinsson was the recipient of the Queen's Graduate Awards from Queen's University. He was a corecipient of the 2013 IEEE Geoscience and and Remote Sensing Society Highest Impact Paper Award.

Automatic Framework for Spectral–Spatial Classification Based on Supervised Feature Extraction and Morphological Attribute Profiles

Pedram Ghamisi, *Student Member*, *IEEE*, Jón Atli Benediktsson, *Fellow*, *IEEE*,
Gabriele Cavallaro, and Antonio Plaza, *Senior Member*, *IEEE*

***Abstract*—Supervised classification plays a key role in terms of accurate analysis of hyperspectral images. Many applications can greatly benefit from the wealth of spectral and spatial information provided by these kind of data, including land-use and land-cover mapping. Conventional classifiers treat hyperspectral images as a list of spectral measurements and do not consider spatial dependencies of the adjacent pixels. To overcome these limitations, classifiers need to use both spectral and spatial information. In this paper, a framework for automatic spectral–spatial classification of hyperspectral images is proposed. In order to extract the spatial information, Extended Multi-Attribute Profiles (EMAPs) are taken into account. In addition, in order to reduce the redundancy of features and address the so-called curse of dimensionality, different supervised feature extraction (FE) techniques are considered. The final classification map is provided by using a random forest classifier. The proposed automatic framework is tested on two widely used hyperspectral data sets; Pavia University and Indian Pines. Experimental results confirm that the proposed framework automatically provides accurate classification maps in acceptable CPU processing times.**

***Index Terms*—Extended Multi-Attribute Profile (EMAP), hyperspectral image analysis, random forest classification, supervised feature extraction (FE).**

I. Introduction

HYPERSPECTRAL imaging instruments are now able to capture hundreds of spectral channels from the same area on the surface of the Earth. By providing very fine spectral resolution with hundreds of (narrow) bands, accurate discrimination of different materials is possible. In parallel, due to recent advances in hyperspectral technology, the spatial resolution of the sensors is also becoming finer, which allows for a detailed characterization of spatial structures in the scene.

Supervised classification techniques play a key role in the analysis of hyperspectral images, and a wide variety of applications can be handled by successful classifiers in the literature [1], including: land-use and land-cover mapping, crop monitoring, forest applications, urban development, mapping, tracking, and risk management.

In the spectral domain, each spectral channel is considered as one dimension. By increasing the dimensionality in the spectral domain, theoretical and practical problems arise. Some of these problems are related to the curse of dimensionality, which is related to the unbalance between the (high) dimensionality of the input data and the (often limited) number of training samples used in the supervised classification process [2]. In [3], Landgrebe shows that too many spectral bands can be undesirable from the standpoint of expected classification accuracy. In other words, when the number of spectral bands (dimensionality) increases, with a constant number of training samples, the accuracy of the statistics estimation decreases. The aforementioned issue demonstrates that there is an optimal number of bands and that (given an available set of training samples) more features do not necessarily lead to better results. Therefore, feature reduction techniques may lead to better classification accuracies [4].

Conventional classifiers treat hyperspectral images as a list of spectral measurements with no particular arrangement [5] and do not consider spatial dependencies of adjacent pixels. In other words, conventional techniques classify images only based on their spectral information alone. Therefore, these approaches discard information associated with the spatial correlations among distinct pixels in the image. In order to address the aforementioned issue, the consideration of both spectral and spatial information has been widely explored in the literature [6]. In addition, spatial information can provide additional information related to the shape and size of different structures, which generally leads to better classification accuracies and classification maps.

Two strategies are commonly used in order to characterize spatial information: crisp neighborhood system and adaptive neighborhood system. Although the first one mostly considers spatial and contextual dependencies in a predefined

Manuscript received September 19, 2013; revised December 13, 2013; accepted January 06, 2014. This work was supported in part by the Icelandic Research Fund for Graduate Students.

P. Ghamisi and G. Cavallaro are with the Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavik 110, Iceland (e-mail: p.ghamisi@gmail.com; cavallaro.gabriele@gmail.com).

J. A. Benediktsson is with the Engineering Research Institute, University of Iceland, Reykjavik 107, Iceland (e-mail: benedikt@hi.is).

A. Plaza is with the Technology of Computers and Communications, Escuela Politecnica de Caceres, University of Extremadura, E-10071 Caceres, Spain (e-mail: aplaza@unex.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2014.2298876

neighborhood system, the latter is more flexible and it is not confined to a given neighborhood system. One way for extracting spatial information with crisp neighborhood is to consider Markov random field (MRF) modeling. MRF is a family of probabilistic models that can be described as a 2-D stochastic process over discrete pixel lattices [7]. There is extensive literature on the use of MRFs in classification, such as [8], [9]. However, the main disadvantages of considering a set of crisp neighbors are that 1) the standard neighborhood system may not contain enough samples, which decreases the effectiveness of the classifier (in particular, when the input data set is of high resolution and the neighboring pixels are highly correlated [6]) and 2) a larger neighborhood system leads to computationally intractable problems [6]. In order to address the aforementioned issues, adaptive neighborhood systems can be taken into consideration. A possible way to develop adaptive neighborhood systems is to use different types of segmentation methods. Image segmentation is a procedure which can be used to modify the accuracy of classification maps [10]. To make such an approach effective, an accurate segmentation of the image is needed [11]. Several works have previously explored the extraction of spatial information using segmentation techniques (e.g., [12]–[14]). Another set of methods which can extract spatial information by using adaptive neighborhood systems relies on morphological filters. Pesaresi and Benediktsson [15] used morphological transformations to build a so-called Morphological Profile (MP). In [16], the MP was used to handle hyperspectral images and named Extended Morphological Profile (EMP) in this context. Attribute Profiles (APs) constitute another extension of the concept of MP and provide a multilevel characterization of an image by the sequential application of morphological attribute filters, which model different specifications of the structural information contained in the scene [17]. APs provide a powerful tool to increase the discrimination of different classes [17], [18]. However, there are two main difficulties associated with the concept of Extended Multi-Attribute Profiles (EMAP), including: 1) how to establish which attributes lead to a better discrimination for different classes and 2) how to determine which values should be considered in order to initialize each AP.

In this paper, a new fully automatic approach is proposed for accurate classification of hyperspectral images. Although the presented framework can also be used for classification of multispectral images with coarser spectral resolution, it is used here for spectral–spatial classification of hyperspectral images. In order to extract the spatial information, EMAP [17] are automatically generated by the proposed framework. In order to reduce the redundancy of the data and address the so-called curse of dimensionality, different supervised feature extraction (FE) techniques are also included in the proposed framework. The final classification map is provided by a Random Forest (RF) classifier [19], [20]. In order to handle high-dimensional data, RF and SVM have been widely considered as the most powerful classifiers since they are robust when handling high-dimensional data with a limited number of training samples. Both the SVM and RF classifiers are comparable in terms of classification accuracies and have been widely used for the purpose of hyperspectral image classification. However, while both methods are shown to be effective classifiers for nonlinear classification problems, SVM requires a computationally demanding parameter tuning process in order to achieve optimal results, whereas RF does not require such a tuning process. In this sense, RF is faster than SVM. In this paper, our main objective is obtaining good classification accuracies in an acceptable CPU processing time. In addition, several studies such as [21] have reported that Hughes phenomenon [2] is more evident when the number of dimensions is high and the data are classified by SVM instead of RF. The proposed approach is tested using two well-known data sets collected by the reflective optics spectrographic imaging system (ROSIS) over the city of Pavia, Italy, and by the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) over the Indian Pines region in northwestern Indiana. The experimental results confirm that the presented framework is able to classify hyperspectral images efficiently both in terms of classification accuracies and CPU processing time. It should be noted that the proposed approach is fully automatic and there is no need to initialize any parameters empirically. The main contribution of this paper compared to other works on EMAP is that, in most of previous works, the EMAP is built using an unsupervised FE approach such as principal component analysis (PCA), independent component analysis (ICA), and Kernel PCA, while in this work, we explore the use of supervised FE for this purpose. Another main difference is concerned with the automatic nature of our approach in the sense that, in previous works, threshold values for making EMAP needed to be initialized manually, while, in the present framework, a general range of parameter values is used to make the parameter selection automatical. Another important difference is that, in most previous works, the outcome of the AP is directly used for classification, while in our framework, we use a second FE strategy prior to classification. As shown by the present contribution, the results of the second FE step are used for classification purposes by our proposed framework, with concatenating features of both FE steps in one vector and then performing the classification step.

This remainder of the paper is organized as follows: the proposed framework is discussed in Section II. Section III is devoted to validating the framework via extensive experimental results. Section IV outlines the main conclusions and provides hints at plausible future research lines.

II. Framework

In the proposed framework, supervised FE is first performed on the input data and the first features with cumulative eigenvalues above 99% are retained. In the case of discriminant analysis FE (DAFE), the criterion is related to the size of the eigenvalues of the scatter matrices. In the case of decision boundry FE (DBFE), it is related to the size of the eigenvalues of the decision boundary feature matrix (DBFM). Let us consider ζ_φ as the output of this step. Then, EMAPs are built on the first few features and the resulting features are concatenated into one stacked vector. In order to reduce the redundancy of the stacked vector, a supervised FE step is performed once again. Let ζ_ω be the output of this step. The final classification map is provided by performing RF classification on the stacked vector, $\zeta = [\zeta_\varphi, \zeta_\omega]^T$. Fig. 1 illustrates the proposed framework by a

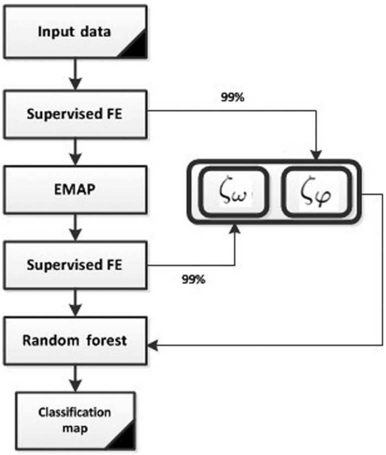


Fig. 1. Flowchart of the proposed framework.

flowchart. In the following, the individual parts of the proposed framework will be discussed in detail.

A. Feature Extraction (FE)

FE consists of finding a set of vectors that represent an observation while reducing its dimensionality. FE techniques can be grouped into two categories: unsupervised approaches and supervised approaches, where the former is used for the purpose of data representation and the latter is considered for overcoming the Hughes phenomena and reducing the redundancy of data in order to improve classification accuracies. PCA is an example of unsupervised FE. PCA does not find optimum feature sets in the sense of class discrimination and discards class specific information. Therefore, for image classification, supervised FE may lead to higher classification accuracies. From one point of view, supervised FE techniques can be split into two categories: parametric and nonparametric. The main disadvantages of nonparametric FE techniques are that they do not have an assumption on the underlying density functions in the data. Therefore, FE for nonparametric classifiers is often not feasible or very time consuming. On the contrary, although the computation cost of nonparametric classifiers is often much larger than that of parametric classifiers, there are some cases where the use of nonparametric FE is desirable. For example, if the underlying densities are unknown or problems involve complex densities, which cannot be approximated by a common parametric density functions, the use of a nonparametric classifier is important [22]. In this work, two approaches are considered for supervised FE: DAFE and DBFE. Below, the considered supervised FE techniques are briefly explained.

1) Discriminant Analysis FE (DAFE): This approach is widely used for dimension reduction in classification problems [23]. Since, DAFE uses the mean vector and the covariance matrix of each class, it is considered as supervised FE. In DAFE, within-class, between-class, and mixture scatter matrices are usually considered as the criteria for class separability. DAFE is fast and works well when the distribution of the data is normal. Otherwise, the performance of DAFE may not be satisfactory. Another problem associated with this method is that if the difference in the class-mean vectors is small, the feature chosen may not be reliable. Similarly, if one class-mean vector is very different from others, its class will dominate the others in the computation of the between-class covariance matrix [24]. As a consequence, the FE process may be ineffective. In addition, DAFE performs the computations with full dimensionality, which requires a large number of training samples in order to accurately estimate parameters. A main shortcoming of DAFE is that DAFE is not full rank and its rank is at maximum $L-1$ where L is the number of classes. Let us assume that the rank of the within-class scatter matrix is u, in this case, only $\min(L-1,u)$ features are selected by using DAFE. Since the complexity of the data in real scenarios could be quite high, using only $L-1$ features may not be enough to fully characterize the data.

2) Decision Boundary Feature Matrix (DBFE): This method was proposed in [25] where it was shown that both discriminantly informative and redundant features can be extracted from the decision boundary between two classes. The features are extracted from the DBFM. In order to obtain the same classification accuracy as in the original space, keeping the eigenvectors of the DBFM corresponding to nonzero eigenvalues is crucial. The performance of this method does not deteriorate even when there is no difference in the mean vectors or covariance matrices. It should be noticed that this approach does not rely on the number of classes in the same way as DAFE. The efficiency of DBFE is highly dependent on the quality and number of training samples, which is not desirable. Another shortcoming of DBFE is that it can be computationally intensive.

B. Extended Multi-Attribute Profile (EMAP)

Mathematical morphology [26]–[29] is a well-established framework which provides operators able to high-quality spatial features. Fundamental mathematical morphology operators, such as Erosion and Dilation (and their combinations: opening and closing), examine the geometrical structures in the image by matching them to small patterns called structuring elements. Depending on the shape and size of the structuring element, undesirable effects can occur in the filtered image; in particular, geometrical characteristics of the structures can be distorted or completely lost. In this work, morphological operators are considered which perform transformations by reconstruction, a class of connected filters [30]. Specifically, they act on connected components, i.e., flat regions of a gray scale image, which are either completely removed or preserved according to their interaction with the structuring element adopted by the transformation.

1) Attribute Filters Based on Tree Representation: Attribute filters [31] are flexible operators that can perform simplification of a grayscale image driven by an arbitrary measure which can be related to characteristics of regions in the scene such as the scale, shape, contrast, etc. Improvements in terms of capability in modeling the spatial information are achievable since these operators are not based on fixed structuring elements, and the image transformation is only computed by merging its connected components. The idea is to extract different types of information, represented by the attributes, from different flat regions, i.e., parts of the scene with the same gray levels. Attribute filters are efficiently implemented with an equivalent representation of the image as a tree [32].

In particular, a thresholding operation of all the mapped values present in the image f, results in upper and lower level sets which are connected components (i.e., flat zones) that can be grouped in the following sets:

$$\mathcal{U}(\mathrm{f}) = \{X : X \in \mathcal{CC}([f \geq \lambda]), \lambda \in \mathbb{Z}\}$$
$$\mathcal{L}(\mathrm{f}) = \{X : X \in \mathcal{CC}([f < \lambda]), \lambda \in \mathbb{Z}\}$$

where $\mathcal{CC}(f)$ being the connected components of the generic image f. There is an inclusion relationship [33] between the connected components extracted by both the upper or lower level sets [belonging to $\mathcal{U}(\mathrm{f})$ or $\mathcal{L}(\mathrm{f})$, respectively]. This property allows for the association of a node in the tree to each connected component and thus represent the image as a hierarchical structure: the max-tree and min-tree [32] structures represent, respectively, the components in $\mathcal{U}(\mathrm{f})$ and $\mathcal{L}(\mathrm{f})$ with their inclusion relations by the thresholding operations. Attribute filters are shape preserving, since they never introduce new edges in an image [32], and operate on regions according to the result of a binary predicate P. In particular, the filtering criteria usually determine whether the value of an attribute α of a given connected component CC verifies a predicate: $P = \alpha(CC) \geq \lambda$ with $\{\alpha(CC), \lambda\} \in \mathbb{R}$ or $\mathbb{Z}$, where λ is a threshold value. When attribute filters are applied to the tree representation of the image, the operator leads to a pruning of the tree by removing the nodes whose associated regions do not fulfill P. Two different filtering approaches have been proposed: pruning the tree by removing whole branches and pruning by not removing all the branches [34]. Attributes can be purely geometric (e.g., area, length of the perimeter, and moment of inertia) or textural (e.g., standard deviation and entropy). A very detailed characterization of features is usually obtained.

2) Attribute Profiles: The spatial features can be derived in different ways such as with Gray-Level Co-occurrence Matrix (GLCM), Differential Morphology Profiles (DMPs), or Urban Complexity Index (UCI) [35]. Here, we propose to use EMAPs instead of GLCM, DMPs, and UCI. The use of EMAPs based on mathematical morphology concepts exhibits some desirable features in the context of hyperspectral image classification. Specifically, they offer a very flexible approach since they can perform the processing based on many different types of attributes. In fact, the attributes can be of any type. For example, they can be purely geometric, or related to the spectral values of the pixels, or on different characteristics. Furthermore, an efficient implementation based on tree representation has been used. In summary, EMAP offers a different strategy to include spatial information when compared to GLCM or UCI.

The spatial information belonging to different features present in very high-resolution data can be efficiently exploited by considering a multilevel approach based on morphological attribute filters. In particular, APs define a general set of profiles which take advantage of the flexibility of the attribute filters in order to better investigate the scene. According to the type of the criteria (increasing, nonincreasing), APs are defined differently. In the case of increasing attributes, the AP is a sequence of attribute openings and closings which include morphological opening and closing profiles by reconstruction [36]. On the other hand, when dealing with increasingness criteria, attribute thinning and thickening over a multilevel approach is applied. The result is the obtained attribute thinning and thickening profiles, which perform a multilevel analysis of the image based on attributes (represented by ordered criteria) not necessary related to the scale of the structures of the image.

APs can be, therefore, regarded as more effective filters than MPs, this is because the latter perform a partial characterization of the objects in the scene as a consequence of the fact that structuring elements are intrinsically unsuitable to describe features related to the graylevel characteristic of the region. Another considerable advantage is that APs are computed according to an effective implementation based on max-tree and min-tree representations, which lead to a reduction of the computational load when compared with conventional profiles built with operators by reconstruction.

3) Extended Attribute Profiles: Since hyperspectral sensors collect information in several spectral bands, Extended Attribute Profiles (EAPs) which are based on morphological attribute filters are adopted in order to perform the analysis of hyperspectral high-resolution images. The extension to multivalued images is not a trivial task, and morphological operators compute their function in a different domain which becomes a subset of the multivariate domain, where the ordering of the mapped vector values is not defined anymore. The EAPs rely on the application of the APs to hyperspectral data and they are simply defined as [36]

$$EAP = \{AP(PC_1), AP(PC_2), \ldots, AP(PC_c)\} \qquad (1)$$

where PC name denotes a principal component obtained after applying PCA [37]. As mentioned before, PCA does not find optimum feature sets in the sense of class discrimination and discards class specific information. Therefore, for image classification, supervised FE leads to higher classification accuracies since such approaches provide optimal features with respect to class specific information. The EAP includes in its definition the EMP since the operators by reconstruction can be viewed as a particular set of morphological attribute. Since the modeling of spatial features is performed by attribute filters, this approach leads to a great flexibility, and the computation of the filters on the max-tree structure reduces the computational complexity with respect to EMP since the tree is built once for each principal component and filtered multiple times, according to the required number of levels.

4) Extended Multi-Attribute Profiles: APs extract efficiently spatial features by considering different attributes; for this reason, EMAPs merge different EAPs in a single data structure [36]

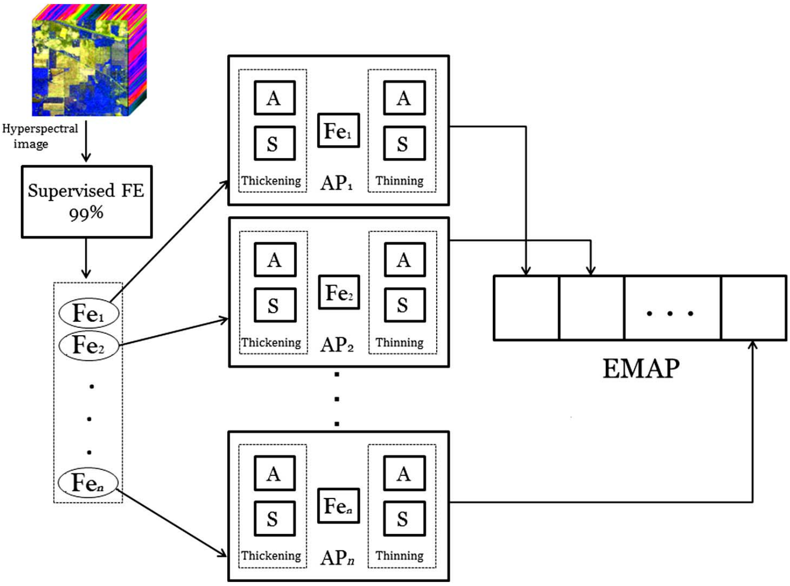


Fig. 2. Automatic framework for the construction of EMAPs. First, a supervised FE step is performed on the input data and the first features with cumulative eigenvalues above 99% are kept. Then, EMAPs are built for the first few features and the output features are concatenated into one stacked vector.

$$EMAP = \{EAP_{a_1}, EAP'_{a_2}, \ldots, EAP'_{a_m})\}. \qquad (2)$$

Since the dimensionality of the features is increased, the EMAP has much greater capabilities in extracting spatial information than a single EAP but, at the same time, the computational cost of processing these features is slightly higher since the max-tree and min-tree are computed only once for each PC and they are filtered with different attributes at different levels.

5) Automatic Framework: Now, an automatic framework is introduced in order to solve issues such as the automatic selection of the attributes that lead to a best possible discrimination between the classes or the automatic identification of the most appropriate values to initialize each AP. Fig. 2 shows the general idea of the automatic framework for the construction of EMAPs. Although the APs can be constructed by using a wide variety of attributes, in the automatic framework, only the area and standard deviation attributes are used, since the aforementioned attributes can be adjusted in an automatic way and are well related to the object hierarchy in the images. The standard deviation is adjusted with respect to the mean of the individual features, since the standard deviation shows dispersion from the mean [21]. Therefore, λ_s is initialized so as to cover a reasonable amount of deviation in the individual feature, which is mathematically given by

$$\lambda_s(Fe_i) = \frac{\mu_i}{100}\{\sigma_{min}, \sigma_{min} + \delta_s, \sigma_{min} + 2\delta_s, \ldots, \sigma_{max}\} \quad (3)$$

where Fe_i denotes the ith feature obtained by a supervised FE. μ_i is the mean of the ith feature and σ_{min}, σ_{max}, and δ_s are 2.5%, 27.5%, and 2.5%, respectively, which leads to 11 thinning and 11 thickening operations. It should be noticed that the abovementioned parameters have been tested on other well-known data sets with different spatial resolutions in [18] and results confirm that these parameters are data set distribution independent and can provide excellent results in terms of classification accuracies.

With regard to the adjustment of λ_a for the area attribute, the resolution of the image should be taken into account in order to construct the EAP [18]. The automatic construction of the attribute area is accomplished by the following expression:

$$\lambda_a(Fe_i) = \frac{1000}{\upsilon}\{a_{min}, a_{min} + \delta_a, a_{min} + 2\delta_a, \ldots, a_{max}\} \quad (4)$$

where a_{min} and a_{max} are initialized by 1 and 14, respectively, with a stepsize increase of δ_a equal to 1. The EAP for the area attribute includes 14 thinning and 14 thickening operations for each feature. Each level is provided in square meters by considering the resolution of the image υ in meters. Each profile covers structures in the range of $1000 - 14\ 000\ \mathrm{m}^2$, which might be a reasonable range of sizes for different structures in both urban and rural cases in remote sensing images [18]. However, different ranges can be considered for different applications.

Regarding (3) and (4), the used parameters have been tested on other well-known data sets with different spatial resolutions in [18] and results confirm that these parameters are data set distribution independent and can provide excellent results in terms of classification accuracies. In other words, those parameters do not need to be tuned for different data sets with different spatial resolutions. In the introduced framework, one only needs to establish a range of parameter values in order to automatically obtain a classification result with high accuracy for different data sets. It turns out that the used parameter ranges have been tested on other well-known data sets with different spatial resolutions, such as the ones described in [18], and the obtained results confirm that these parameters are data set independent. In other words, those parameter ranges can be fixed for different data sets with different spatial resolutions. In [38], it was shown that the automatic scheme with only two attributes (area and standard deviation) can provide results comparable with a manual scheme with four attributes in terms of classification accuracy and CPU processing time.

C. Fusion of Extracted Features Via Vector Stacking

As indicated in Fig. 1, the input data are transformed by a supervised FE and only the first few features are used in order to reduce redundancy in the data while keeping most of the data variance. Then, the EMAP is computed by using only the first effective features that correspond to 99% of the eigenvalues.

Let ζ_φ be the set of features retained. Then, MAP is performed on each feature of ζ_φ and the output features are concatenated into one stacked vector. In order to address the so-called curse of dimensionality and reduce the redundancy of the stacked vector, a supervised FE step is performed once again. Let ζ_ω be the output of this step consisting of the features with cumulative eigenvalues above 99%. The final classification map is achieved by performing RF classification on the stacked vector; $\zeta = [\zeta_\varphi, \zeta_\omega]^T$.

D. Random Forest (RF)

RF was first introduced in [19]. It is an ensemble method for classification and regression. Ensemble classifiers get their name from the fact that several classifiers are trained and their individual results are then combined through a voting process. For the classification of an object from an input vector, the input vector is run down each tree in the forest. Each tree provides a unit vote for a particular class and the forest chooses the classification having the most votes. Based on [20], the computational complexity of the RF algorithm is $cT\sqrt{MN}\log(N)$ where c is a constant, T denotes the number of trees in the forest, M is regarded as the number of variables, and N is the number of samples in the data set. It is easy to infer that RF is not computationally intensive but demands a considerable amount of memory, since it is necessary to store an $N \times T$ matrix in the process. RF has several advantages, such as the capacity to provide good classification accuracies and to handle many variables. Another advantage of the RF classifier is that it is insensitive to noise in the training samples. In addition, RF provides an unbiased estimate of the test set error as trees are added to the ensemble, with almost no sensitivity to overfitting issues.

III. Experimental Results

A. Data Description

Two hyperspectral data sets were used in experiments. They are described as follows.

1) Pavia University: The first test case is a hyperspectral data set captured on the city of Pavia, Italy by the Reflective Optics Spectrographic Imaging System (ROSIS-03) airborne instrument. The ROSIS-03 sensor has 115 data channels with a spectral coverage ranging from 0.43 to 0.86 μm. The data have been corrected atmospherically, but not geometrically. The spatial resolution is 1.3 m per pixel. The data set covers the Engineering School at the University of Pavia and consists of different classes including: trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow, and soil. In our experiments, 12 noisy data channels were eliminated and 103 data channels used for processing. The original data set comprises 640×340 pixels. Fig. 3(a) shows a false color composite of Pavia University and Fig. 3(b) shows a fixed training set that will be used for training purposes in this paper. Fig. 3(c) shows the available reference data for the scene. The number of available test and training samples is listed in Table I.

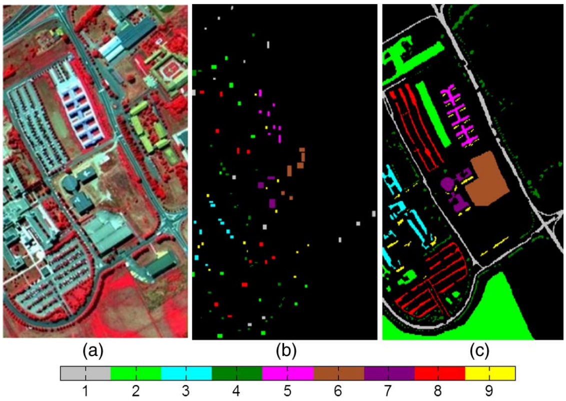


Fig. 3. The ROSIS-03 Pavia University data set: (a) false color image, (b) training samples, and (c) test samples, where each color represents a specific information class. The information classes are listed in Table I.

2) Indian Pines Data: The second data set used in experiments is the well-known data set captured on Indian Pines (NW Indiana) in 1992 comprising 16 classes (see Fig. 4), mostly related to different land covers. The data set consists of 145×145 pixels with spatial resolution of 20 m/pixel. In this work, 200 data channels are used, i.e., after the removal of the spectral bands affected by atmospheric absorption. The number of training and test samples is displayed in Table II.

It should be noted that, in addition to selecting widely used data sets in the hyperspectral imaging community, we have used exactly the same training and test samples that have been considered in most works related to spectral–spatial classification of hyperspectral images. Some of the works that have considered exactly the same training and test samples are those in [9], [39], and [40]. In other words, we not only used the same number of training and test samples adopted by other state-of-the-art methods, but also these samples have exactly the same spatial locations in the data. This way of using the training and test samples makes this work fully comparable with other spectral and spatial classification techniques reported in the literature. In order to keep consistency with previous results, each method was run only once since we have not used different training and test samples, but instead used exactly the same samples as adopted in the previous studies.

TABLE I
PAVIA UNIVERSITY: NUMBER OF TRAINING AND TEST SAMPLES ALONG WITH CLASSIFICATION ACCURACIES FOR THE RAW SPECTRAL DATA IN PERCENTAGE

Class		Number of samples		Raw
Number	Name	training	test	(103)
1	Asphalt	548	6631	80.8
2	Meadows	540	18 649	56.1
3	Gravel	392	2099	53.5
4	Trees	524	3064	98.7
5	Metal sheets	256	1345	99.1
6	Soil	532	5029	78.1
7	Bitumen	375	1330	84.3
8	Bricks	514	3682	91.0
9	Shadows	231	947	98.3
Kappa	–	–	–	0.6511
OA	–	–	–	71.64
AA	–	–	–	82.25

The number of features is given in the parentheses.

TABLE II
INDIAN PINES: NUMBER OF TRAINING AND TEST SAMPLES ALONG WITH CLASSIFICATION ACCURACIES FOR THE RAW SPECTRAL DATA IN PERCENTAGE

Class		Number of samples		Raw
Number	Name	training	test	(200)
1	Corn-notill	50	1384	57.5
2	Corn-mintill	50	784	58.6
3	Corn	50	184	85.8
4	Grass-pasture	50	447	85.6
5	Grass-trees	50	697	79.9
6	Hay-windrowed	50	439	94.7
7	Soybean-notill	50	918	78.5
8	Soybean-mintill	50	2418	58.8
9	Soybean-clean	50	564	62.9
10	Wheat	50	162	96.3
11	Woods	50	1244	88.5
12	Bldg-grass-tree-drives	50	330	57.5
13	Stone-steel-towers	50	45	93.3
14	Alfalfa	15	39	53.8
15	Grass-pasture-mowed	15	11	81.8
16	Oats	15	5	100
Kappa	–	–	–	0.6642
OA	–	–	–	70.24
AA	–	–	–	76.98

The number of features is given in the parentheses.

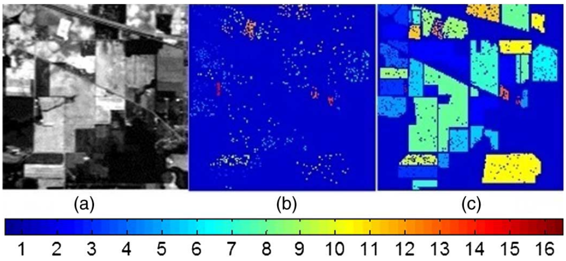


Fig. 4. AVIRIS Indian Pines data set: (a) spectral band number 27 ($\lambda = 646.72$ nm), (b) training samples, and (c) test samples, where each color represents a specific information class. The information classes are listed in Table II.

B. Experimental Setting

In experiments, the input image is transformed by a supervised FE and the first features with cumulative eigenvalues above 99% are retained, since they are expected to contain most of the variance in the original data sets. In the proposed framework, there is a second FE step which is conducted using the same criterion. For simplicity, the names of the different classifiers will be referred hereinafter as follows:

1) *Raw*: when the input data are classified by RF.
2) *Spec*: when only Spectral information resulting from the first supervised FE is classified by RF.
3) *AP*: when the selected features are used in order to produce the EMAP and classified by the RF.
4) n_m: when a supervised FE is performed for the second time and the output is classified by the RF. We have decided to use the name n_m in this case, where n FE approach means that the input data set first is transformed by n FE approach and the EMAP is then transformed by m FE approach. As an example, DB_DA means that the raw data were transformed by DBFE and the EMAP by DAFE.
5) ζ_{DA}: stacked vector consisting all features resulting from the first and second supervised FE. The suffix DA refers to the second FE technique.
6) ζ_{DB}: stacked vector consisting of all features resulting from the first and second supervised FE. The suffix DB refers to the second FE technique.

In the following, the number of features for Spec indicates the number of features with cumulative eigenvalues of more than 99% after performing DAFE or DBFE on the raw data. For example, it can be seen from Table III that six features are kept for Spec. It means that, first, the input data are transformed by DAFE and the first features with cumulative eigenvalues of more than 99% were kept (six features). These six features are used as a baseline for constructing the EMAP. Then, 14 thinning and 14 thickening are produced for the area attribute and 11 thinning and 11 thickening are produced for the standard deviation attribute. Therefore, each feature was used to produce 50 attributes and, by considering the feature itself in that vector, we have 51 features for each feature obtained by DAFE ($6 \times 51 = 306$ features for AP). Then, the second FE was performed and the first features with cumulative eigenvalues of more than 99% were kept. In this way, for Table III, DA_DA and DA_DB consist of 8 and 24 features, respectively. ζ_{DA} is the combination of Spec and DA_DA ($6 + 8 = 14$) and ζ_{DB} is the combination of Spec and DA_DB ($6 + 24 = 30$).

The way we calculate the CPU processing of each method is listed as follows:

1) *Spec*: CPU processing time of the first FE plus the CPU processing time of the corresponding RF classification.
2) *AP*: CPU processing time of the first FE plus the CPU processing time of producing EMAP plus the CPU processing time of the corresponding RF classification.
3) n_m: CPU processing time of the first FE plus the CPU processing time of producing EMAP plus the CPU processing time of the second FE plus the CPU processing time of the corresponding RF classification.
4) $\zeta_{\mathrm{DA\ or\ DB}}$: CPU processing time of the first FE plus the CPU processing time of producing EMAP plus the CPU

TABLE III
PAVIA UNIVERSITY: CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT METHODS FOR THE ROSIS PAVIA UNIVERSITY SCENE AFTER APPLYING DAFE

Class Number	Class Name	Spec (6)	AP (306)	DA_DA (8)	DA_DB (24)	ζ_{DA} (14)	ζ_{DB} (30)
1	Asphalt	82.9	98.0	98.2	97.1	**98.3**	97.0
2	Meadows	71.4	92.6	70.1	82.5	69.0	**96.5**
3	Gravel	69.5	81.0	**93.4**	89.7	91.6	86.4
4	Trees	92.2	97.8	**99.5**	99.3	**99.5**	**99.5**
5	Metal sheets	99.9	99.8	99.9	99.8	**100**	99.8
6	Soil	87.8	98.6	99.8	**99.9**	99.7	**99.9**
7	Bitumen	84.5	**100**	99.7	**100**	99.7	99.8
8	Bricks	85.1	96.1	**99.4**	99.3	**99.4**	**99.4**
9	Shadows	**97.7**	94.5	92.5	91.0	92.4	91.5
Kappa	–	0.7426	0.9317	0.8258	0.8862	0.8187	**0.9619**
OA	–	79.63	94.77	86.13	91.13	85.54	**97.11**
AA	–	85.71	95.41	94.75	95.42	94.43	**96.68**

The number of features used for classification purposes is reported in the parentheses.

TABLE IV
PAVIA UNIVERSITY: CPU PROCESSING TIME (IN SECONDS) OF DIFFERENT METHODS AFTER APPLYING DAFE

Spec	AP	DA_DA	DA_DB	ζ_{DA}	ζ_{DB}
7	131	51	161	48	165

processing time of the second FE plus the CPU processing time of corresponding the RF classification.

The following measures are used in order to evaluate the performance of different classification methods.

1) *Average accuracy (AA):* this metric shows the average value of the class classification accuracy.
2) *Overall accuracy (OA):* this metric refers to the number of samples, which are classified correctly divided by the number of test samples.
3) *Kappa coefficient:* this metric provides information regarding the agreement corrected by the level of agreement that could be expected due to chance alone.
4) *CPU processing time:* this metric shows the speed of different algorithms. It should be noted that, since in all algorithms (except Spectral), EMAP is carried out, the CPU processing time of this step is discarded from all methods. Hence, the CPU processing time is only provided for AP, $\mathrm{AP} + \mathrm{spectral}$, DAFE, and NWFE. All methods were implemented in MATLAB on a computer having Intel(R) Pentium(R) 4 CPU 3.20 GHz and 4 GB of memory.

C. Experimental Results

1) Pavia University: Table III gives information related to the classification accuracies of different methods after applying DAFE, with the corresponding CPU processing times listed in Table IV. As it can be observed from Tables I and III, the Spectral classification with only six features improves the OA of the Raw, data with 103 bands by 8%. Also, the class accuracies of Meadows and Gravel classes can be improved. Specifically, many samples of Meadows are misclassified as belonging to soil. Moreover, many samples of Gravel are misclassified as belonging to Asphalts and Bricks.

As can be seen from Table III, ζ_{DB} (consisting of 30 features) outperforms other methods significantly. ζ_{DB} improves the OA of Spectral, AP, $\mathrm{DA\_DA}$, $\mathrm{DA\_DB}$, and ζ_{DA} by 18%, 2.5%, 11, 6%, and 12%, respectively.

As it was already observed for the AVIRIS Indian Pines data set, AP achieves the best OA after ζ_{DB} since AP can model spatial dependencies of different objects by considering an adaptive neighborhood system. As can be seen from Table V, the OA of Spectral with 29 features improves the OA of Raw (with 103 bands) by 8%. Another observation is that AP, $\mathrm{DB\_DA}$, ζ_{DA}, and ζ_{DB} provide good performance. However, $\mathrm{DB\_DA}$ with only seven features provides the best results in terms of classification accuracies and CPU processing time.

By comparing the results reported in Tables III and V, it is easy to infer that DBFE works better than DAFE. The main reason behind this may be closely related to the fact that DAFE is not full rank (its rank is at most equal to $L - 1$ where L is the number of classes). Sometimes, the aforementioned number of features is not enough in order to discriminate between different classes of interest. However, DAFE is faster than DBFE. This fact can also be observed in Tables IV and VI.

Based on our experimental results, the proposed framework improves all methods in terms of classification accuracies for Pavia University data set. For example, the proposed method improves the classification accuracy of the classification technique proposed in [39] by almost 11%. Based on the results reported in [41], the proposed method improves the OA of the previous method with PCA by almost 21% and ICA by 3.5%. These are quite important achievements from the viewpoint of classification accuracy (in this regard, our framework provides some of the best classification results ever reported in the literature for the considered scene). The main disadvantage of the proposed method is the fact that the final result is dependent on the second FE and it is difficult to anticipate which one of ζ_{DA} or ζ_{DB} works better. The investigation of these aspects will be a subject for our future research efforts.

TABLE V
PAVIA UNIVERSITY: CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT METHODS FOR THE ROSIS PAVIA UNIVERSITY SCENE AFTER APPLYING DBFE

Class		Spec	AP	DB_DA	DB_DB	ζ_{DA}	ζ_{DB}
Number	Name	(29)	(1479)	(7)	(30)	(36)	(59)
1	Asphalt	85.0	**98.1**	97.8	97.3	97.6	96.7
2	Meadows	68.0	94.4	**97.4**	88.8	96.9	95.8
3	Gravel	69.4	98.0	97.2	74.7	**98.3**	87.0
4	Trees	95.2	87.3	98.3	95.2	98.5	**99.3**
5	Metal sheets	99.8	99.6	99.5	**99.9**	**99.9**	99.8
6	Soil	93.6	**100**	99.9	99.4	**100**	99.9
7	Bitumen	86.1	**100**	99.6	99.7	99.5	99.9
8	Bricks	87.3	98.1	99.2	98.9	99.3	**99.4**
9	Shadows	**97.5**	97.1	94.0	95.9	95.7	91.8
Kappa	–	0.7441	0.9481	**0.9746**	0.9078	0.9732	0.9576
OA	–	79.56	96.04	**98.07**	92.91	97.97	96.78
AA	–	86.91	96.98	98.15	94.47	**98.46**	96.67

The number of features used for classification purposes is reported in the parentheses.

TABLE VI
PAVIA UNIVERSITY: CPU PROCESSING TIME (IN SECONDS) OF DIFFERENT METHODS AFTER APPLYING DBFE

Spec	AP	DB_DA	DB_DB	ζ_{DA}	ζ_{DB}
36	478	201	803	833	827

TABLE VII
INDIAN PINES: CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT METHODS FOR THE AVIRIS INDIAN PINES SCENE AFTER APPLYING DAFE

Class		Spec	AP	DA_DA	DA_DB	ζ_{DA}	ζ_{DB}
Number	Name	(13)	(663)	(13)	(45)	(26)	(58)
1	Corn-notil	54.3	82.7	82.8	76.1	**88.5**	76.2
2	Corn-mintill	52.8	**96.0**	95.5	90.0	95.1	88.2
3	Corn	67.9	92.9	**99.4**	96.7	98.9	95.1
4	Grass-pasture	89.9	93.7	**95.7**	94.4	94.6	94.6
5	Grass-trees	87.9	96.1	**97.5**	95.9	97.1	95.4
6	Hay-windrowed	97.0	**99.7**	98.6	99.3	98.6	99.3
7	Soybean-notill	63.8	**91.6**	81.0	87.0	86.6	83.5
8	Soybean-mintill	44.8	85.1	76.7	73.4	**91.3**	73.0
9	Soybean-clean	64.3	87.7	**89.8**	89.3	89.7	86.7
10	Wheat	98.1	99.3	99.3	**100**	99.3	**100**
11	Woods	85.2	99.3	**99.9**	91.8	99.4	93.8
12	Bldg-grass-tree-drives	80.9	99.0	**99.3**	98.7	**99.3**	97.2
13	Stone-steel-towers	93.3	**100**	**100**	**100**	**100**	**100**
14	Alfalfa	56.4	**97.4**	94.8	**97.4**	94.8	**97.4**
15	Grass-pasture-mowed	**100**	**100**	**100**	90.9	**100**	**100**
16	Oats	80.0	**100**	**100**	**100**	**100**	**100**
Kappa	–	0.6118	0.8987	0.8683	0.8359	**0.9227**	0.8295
OA	–	65.47	91.13	88.47	85.53	**93.27**	84.99
AA	–	76.07	95.07	94.44	92.59	**95.86**	92.56

The number of features used for classification purposes is reported in the parentheses.

TABLE VIII
INDIAN PINES: CPU PROCESSING TIME (IN SECONDS) OF DIFFERENT METHODS AFTER APPLYING DAFE

Spec	AP	DA_DA	DA_DB	ζ_{DA}	ζ_{DB}
2	17	13	64	13	65

Fig. 5 shows classification maps for different methods started by DAFE applied on Pavia University.

2) Indian Pines: The low spatial resolution of this data set adds more complexity, since it leads to the presence of highly mixed pixels (which are mainly due to the early growth cycle of most of the agricultural features in the scene). In this case, the classification results may be degraded by the presence of mixed pixels in the scene. In addition, the significant

TABLE IX
INDIAN PINES: CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT METHODS FOR THE AVIRIS INDIAN PINES SCENE AFTER APPLYING DBFE

	Class	Spec	AP	DB_DA	DB_DB	ζ_{DA}	ζ_{DB}
Number	Name	(16)	(816)	(13)	(43)	(29)	(59)
1	Corn-notil	51.4	79.9	71.6	75.2	**86.2**	73.7
2	Corn-mintill	56.8	96.4	96.5	90.0	**96.6**	90.4
3	Corn	72.8	88.5	**94.0**	**94.0**	**94.0**	**94.0**
4	Grass-pasture	85.2	93.5	**95.5**	93.7	95.0	93.2
5	Grass-trees	87.9	**99.0**	95.9	98.4	94.8	98.5
6	Hay-windrowed	94.3	**99.3**	99.0	99.0	99.0	99.0
7	Soybean-notill	63.6	**87.2**	78.1	80.0	86.8	77.8
8	Soybean-mintill	46.9	82.0	75.2	73.8	**85.9**	70.3
9	Soybean-clean	65.4	84.9	**87.2**	78.1	84.9	77.6
10	Wheat	99.3	**100**	98.7	**100**	**100**	**100**
11	Woods	81.1	**99.6**	99.5	94.6	99.4	93.1
12	Bldg-grass-tree-drives	75.4	98.7	**99.7**	99.3	**99.7**	98.4
13	Stone-steel-towers	91.1	**100**	**100**	**100**	**100**	**100**
14	Alfalfa	69.2	**97.4**	**97.4**	**97.4**	**97.4**	**97.4**
15	Grass-pasture-mowed	81.8	**100**	**100**	**100**	**100**	**100**
16	Oats	40.0	**100**	**100**	80.0	**100**	80.0
Kappa	–	0.6058	0.8808	0.8388	0.8254	**0.9001**	0.8083
OA	–	64.99	89.56	85.90	84.64	**91.27**	83.11
AA	–	72.67	94.18	93.05	90.88	**95.01**	90.25

The number of features used for classification purposes is reported in the parentheses.

TABLE X
INDIAN PINES: CPU PROCESSING TIME (IN SECONDS) OF DIFFERENT METHODS AFTER APPLYING DBFE

Spec	AP	DB_DA	DB_DB	ζ_{DA}	ζ_{DB}
26	45	44	132	45	133

differences in the number of pixels in the reference data for different classes make the classification task even more complicated.

In these data, there is a high confusion between classes Soybean-mintill and corn-notill which degrades the class accuracies of both of them. By comparing Tables II and VII, it is easy to infer that, by performing DAFE on the input data and choosing the first features with cumulative eigenvalues above 99%, OA is reduced from 70.24% (Raw) to 65.47% (Spectral). This reveals that of only 13 features are not sufficient to discriminate between different classes, as compared to hundreds of spectral bands from the input data.

As it can be seen from Table VII, AP improves the overall accuracy of Raw in more than 25%. The main reason behind this significant improvement is that AP not only considers the spectral information, but also can model the spatial information contained in the input data. The best classification accuracies in Table VII are achieved by the proposed method; ζ_{DA}, which improves the overall accuracy of Spectral, AP, DA_DA, DA_DB, and ζ_{DB} by almost 28, 2, 5%, 8%, and 9%, respectively. It should be noted that the new method can discriminate different classes by considering only 26 features. Moreover, the CPU processing time of the proposed method is acceptable and takes only 13 s to classify the input data set in the considered computing environment. Table VIII gives information regarding the CPU processing time of different methods after applying DAFE.

After ζ_{DA}, AP exhibits the best performance among other techniques in terms of classification accuracies. This confirms that the consideration of spatial information has a significant influence on the discrimination of different classes. By including a second FE step, although classification accuracy for some classes such as classes 3 and 4 are improved, the overall accuracy of AP is reduced from 91.13% to 88.47% (DA_DA) and 85.53% (DA_DB).

Table IX gives information related to the classification accuracies of different methods after DBFE. The corresponding CPU processing times are listed in Table X. By comparing Tables II and IX, one can infer that the OA of the Raw classification decreases when DBFE is performed. Again, the proposed method outperforms other techniques with acceptable CPU processing time (45 s) in this particular case.

In contrast, it is also important to emphasize AP exhibits an acceptable performance in terms of classification accuracies when compared to other classifiers (its performance is only slightly lower than ζ_{DA}). AP provides 720 features. This reveals that RF is a robust classifier when dealing with very high-dimensional data. Also, it is worth mentioning that ζ_{DA} provides the best performance overall, and improves the OA of Spectral, AP, DB_DA, DB_DB, and ζ_{DB} by more than 26%, 1.7%, 6.2%, 6.5%, and 7.9%, respectively.

As can be seen from Tables VII and IX, AP provides 585 and 720 features, respectively. The table shows that RF can properly handle classification problems consisting of high-dimensional input features and limited training samples, with acceptable CPU processing time. In almost all cases, DAFE outperforms DBFE in terms of classification accuracies and CPU processing time. A possible reason for this may be

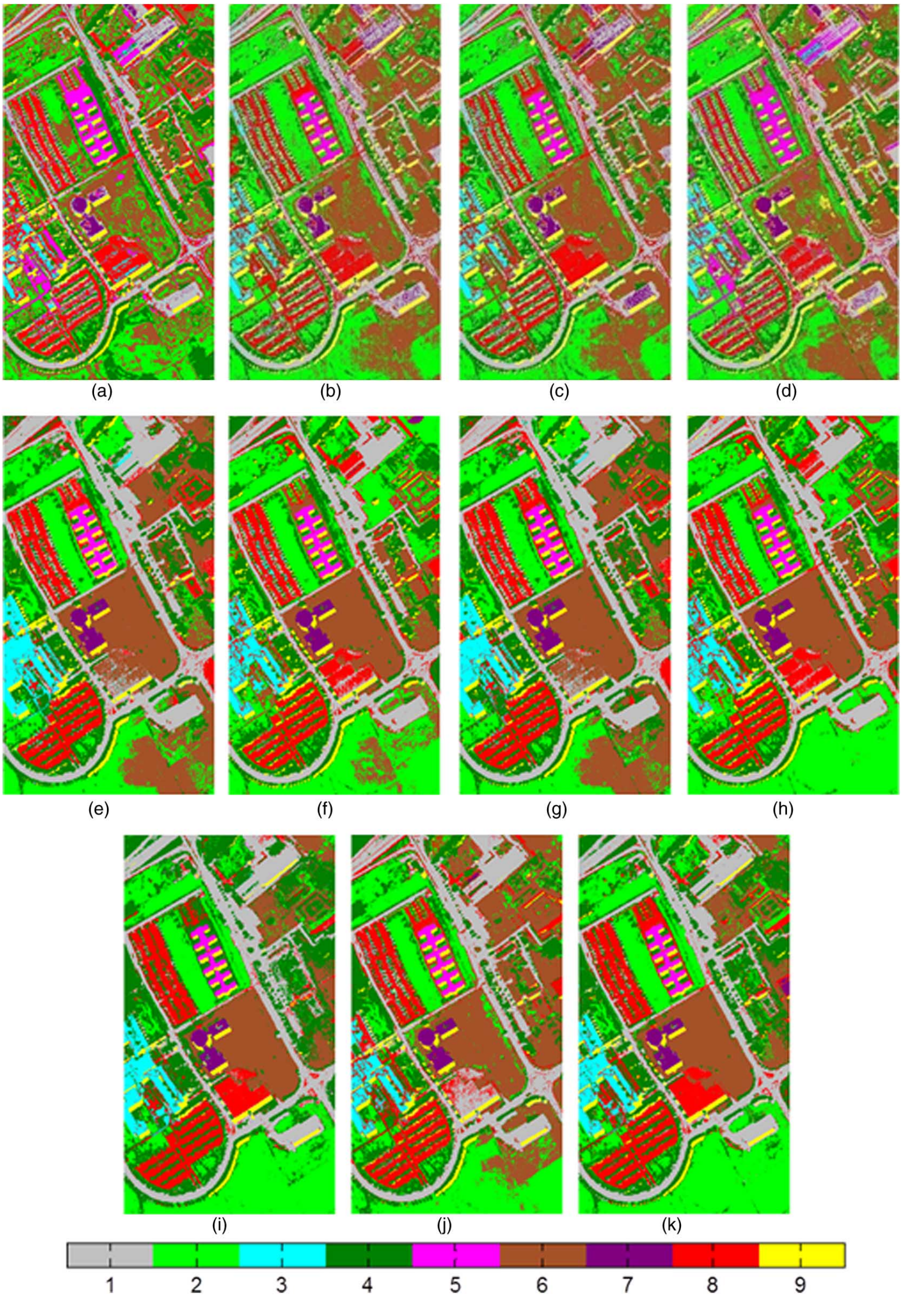


Fig. 5. Pavia University: (a)–(f), classification maps for different methods started by DAFE: (a) Raw, (b) AP, (c) DA_DA, (d) DA_DB, (e) ζ_{DA}, and (f) ζ_{DB}. (g)–(k), classification maps of different methods started by DBFE: (g) AP, (h) DB_DA, (i) DB_DB, (j) ζ_{DBA}, and (k) ζ_{DB}.

the fact that the number of selected features used by DBFE is not sufficient. As a result, more features need to be considered in order to provide more consistent results in the case of DBFE, which can be computationally intensive and its performance is highly dependent on the training samples.

Fig. 6 shows classification maps for different methods started by DAFE applied on Indian Pines.

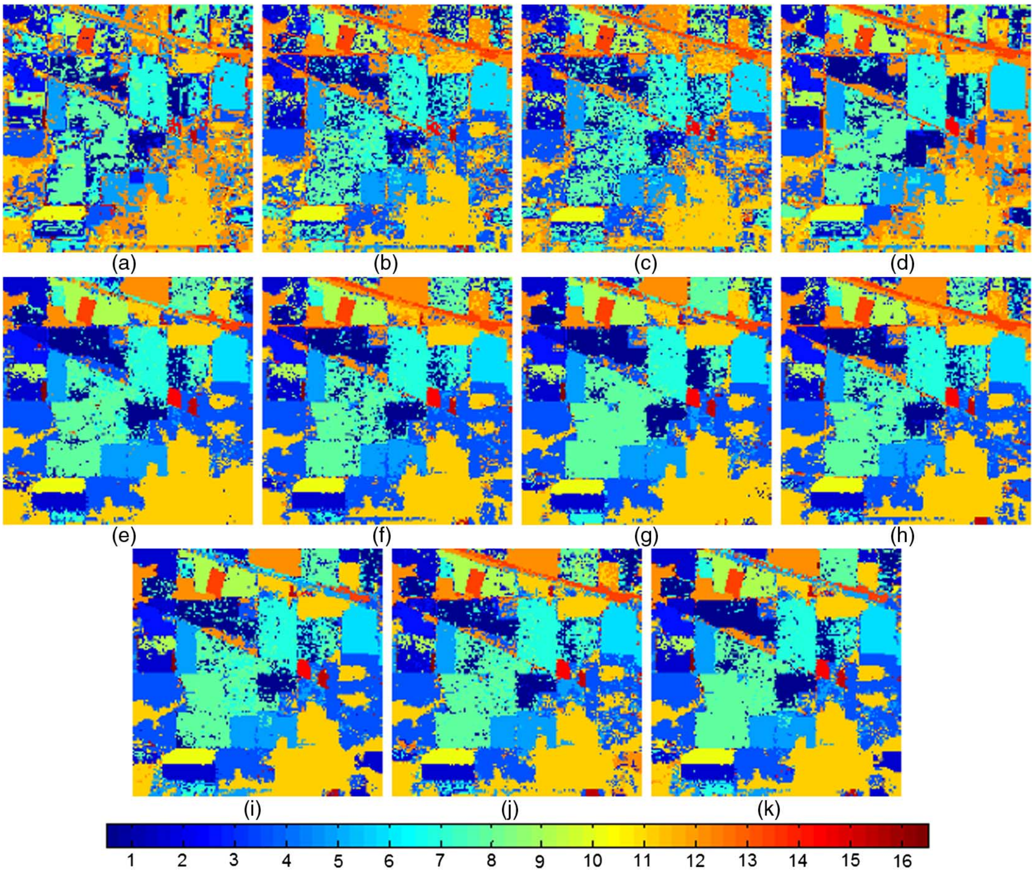


Fig. 6. AVIRIS (a)–(f), classification maps for different methods after applying DAFE: (a) Raw, (b) AP, (c) DA_DA, (d) DA_DB, (e) ζ_{DA}, and (f) ζ_{DB}. (g)–(k), classification maps of different methods after applying DBFE: (g) AP, (h) DB_DA, (i) DB_DB, (j) ζ_{DBA}, and (k) ζ_{DB}.

IV. Conclusion

In this paper, we have developed a new automatic framework for the classification of hyperspectral images. Our framework uses both spectral and spatial information. In order to include the spatial information, morphological APs are taken into account. For reducing the redundancy of the extracted features and deal with the curse of dimensionality introduced by the Hughes effect, supervised FE methods (DAFE and DBFE) are considered. The proposed framework is extensively tested on two widely used hyperspectral data sets, i.e., the ROSIS-03 Pavia University scene and the AVIRIS Indian Pines. Different methods have been used to implement the presented framework, and the results provided have been compared in terms of classification accuracies and CPU processing time.

It should be noted that the two selected hyperspectral data sets represent very different case studies collected by different instruments. The former is related to urban area problems and presents high spatial resolution. In turn, the latter has medium-size spatial resolution and is related to agricultural land-cover classification problems. The good classification accuracies obtained in both case studies indicate the good generalization properties of the presented framework. In addition, the new approach achieves better classification accuracies than other widely used classification techniques, with acceptable CPU processing time. We emphasize that the proposed procedure is fully automatic, which is a highly desirable feature.

A topic of future investigation is the optimal selection (in terms of classification accuracies) of the FE method in the second stage of the proposed approach. Another topic deserving future research is the development of parallel implementations of the presented approach in high-performance computing architectures, although the processing times reported in our experiments (measured in a standard desktop CPU) are quite fast for the considered data sets.

Acknowledgment

The ROSIS data and corresponding reference information were kindly provided by Prof. P. Gamba from the University of Pavia, Italy.

References

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, no. 1, pp. S110–S122, 2009.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

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

[4] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson. "Classification of hyperspectral images with binary fractional order Darwinian PSO and random forests," in *Proc. SPIE 8892, Image Signal Process. Remote Sens. XIX,* Oct. 17, 2013, p. 88920S, doi: 10.1117/12.2027641.

[5] S. Tadjudin, "Classification of high dimensional data with limited training samples," Ph.D. thesis, School of Electrical and Computer Engineering, Purdue Univ., West Lafayette, IN, 1998, 123pp. [Online]. Available: http://dynamo.ecn.purdue.edu/~landgreb/publications.html.

[6] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[7] H. Derin and P. A. Kelly, "Discrete-index Markov-type random processes," *Proc. IEEE*, vol. 77, no. 10, pp. 1485–1510, Oct. 1989.

[8] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," in *Proc. Pattern Recogn. Mach. Intell.,* 2005, pp. 260–265.

[9] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral–spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. PP, no. 99, pp. 1–10, Jun. 2013 [Online]. Available: http://dx.doi.org/10.1109/TGRS.2013.2263282.

[10] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, "Use of Darwinian particle swarm optimization technique for the segmentation of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.,* 2012, pp. 4295–4298.

[11] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, "An efficient method for segmentation of images based on fractional calculus and natural selection," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12 407–12 417, 2012.

[12] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.

[13] P. Ghamisi, M. S. Couceiro, F. M. L. Martins, and J. A. Benediktsson, Multilevel Image Segmentation Based on Fractional-Order Darwinian Particle Swarm Optimization, *IEEE Trans. Geosci. Remote Sens.*, vol. PP, no. 99, pp. 1–13, Jun. 2013, doi: 10.1109/TGRS.2013.2260552 [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6524014&isnumber=4358825.

[14] P. Ghamisi, M. S. Couceiro, M. Fauvel, and J. A. Benediktsson, "Integration of segmentation techniques for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 342–346, Jan. 2014.

[15] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[16] J. A. Palmason, J. A. Benediktsson, J. R. Sveinsson, and J. Chanussot, "Classification of hyperspectral data from urban areas using morphological preprocessing and independent component analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, no. 3, 2005, pp. 176–179.

[17] M. Dalla Mura, J. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[18] M. Pedergnana, P. Marpu, M. Dalla Mura, J. Benediktsson, and L. Bruzzone, "A novel technique for optimal feature selection in attribute profiles based on genetic algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 6, pp. 3514–3528, Jun. 2013.

[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, p. 532, 2001.

[20] L. Breiman, "RF tools a class of two eyed algorithms," in *Proc. SIAM Workshop*, San Francisco, CA, USA, 2003.

[21] P. 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," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 293–297, Mar. 2013.

[22] C. Lee and D. A. Landgrebe, "Decision boundary feature extraction for nonparametric classification," *IEEE Trans. Syst. Man Cybern.*, vol. 23, no. 2, pp. 433–444, Mar./Apr. 1993.

[23] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, NY, USA: Academic Press, 1974.

[24] L. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, Nov. 1999.

[25] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 388–400, Apr. 1993.

[26] J. Serra, *Mathematical Morphology, Theoretical Advances*, vol. 2, New York, NY, USA: Academic Press, 1988.

[27] J. Serra, *Image Analysis and Mathematical Morphology.* London, U.K.: Academic Press, 1982.

[28] P. Soille, "Morphological Image Analysis, Principles and Applications," vol. 2, Berlin, Germany: Springer Verlag, 2003.

[29] L. Najman and H. Talbot, *Mathematical Morphology*, Hoboken, NJ, USA: Wiley, 2010.

[30] E. J. Breen and R. Jones, "Attribute openings, thinnings, and granulometries," *Comput. Vision Image Understand.*, vol. 64, no. 3, pp. 377–389, 1996.

[31] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute filters for the analysis of very high resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.* (IGARSS 2009), vol. 3, 2009, pp. III-97–III-100.

[32] P. Salembier, A. Oliveras, and L. Garrido, "Antiextensive connected operators for image and sequence processing," *IEEE Trans. Image Process.*, vol. 7, no. 4, pp. 555–570, Apr. 1998.

[33] V. Caselles and P. Monasse, *Geometric Description of Images as Topographic Maps*, 1st ed. New York, NY, USA: Springer Publishing Company, Incorporated, 2009.

[34] E. R. Urbach, J. B.T.M. Roerdink, and M. H. F. Wilkinson, "Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 272–285, Feb. 2007.

[35] X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 257–272, Jan. 2013.

[36] M. Dalla Mura, "Advanced techniques based on mathematical morphology for the analysis of remote sensing images," Ph.D. dissertation, Univ. Trento, Trento, Italy, 2011.

[37] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis*, 4th ed. Berlin, Germany: Springer Heidelberg, 2006.

[38] P. Ghamisi, J. A. Benediktsson, and J. R. Sveinsson, "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction," *IEEE Trans. Remote Sens. Geosci.*, vol. PP, no. 99, pp. 1–12, Dec. 2013.

[39] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[40] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[41] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

Pedram Ghamisi (S'13) graduated with the B.Sc. degree in civil (survey) engineering from the Tehran South Campus, Azad University, Tehran, Iran. Then, he recieved the M.Sc. degree in remote sensing from K.N. Toosi University of Technology, Tehran, Iran, in 2012. He is currently a Ph.D. student in electrical and computer engineering at the University of Iceland, Reykjavík, Iceland.

His research interests are in remote sensing and image analysis with the current focus on spectral and spatial techniques for hyperspectral image classification.

Mr. Ghamisi received the Best Researcher Award for M.Sc. students in K. N. Toosi University of Technology in the academic year 2010–2011. He was the recipient of the IEEE Mikio Takagi Prize, which was awarded for the first place in the Student Paper Competition at the 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Melbourne, Australia, July 2013. He serves as a Reviewer for a number of journals including IEEE TGRS, IEEE Trans. Image Processing, IEEE JSTARS and IEEE GRSL.

Jón Atli Benediktsson (S'84–M'90–SM'99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984, and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

Currently, he is a Pro Rector for Academic Affairs and Professor of electrical and computer engineering with the University of Iceland. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in those fields.

From 2011 to 2012, he was a President of the IEEE Geoscience and Remote Sensing Society (GRSS) and has been on the GRSS AdCom since 2000. From 2003 to 2008, he was an Editor of the IEEE Transactions on Geoscience and Remote Sensing (TGRS) and has served as an Associate Editor of TGRS since 1999, the IEEE Geoscience and Remote Sensing Letters since 2003, and IEEE Access since 2013. From 2007 to 2010, he is in the International Editorial Board of the *International Journal of Image and Data Fusion* and was the Chairman of the Steering Committee of IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS). He is a Co-founder of the biomedical start up company Oxymap (www.oxymap.com). He is a Fellow of SPIE.

Dr. Benediktsson received the Stevan J. Kristof Award from Purdue University, in 1991 for outstanding graduate student in remote sensing. In 1997, he was the recipient of the Icelandic Research Council's Outstanding Young Researcher Award; in 2000, he was granted the IEEE Third Millennium Medal; in 2004, he was a co-recipient of the University of Iceland's Technology Innovation Award; in 2006, he received the yearly research award from the Engineering Research Institute of the University of Iceland; and in 2007, he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society. He is co-recipient of the 2012 IEEE Transactions on Geoscience and Remote Sensing Paper Award. He received the 2013 IEEE/VFI Electrical Engineer of the Year Award and in 2013, he was a co-recipient of the IEEE GRSS Highest Impact Paper Award. He is a Member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica and Tau Beta Pi.

Gabriele Cavallaro received the B.S. and M.S. degrees in telecommunications engineering from the University of Trento, Trento, Italy, in 2011 and 2013, respectively. He did the Master thesis at the University of Iceland, Reykjavik, Iceland, in remote sensing field on morphological attribute filters based on the Inclusion Tree for the analysis of very high-resolution remote sensing images. At present, he is a Ph.D. student at the University of Iceland and at the University of Extremadura, Cáceres, Spain.

Antonio Plaza (M'05–SM'07) received the M.S. and Ph.D. degrees in computer engineering from the University of Extremadura, Cáceres, Spain.

He is a Associate Professor (with accreditation for Full Professor) with the Department of Technology of Computers and Communications, University of Extremadura, where he is the Head of the Hyperspectral Computing Laboratory (HyperComp). From 2007 to 2011, he was the Coordinator of the Hyperspectral Imaging Network, a European project with total funding of 2.8 MEuro. He authored more than 370 publications, including more than 100 JCR journal papers (60 in IEEE journals), 20 book chapters, and over 230 peer-reviewed conference proceeding papers (90 in IEEE conferences). He has guest edited seven special issues on JCR journals (three in IEEE journals). He has been a Chair for the IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (2011).

Dr. Plaza is a recipient of the recognition of Best Reviewers of the IEEE Geoscience And Remote Sensing Letters, in 2009, and a recipient of the recognition of Best Reviewers of the IEEE Transactions on Geoscience and Remote Sensing, in 2010 and a journal for which he has served as Associate Editor in 2007–2012. From 2011 to 2012, he was a Member of the Editorial Board of the IEEE Geoscience and Remote Sensing Newsletter and a member of the steering committee of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing in 2012. He is also an Associate Editor for the IEEE Geoscience and Remote Sensing Magazine. From 2011 to 2012, he served as the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) and is currently serving as a President of the Spanish Chapter of IEEE GRSS (since November 2012). He is currently serving as the Editor-in-Chief of the IEEE Transactions on Geoscience and Remote Sensing journal, since January 2013. Additional information: http://www.umbc.edu/rssipl/people/aplaza.

CHAPTER 5

Supervised Feature Reduction

A Novel Feature Selection Approach Based on FODPSO and SVM

Pedram Ghamisi, *Student Member, IEEE*, Micael S. Couceiro, *Member, IEEE*, and Jon Atli Benediktsson, *Fellow, IEEE*

***Abstract*—A novel feature selection approach is proposed to address the curse of dimensionality and reduce the redundancy of hyperspectral data. The proposed approach is based on a new binary optimization method inspired by *fractional-order Darwinian particle swarm optimization* (FODPSO). The overall accuracy (OA) of a *support vector machine* (SVM) classifier on validation samples is used as fitness values in order to evaluate the informativity of different groups of bands. In order to show the capability of the proposed method, two different applications are considered. In the first application, the proposed feature selection approach is directly carried out on the input hyperspectral data. The most informative bands selected from this step are classified by the SVM. In the second application, the main shortcoming of using attribute profiles (APs) for spectral–spatial classification is addressed. In this case, a stacked vector of the input data and an AP with all widely used attributes are created. Then, the proposed feature selection approach automatically chooses the most informative features from the stacked vector. Experimental results successfully confirm that the proposed feature selection technique works better in terms of classification accuracies and CPU processing time than other studied methods without requiring the number of desired features to be set *a priori* by users.**

***Index Terms*—Attribute profile (AP), automatic classification, feature extraction, hyperspectral image analysis, random forest (RF) classifier, spectral–spatial classification.**

I. Introduction

HYPERSPECTRAL remote sensors acquire a massive amount of data by obtaining many measurements, not knowing which data are relevant for a given problem. The trend for hyperspectral imagery is to record hundreds of channels from the same scene. The obtained data can characterize the chemical composition of different materials and potentially be helpful in analyzing different objects of interest.

In the spectral domain, each spectral channel is considered as one dimension, and each pixel is represented as a point in that domain. By increasing the number of spectral channels in the spectral domain, theoretical and practical problems may arise, and conventional techniques that are applied on multispectral data are no longer appropriate for the processing of high dimensional data [1]–[3].

The aforementioned characteristics show that conventional techniques based on the computation of fully dimensional space may not provide accurate classification results when the number of training samples is not substantial. For instance, while keeping the number of samples constant, after a few number of bands, the classification accuracy actually decreases as the number of features increases [1]. For the purpose of classification, these problems are related to the *curse of dimensionality* [4]. In order to tackle this issue and use a smaller number of training samples, the use of feature selection and extraction techniques would be of importance.

From one point of view, feature selection techniques can be split into two categories: unsupervised and supervised. Supervised feature selection techniques aim at finding the most informative features with respect to the available prior knowledge and lead to better identification and classification of different classes of interest. On the contrary, unsupervised methods are used in order to find distinctive bands when prior knowledge of the classes of interest is not available. Information entropy [5], first spectral derivative [6], and uniform spectral spacing [7] can be considered as unsupervised feature selection techniques, whereas supervised feature selection techniques usually try to find a group of bands achieving the largest class separability. Class separability can be calculated by considering several approaches such as divergence [8], transformed divergence [8], Bhattacharyya distance [9], and Jeffries–Matusita distance [8]. A comprehensive overview of different feature selection and extraction techniques is provided in [10]. However, these metrics usually suffer from the following shortcomings.

1) They are usually based on the estimation of the second-order statistics (e.g., covariance matrix), and in this case, they demand many training samples in order to estimate the statistics accurately. Therefore, in a situation when the number of training samples is limited, that may lead to the singularity of the covariance matrix. In addition, since the bands in hyperspectral data usually have some redundancy, the probability of singularity will even increase.
2) In order to select informative bands, corrupted bands (e.g., water absorption bands and bands with a low signal-to-noise ratio) are usually preremoved, which is a time-consuming task. Furthermore, conventional feature selection methods can be computationally demanding. To select an m feature subset out of a total of n features, $n!/(n-m)!m!$ operations must be calculated, which is

Manuscript received March 24, 2014; revised July 27, 2014 and October 13, 2014; accepted October 31, 2014. This work was supported in part by the Icelandic Research Fund for Graduate Students.

P. Ghamisi and J. A. Benediktsson are with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland.

M. S. Couceiro is with Ingeniarius, Lda., 3050-381 Mealhada, Portugal, and also with the Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2014.2367010

a laborious task and demands a significant amount of computational memory. In other words, the conventional feature selection techniques are only feasible in relatively low dimensional cases.

In order to address the aforementioned shortcomings of the conventional feature selection techniques, the use of stochastic and bioinspired optimization-based feature selection techniques, such as genetic algorithms (GAs) and particle swarm optimization (PSO), is considered attractive. The main reasons behind this trend are that 1) in evolutionary feature selection techniques, there is no need to calculate all possible alternatives in order to find the most informative bands, and furthermore, 2) in the evolutionary approaches, usually a metric is chosen as the fitness function, which is not based on the calculation of the second-order statistics, and therefore, the singularity of the covariance matrix is not a problem. In the literature, there is an extensive number of works related to the use of evolutionary optimization-based feature selection techniques. These methods are mostly based on the use of GA and PSO. For example, in [11], Bazi and Melgani proposed an SVM classification system that allows the detection of the most distinctive features and the estimation of the SVM parameters (e.g., regularization and kernel parameters) by using a GA. In [12], Daamouche *et al.* proposed to use PSO in order to select the most informative features obtained by morphological profiles for classification. In [13], in order to address the main shortcomings of GA- and PSO-based feature selection techniques and to take the advantage of their strength, a new feature selection approach is proposed, which is based on the hybridization of GA and PSO. In [14], a method was introduced, which allows to simultaneously solve problems of clustering, feature detection, and class number estimation in an unsupervised way. However, this method suffers from the computational time required by the optimization process.

In GA, if a chromosome is not selected for mating, the information contained by that individual is lost, since the algorithm does not have a memory of its previous behaviors. Furthermore, PSO suffers from the premature convergence of a swarm, because 1) particles try to converge to a single point, which is located on a line between the global best and the personal best positions, which this point does not guarantee to be a local optimum [15], and 2) furthermore, the fast rate of information flow between particles can lead to the creation of similar particles. This results in a loss in diversity [16].

In this paper, a novel feature selection approach is proposed, which is based on a new binary optimization technique and the SVM classifier. The new approach is capable of handling very high dimensional data even when only a limited number of training samples is available (ill-posed situation) and when conventional techniques are not able to proceed. In addition, despite the conventional feature selection techniques for which the number of desired features needs to be initialized by the user, the proposed approach is able to automatically select the most informative features in terms of classification accuracy within an acceptable CPU processing time without requiring the number of desired features to be set *a priori* by users. The new feature selection technique is, at first, compared with the traditional PSO-based feature selection in terms of classification accuracy on validation samples and CPU processing time. Then, the new method is compared with a few well-known feature selection and extraction techniques. Furthermore, the new method will be taken into account in order to overcome the main shortcomings of using attribute profiles (APs).

The rest of this paper is organized as follows: First, the new feature selection approach is described in Section II. Then, Section III briefly describes SVM. Section IV is devoted to the methodology of the proposed approach. Section V is on experimental results, and main concluding remarks are furnished in Section VI.

II. Fractional-Order Darwinian Particle Swarm Optimization (FODPSO)-Based Feature Selection

In brief, the goal is to overcome the curse of dimensionality [4] by selecting the optimal l_E bands for the classification, i.e., $l_E \leq l$, wherein l is the total number of bands in a given image I. Selecting the most adequate bands is a complex task as the classification overall accuracy (OA) first grows and then declines as the number of spectral bands increases [4]. Hence, this paper tries to find the optimal l_E bands that maximize the OA obtained as

$$\mathrm{OA} = \frac{\sum_{i}^{N_c} C_{ii}}{\sum_{ij}^{N_c} C_{ij}} \times 100 \tag{1}$$

wherein C_{ij} is the number of pixels assigned to class j, which belongs to class i. C_{ii} denotes the number of pixels correctly assigned to class i, and N_c is the number of classes.

In this paper, optimal features are selected through an optimization procedure in such a way that each solution gets its fitness value from the SVM classifier over validation samples. The optimization procedure is handled with PSO algorithms.

In 1995, Eberhart and Kennedy proposed the PSO algorithm for the first time [17]. The stochastic optimization ability of the algorithm is enhanced due to its cooperative simplistic mechanism, wherein each particle presents itself as a possible solution of the problem, e.g., the best l_E bands. These particles travel through the search space to find an optimal solution, by interacting and sharing information with other particles, namely, their individual best solution (personal best), and computing the global best [18].

The success of this algorithm has given rise to a chain of PSO-based alternatives in recent years, so as to overcome its drawbacks, namely, the stagnation of particles around suboptimal solutions. One of the proposed methods was denoted as Darwinian PSO (DPSO) [19]. The idea is to run many simultaneous parallel PSO algorithms, each one as a different swarm, on the same test problem, and then, a simple natural selection mechanism is applied. When a search tends to a suboptimal solution, the search in that area is simply discarded, and another area is searched instead. In this approach, at each step, swarms that get better are rewarded (extend particle life or spawn a new descendent), and swarms that stagnate are punished (reduce swarm life or delete particles). For more information regarding how these rewards and punishments can be applied, please see [20]. DPSO has been investigated for the segmentation of remote sensing data in [21].

Despite the positive results obtained by Tillett *et al.* [19], this *coopetitive* approach also increases the computational complexity of the optimization method. As many swarms of cooperative test solutions (i.e., particles) simultaneously run in a competitive fashion, the computational requirements increase, and, as a consequence, the convergence time also increases. Therefore, and to further improve the DPSO algorithm, an extended version denoted as FODPSO was presented in [22], in which fractional calculus is used to control the convergence rate of the algorithm. This method has been further investigated for gray-scale and hyperspectral image segmentation in [20] and [23]. An important property revealed by fractional calculus is that, while an integer-order derivative just implies a finite series, the fractional-order derivative requires an infinite number of terms. In other words, integer derivatives are "local" operators, whereas fractional derivatives have, implicitly, a "memory" of all past events. The characteristics revealed by fractional calculus make this mathematical tool well suited to describe phenomena, such as the dynamic phenomena of particles' trajectories.

Therefore, supported on the FODPSO previously presented in [22], and based on the *Grunwald Letnikov* definition of fractional calculus, in each step t, the fitness function represented by (1) is used to evaluate the success of particles (i.e., OA). To model the swarm, each particle n moves in multidimensional space according to the position $(x_n[t])$, and velocity $(v_n[t])$, values that are highly dependent on local best $(\breve{x}_n[t])$ and global best $(\breve{g}[t])$ information, i.e.,

$$\begin{aligned} v_n^s[t+1] =& w_n^s[t+1] + \rho_1 r_1 \left(\breve{g}^s[t] - x_n^s[t]\right) \\ &+ \rho_2 r_2 \left(\breve{x}_n^s[t] - x_n^s[t]\right) \end{aligned} \tag{2}$$

$$\begin{aligned} w_n^s[t+1] =& \alpha v_n^s[t] + \frac{1}{2}\alpha(1-\alpha)v_n^s[t-1] \\ &+ \frac{1}{6}\alpha(1-\alpha)(2-\alpha)v_n^s[t-2] \\ &+ \frac{1}{24}\alpha(1-\alpha)(2-\alpha)(3-\alpha)v_n^s[t-3]. \end{aligned} \tag{3}$$

Since the proposed FODPSO-based feature selection approach is based on running many simultaneous swarms in parallel over the search space, s shows the number of each swarm.

The coefficients ρ_1 and ρ_2 are assigned weights, which control the inertial influence of "the globally best" and "the locally best", respectively, when the new velocity is determined. Typically, ρ_1 and ρ_2 are constant integer values, which represent "social" and "cognitive" components with $\rho_1 + \rho_2 < 2$ [24]. However, different results can be obtained by assigning different values for each component.

The fractional coefficient α will weigh the influence of past events on determining a new velocity, i.e., $0 < \alpha < 1$. With a small α, particles ignore their previous activities, thus ignoring the system dynamics and becoming susceptible to get stuck in local solutions (i.e., exploitation behavior). On the other hand, with a large α, particles will have a more diversified behavior, which allows exploration of new solutions and improves the long-term performance (i.e., exploration behavior). However, if the exploration level is too high, then the algorithm may take longer to find the global solution. Based on [24], a good α value can be selected in the range of 0.6–0.8.

The parameters r_1 and r_2 are random vectors with each component generally a uniform random number between 0 and 1.

In order to investigate FODPSO for the purpose of feature selection, the dimension of each particle should be equal to the number of features. This way, the velocity dimension $(\dim v_n[t])$ and the position dimension $(\dim x_n[t])$ correspond to the total number of bands of the image, i.e., $\dim v_n[t] = \dim x_n[t] = l$. In other words, each particle's velocity will be represented as an l-dimensional vector. In addition, as one wishes to use the algorithm for band selection, each particle represents its position in binary values, i.e., 0 or 1, where 0 demonstrates the absence of the corresponding feature, and 1 has a dual meaning. In this case, as proposed by Khanesar *et al.* [25], the velocity of a particle can be associated to the probability of changing its state as

$$\Delta x_n^s[t+1] = \frac{1}{1 + e^{-v_n^s[t+1]}}. \tag{4}$$

Nevertheless, as one wishes to use the algorithm for band selection, each particle represents its position in binary values, i.e., 0 or 1. This may be represented as

$$x_n^s[t+1] = \begin{cases} 1, & \Delta x_n^s[t+1] \geq r_x \\ 0, & \Delta x_n^s[t+1] < r_x \end{cases} \tag{5}$$

wherein r_x is a random l-dimensional vector with each component generally a uniform random number between 0 and 1. Therefore, each particle moves in multidimensional space according to its position $x_n^s[t]$ from the discrete-time system represented by (2)–(5). In other words, each particle's position will be represented as an l-dimensional binary vector.

To make it easier to understand the proposed strategy, an example is given in Fig. 1. As the figure shows, the image has only five bands, i.e., $l = 5$. This means that each particle will be defined by its current velocity and position in 5-D space, i.e., $\dim v_n[t] = \dim x_n[t] = 5$. In this example, and to allow a straightforward understanding, only a swarm of two particles was considered. As it is possible to observe at time/iteration $t = 1$, particle 1 is positioned in such a way that it ignores the fourth band, i.e., $x_1[1] = [1\,1\,1\,0\,1]$, whereas particle 2 ignores the first and third bands, i.e., $x_2[1] = [0\,1\,0\,1\,1]$. Computing (1) under those conditions returns an OA of $\mathrm{OA}_1 = 60\%$ and $\mathrm{OA}_2 = 64\%$ for particles 1 and 2, respectively. Considering only those two particles, particle 2 is considered as the best performing one from the swarm, thus attracting particle 1 toward itself. Such attraction induces the velocity of particle 1 for iteration 2 and, consequently, its position.[1]

[1] The MATLAB code for the PSO- and FODPSO-based feature selection approaches will be provided on a request by sending an e-mail to the authors.

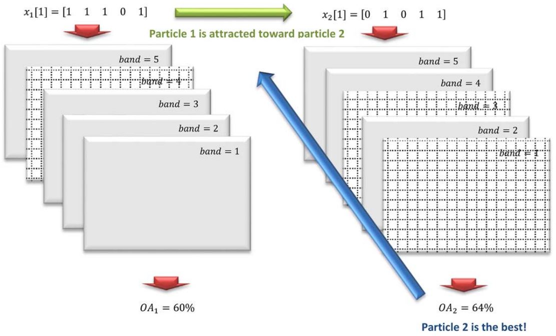


Fig. 1. Band coding of two particles for an image with five bands. Gridded bands are ignored in the classification process.

III. SVM

As discussed before, in the proposed method, the OA of SVM over validation samples is considered as the fitness value. SVM has attracted much attention due to its capability of handling the curse of dimensionality in comparison with conventional classification techniques. The main reasons behind the success of the approach are that 1) SVM is based on a margin maximization principle that helps avoid estimating the statistical distributions of different classes in hyperdimensional feature space and that 2) SVM takes advantage of the strong generalization capability obtained by its sparse representation of the decision function [11].

In hyperspectral image analysis, the random forest (RF) classifier and SVM play a key role since they can handle high-dimensional data even with a limited number of training samples. In this paper, we prefer to use SVM rather than RF due to its susceptibility to noise. Due to its sensibility, corrupted and noisy bands may significantly influence the classification accuracies. As a result, when RF is considered as the fitness function, due to its capability to handle different types of noises, corrupted bands cannot be eliminated even after a high number of iterations. On the contrary, since SVM is more sensible than RF against noise, it can detect and eliminate corrupted bands after a few iterations, which can be considered as a privilege for the final classification step.

The general idea behind SVM is to separate training samples belonging to different classes by tracing maximum margin hyperplanes in the space where the samples are mapped [26]. SVMs were originally introduced for solving linear classification problems. However, they can be generalized to nonlinear decision functions by considering the so-called *kernel trick* [27]. A kernel-based SVM is being used to project the pixel vectors into higher dimensional space and estimate maximum margin hyperplanes in this new space, in order to improve the linear separability of data [27]. The sensitivity to the choice of the kernel and regularization parameters can be considered as the most important disadvantages of SVM. The latter is classically overcome by considering cross-validation techniques using training data [28]. The Gaussian radial basis function is widely used in remote sensing [27]. More information regarding SVM can be found in [29] and [30].

IV. Methodology

In order to show the different capabilities of the proposed feature selection technique, two different scenarios have been taken into consideration.

A. First Scenario

In the first scenario, the new feature selection approach is directly performed on raw data sets in order to select the most informative bands from the whole data set. The main work flow of the proposed method for this scenario is listed as follows.

1) Training samples are split into two categories: training and validation samples.
2) FODPSO-based feature selection is performed on raw data. SVM is chosen as the fitness function, and its corresponding OA on validation samples is considered as the fitness value.
3) The selected bands are classified by SVM with the whole training and test samples, and final classification map will be achieved.

B. Second Scenario

In the second scenario, an application of the proposed FODPSO-based feature selection technique will be shown. In this scenario, we address the main shortcomings of using AP: 1) which attributes should be taken into account and 2) which values should be opted as threshold values. A comprehensive discussion related to AP and all its modifications and alternatives can be found in [31]. In this scenario, different types of attributes with the wide ranges of threshold values will be constructed for building a feature bank, and then, we let the proposed feature selection technique choose the most informative features from the bank with respect to the classification accuracy for the validation samples. In other words, the new feature selection technique not only solves the main shortcomings associated with the concept of AP, but also reduces the redundancy of the features and addresses the curse of dimensionality.

Fig. 2 illustrates the flowchart of the proposed method based on the FODPSO feature selection technique for the second scenario. The main work flow of this method is listed as follows.

1) A feature bank is made, consisting of raw input data and an AP obtained with four attributes with a wide range of threshold values.
   - The raw input data are transformed by principal component analysis (PCA).
   - The most important principal components (PCs), i.e., components with cumulative variance of more than 99%, are kept and used as base images for the extended multi-AP (EMAP).
   - The obtained EMAP and the raw input data are concatenated into a stacked vector (let us call the output of this step $\wp$).
2) Training samples are split into two categories: training and validation samples.

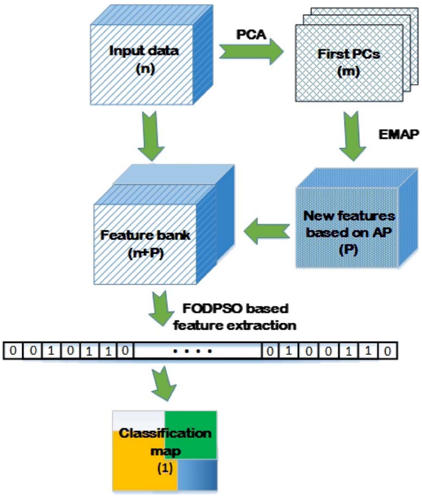


Fig. 2. General idea of the proposed classification framework based on the FODPSO-based feature selection technique for the second scenario.

3) FODPSO-based feature selection is performed on $\wp$. The fitness of each particle is evaluated by the OA of SVM for the validation samples. After a few iterations, the FODPSO-based feature selection approach finds the most informative bands with respect to the OA of SVM over the validation samples (the output of this step will be called $\Im$).
4) $\Im$ is classified by SVM by considering the whole set of training and test samples, and a final classification map will be achieved.

It should be noted that, in this work, the PCA can be replaced by other feature extraction techniques (in particular, kernel PCA and nonparametric weighted feature extraction (NWFE), which have shown promising results in order to produce APs [31]). Now, a brief discussion on EMAP is given.

1) EMAP: In order to overcome the shortcomings of the morphological profile, AP was introduced in [32] for extracting spatial information of the input data, which is based on attribute filters. APs can be regarded as more effective filters than morphological profiles because the concept of the attribute filters is not limited to only the size of different objects, and the APs are able to characterize other characteristics such as shape of existing objects in the scene. In addition, APs are computed according to an effective implementation based on max-tree and min-tree representations, which lead to a reduction in the computational load when compared with conventional profiles built with operators by reconstruction. An AP is built by the sequence of attribute thinning and thickening transformations defined with a sequence of progressively stricter criteria [32]. To handle hyperspectral images, the extension of AP was proposed in [33]. Extended AP is a stacked vector of different APs computed on the first C features extracted from the original data set. When different attributes $a_1, a_2, \ldots, a_M$ are concatenated into a stacked vector, the EMAP is obtained. More information regarding AP and its different variations can be found in [2], [3], [31], and [32]. In this paper, the following attributes have been taken into account:

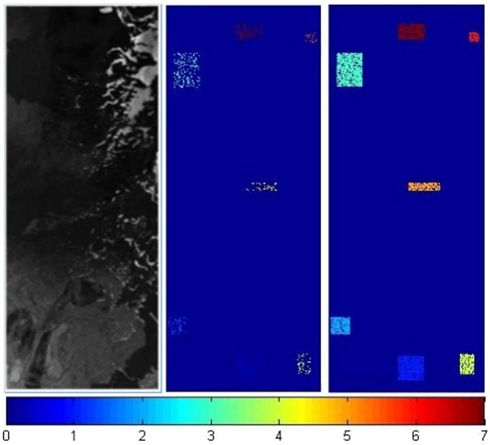


Fig. 3. AVIRIS Hekla data set. (a) Spectral band number 50. (b) Training samples and (c) test samples, where each color represents a specific information class. The information classes are listed in Table I.

1) (a) area of the region (related the size of the regions);
2) (s) standard deviation (as an index for showing the homogeneity of the regions);
3) (d) diagonal of the box bounding the regions;
4) (i) moment of inertia (as an index for measuring the elongation of the regions).

V. Experimental Results

A. Data Description

1) Hekla Data: The first hyperspectral data set used was collected on June 17, 1991 by AVIRIS (having a spatial resolution of 20 m) from the volcano Hekla in Iceland (see Fig. 3). Sixty four bands (from 1.84 μ to 2.4 μ) were removed due to the technical problem with the fourth spectrometer in 157 bands. An image of size 500×200 was used in this paper for the real experiments. For this data set, from the total number of training samples, which is equal to 966, 50% was chosen for training and the rest as validation samples, in order to perform the PSO- and FODPSO-based feature selection approaches. After finding the most informative bands with respect to the OA of SVM over validation samples, all 966 samples are used for training in order to perform SVM on the selected bands. The number of training, validation, and test samples is displayed in Table I.

2) Indian Pines Data: The second data set used in experiments is the well-known data set captured on Indian Pines (NW Indiana) in 1992 comprising 16 classes (see Fig. 4), mostly related to different land covers. The data set consists of 145×145 pixels with spatial resolution of 20 m/pixel. In this paper, 220 data channels (including all noisy and atmospheric absorbed bands) are used. In the same way, for Indian Pines, from the total number of training samples, which is equal to 695, 50% of the samples were chosen for training and the rest

TABLE I
HEKLA: NUMBER OF TRAINING, VALIDATION, AND TEST SAMPLES. FOR THE FINAL CLASSIFICATION STEP, THE TOTAL OF TRAINING AND VALIDATION SAMPLES IS USED TO TRAIN THE SVM

Class		Number of Samples		
Number	Name	Training	Validation	Test
1	Andesite lava 1970	97	97	829
2	Andesite lava 1980 I	54	54	442
3	Andesite lava 1991 I	150	150	1196
4	Andesite lava moss cover	49	49	370
5	Hyaloclastite formation	38	38	334
6	Rhyolite	20	19	143
7	Firn-glacier ice	76	75	549

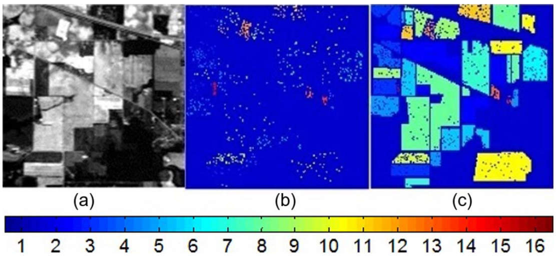


Fig. 4. AVIRIS Indian Pines data set. (a) Spectral band number 27 ($\lambda = 646.72\ \mu$). (b) Training samples and (c) test samples, where each color represents a specific information class. The information classes are listed in Table II.

TABLE II
INDIAN PINES: NUMBER OF TRAINING, VALIDATION, AND TEST SAMPLES. FOR THE FINAL CLASSIFICATION STEP, THE TOTAL OF TRAINING AND VALIDATION SAMPLES IS USED TO TRAIN THE SVM

Class		Number of Samples		
Number	Name	Training	Validation	Test
1	Corn-notill	25	25	1384
2	Corn-mintill	25	25	784
3	Corn	25	25	184
4	Grass-pasture	25	25	447
5	Grass-trees	25	25	697
6	Hay-windrowed	25	25	439
7	Soybean-notill	25	25	918
8	Soybean-mintill	25	25	2418
9	Soybean-clean	25	25	564
10	Wheat	25	25	162
11	Woods	25	25	1244
12	Bldg-Grass-Tree-Drives	25	25	330
13	Stone-Steel-Towers	25	25	45
14	Alfalfa	8	7	39
15	Grass-pasture-mowed	8	7	11
16	Oats	8	7	5

for the validation samples, in order to perform the PSO- and FODPSO-based feature selection approaches. After performing PSO- and FODPSO-based feature selection approaches, all 695 samples are used for training in order to perform SVM on the selected bands. The number of training, validation, and test samples is displayed in Table II.

In this paper, in addition to selecting data sets that are widely used in the hyperspectral imaging community, we have used exactly the same training and test samples that have been considered in most works related to the classification of hyperspectral images. Some of the works that have considered exactly the same training and test samples are given in [3] and [34]. In other words, we not only used the same number of training and test samples adopted by other state-of-the-art methods, but the samples also have exactly the same spatial locations in the data. This way of using the training and test samples makes this work fully comparable with other spectral and spatial classification techniques reported in the literature.

B. General Information

The following measures are used in order to evaluate the performance of different classification methods.

1) *Average Accuracy (AA)*: This index shows the average value of the class classification accuracy.
2) *OA*: This index represents the number of samples, which is classified correctly, divided by the number of test samples.
3) *Kappa Coefficient (k)*: This index provides information regarding the amount of agreement corrected by the level of agreement that could be expected due to chance alone.
4) *CPU Processing Time*: This measure shows the speed of different algorithms. It should be noted that, since in all algorithms (except Raw) EMAP is carried out, the CPU processing time of this step is discarded from all methods. Hence, the CPU processing time is only provided for AP, Raw + AP, decision boundary feature extraction (DBFE), and NWFE. Except DBFE and NWFE, which have been used in the *MultiSpec* software, all methods used were programmed in MATLAB on a computer having Intel Pentium 4 CPU 3.20 GHz and 4 GB of memory.

The number of iterations in each run for PSO- and FODPSO-based feature selection techniques is equal to 10. Since PSO- and FODPSO-based feature selection techniques are randomized methods, which are based on different first populations, each algorithm has here been run 30 times, and results are shown in different histograms and compared with different indexes in order to examine the capabilities of PSO- and FODPSO-based feature selection techniques.

It should be noted that, in this paper, in order to compare PSO and FODPSO, both data sets (Hekla and Indian Pines) have been taken into account. However, for the first and second scenarios, we preferred to use only Indian Pines since this data set is more complex for classification. Therefore, the capability of the proposed method can be shown more clearly by using Indian Pines instead of Hekla. In this case, 25 PCs with a cumulative variance of more than 99% were selected as the base images for producing EMAP for Indian Pines.

Based on [24], the sum of all ρ's should be inferior to 2 and alpha should be near 0.632. Therefore, the parameters ρ_1, ρ_2, and α are initialized by 0.8, 0.8, and 0.7, respectively. It should be noted that the same set of parameters has been used for both data sets and both scenarios, in order to show that the proposed technique is data set distribution independent, and with the same set of parameters, for all different data sets and scenarios, the proposed method can lead to an acceptable results in terms of accuracies and CPU processing time.

After comparing PSO- and FODPSO-based feature selection techniques in terms of OA over validation samples, the best method will be chosen for further evaluation. Then, the best method will be compared with the other well-known feature selection and extraction techniques. In order to have a fair comparison, among 30 runs, four runs have been chosen, and their classification accuracies are compared with obtained results from the other feature selection and extraction techniques. This way, the results of 30 runs have been sorted in an increasing order with respect to their OA over validation samples. Then, for a fair comparison with the other methods, the four runs are selected as follows.

1) Min: SVM classification is applied on the bands selected with the least OA over the validation samples among 30 runs (the first group of the most informative bands when the results of 30 runs have been sorted in an increasing order).
2) $Median^1$: SVM classification is applied on the bands selected with the median OA over the validation samples among 30 runs (the 15th group of the most informative bands when the results of 30 runs have been sorted in an increasing order).
3) $Median^2$: SVM classification is applied on the most informative bands selected with the median OA over the validation samples among 30 runs (the 16th group of bands when the results of 30 runs have been sorted in an increasing order).
4) Max: SVM classification is applied on the bands selected with the highest OA over the validation samples among 30 runs (the 30th group of the most informative bands when the results of 30 runs have been sorted in an increasing order).

Other methods for the purpose of comparison are listed as follows.

1) *Raw*: The input data are directly classified with SVM without performing any feature selection or extraction technique.
2) *Div*: Divergence feature selection is performed on the input data and the selected bands are classified by SVM.
3) *TD*: Transformed divergence feature selection is performed on the input data, and the selected bands are classified by SVM.
4) *Bhathacharyya*: Bhathacharyya distance feature selection is performed on the input data, and the selected bands are classified by SVM.
5) *DBFE*: DBFE is performed on the input data, and the selected bands are classified by SVM.
6) *NWFE*: NWFE is performed on the input data, and the selected bands are classified by SVM.

For the second scenario:

1) *AP*: The feature bank including all four attributes with a wide range of threshold values is classified by SVM.
2) $Raw + AP$: The Raw and AP are concatenated into a stacked vector and classified by SVM.

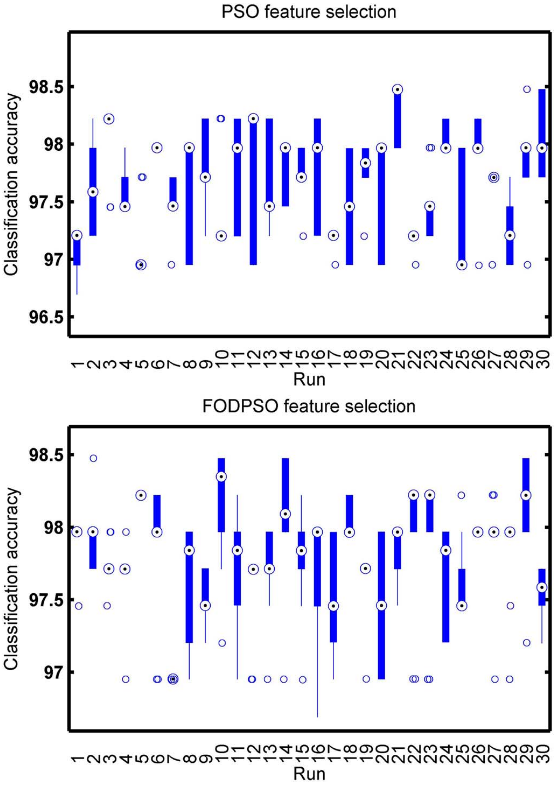


Fig. 5. Hekla: Box plots for OA over 30 runs for (top) PSO-based feature selection and (bottom) FODPSO-based feature selection.

The following ranges for different attributes have been taken into account in order to build the feature bank:

$$a = \left(\frac{1000}{phi}\right) \times \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$$

$$s = \left(\frac{\mu}{100}\right) \times \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$$

$$d = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$

$$i = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$$

where phi and μ are the resolution of the image in meters and the mean value of a feature, respectively.

Again, in order to secure the fairness of the comparison, the number of features for DBFE and NWFE has been chosen in two different ways: 1) the number of selected features is equal to the number of features, which provides Max (see a few lines above), and 2) the top few eigenvalues, which account for 99% of the total sum of the eigenvalues, were selected.

The data sets have been classified with SVM and a Gaussian kernel. Fivefold cross validation is taken into account in order to select the hyperplane parameters when SVM is used for the last step (for the classification of informative bands).

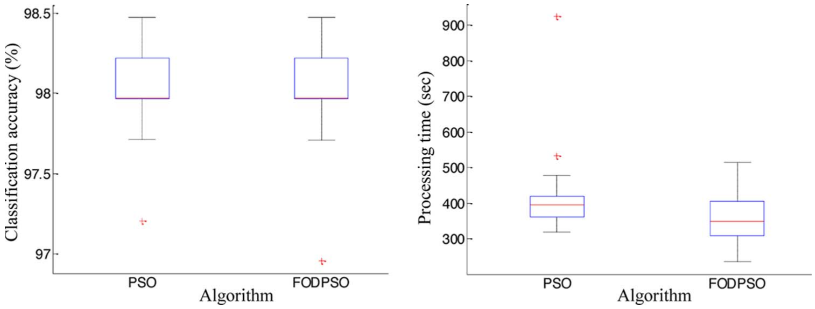


Fig. 6. Hekla: Final classification accuracy in percentage and processing time in seconds for the PSO- and FODPSO-based feature selection approaches.

C. FODPSO Versus PSO

1) Hekla: Fig. 5 shows the box plots for the OA of the classification for 30 runs (ten iterations within each run) for PSO- and FODPSO-based feature selection approaches, respectively. As can be observed in Fig. 5, although one cannot perceive any major differences between both methods, the FODPSO-based feature selection approach presents an overall smaller interquartile range, i.e., the OA of the classification, at each run, has less dispersion regardless of the number of trials. On the other hand, the average value of the OA is slightly larger than for the PSO-based method.

One-way MANOVA analysis was carried out to assess whether both the PSO- and FODPSO-based algorithms have a statistically significant effect on the classification performance. The significance of the different types of algorithm used (independent variable) on the final OA and the CPU processing time (dependent variables) was analyzed using one-way MANOVA after checking the assumptions of multivariate normality and homogeneity of variance/covariance, for a significance level of 5%.

The assumption of normality of each of the univariate dependent variables was examined using a paired-sample *Kolmogorov–Smirnov* $(p\text{-}value < 0.05)$ [35]. Although the univariate normality of each dependent variable has not been verified, since $n \geq 30$, and this was assumed by benefiting from using the central limit theorem [36], [37]. Consequently, the assumption of multivariate normality was validated [37], [38]. Note that MANOVA makes the assumption that the within-group covariance matrices are equal. Therefore, the assumption about the equality and homogeneity of the covariance matrix in each group was verified with the Box's M Test $(M = 72.7642,\ F(3; 720) = -2.0706;\ p\text{-}value = 1.0000)$ [38].

The MANOVA analysis revealed that the type of algorithm did not lead to a statistically significant different outcome for the multivariate composite $(F(1; 58) = 3.5830;\ p\text{-}value = 0.1667)$. In this situation, the FODPSO-based solution produces slightly better solutions than the PSO but is considerably faster than the latter. To easily assess the differences between both algorithms, let us graphically show the outcome of each trial using box plot charts (see Fig. 6). The ends of the blue boxes and the horizontal red line in between correspond to the first and third quartiles and the median values, respectively. As one may observe, by benefiting from the fractional version of the algorithm, one is able to slightly increase the OA while, at the same time, slightly decrease the CPU processing time.

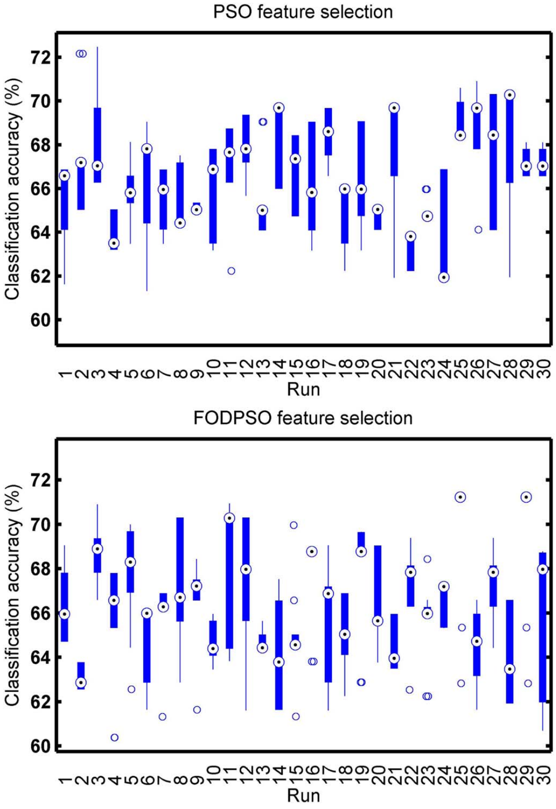


Fig. 7. Indian Pines: Box plots for OA in percentage over 30 runs for (top) PSO-based feature selection and (bottom) FODPSO-based feature selection.

2) Indian Pines: Fig. 7 shows the box plots for the OA of the classification for 30 runs (ten iterations within each run) for PSO- and FODPSO-based feature selection approaches, respectively. Similarly as before, despite the lack of major differences, the FODPSO-based feature selection approach presents an overall smaller interquartile range and a larger average value of the OA.

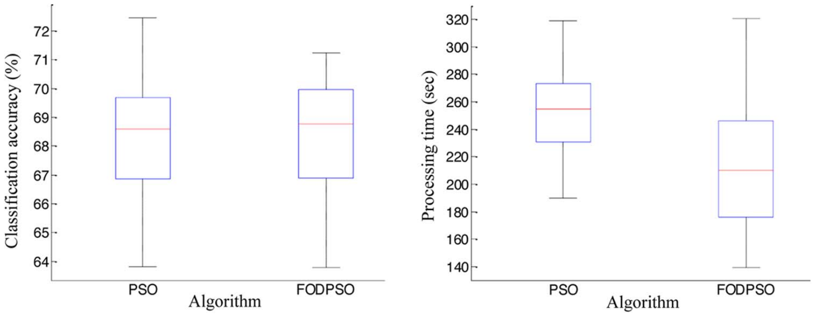


Fig. 8. Indian Pines: Final classification accuracy in percentage of the PSO- and FODPSO-based feature selection approaches.

Once again, one-way MANOVA analysis was carried out to assess whether both PSO- and FODPSO-based algorithms have a statistically significant effect on the classification performance.

The assumption about the equality and homogeneity of the covariance matrix in each group was verified with the Box's M Test $(M = 72.5156,\ \ F(3; 720) = -2.0636;\ \ p\text{-}value = 1.0000)$.

For this data set, the MANOVA analysis revealed that the type of the feature selection algorithm led to a statistically significant different outcome on the multivariate composite $(F(1; 58) = 14.6338;\ \ p\text{-}value < 0.0001)$. As the MANOVA detected significant statistical differences, we proceeded to the commonly used ANOVA for each dependent variable. By carrying an individual test on each dependent variable, it was possible to observe that the OA does not present statistically significant differences $(F(1; 58) = 0.0116;\ \ p\text{-}value = 0.9145)$. On the other hand, it is in the CPU processing time that both algorithms diverge the most, thus resulting in statistically significant differences between them $(F(1; 58) = 16.7499;\ \ p\text{-}value < 0.0001)$. As expected, the FODPSO-based solution produces slightly better solutions than the PSO considerably faster than the latter.

To easily assess the differences between both algorithms, the outcome of each trial is graphically shown using box plot charts (see Fig. 8). As one may observe, by benefiting from the fractional version of the algorithm, one is able to slightly increase the OA (slightly higher median value) and, at the same time, considerably decrease the CPU processing time.

D. First Scenario

Fig. 9 shows the list of the selected bands by the proposed method in 30 different runs. Runs 2, 16, 30, and 29 are selected as Min, $Median^1$, $Median^2$, and Max, respectively. As it was mentioned before, since the FODPSO-based feature selection technique is a randomized method that is based on different first populations, the selected bands are different in different runs.

As can be seen from Table III, the proposed method (Max) provides the best results in terms of OA, followed by $Median^1$ and $Median^2$ (other runs of the proposed technique). This shows that different alternatives of the proposed method (except Min) demonstrate the best performance and improve the other techniques in terms of classification accuracies.

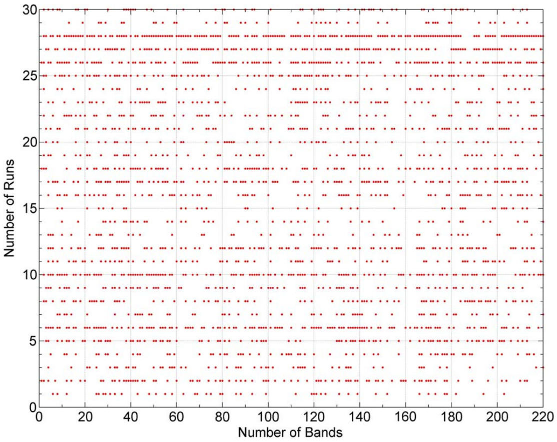


Fig. 9. First scenario: Selected bands by the proposed method in 30 different runs. Runs 2, 16, 30, and 29 are selected as Min, $Median^1$, $Median^2$, and Max, respectively.

Some algorithms, such as the originally proposed DBFE [39], require the use of the second-order statistics (e.g., the covariance matrix) to characterize the distribution of training samples with respect to the mean. In hyperspectral image analysis, the number of available training samples is usually not sufficient to make a good estimate of the covariance matrix. In this case, the use of sample covariance, or common covariance [1], may not be successful. As an example, either when the sample or the common covariance approach is chosen to estimate the statistics for each available class for DBFE, if the number of pixels in the classes is not, one more than the total number of features being used (at least), the DBFE stops working. In this case, the leave-one-out covariance (LOOC) [1] estimator can be used as an alternative to estimate the covariance matrix. The normal minimum number of required samples for a sample class covariance matrix is $l + 1$ samples for l-dimensional data. For the LOOC estimator, only a few samples are all that is needed. In general, this covariance estimator is nonsingular when at least three samples are in hand regardless of the dimensions of the data, and so it can be used even though the sample covariance or common covariance estimates are singular.

TABLE III
FIRST SCENARIO: THE CLASSIFICATION OF DIFFERENT TECHNIQUES IN PERCENTAGE FOR INDIAN PINES. THE NUMBER OF FEATURES IS SHOWN IN BRACKETS. THE BEST ACCURACY IN EACH ROW IS SHOWN IN BOLD. TIME IN SECONDS

Class	Raw	Min	$Median^1$	$Median^2$	Max	DBFE	DBFE-99%	NWFE	NWFE-99%
No.	(220)	(84)	(79)	(57)	(44)	(44)	(17)	(44)	(120)
1	54.1	67.1	69.5	67.1	**70.9**	54.4	54.1	59.6	49.1
2	57.5	67.6	66.8	**69.6**	68.6	62.2	53.8	51.2	46.3
3	80.4	82.0	84.2	**90.7**	88.5	73.3	70.1	67.3	64.6
4	88.3	92.6	92.6	88.1	**92.8**	86.5	88.1	86.5	89.0
5	81.4	85.2	**88.5**	82.6	88.3	88.3	87.2	87.5	84.3
6	92.2	92.9	94.9	94.0	95.8	94.5	96.3	**97.2**	**97.2**
7	68.0	73.7	76.6	76.4	**78.7**	61.1	62.2	61.0	54.0
8	49.1	47.7	**63.6**	58.4	61.7	47.2	43.9	34.2	35.7
9	64.1	73.7	82.0	79.7	**83.5**	72.3	67.7	62.2	63.4
10	95.6	96.2	97.5	**98.7**	**98.7**	**98.7**	99.3	**98.7**	98.1
11	79.0	82.7	87.1	87.5	**91.1**	86.6	85.3	80.9	84.4
12	64.5	70	73.0	73.3	**77.5**	76.9	73.0	69.6	72.1
13	95.5	97.7	97.7	**100**	97.7	91.1	93.3	95.5	91.1
14	64.1	79.4	**92.3**	84.6	89.7	61.5	58.9	71.7	61.5
15	81.8	**100**	81.8	90.9	90.9	81.8	**100**	81.8	63.6
16	**100**	**100**	**100**	60	**100**	40	40	40	40
AA	76.02	81.82	84.29	81.39	**85.94**	73.57	73.36	71.61	68.44
OA	65.41	70.11	76.22	74.17	**77.18**	66.95	64.96	61.98	60.13
K	0.6119	0.6646	0.7306	0.7088	**0.7418**	0.6273	0.6055	0.5749	0.5533
Time	94	165	192	276	223	105	72	89	132

As discussed before, the conventional feature selection techniques are only feasible in relatively low dimensional cases. This way, as the number of bands increases, the required statistical estimation becomes unwieldy. In our case, the methods divergence, transformed divergence, and Bhattacharyya distance stopped working since in our data sets, the corrupted bands have not been eliminated, and also, the dimensionality of the data sets is high. However, since the proposed method is based on the evolutionary technique, there is no need to calculate all possible alternatives in order to find the most informative bands. Another advantage of using the proposed method is that there is no need to estimate the second-order statistics, and in this manner, the singularity of the covariance matrix is not a problem. Therefore, the FODPSO-based feature selection technique can find the most informative bands in a very reasonable CPU processing time when the other techniques stop and cannot lead to a conclusion.

E. Second Scenario

Fig. 10 depicts the box plots for the OA of the classification for 30 runs (ten iterations within each run) for PSO- and FODPSO-based feature selection approaches, respectively. As before, Fig. 10 shows the advantage of the FODPSO-based approach over the alternative. In this case, one can easily perceive the differences, wherein both the interquartile range and the average value of the OA are considerably improved. In other words, the FODPSO-based feature selection technique is able to find a better and more stable solution than the PSO-based feature selection technique.

The significance of the different types of algorithm used (independent variable) on the final OA and the CPU processing time (dependent variables) was analyzed using one-way MANOVA.

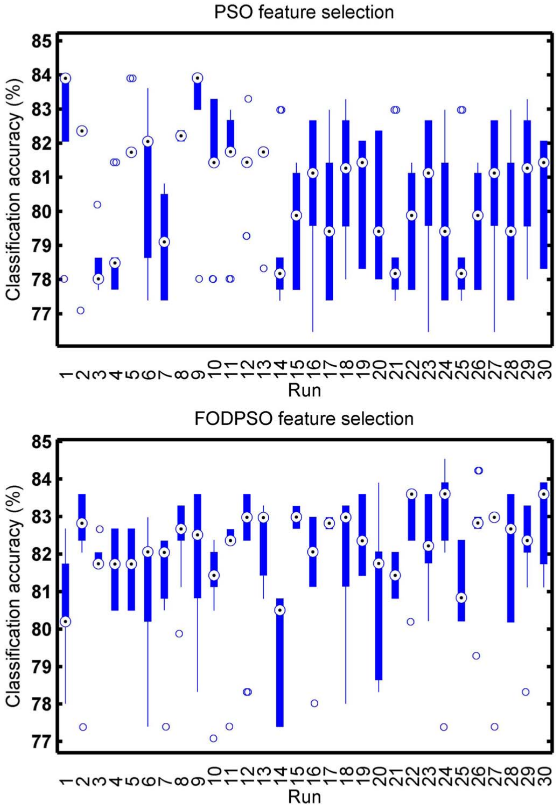


Fig. 10. Indian Pines: Box plots for OA in percentage over 30 runs for (top) PSO-based feature selection and (bottom) FODPSO-based feature selection.

The assumption about the equality and homogeneity of the covariance matrix in each group was verified with the Box's M Test $(M = 72.8921,\ F(3; 720) = -2.07424;\ p\text{-}value = 1.0000)$. This suggests that the design is balanced and, since there is an equal number of observations in each cell $(n = 30)$, the robustness of the MANOVA tests is guaranteed.

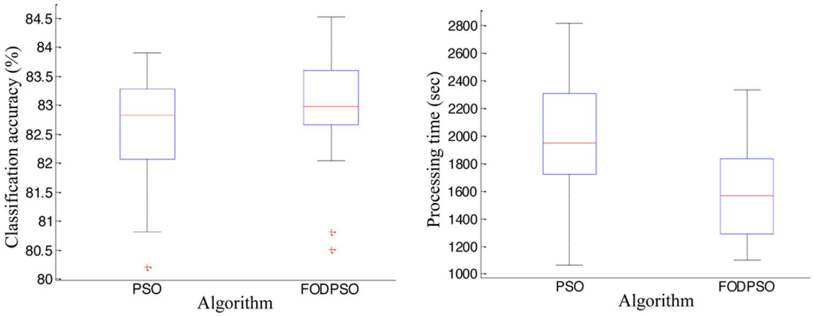


Fig. 11. Indian Pines in the second scenario: Final classification accuracy in percentage of the PSO- and FODPSO-based feature selection approaches.

The MANOVA analysis revealed that the type of algorithm led to a statistically significant different outcome on the multivariate composite $(F(1;58) = 18.8030;\ p\text{-}value < 0.0001)$. As the MANOVA detected significant statistical differences, we proceeded to the commonly used ANOVA for each dependent variable. By carrying an individual test on each dependent variable, it was possible to observe that the OA does not present statistically significant differences $(F(1;58) = 3.3804;$ $p\text{-}value = 0.0711)$. On the other hand, it is once again in the CPU processing time that both algorithms diverge the most, presenting statistically significant differences $(F(1;58) = 20.7238;\ p\text{-}value < 0.0001)$. As expected, the FODPSO-based approach (higher median value) produces slightly better solutions than the PSO and is considerably faster than the latter.

To easily assess the differences between both algorithms, the outcome of each trial is graphically shown using box plot charts (see Fig. 11). In the second scenario, one can easily observe the benefits of the fractional version of the algorithm, achieving a high level of OA in a short period of time.

Table IV gives information regarding the number of selected features in Min, $Median^1$, $Median^2$, and Max for different attributes, area (a) with 725 features, standard deviation (s) with 500 features, moment of inertia (i) with 450 features, and diagonal of the box bounding the regions (d) with 500 features. As can be inferred from the table, the proposed method selects different number of features for different attributes in different runs. Therefore, it is difficult to conclude which attribute leads to better classification accuracies. However, it seems that the proposed methodology selects the highest number of features for the area attribute. The reason behind this might be that the area attribute is well related to the object hierarchy in the images, and it generally can model the spatial information of images in a good way.

As can be seen from Table V, the proposed feature selection technique has the best performance when the other feature selection and extraction techniques are not able to process the data due to very high dimensionality and the limited number of training samples. All alternatives of the proposed method have almost the same performance in terms of OA and significantly improve on AP and Raw + AP in terms of classification accuracies. Both AP and Raw + AP dramatically suffer by the curse of dimensionality and the high redundancy of available features in the feature bank.

TABLE IV

SECOND SCENARIO: THE NUMBER OF SELECTED FEATURES IN Min, $Median^1$, $Median^2$, AND Max FOR DIFFERENT ATTRIBUTES, AREA (A) WITH 725 FEATURES, STANDARD DEVIATION (S) WITH 500 FEATURES, MOMENT OF INERTIA (I) WITH 450 FEATURES, AND DIAGONAL OF THE BOX BOUNDING THE REGIONS (D) WITH 500 FEATURES

Group	a	s	i	d	Input data	Total number of selected features
Min	152	106	85	107	47	497
$Median^1$	124	79	76	85	34	398
$Median^2$	153	118	92	112	42	517
Max	83	74	65	53	31	306
Total features	725	500	450	500	220	–

TABLE V

SECOND SCENARIO: THE CLASSIFICATION OF DIFFERENT TECHNIQUES IN PERCENTAGE FOR INDIAN PINES. THE NUMBER OF FEATURES IS SHOWN IN BRACKETS. THE BEST ACCURACY IN EACH ROW IS SHOWN IN BOLD

Class No.	Raw (220)	AP (2175)	Raw+AP (2395)	Min (497)	$Median^1$ (398)	$Median^2$ (517)	Max (306)
1	54.1	68.3	67.7	**78.1**	76.0	76.3	76.6
2	57.5	79.3	78.5	88.2	**90.1**	87.7	89.5
3	80.4	82.6	82.6	94.0	92.9	**95.6**	93.4
4	88.3	83.4	81.8	93.9	**94.4**	**94.4**	**94.4**
5	81.4	81.2	80.4	90.2	90.2	90.1	**90.3**
6	92.2	94.9	94.5	98.8	98.8	98.6	**99.0**
7	68.0	75.8	75.8	82.7	82.0	**84.8**	77.4
8	49.1	68.5	67.7	84.9	78.4	**85.1**	82.4
9	64.1	75.5	73.7	87.2	84.5	86.8	**89.1**
10	95.6	91.9	90.7	98.7	98.1	**99.3**	98.7
11	79.1	88.2	87.9	95.8	95.5	94.2	**95.9**
12	64.5	97.8	**98.1**	94.5	93.0	94.8	94.2
13	95.5	88.8	88.8	**100**	**100**	**100**	**100**
14	64.1	51.2	43.5	92.3	89.7	89.7	**97.4**
15	81.8	81.8	81.8	81.8	90.9	90.9	**100**
16	**100**	**100**	80	**100**	**100**	**100**	**100**
AA	76.02	81.87	79.64	91.35	90.94	91.80	**92.44**
OA	65.41	77.54	76.85	**87.83**	85.73	87.60	86.78
K	0.611	0.7465	0.739	**0.8613**	0.8377	0.8587	0.8494
Time	–	1295	1434	1556	1529	1567	1498

In addition, the proposed method can be considered as a good solution to overcome the shortcomings of APs. As can be seen, the proposed method can automatically find the most informative features from the feature bank including highly redundant features.

VI. Conclusion

In this paper, a novel feature selection approach has been proposed, which is based on a new binary optimization technique named binary FODPSO and SVM. The proposed approach was compared with commonly used feature selection and feature extraction approaches in experiments using standard AVIRIS hyperespectral data sets. Based on the experiments, the following points can be concluded.

- Binary FODPSO exploits many swarms in which each swarm individually performs similar to an ordinary PSO algorithm with rules governing the collection of swarms that are designed to simulate natural selection. Moreover, the concept of fractional derivative is used to control the convergence rate of particles. The aforementioned reasons lead to a better performance than binary PSO in terms of CPU processing time and OA for the cross-validation samples.
- In the novel feature selection approach, there is no need to set the number of output features, and the proposed approach can automatically select the most informative features in terms of classification accuracies.
- Since the new approach is based on an evolutionary method, it is much faster than other well-known feature selection techniques that demand an exhaustive process to select the most informative bands. In this sense, the new approach can work appropriately in a situation when other feature selection techniques are not applicable.
- Since the new feature selection approach is based on an SVM classification that is capable of handling high-dimensional data with a limited number of training samples, it can proceed to select the most informative features in an ill-posed situation when other feature selection/extraction techniques cannot proceed without a powerful technique for estimating the statistics for each class. As an example, when the original way is opted to estimate the statistics for each class, DBFE based on original statistics cannot proceed, since the number of pixels in the following classes needs to be at least one more than the total number of features being used, and LOOC statistics must be taken into account to handle this issue [1]. However, the new method can handle this problem effectively.
- The new approach can solve the main shortcomings of using AP for classification.

As a possible future work, we aim at finding the best SVM parameters (i.e., regularization and kernel parameters) by using the proposed binary FODPSO in an automatic way instead of adjusting the parameters by using a cross-validation procedure after performing binary FODPSO. In addition, in the second scenario, the proposed feature selection approach can be performed on each AP separately, which generally leads to higher classification accuracy but in a higher CPU processing time. Therefore, another topic deserving future research is the development of parallel implementations of the presented approach in high-performance computing architectures, although the processing times reported in our experiments (measured in a standard desktop CPU) are quite small for the considered data sets.

References

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

[2] P. Ghamisi, J. A. Benediktsson, and J. R. Sveinsson, "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5771–5782, Sep. 2014.

[3] P. Ghamisi, J. A. Benediktsson, G. Cavallaro, and A. Plaza, "Automatic framework for spectral–spatial classification based on supervised feature extraction and morphological attribute profiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2147–2160, Jun. 2014.

[4] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[5] J. C. Russ, *The Image Processing Handbook*, 3rd ed. Boca Raton, FL, USA: CRC Press, 1999.

[6] D. J. Wiersma and D. A. Landgrebe, "Analytical design of multispectral sensors," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-18, no. 2, pp. 180–189, Apr. 1980.

[7] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 7, pp. 793–802, 2004.

[8] P. H. Swain, "Fundamentals of pattern recognition in remote sensing," in *Remote Sensing-The Quantitative Approach*, P. H. Swain and S. Davis, Eds., New York, NY, USA: McGraw-Hill, 1978.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, San Diego, CA, USA: Academic, 1990.

[10] X. Jia, B. Kuo, and M. M. Crawford, "Feature mining for hyperspectral image classification," *Proc. IEEE*, vol. 101, no. 3, pp. 676–697, Mar. 2013.

[11] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, Nov. 2006.

[12] A. Daamouche, F. Melgani, N. Alajlan, and N. Conci, "Swarm optimization of structuring elements for VHR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1334–1338, Nov. 2013.

[13] P. Ghamisi and J. A. Benediktsson, "Feature selection based on hybridization of genetic algorithm and particle swarm optimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 309–313, Feb. 2015.

[14] A. Paoli, F. Melgani, and E. Pasolli, "Clustering of hyperspectral images based on multiobjective particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 12, pp. 4175–4188, Dec. 2009.

[15] F. vandenBergh and A. P. Engelbrecht, "A cooperative approach to particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 225–239, Jun. 2004.

[16] K. Premalatha and A. Natarajan, "Hybrid PSO and GA for global maximization," *Int. J. Open Problems Comput. Math.*, vol. 2, no. 4, pp. 597–608, 2009.

[17] J. Kennedy and R. Eberhart, "A new optimizer using particle swarm theory," *Proc. IEEE 6th Int. Symp. Micro Mach. Human Sci.*, 1995, pp. 39–43.

[18] Y. D. Valle, G. K. Venayagamoorthy, S. Mohagheghi, J. C. Hernandez, and R. Harley, "Particle swarm optimization: Basic concepts, variants and applications in power systems," *IEEE Trans. Evol. Comput.*, vol. 12, no. 2, pp. 171–195, Apr. 2008.

[19] J. Tillett, T. M. Rao, F. Sahin, R. Rao, and S. Brockport, "Darwinian particle swarm optimization," *Proc. 2nd Indian Int. Conf. Artif. Intell.*, 2005, pp. 1–14.

[20] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, "An efficient method for segmentation of images based on fractional calculus and natural selection," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12 407–12 417, Nov. 2012.

[21] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, "Use of Darwinian particle swarm optimization technique for the segmentation of remote sensing images," *Proc. IGARSS*, 2012, pp. 4295–4298.

[22] M. S. Couceiro, R. P. Rocha, N. M. F. Ferreira, and J. A. T. Machado, "Introducing the fractional order Darwinian PSO," *Signal, Image Video Process., Springer, no. Fractional Signals and Systems Special Issue*, vol. 6, no. 3, pp. 343–350, Sep. 2012.

[23] P. Ghamisi, M. S. Couceiro, F. M. Martins, and J. A. Benediktsson, "Multilevel image segmentation approach for remote sensing images based on fractional-order Darwinian particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2382–2394, May 2014.

[24] M. S. Couceiro, F. M. Martins, R. P. Rocha, and N. M. Ferreira, "Mechanism and convergence analysis of a multi-robot swarm approach based on natural selection," *J. Intell. Robot. Syst.*, vol. 76, no. 2, pp. 353–381, Nov. 2014.

[25] M. A. Khanesar, M. Teshnehlab, and M. A. Shoorehdeli, "A novel binary particle swarm optimization," *Proc. IEEE Mediterranean Conf. Control Autom.*, pp. 1–6, 2007.

[26] V. N. Vapnic, *Statistical Learning Theory*, Hoboken, NJ: Wiley, 1998.

[27] B. Scholkopf and A. J. Smola, *Learning with Kernels*, Cambridge, MA, USA: MIT Press, 2002.

[28] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for the classification of hyperspectral remote-sensing data over urban areas," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 1–14, 2009.

[29] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Evaluation of kernels for multiclass classification of hyperspectral remote sensing data," *Proc. ICASSP*, vol. 2, pp. 813–816, 2006.

[30] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, May 2002.

[31] P. Ghamisi, M. Dalla Mura, and J. A. Benediktsson, "A survey on spectral–spatial classification techniques based on attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[32] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[33] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, Jul. 2010.

[34] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[35] J. A. Peacock, "Two-dimensional goodness-of-fit testing in astronomy," *Mon. Not. R. Astronom. Soc.*, vol. 202, no. 3, pp. 615–627, 1983.

[36] A. C. Pedrosa and S. M. A. Gama, *Introducao Computacional a Probabilidade e Estatistica*, Porto, Portugal: Porto Editora, 2004, p. 348.

[37] J. Maroco, *Anlise Estatstica com Utilizao do SPSS*, Lisboa, Portugal: Edies Silabo, 2010,p. 199.

[38] J. Pallant, *SPSS Survival Manual*, 4th ed. Kindle Edition Buckingham, U.K.: Open Univ. Press, 2011.

[39] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 388–400, Apr. 1993.

Pedram Ghamisi (S'12) received the B.Sc. degree in civil (survey) engineering from the Tehran South Campus of Azad University, Tehran, Iran, and the M.Sc. degree in remote sensing from K.N. Toosi University of Technology, Tehran, in 2012. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Iceland, Reykjavík, Iceland.

His research interests include remote sensing and image analysis with the current focus on spectral and spatial techniques for hyperspectral image classification and the integration of LiDAR and hyperspectral data for land cover assessment.

Mr. Ghamisi serves as a Reviewer for a number of journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and *Pattern Recognition Letters*. He received the Best Researcher Award for M.Sc. students from K.N. Toosi University of Technology during the academic year 2010–2011. At the 2013 IEEE International Geoscience and Remote Sensing Symposium, Melbourne, July 2013, he was awarded the IEEE Mikio Takagi Prize, for winning the Student Paper Competition at the conference between approximately 70 people.

Micael S. Couceiro (M'12) received the B.Sc., Teaching Licensure, and Master's degrees in electrical engineering (automation and communications) from the Engineering Institute of Coimbra, Polytechnic Institute of Coimbra (ISEC-IPC), Coimbra, Portugal, and the Ph.D. degree in electrical and computer engineering (automation and robotics) from the Faculty of Sciences and Technology, University of Coimbra, Coimbra, in 2014.

He is currently the CEO of Ingeniarius, Lda., Mealhada, Portugal. Over the past six years, he has been conducting scientific research on several areas besides robotics, namely, computer vision, sports engineering, economics, sociology, digital media, and others, all at the Institute of Systems and Robotics (ISR-UC) and in the RoboCorp research group from IPC. This resulted in more than 20 scientific articles in international impact factor journals and more than 40 scientific articles at international conferences. In addition to research, he has been invited for lecturing, tutoring, and organization of events (e.g., professional courses, national and international conferences, among others), both in the public and private domains.

Jon Atli Benediktsson (S'84–M'90–SM'99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984 and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is currently the Pro-Rector of Science and Academic Affairs and a Professor of electrical and computer engineering with the University of Iceland. He has extensively published in his areas of interest. His research interests include remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing.

Dr. Benediktsson was the 2011–2012 President of the IEEE Geoscience and Remote Sensing Society (GRSS) and has been on the GRSS Administrative Committee since 2000. He was Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008 and has served as an Associate Editor of TGRS since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and the IEEE ACCESS since 2013. He is on the Editorial Board of the PROCEEDINGS OF THE IEEE and the International Editorial Board of the *International Journal of Image and Data Fusion* and was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING during 2007–2010. He is a cofounder of the biomedical startup company Oxymap. He is a Fellow of the International Society for Optics and Photonics (SPIE). He is a member of the 2014 IEEE Fellow Committee. He received the Stevan J. Kristof Award from Purdue University, West Lafayette, IN, USA, in 1991 as an outstanding graduate student in remote sensing. In 1997, he was a recipient of the Icelandic Research Council's Outstanding Young Researcher Award; in 2000, he was granted the IEEE Third Millennium Medal; in 2004, he was a corecipient of the University of Iceland's Technology Innovation Award; in 2006, he received the yearly research award from the Engineering Research Institute of the University of Iceland; and in 2007, he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society. He was a corecipient of the 2012 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Paper Award, and in 2013, he was a corecipient of the IEEE GRSS Highest Impact Paper Award. In 2013, he received the IEEE/VFI Electrical Engineer of the Year Award. In 2014, he was a corecipient of the 2012–2013 Best Paper Award from the International Journal of Image and Data Fusion. He is a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica, and Tau Beta Pi.

Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization

Pedram Ghamisi, *Student Member, IEEE*, and Jon Atli Benediktsson, *Fellow, IEEE*

***Abstract*—A new feature selection approach that is based on the integration of a *genetic algorithm* and *particle swarm optimization* is proposed. The overall accuracy of a *support vector machine* classifier on validation samples is used as a fitness value. The new approach is carried out on the well-known Indian Pines hyperspectral data set. Results confirm that the new approach is able to automatically select the most informative features in terms of classification accuracy within an acceptable CPU processing time without requiring the number of desired features to be set *a priori* by users. Furthermore, the usefulness of the proposed method is also tested for road detection. Results confirm that the proposed method is capable of discriminating between road and background pixels and performs better than the other approaches used for comparison in terms of performance metrics.**

***Index Terms*—Attribute profile, feature selection, hybridization of genetic algorithm (GA) and particle swarm optimization (PSO), hyperspectral image analysis, road detection, support vector machine (SVM) classifier.**

I. Introduction

SUPERVISED classification techniques classify the input data by partitioning the feature space into decision regions, by using a set of training samples for each class. These samples are usually obtained by manual labeling of a small number of pixels in an image or based on some field measurements. Thus, the collection of these samples is expensive and time demanding. As a result, the number of available training samples is usually limited, which is a challenging issue in supervised classification.

In [1], it was shown that, after a few features, while the number of training samples is kept constant, the classification accuracy actually decreases as the number of features increases. For the purpose of classification, this is referred to as the *curse of dimensionality* [2]. To address this issue, the use of feature selection/extraction techniques is of importance.

Feature extraction/selection techniques can be grouped into two categories: unsupervised and supervised approaches. For the purpose of image classification, the latter techniques are preferred since they try to reduce the dimensionality of the data while maximizing the separability between classes. Nonparametric weighted feature extraction (NWFE) and parametric decision boundary feature extraction (DBFE) have been extensively used for this purpose. On the other hand, divergence, transformed divergence, Bhattacharyya distance, and Jeffries–Matusita distance are well-known feature selection techniques that have been widely used in remote sensing. For more information on the aforementioned techniques, please see [1].

Conventional feature selection techniques usually demand many samples to estimate statistics accurately. In addition, they are usually based on an exhaustive process for finding the best set of features, and in this case, they are time demanding, and their CPU processing time exponentially increases as the number of bands (features) increases. To this extent, a new generation of feature selection techniques is based on evolutionary optimization methods, since they are not based on an exhaustive process and can lead to a conclusion in a faster way. In addition, by considering an efficient fitness function for these methods, they can handle high-dimensional data with even a limited number of training samples (ill-posed situations). In particular, the *genetic algorithm* (GA) and *particle swarm optimization* (PSO) have gained significant attention from researchers. There is an extensive literature regarding the use of the GA and PSO for the purpose of feature selection. For example, in [3], Bazi and Melgani proposed a support vector machine (SVM) classification system that allows for detecting the most distinctive features and estimating the SVM parameters by using a GA. In [4], Daamouche *et al.* proposed the use of PSO to select for classification the most informative features obtained by morphological profiles. However, both PSO and the GA suffer from a few shortcomings. The main shortcoming of PSO is the premature convergence of a swarm. The key reason behind this shortcoming is that particles try to converge to a single point, which is located on a line between the global best and the personal best positions. This point is not guaranteed for a local optimum [5]. Another reason could be the fast rate of information flow between particles, which leads to the creation of similar particles. This results in a loss in diversity. Furthermore, the possibility of being trapped in local optima is increased [6]. The main advantage of using PSO is its simple concept along with the fact that it can be implemented in a few lines of code. Furthermore, PSO also has a memory of past iterations. On the other hand, in the GA, if a chromosome is not selected, the information contained by it is lost. However, without a selection operator as in the GA, PSO may waste resources on inferior individuals [6]. PSO may enhance the search capability for finding an optimal solution. However, the GA has difficulty in finding an exact solution [7].

In this letter, to address the main shortcomings of GA- and PSO-based feature selection techniques and to take the advantage of their strength, a new feature selection approach is proposed, which is based on the integration of the GA

Manuscript received May 12, 2014; revised June 17, 2014; accepted July 7, 2014. This work was supported in part by the Icelandic Research Fund for Graduate Students.

The authors are with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavik, Iceland (e-mail: peg6@hi.is; benedikt@hi.is).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2014.2337320

and PSO. To find the most discriminative features in terms of classification accuracies, the overall accuracy (OA) of the SVM classifier over validation samples is investigated as a fitness value. The SVM is selected due to the fact that it is capable of providing acceptable classification accuracies for high-dimensional data when even a limited number of training samples is available. To evaluate the efficiency of the proposed method, two different scenarios are drawn.

i) In the first scenario, the proposed feature selection approach is performed on a well-known hyperspectral data set, i.e., the AVIRIS Indian Pines. Results demonstrate that the new method can significantly increase the classification accuracy of the raw data in an acceptable CPU processing time.

ii) In the second scenario, the proposed feature selection technique is applied on a set of features derived by attribute profiles [8], to select the most discriminative features for detecting roads from a background. Results infer that the new feature selection approach is able to detect roads in a complex urban image with acceptable accuracy.

This letter is organized as follows: The proposed methodology is discussed in Section II. Section III is devoted to experimental results. Finally, Section IV outlines the main conclusions.

II. Methodology

Here, first, the concept of two well-known optimization techniques, namely, GA and PSO, will be recalled. Then, the proposed feature selection technique, which is based on the hybridization of the GA and PSO (HGAPSO + SVM), will be described.

A. GA

The GA is inspired by the genetic process of biological organisms. The GA consists of several solutions called chromosomes or individuals. Each chromosome in a binary GA includes several genes with binary values 0 and 1, which determines the attributes for each individual. A set of the chromosomes is made up to form a population. The merit of each chromosome is evaluated by using a fitness function. Fit chromosomes are selected for the generation of new chromosomes. In that step, two fit chromosomes are selected and combined through a crossover step to produce a new offspring (or solution). Then, mutation is applied on the population to increase the randomness of individuals for decreasing the possibility of getting stuck in local optimum [9].

B. PSO

PSO is a biologically inspired technique derived from the collective behavior of bird flocks, first introduced by Kennedy and Eberhart [10]. PSO consists of a set of solutions (particles) called population. Each solution consists of a set of parameters and represents a point in multidimensional space. A group of particles (population) makes up a swarm. Particles move through the search space with a specified velocity for finding the optimal solution. Each particle has a memory that helps it in keeping the track of its previous best position. The positions of the particles are distinguished as *personal best* and *global best*. The velocities of particles are adjusted according to the historical behavior of each particle and its neighbors, while they fly through the search space. Each move of particles is deeply influenced by its current position, its memory of previous useful parameters, and the group knowledge of the swarm [10]. Therefore, the particles have a tendency to fly toward improved search areas over the course of the search process.

The velocity of the ith particle in the $(k+1)$th iteration is mathematically defined as

$$V_i^{k+1} = WV_i^k + C_1 r_1 \left(pb_i^k - X_i^k\right) + C_2 r_2 \left(gb_d^k - X_i^k\right) \quad (1)$$

where C_1 and C_2 are acceleration constants, r_1 and r_2 are random values in the range of 0 and 1, W is the inertia weight (predefined by the user), X_i^k shows the position of each particle in d-dimensional search space, pb_i^k is the best previous position of each particle named particle best position, and gb^k is the best position of all the particles (called the global best particle). The position of the ith particle is updated by

$$X_i^{k+1} = X_i^k + V_i^{k+1}. \quad (2)$$

The PSO was originally introduced for the optimization of problems in continuous multidimensional search space. To extend that concept to feature selection, it needs to be developed to deal with binary data, in which 0 and 1 demonstrate the absence and presence of a band, respectively. In [10], *Kennedy* and *Eberhart* applied the sigmoid transformation on the velocity component to develop a binary discrete PSO to control the range of velocity between 0 and 1 according to

$$\Delta X_i^{k+1} = \frac{1}{1+\exp\left(-V_i^{k+1}\right)}. \quad (3)$$

For updating the position of each particle, ΔX_i^{k+1} is compared with r_x, which is a random d-dimensional vector in which each component is, in general, a uniform random number between 0 and 1 according to

$$X_i^{k+1} = \begin{cases} 1, & \Delta X_i^{k+1} \geq r_x \\ 0, & \Delta X_i^{k+1} < r_x. \end{cases} \quad (4)$$

C. HGAPSO + SVM

GA and PSO can be combined in different ways. However, in the proposed feature selection approach, hybridization is obtained through integrating the standard velocity and update rules of PSO with selection, crossover, and mutation from the GA.

Fig. 1 shows the block diagram of the proposed approach. To investigate the hybridization of the GA and PSO for the purpose of feature selection, the dimension of each particle needs to be equal to the number of features. In this case, that velocity dimension, i.e., $\dim V_i^k$, as well as the position dimension, i.e., $\dim X_i^k$, correspond to the total number of bands (l bands) in the input data $(\dim V_i^k = \dim X_i^k = l)$. In that case, each particle's velocity is represented as a l-dimension vector. In addition, as one wishes to use the algorithm for band selection, each particle represents its position in binary values, i.e., 0 or 1, where 0 and 1 demonstrate the absence and the presence of the corresponding feature, respectively.

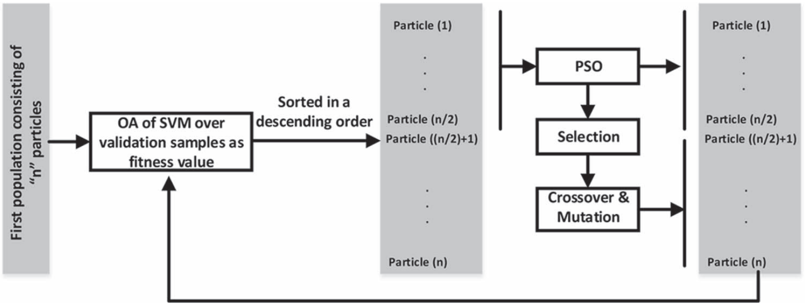


Fig. 1. Flowchart of the proposed method.

In this letter, a random population is initially generated. The individuals in the population may be regarded as chromosomes with respect to the GA or as particles with respect to PSO. Then, a new population for the next generation is produced through enhancement, crossover, and mutation as described below.

Enhancement: In each generation, after the fitness values for all the individuals in the same population are calculated (the OA of SVM on validation samples), the top half of the best-performing particles is selected. These individuals are regarded as elites. Then, the elites are enhanced by PSO. By using these enhanced elites as parents, the generated offsprings usually achieve better performance than using the elites directly [11]. Furthermore, (1) is applied to the elites. In each iteration, the range of velocity is regulated between 0 and 1 with the sigmoid function [see (3)] and compared with a random chromosome between 0 and 1 to update the position in binary format [see (4)]. By performing PSO on the elites, the search ability of the algorithm may increase. Half of the population in the next generation consists of the enhanced individuals, and the rest is generated by the crossover operation.

Crossover: To produce well-performing individuals, the crossover operation is only performed on selected individuals produced by PSO. To select parents for the crossover operation, a tournament selection scheme is used, in which two enhanced elites are selected at random, and their fitness values are compared. The individual with the better fitness value is selected as a parent and inserted into the mating pool. Then, the two individuals are moved back to the population. In the same way, the other parent is chosen and moved to the mating pool. Two offsprings are created by performing crossover on the selected parents. A two-point crossover operation is used. The produced offsprings make up the other half of the population in the next generation.

Mutation: This operation occurs along with the crossover operation. Here, uniform mutation is adopted. In our case, a constant mutation probability that is equal to 0.01 is used.

III. Experimental Results

A. Description of Data Sets

1) Indian Pines: The hyperspectral data set used in experiments is the well-known AVIRIS data captured of Indian Pines (NW Indiana) in 1992 comprising 16 classes, mostly related to different land covers. The data set consists of 145 × 145 pixels with a spatial resolution of 20 m/pixel. In this letter, 220 data channels (including all noisy and atmospheric absorption bands) are used. Training samples are available for 16 classes, and the total number of training and test samples is 695 and 9691, respectively. The same training and test samples for all 16 classes as in [12] are chosen, and a half of the training samples is selected for validation.

2) Toronto: The RGB Toronto Roads data set is captured at the resolution of 1.2 m/pixel. This data set contains three bands consisting of 1500 × 1500 samples. Fig. 2(a) and (b) shows this data set and its corresponding digitized samples [13]. For this data set, 0.01 of the total samples are randomly chosen as training samples (1052 samples for class Road and 21448 for class No-road) and the rest as test samples (10 2007 samples for class Road and 2 125 493 for class No-road). Then, a half of the training samples is chosen for validation.

B. General Information

The proposed method was implemented in *MATLAB*, on a computer having Intel(R) Core(TM) i7 CPU 2.40 GHz and 16 GB (15.9 GB usable) of memory.

The number of populations in the first and second scenarios was set as 20 and 10, respectively. The same set of parameters for both data sets was chosen, which infers that the proposed method is data set distribution independent, and there is no need to set any parameters for it, and the method can automatically choose the most informative bands in terms of classification accuracies.

The hybridization of GA-PSO will automatically stop, when the difference between the OA of the best solution and the average value of fitness values in a swarm is less than a predefined threshold value.

For the first scenario, to compare the capability of the proposed methodology, four well-known feature selectors, namely, divergence, transformed divergence, Bhattacharyya distance, and Jeffries–Matusita distance, have been taken into account. In addition, two frequently used supervised feature extraction techniques, namely, DBFE and NWFE, have been considered. In the case of NWFE and DBFE, features with cumulative eigenvalues above 99% are retained and classified with SVM. This way of choosing features has been widely used in the literature (e.g., in [14] and [15]). In addition to the aforementioned techniques, GA + SVM and PSO + SVM are investigated to be compared with the proposed approach. Since the first scenario is related to feature selection and image classification,

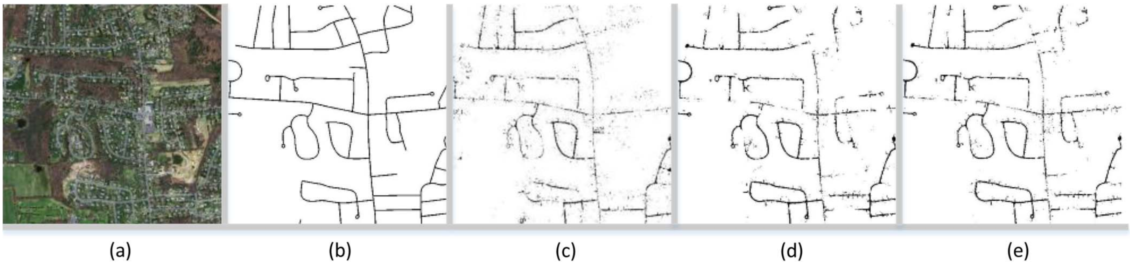


Fig. 2. (a) Input data. (b) Manually produced reference map. (c) Map obtained by $\mathrm{SVM}_{\mathrm{RGB}}$. (d) Map obtained by $\mathrm{SVM}_{\mathrm{AP}}$. (e) Map obtained by HGAPSO + $\mathrm{SVM}_{\mathrm{AP}}$.

TABLE I

FIRST SCENARIO: CLASSIFICATION ACCURACIES AND CPU PROCESSING TIME IN SECONDS. THE BEST RESULT IN EACH ROW IS SHOWN IN BOLDFACE. THE NUMBER OF FEATURES IS SHOWN IN BRACKET. SINCE PSO + SVM, GA + SVM, AND HGAPSO + SVM ARE THE AVERAGE OF TEN RUNS, THE NUMBER OF FEATURES IS NOT GIVEN

Index	SVM (220)	DBFE (17)	NWFE (120)	PSO +SVM	GA +SVM	HGAPSO +SVM
AA	76.02	73.36	60.13	74.94±2.54	73.96±4.37	**77.92**±1.42
Kappa	0.6119	0.6055	0.5533	0.7281±0.025	0.7141±0.040	**0.7495**±0.0069
OA	65.41	64.96	68.44	74.69±2.38	73.39±3.81	**76.68**±0.64
Time(s)	70	72	132	293±39	121±19	201±58

OA, average accuracy, kappa coefficient, and CPU processing time are considered for the evaluation of the final results. Since the PSO+SVM, GA+SVM, and HGAPSO+SVM are based on evolutionary techniques and their results can be different in different runs, all aforementioned approaches have been run ten times, and the average results are reported in Table I.

For the second scenario, since it is related to road detection, the root mean square error (RMSE) is taken into account as it was suggested that the RMSE is the most solid index [16]. For this scenario, since the Toronto data consist of only three components (RGB), to produce extra features, an attribute profile is used. A morphological attribute profile is considered as the generalization of morphological profile, which simplifies the input image by using the sequential stricter thresholds to model spatial information of the input image. For a detailed description of the attribute profile, refer to [8] and [14]. In this letter, three attributes, i.e., area ($\lambda_a = (1000/v)\ \{1, 3, 5, 7\}$, where v is the resolution of the input data), standard deviation ($\lambda_s = (\mu_i/100)\ \{30, 40\}$, where μ is the mean of the ith feature), and the diagonal of the box bounding the regions ($\lambda_d = \{25, 50, 100\}$), are used. However, other types of attributes with different ranges can be used. In this case, 19 features for each component (including itself) were produced. Since we have three components (R, G, and B), the total number of produced features is 57, which was considered as the input for the proposed methodology. Then, HGAPSO + SVM is applied on the features obtained by the attribute profile (and named as HGAPSO + $\mathrm{SVM}_{\mathrm{AP}}$) and compared with 1) the result of SVM performed on the RGB data (named as $\mathrm{SVM}_{\mathrm{RGB}}$) and 2) the result of SVM performed on the features produced by the attribute profile (named as $\mathrm{SVM}_{\mathrm{AP}}$).

The data sets have been classified with SVM using a Gaussian kernel. Fivefold cross validation is taken into account to select the hyperplane parameters when SVM is used for the last step (for the classification of informative bands).

C. *First Scenario*

The result of classification with different techniques is listed in Table I. These results have been obtained when conventional feature selection techniques, including divergence, transformed divergence, and Bhattacharyya distance, cannot work due to the singularity of the covariance matrix. The main reasons behind this shortcoming are that the conventional feature selectors cannot eliminate the corrupted bands automatically, and this step should be done manually, which is time consuming. In addition, when there is not a balance between the number of bands and the number of training samples, the aforementioned conventional feature selection techniques will not perform well. Furthermore, almost all of the conventional feature selection methods are computationally time demanding. For those approaches, to select a subset of m features out of a total of n features, $n!/(n-m)!m!$ alternatives must be calculated, which is a laborious task and demands a lot of computational memory. In other words, the feature selection techniques are only feasible in relatively low-dimensional cases. Another shortcoming of most of the conventional methods (particularly divergence, transformed divergence, and Bhattacharyya distance) is that the number of desired features must be initialized *a priori*. In contrast, since evolutionary-based feature selection techniques (e.g., PSO + SVM, GA + SVM, and HGAPSO + SVM) are not based on the calculation of the second-order statistics, the singularity of the covariance matrix is not a problem. In addition, when an evolutionary technique is taken into consideration, there is no need to calculate all different alternatives to find the most informative bands, and, therefore, these methods are usually faster than the conventional ones. Furthermore, in the proposed method, there is no need to initialize the number of desired features, and the approach can find the most informative bands with respect to the OA of SVM over the validation samples.

Some algorithms, such as the originally proposed DBFE, demand the use of second-order statistics (i.e., the covariance matrix) to characterize the distribution of training samples with respect to the mean. In that case, if the number of available training samples is not sufficient, a good estimation of the covariance matrix might be impossible. For this purpose, the use of a sample covariance or a common covariance [1] may not be successful. As an example, either when the sample covariance or the common covariance is taken into account to estimate the statistics for each available class for DBFE, if the number of pixels in the classes is not, at least one, greater than the total number of features being used, the DBFE stops working. To handle this issue, the leave-one-out covariance [1] estimator can be used as an alternative to estimate the covariance matrix. However, this is not a problem for evolutionary-based feature selectors since they are nonparametric and do not need to estimate class conditional densities. In addition, they can efficiently handle high-dimensional data with a very limited number of training samples due to the generalization of the SVM, which has been considered as the fitness function. As can be seen from Table I, the proposed method outperforms NWFE and DBFE in terms of classification accuracies and improves the OA of DBFE and NWFE by almost 12 and 8 percent, respectively.

As can be seen from Table I, HGAPSO + SVM outperforms the other evolutionary-based feature selection techniques (e.g., GA + SVM and PSO + SVM) in terms of classification accuracy. On the other hand, PSO + SVM has the highest CPU processing time among other evolutionary-based feature selectors. The main reason of this shortcoming is that although PSO is a fast optimization method, it converged after a higher number of iterations. On the contrary, although the convergence of GA + SVM is faster than that of PSO + SVM and HGAPSO + SVM, it has the worst classification accuracies due to the premature convergence of the chromosomes.

Since all the evolutionary-based optimization methods are based on a random process, the selected features are different in different trials. In the experiments, the proposed approach selected 73–94 features in ten different trials. It should be noted that the proposed approach allows for the detection of the best distinctive features without requiring the number to be set *a priori* by the user.

D. Second Scenario

The obtained RMSE for $\mathrm{SVM}_{\mathrm{RGB}}$, $\mathrm{SVM}_{\mathrm{AP}}$, and $\mathrm{HGAPSO}+\mathrm{SVM}_{\mathrm{AP}}$ are 0.7669, 0.6461, and 0.6049, respectively. $\mathrm{HGAPSO}+\mathrm{SVM}_{\mathrm{AP}}$ provides the smallest RMSE among all techniques, which confirms the capability of the proposed method to detect the classes of interest. The main reason that the proposed approach outperforms $\mathrm{SVM}_{\mathrm{AP}}$ is that although attribute profiles are a powerful technique to model spatial information of an image, they produces redundant features that can reduce the classification accuracies. However, by using the proposed technique, the most informative features can be selected leading to higher classification accuracies. Fig. 2 shows the input data, the manually produced reference map, and the maps obtained by $\mathrm{SVM}_{\mathrm{RGB}}$, $\mathrm{SVM}_{\mathrm{AP}}$, and $\mathrm{HGAPSO}+\mathrm{SVM}_{\mathrm{AP}}$, respectively. As can be seen, the proposed method detects more details from the road network as compared with the other approaches and outperforms $\mathrm{SVM}_{\mathrm{RGB}}$ and $\mathrm{SVM}_{\mathrm{AP}}$.

IV. Conclusion

In this letter, a new feature selection technique, which does not need to set the number of desired features *a priori*, has been introduced, based on the integration of GA and PSO. According to the experiments, the following can be concluded.

1) The proposed method can find informative bands in terms of classification accuracies in an acceptable CPU time.
2) The proposed method can be used for road detection.
3) In the novel feature selection approach, there is no need to set the number of output features since the proposed approach can automatically select the most useful features in terms of classification accuracies.
4) The proposed method is data set distribution independent, and for it, there is no need to initialize any parameters.
5) Since the proposed algorithm is based on evolutionary techniques, it is much faster than other well-known feature selection techniques that require an exhaustive process to select the most informative bands. Therefore, the new approach can work appropriately in a situation in which other feature selection techniques are not applicable.
6) Since SVM is considered as a fitness function in the proposed method, it can handle high-dimensional data with a limited number of training samples, when other feature selection techniques cannot proceed due to singularity problems of covariance matrices.

References

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

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[3] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, Nov. 2006.

[4] A. Daamouche, F. Melgani, N. Alajlan, and N. Conci, "Swarm optimization of structuring elements for VHR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1334–1338, Nov. 2013.

[5] F. Bergh and A. P. Engelbrecht, "A cooperative approach to particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 225–239, Jun. 2004.

[6] K. Premalatha and A. M. Natarajan, "Hybrid PSO and GA for global maximization," *Int. J. Open Probl. Comput. Math.*, vol. 2, no. 4, pp. 597–608, Dec. 2009.

[7] R. C. Eberhart and Y. Shi, "Comparison between genetic algorithms and particle swarm optimization," in *Evolutionary Programming VII*, V. W. Porto, Ed. Berlin, Germany: Springer-Verlag, 1998.

[8] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[9] D. Beasley, D. Bull, and R. Martin, "An overview of genetic algorithms," *Univ. Comput.*, vol. 15, no. 2, pp. 58–69, 1993.

[10] J. Kennedy and R. Eberhart, *Swarm Intelligence*. San Francisco, CA, USA: Morgan Kaufmann, 2001.

[11] P. Ghamisi, F. Sepehrband, J. Choupan, and M. Mortazavi, "Binary hybrid GA-PSO based algorithm for compression of hyperspectral data," presented at the 5th ICSPCS, Dec. 2011, pp. 1–8.

[12] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral–spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2565–2574, May 2014.

[13] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 2013.

[14] P. Ghamisi, J. A. Benediktsson, and J. R. Sveinsson, "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5771–5782, Sep. 2014.

[15] P. Ghamisi, J. A. Benediktsson, G. Cavallaro, and A. Plaza, "Automatic framework for spectral–spatial classification based on supervised feature extraction and morphological attribute profiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, to be published.

[16] M. Mokhtarzade and M. V. Zoej, "Road detection from high-resolution satellite images using artificial neural networks," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 9, no. 1, pp. 32–40, Feb. 2007.

CHAPTER 6

Conclusion and Future Works

6.1 Conclusions

The main objective of this thesis was the proposal of robust spectral-spatial classification approaches for hyperspectral data. In addition, special emphasis was given to the accuracy and speed of the proposed approaches. Furthermore, we tried to make the spectral-spatial classification approaches automatic in order to reduce users' efforts and the exhaustive time in order to handle high volumetric hyperspectral data for real-time applications. As the main concluding remarks, the following points can be mentioned:

- In the second chapter of the thesis, a fully automatic framework was introduced for the spectral-spatial classification of hyperspectral images. In the framework, SVM was used for the extraction of spectral information and HMRF was taken into account for the extraction of spatial information. In the final step, the outputs of SVM and HMRS were combined by using the majority voting. The efficiency of the proposed method has been tested in both situations with and without considering the gradient step. The proposed method was evaluated on two data sets (Indian Pines and Salinas). In both cases, the new approach has been provided good results in terms of classification accuracies in an automatic way. In that work, the concept of HMRF was used for the first time in the field of remote sensing, and the efficiency of that for the segmentation of hyperspectral images was demonstrated.

- In the third chapter, two novel multilevel thresholding segmentation methods have been proposed for grouping the pixels of benchmark image processing images as well as multispectral and hyperspectral images into different homogenous regions. Those methods were based on DPSO and FODPSO, which can be used in finding the optimal set of threshold values and use many swarms of test solutions which may exist at any time. In those approaches, each swarm individually performs just like an ordinary (PSO) algorithm with a set of rules governing the collection of swarms that are designed to simulate natural selection. In addition, for FODPSO, the concept of fractional derivative was used to control the convergence rate of particles. With respect to the obtained results, the FODPSO outperformed the classical PSO and DPSO within multilevel segmentation problems on benchmark image processing images and remote sensing data from different points of view such as CPU processing time and corresponding fitness value. Experimental results also indicated that the FODPSO was more robust than the two other methods and had a higher potential for finding the optimal set of thresholds with more between-class variance in less computational time, especially for higher segmentation levels and for images with a wide variety of intensities. In addition, to show the efficiency of the proposed segmentation method on the result of classification, a novel classification approach based on the new segmentation method and SVM was proposed. Results confirmed that the FODPSO-based segmentation method improved the SVM in terms of classification accuracies when compared to the standard SVM classification of the raw image data. Furthermore, a combination of MSS and FODPSO has been taken into account for the spectral-spatial classification of hyperspectral images. In that work, results indicated that the use of both segmentation methods can overcome the shortcomings of each other and the combination can improve the result of classification significantly. It should be noted that, in that chapter, the concepts of DPSO- and FODPSO-base segmentation techniques were used for the first time in the pattern recognition community.

- In the fourth chapter, the usefulness of AP and its extensions and modifications have been taken into account for the classification of hyperspectral data sets. In that chapter, two novel spectral-spatial classification frameworks have been introduced, which are able to automatically classify input hyperspectral data sets very accurately within a short period of time. In all the above-mentioned cases, the use of AP and its extensions demonstrated its effectiveness for modeling some regional characteristics (e.g., scale, shape, and contrast), provided a multi-level decomposition of an image. As shown in that chapter, AP and its extensions can be considered as simple yet effective approach for the classification of hyperspectral images.

- In the fifth chapter, two novel approaches have been proposed for the purpose of feature selection addressing the curse of dimensionality. The first approach was based on a new optimization technique named BFODPSO as well as SVM. The second approach was based on the hybridization of GA and PSO as well as SVM. For the both approaches, there is no need to set the number of output features as *a priori* and the proposed approaches can automatically select the most informative features in terms of classification accuracies. In addition, since the both approaches are based on an evolutionary method, it is much faster than other well-known feature selection techniques which demand an exhaustive process to select the most informative bands. In this sense, the proposed approaches can work appropriately in a situation when other feature selection techniques are not applicable. Since the new feature selection approaches are based on a SVM classification which is capable of handling high dimensional data with a limited number of training samples, they can proceed to select the most informative features in ill-posed situations when other feature selection/extraction techniques cannot proceed without a powerful technique for estimating the statistics for each class.

6.2 Perspectives

- Due to the speed and efficiency of the FODPSO-based segmentation, it can be evaluated in image segmentation applications for the real-time autonomous deployment and distributed localization of sensor nodes. The objective is to deploy the nodes only in the terrains of interest, which are identified by segmenting the images captured by a camera onboard an unmanned aerial vehicle using the FODPSO algorithm. Such a deployment has importance for emergency applications, such as disaster monitoring and battlefield surveillance. In addition, finding a way for the estimation of the number of thresholds in FODPSO-based segmentation and joint multichannel segmentation instead of segmenting data set band by band would be of interest.

- The selection of attributes and their related thresholds is also another area, which demands a further improvement. In addition, although few strategies for the authomatic selection of the attribute thresholds have been proposed, they are limited to some attributes (i.e., area and standard deviation) and might not be applicable to others, thus opening the need for developing more generic selection strategies for the filter parameters.

- It would be interesting to further improve and adapt the proposed approaches for a wide variety of applications such as land-cover mapping, urban management and modeling, species identification in forested areas and so on.

- In order to manage and monitor many predictable and unpredictable natural disasters (including but not limited to earthquakes, floods, weather events, landslides and wildfires), we are particularly in need of developing fast, simple, automatic and efficient methods for disaster management. Undoubtedly, the aforementioned points have a major participate in the economics of different countries. As a result, assessing the usefulness of the proposed approaches in real-time applications where a rapid and accurate response is needed would be very interesting.

- Another topic deserving future research is the development of parallel implementations of the presented approach in high-performance computing architectures, although the processing times reported in our experiments throughout the thesis (measured in a standard desktop CPU) are quite fast for the considered data sets.

APPENDIX A

Accuracy assessment

This chapter provides a rough idea regarding the assessment matrices, which have been extensively used in order to evaluate the result of the output classification map. In general, almost all metrics for the assessment of the final classification map are based on the *confusion matrix*. This matrix provides a possibility for evaluating the exactitude of a given classification map with respect to the reference map. In this appendix, the confusion matrix will firstly be described, and then, several specific and global estimators are extracted from the confusion matrix.

A.1 Confusion Matrix

In pattern recognition, a confusion matrix is considered as a visualization tool typically used in supervised learning. In this matrix, each column infers the instances in a predicted class, while each row represents the instances in an actual class. This matrix is also be able to infer where the classification technique lead to confusion (i.e., commonly mis-labeling one class as another). Table A.1 represents an example of confusion matrix for a 3-class classification problem. The term C_i represents the class i and the term C_{ij} refers to the number of pixels which are wrongly assigned to the class j, which are referenced as class i. N_c represents the number of classes in the referenced map.

Table A.1: Confusion Matrix for a 3-class classification problem.

Percentage	Classification data				
Reference data	C_1	C_2	C_3	Row total	Producer's accuracy
C_1	C_{11}	C_{12}	C_{13}	$\sum_i^{N_c} C_{1i}$	$\frac{C_{11}}{\sum_i^{N_c} C_{1i}}$
C_2	C_{21}	C_{22}	C_{23}	$\sum_i^{N_c} C_{2i}$	$\frac{C_{22}}{\sum_i^{N_c} C_{2i}}$
C_3	C_{31}	C_{32}	C_{33}	$\sum_i^{N_c} C_{3i}$	$\frac{C_{33}}{\sum_i^{N_c} C_{3i}}$
Column total	$\sum_i^{N_c} C_{i1}$	$\sum_i^{N_c} C_{i2}$	$\sum_i^{N_c} C_{i3}$	N	
User's accuracy	$\frac{C_{11}}{\sum_i^{N_c} C_{i1}}$	$\frac{C_{22}}{\sum_i^{N_c} C_{i2}}$	$\frac{C_{33}}{\sum_i^{N_c} C_{i3}}$		

A.1.1 Overall Accuracy (OA)

The OA is the percentage of correctly classified pixels, which can be estimated as follows:

$$\mathrm{OA} = \frac{\sum_i^{N_c} C_{ii}}{\sum_{i,j}^{N_c} C_{ij}} \times 100.$$

A.1.2 Class Accuracy (CA)

The CA (or producer's accuracy) is regarded as the percentage of correctly classified pixels for each class. This metric infers how well a certain area was classified. This metric includes the error of omission in which the more errors of omission, the lower the producer accuracy. This metric is calculated by dividing the number of correct pixels in one class by the total number of pixels as derived from reference data as follows:

$$\mathrm{CA}_i = \frac{C_{ii}}{\sum_j^{N_c} C_{ij}} \times 100.$$

A.1.3 Average Accuracy (AA)

The AA is the mean of class accuracies for all the classes, which can be estimated as follows:

$$\mathrm{AA} = \frac{C_{ii}}{\sum_j^{N_c} CA_i} \times 100.$$

It should be noted that either OA or AA is closed to 100%, it infers that the classification accuracy is more accurate. The problem associated with the concept of OA occurs when a referenced set is unbalanced. In this case, the OA may not be a good representer for the true performance of the classifier. For example, if a class has very few number of referenced pixels, its influence will be very low on the OA, while it will be more influence in the AA since the mean is done the number of classes rather than the whole number of pixels. Strong difference between OA and AA may indicate that a specific class is wrongly classified with a high proportion.

A.1.4 Kappa Coefficient (*k*)

This metric is a statistical measurement of agreement between the final classification map and the reference map. It is the percentage agreement corrected by the level of agreement that could be expected due to chance alone. It is generally thought to be a more robust measure than simple percent agreement calculation since *k* takes into account the agreement occurring by chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where

$$P_o = \text{OA}, \qquad P_e = \frac{1}{N^2}\sum_{i}^{N_c} C_{i+}C_{+i}, \quad C_{i+} = \sum_{j}^{N_c} C_{ij}, \qquad C_{+i} = \sum_{j}^{N_c} C_{ji}.$$

Bibliography

[1] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT, no. 14, pp. 55–63, 1968.

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

[3] D. W. Scott, "Multivariate density estimation.," *in Proc. SPIE Eur. Remote Sens.*, 1992.

[4] E. J. Wegman, "Hyperdimensional data analysis using parallel coordinates,," *Jour. American Stat. Assoc.*, vol. 85, no. 411, pp. 664–675, 1990.

[5] L.O. Jimenez and D.A. Landgrebe, "Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Trans. Sys., Man, and Cyber., Part C: Applications and Reviews*, vol. 28, no. 1, pp. 39–54, Feb 1998.

[6] K. Fukunaga, *Intr. Statistical Pattern Recog.*, Academic Press, Inc., San Diego, California, 1990.

[7] P. Diaconis and D. Freedman, "Asymptotics of graphical projection pursuit.," *The Annals of Statistics*, vol. 12, no. 3, pp. 793–815, 1984.

[8] E. Magli, G. Olmo, and E. Quacchio, "Optimized onboard lossless and near-lossless compression of hyperspectral data using calic," *IEEE Geosci. and Remote Sens. Lett.*, vol. 1, no. 1, pp. 21–25, Jan 2004.

[9] M. Fauvel, J.A. Benediktsson, J. Chanussot, and J.R. Sveinsson, "Spectral and spatial classification of hyperspectral data using svms and morphological profiles," *IEEE Trans. Geos. and Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, 2008.

[10] I.T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics. Springer, 2002.

[11] K. Fukunaga, *Intr. Statistical Pattern Recog.*, Academic Press, 1974.

[12] L.O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosc. and Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, 1999.

[13] B. C. Kuo and D. A. Landgrebe, "A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction," *Comput.Vis. ImageUnderst.*, vol. 64, no. 3, pp. 377–389, 1996.

[14] Chulhee Lee and David A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pat. Analysis Mach. Intel.*, vol. 15, no. 4, pp. 388–400, 1993.

[15] J. A. Benediktsson and I. Kanellopoulos, "Classification of multisource and hyperspectral data based on decision fusion," *IEEE Trans. Geos. and Remote Sens.*, vol. 37, no. 3, pp. 1367–1377, May 1999.

[16] J. C. Russ, *The Image Processing Handbook*, CRC Press LLC, 3rd edition, 1999.

[17] D. J. Wiersma and D. A. Landgrebe, "Analytical design of multispectral sensors," *IEEE Trans. Geos. and Remote Sens.*, vol. GE, no. 18, pp. 180–189, 1980.

[18] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 7, pp. 793–802, 2004.

[19] P. H. Swain, *Fundamentals of pattern recognition in remote sensing*, Remote Sensing—The Quantitative Approach, P. H. Swain and S. Davis, Eds. New York: McGraw-Hill, New York,, 1978.

[20] Y. Bazi and F. Melgani, "Toward an optimal svm classification system for hyperspectral remote sensing images," *IEEE Trans. Geos. and Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, 2006.

[21] A. Daamouche, F. Melgani, N. Alajlan, and N. Conci, "Swarm optimization of structuring elements for vhr image classification," *IEEE Geos. and Remote Sens. Lett.*, vol. 10, no. 6, pp. 1334–1338, 2013.

[22] A. Paoli, F. Melgani, and E. Pasolli, "Clustering of hyperspectral images based on multiobjective particle swarm optimization," *IEEE Trans. Geos. and Remote Sens.*, vol. 47, no. 12, pp. 4175–4188, 2009.

[23] P. Ghamisi and J. A. Benediktsson, "Feature selection based on hybridization of genetic algorithm and particle swarm optimization," *IEEE Geos. and Remote Sens. Lett.*, vol. 12, no. 2, pp. 309–313, 2015.

[24] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, "A novel feature selection approach based on fodpso and svm," *IEEE Trans. Geos. Remote Sens.*, vol. 53, no. 5, pp. 2935–2947, May 2015.

[25] D. Beasley, D. R. Bull, and R. R. Martin, *An overview of genetic algorithms*, vol. 15, Univ. Camping, 1993, Part 1.

[26] B. L. Mille and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Comp. Syst.*, vol. 9, no. 1995, pp. 193– 212, 1995.

[27] J.H Holland, *Adaptation in Natural and Artificial Systems*, MIT Press, 2ns edition, 1992.

[28] P. K. Chawdhry, R. Roy, and R. K. Pant, *Soft Computing in Engineering Design and Manufacturing*, Springer, 1998.

[29] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. 5-th Berkeley Symp. Math. Stat. Prob.*, pp. 281–297, 1967.

[30] G. Ball and D. Hall, "ISODATA, a novel method of data analysis and classification," *Stanford Univ., Stanford, CA, Tech. Rep. AD-699616*, 1965.

[31] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy, "Conjugate gradient neural networks in classification of very high dimensional remote sensing data," *Int. Jour. Remote Sens.*, vol. 14, no. 15, pp. 2883–2903, 1993.

[32] H. Yang, F. V. D. Meer, W. Bakker, and Z. J. Tan, "A back—propagation neural network for mineralogical mapping from aviris data," *Int. Jour. Remote Sens.*, vol. 20, no. 1, pp. 97–110, 1999.

[33] J. A. Benediktsson, "Statistical methods and neural network approaches for classification of data from multiple sources," *PhD thesis, Purdue Univ., School of Elect. Eng. West Lafayette*, 1990.

[34] J. A. Richards, "Analysis of remotely sensed data: The formative decades and the future," *IEEE Trans. Geos. and Remote Sens.*, vol. 43, no. 3, pp. 422–432, 2005.

[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[36] L. Breiman, "RF tools a class of two eyed algorithms," *SIAM Workshop*, 2003.

[37] V. N. Vapnik, *Statistical learning theory.*, New York: Wiley, 1998.

[38] B. Scholkopf and A. J. Smola, *Learning with Kernels.*, MIT Press, 2002.

[39] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for the classfication of hyperspectral remote-sensing data over urban areas," *EURASIP Jour. Adv. Signal Proc.*, pp. 1–14, 2009.

[40] S. Tadjudin and D.A. Landgrebe, "Classification of high dimensional data with limited training samples," in *Tech. Rep., School of Electrical and Computer Engineering, Purdue University*, 1998.

[41] Y. Tarabalka, "Classification of hyperspectral data using spectral–spatial approaches," *PhD thesis*, 2010.

[42] H. Derin and P. A. Kelly, "Discrete-index Markov-type random processes," *Proceeding of the IEEE*, vol. 77, no. 10, pp. 1485–1510, 1989.

[43] 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*, vol. 101, March 2013.

[44] Q. Jackson and D. Landgrebe, "Adaptive bayesian contextual classification based on markov random fields," *IEEE Trans. Geos. and Remote Sens.*, vol. 40, no. 11, pp. 2454–2463, 2002.

[45] Yuliya Tarabalka, Mathieu Fauvel, Jocelyn Chanussot, and Jón Atli Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geos. and Remote Sens. Lett.*, 2010.

[46] A. Farag, R. Mohamed, and A. El-Baz, "A unified framework for map estimation in remote sensing image segmentation," *IEEE Trans. Geosci. and Remote Sens.*, vol. 43, no. 7, pp. 1617–1634, 2005.

[47] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," *Proc. PReMI*, pp. 260–265, 2005.

[48] D. Liu, M. Kelly, and P. Gong, "A spatial-temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery," *Remote Sens. Env.*, vol. 101, no. 10, pp. 167–180, 2006.

[49] G. Moser and S. B. Serpico, "Combining support vector machines and Markov random fields in an integrated framework for contextual image classification," *IEEE Trans. Geos. and Remote Sens.*, vol. 51, no. 5, pp. 2734–2752, 2013.

[50] M. Khodadadzadeh, R. Rajabi, and H. Ghassemian, "Combination of region-based and pixel-based hyperspectral image classification using erosion technique and MRF model," *Elec. Eng. (ICEE)*, pp. 294–299, may 2010.

[51] G. Zhang and X. Jia, "Simplified conditional random fields with class boundary constraint for spectral-spatial based remote sensing image classification," *IEEE Geos. and Remote Sens. Lett.*, vol. 9, no. 5, 2012.

[52] B. Tso and R. C. Olsen, "Combining spectral and spatial information into hidden Markov models for unsupervised image classification," *Intern. Jour. Remote Sens.*, vol. 26, no. 10, pp. 2113–2133, 2005.

[53] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral–spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Remote Sens. and Geos.*, vol. 52, no. 5, pp. 2565–2574, May 2014.

[54] P. Ghamisi, J.A. Benediktsson, and M.O. Ulfarsson, "The spectral-spatial classification of hyperspectral images based on hidden Markov random field and its expectation-maximization," in *IGARSS 2013*, July 2013, pp. 1107–1110.

[55] G. G. Hazel, "Multivariate gaussian mrf for multispectral scene segmentation and anomaly detection," *IEEE Trans. Geos. and Remote Sens.*, vol. 38, no. 3, pp. 1199–1211, May.

[56] F. Tsai, C. K. Chang, , and G. R. Liu, "Texture analysis for three dimension remote sensing data by 3d glcm," *27th Asian Conf. Remote Sens.*, pp. 1–6, 2006.

[57] X. Huang and L. Zhang, "A comparative study of spatial approaches for urban mapping using hyperspectral rosis images over pavia city, northern italy," *Inter. Jour. Remote Sens.*, vol. 30, no. 12, pp. 3205–3221, 2009.

[58] 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*, vol. 101, no. 3, pp. 652–675, 2013.

[59] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, "An efficient method for segmentation of images based on fractional calculus and natural selection," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12407–12417, 2012.

[60] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation,," *J. Electron. Imag.*, vol. 13, no. 1, pp. 146–168, 2004.

[61] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, "Use of darwinian particle swarm optimization technique for the segmentation of remote sensing images," *IGARSS 2012*, pp. 4295–4298, July 2012.

[62] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques,," *IEEE Trans. Geos. and Remote Sens.*, vol. 47, no. 5, pp. 2973–2987, 2009.

[63] P. Ghamisi, M. S. Couceiro, F. M.L. Martins, and J. A. Benediktsson, "Multilevel image segmentation approach for remote sensing images based on fractional–order Darwinian particle swarm optimization," *IEEE Trans. on Geos. and Remote Sens.*, vol. 52, no. 5, pp. 2382–2394, May 2014.

[64] P. Ghamisi, M. Couceiro, M. Fauvel, and J. A. Benediktsson, "Integration of segmentation techniques for classification of hyperspectral images," *IEEE Geos. and Remote Sens. Let.*, vol. 11, no. 1, pp. 342–346, Jan 2014.

[65] Y. Tarabalka, J. Chanussot, and J.A. Benediktsson, "Segmentation and classification of hyperspectral images using watershed transformation,," *Pattern Recog.*, vol. 43, no. 7, pp. 2367–2379, 2010.

[66] A. Darwish, K. Leukert, and W. Reinhardt, "Image segmentation for the purpose of object–based classification,," *IGARSS 2003*, vol. 3, pp. 2039–2041, 2003.

[67] J. Tilton, "Analysis of hierarchically related image segmentations,," *IEEE Work. Adv. Tech. Analysis Remotely Sensed Data*, pp. 60–69, 2003.

[68] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray–level picture thresholding using the entropy of the histogram,," *Comp. Vis. Graph. Image Proc.*, vol. 2, pp. 273–285, 1985.

[69] T. Pun, "A new method for grey–level picture thresholding using the entropy of the histogram,," *Computer Vision Graphics Image Processing*, vol. 2, pp. 223–237, 1980.

[70] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. Syst. Man Cybern*, vol. 9, pp. 62–66, 1979.

[71] R. V. Kulkarni and G. K. Venayagamoorthy, "Bio–inspired algorithms for autonomous deployment and localization of sensor,," *IEEE trans. syst.*, vol. 40, no. 6, pp. 663–675, 2010.

[72] Y. Kao, E. Zahara, and I. Kao, "A hybridized approach to data clustering," *Expert Systems with Applications*, vol. 34, no. 2008, pp. 1754–1762, 2008.

[73] D. Floreano and C. Mattiussi, "Bio–inspired artificial intelligence: Theories and methods and technologies," *Cambridge, MA: MIT Press*, 2008.

[74] J. Kennedy and R. Eberhart, "A new optimizer using particle swarm theory," *IEEE Sixth Inte Symp Mic Mach Human Science*, vol. 34, no. 2008, pp. 39–43, 1995.

[75] J. Tillett, T. M. Rao, F. Sahin, R. Rao, and S. Brockport, "Darwinian particle swarm optimization," *Proc. 2nd Ind. Inter. Conf. Art. Intel.*, pp. 1474–1487, 2005.

[76] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, "Extending the fractional order Darwinian particle swarm optimization to segmentation of hyperspectral images," *Proc. SPIE 8537, Image Signal Proc. Remote Sens. XVIII, 85370F*, pp. 85370F–85370F–11, 2012.

[77] P. Ghamisi, A. Ali, M. Couceiro, and J. Benediktsson, "A novel evolutionary swarm fuzzy clustering approach for hyperspectral imagery," *IEEE Jour. Selec. Top. Appl. Earth Observ. Remote Sens.*, vol. PP, no. 99, pp. 1–10, 2015.

[78] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. and Remote Sens.*, vol. 47, no. 9, pp. 2973–2987, 2009.

[79] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Multiple spectral-spatial classification approach for hyperspectral data," *IEEE Trans. on Geosc. and Remote Sens.*, vol. 48, no. 11, pp. 4122–4132, 2010.

[80] J. C. Tilton, Y. Tarabalka, P. M. Montesano, and E. Gofman, "Best merge region growing segmentation with integrated non–adjacent region object aggregation," *IEEE Trans. on Geos. and Remote Sens.*, vol. 50, no. 11, pp. 4454 – 4467, 2012.

[81] P. Soille, *Morphological Image Analysis, Principles and Applications*, Springer, 2nd, edition, 2003.

[82] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. and Remote Sens.*, vol. 39, no. 2, pp. 309–320, 2001.

[83] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geos. and Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, 2010.

[84] M. Chini, N. Pierdicca, and W. Emery, "Exploiting SAR and VHR optical images to quantify damage caused by the 2003 bam earthquake," *IEEE Trans. Geosci. and Remote Sens.*, vol. 47, no. 1, pp. 145–152, 2009.

[85] D. Tuia, F. Pacifici, M. Kanevski, and W. Emery, "Classification of very high spatial resolution imagery using mathematical morphology and support vector machines,," *IEEE Trans. Geosci. and Remote Sens.*, vol. 47, no. 11, pp. 3866–3879, 2009.

[86] H. Akcay and S. Aksoy, "Automatic detection of geospatial objects using multiple hierarchical segmentations,," *IEEE Trans. Geosci. and Remote Sens.*, vol. 46, no. 7, pp. 2097–2111, 2008.

[87] J. Chanussot, J. Benediktsson, and M. Fauvel, "Classification of remote sensing images from urban areas using a fuzzy possibilistic model,," *IEEE Geosci. and Remote Sens. Let.*, vol. 3, no. 1, pp. 40–44, 2006.

[88] J. A. Benediktsson, M. Pesaresi, and K. Arnason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations,," *IEEE Trans. Geosci. and Remote Sens.*, vol. 41, no. 9, pp. 1940–1949, 2003.

[89] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles,," *IEEE Trans. Geosci. and Remote Sens.*, vol. 43, no. 3, pp. 480–491, 2005.

[90] R. Bellens, S. Gautama, L. M. Fonte, W. Philips, J. C. W. Chan, and F. Canters, "Improved classification of VHR images of urban areas using directional morphological profiles,," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 2803–2813, 2008.

[91] P. Soille and M. Pesaresi, "Advances in mathematical morphology applied to geoscience and remote sensing,," *IEEE Trans. Geosci. and Remote Sens.*, vol. 40, no. 9, pp. 2042–2055, 2002.

[92] E. J. Breen and R. Jones, "Attribute openings, thinnings and granulometries," *IEEE Trans. Geosci. and Remote Sens.*, vol. 40, no. 11, pp. 2486–2494, 2013.

[93] N. Bouaynaya and D. Schonfeld, "Theoretical foundations of spatially-variant mathematical morphology part ii: Gray-level images," *IEEE Trans. Pattern Analysis Mach Intel*, vol. 30, no. 5, pp. 837–850, 2008.

[94] M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, "Modeling structural information for building extraction with morphological attribute filters," *Proc. SPIE 7477, Image and Signal Processing for Remote Sensing XV, 747703*, August, Berlin 2009.

[95] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of

hyperspectral data," *Inter. Jour. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, 2010.

[96] P. Ghamisi1, J. A. Benediktsson1, and S. Phinn, "Fusion of hyperspectral and lidar data in classification of urban areas," in *IGARSS 2014.*, July 2014.

[97] P. Ghamisi, M. Dalla Mura, and J. A. Benediktsson, "A survey on spectral–spatial classification techniques based on attribute profiles," *IEEE Trans. Geos. and Remote Sens.*, vol. 53, no. 5, pp. 2335–2353, May 2015.

[98] P. Salembier, A. Oliveras, , and L. Garrido, "Anti-extensive connected operators for image and sequence processing," *IEEE Trans. Image Proc.*, vol. 7, no. 4, pp. 555–570, 1998.

[99] M. Pedergnana, P. R. Marpu, M. D. Mura, J. A Benediktsson, and L. Bruzzone, "A novel technique for optimal feature selection in attribute profiles based on genetic algorithms," *IEEE Trans. Geo. and Remote Sens.*, vol. 51, no. 6, pp. 3514–3528, June 2013.

[100] P. Ghamisi, J. A. Benediktsson, G. Cavallaro, and A. Plaza, "Automatic framework for spectral-spatial classification based on supervised feature extraction and morphological attribute profiles," *IEEE Jour. Selec. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2147–2160, 2014.

[101] P. Ghamisi, J. A. Benediktsson, and J. R. Sveinsson, "Automatic spectral–spatial classification framework based on attribute and supervised feature extraction," *IEEE Trans. on Geos. and Remote Sens.*, vol. 52, no. 9, pp. 5771–5782, Sept 2014.