Augmented Vision and Reality 3

Saurabh Prasad
Lori M. Bruce
Jocelyn Chanussot *Editors*

Optical Remote Sensing

Advances in Signal Processing and Exploitation Techniques

Springer

Augmented Vision and Reality

Series Editors

Riad Hammoud, DynaVox Technologies, Pittsburgh, PA, USA

Lawrence B. Wolff, Equinox Corporation, New York, USA

For further volumes:
http://www.springer.com/series/8612

Saurabh Prasad · Lori M. Bruce
Jocelyn Chanussot
Editors

Optical Remote Sensing

Advances in Signal Processing and Exploitation Techniques

Springer

Editors

Asst. Prof. Saurabh Prasad
Department of Electrical and Computer Engineering
Geosystems Research Institute
Mississippi State University
Box 9652, Mississippi
MS 39762
USA
e-mail: saurabh.prasad@ieee.org

Prof. Dr. Jocelyn Chanussot
Institut Polytechnique de Grenoble
av. Félix Viallet 46
38000 Grenoble CX 1
France
e-mail: jocelyn.chanussot@gipsa-lab.inpg.fr

Prof. Lori M. Bruce
Department of Electrical and Computer Engineering
Geosystems Research Institute
Mississippi State University
Box 9652, Mississippi
MS 39762
USA
e-mail: bruce@bagley.msstate.edu

ISSN 2190-5916 e-ISSN 2190-5924

ISBN 978-3-642-14211-6 e-ISBN 978-3-642-14212-3

DOI 10.1007/978-3-642-14212-3

Springer Heidelberg Dordrecht London New York

Cover design: eStudio Calamar, Berlin/Figueres

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The evolution of optical remote sensing over the past few decades has enabled the availability of rich spatial, spectral and temporal information to remote sensing analysts. Although this has opened the doors to immense possibilities for analysis of optical remotely sensed imagery, it has also necessitated advancements in signal processing and exploitation algorithms to keep up with advances in the quality and quantity of available data. As an example, the transition from multispectral to hyperspectral imagery requires conventional statistical pattern classification algorithms to be modified to effectively extract useful information from the high dimensional hyperspectral feature space. Although hyperspectral imagery is expected to provide a much detailed spectral response per pixel, conventional algorithms developed and perfected for multispectral data would often be sub-optimal for hyperspectral data. At best, they would require a significant increase in the ground-truth (training) data employed for analysis—something that is often hard to come by, and is often far too costly. As a result, signal processing and pattern recognition algorithms for analysis of such data are also evolving to cope with such issues and result in practical applications.

The last decade has seen significant advances in algorithms that represent, visualize and analyze optical remotely sensed data. These advances include new algorithms to effectively compress high dimensional imagery data for efficient storage and transmission; new techniques to effectively visualize remotely sensed data; new analysis and classification techniques to analyze and classify remotely sensed imagery; and techniques to fuse remotely sensed imagery acquired simultaneously from different sensing modalities. This book brings together leading experts in these fields with the goal of bringing the cutting edge in signal processing and exploitation research closer to users and developers of remote sensing technology. This book is not intended to be a textbook for introductory remote sensing analysis. There are existing textbooks that provide a tutorial introduction to signal and image processing methods for remote sensing. This book is intended to be a valuable reference to graduate students and researchers in the academia and the industry who are interested in keeping abreast with the current state-of-the-art in signal and image processing techniques for optical remote sensing. This book consists of 15 chapters. Chapter 1 is an introductory chapter that sets the stage for the remainder of this book. In this chapter, we identify three key broad challenges and open problems associated with the analysis of modern optical remotely sensed imagery, and provide a motivation for each of the 14 chapters that follow within the context of these broad challenges. Chapters 2 through 6 present advances in algorithms for effective representation and visualization of high dimensional remotely sensed optical data, including on-board compressive sensing, coded aperture imaging and visualization techniques. Chapters 7 through 12 cover advances in statistical pattern classification and data analysis techniques, including multi-classifier systems and information fusion, morphological profiles, kernel methods, manifold learning and spectral pixel unmixing. Chapters 13 through 15 cover advances in multi-sensor data fusion techniques.

We would like to acknowledge and sincerely thank all contributors who participated in this collection. This book represents the state-of-the-art in signal and image processing research for optical remote sensing and would not have been possible if these contributors, who are leading experts in the field had not come together to work on these chapters. Their feedback and review of all chapters in this book was instrumental in making this a coherent and complete reference.

Mississippi State University, U.S.A., and
Grenoble Institute of Technology, France,
01-July-2010

Saurabh Prasad
Lori M. Bruce
Jocelyn Chanussot

Contents

Introduction

Saurabh Prasad, Lori M. Bruce and Jocelyn Chanussot

As the name suggests, remote sensing entails the use of sensing instruments for acquiring information remotely about an area of interest on the ground. The term "information" can refer to a wide variety of observable quantities (signals), such as reflected solar radiation across the electromagnetic spectrum and emitted thermal radiation from the earth's surface as measured from handheld [1], airborne [2] or spaceborne imaging sensors [3, 4]; received back-scattered microwave radiation from radio detection and ranging (RADAR), synthetic aperture radar (SAR) [5–8] or light detection and ranging (LIDAR) [9–11] equipment; electrical conductivity as measured from airborne sensors, etc. Availability and effective exploitation of such data has facilitated advances in applied fields such as weather prediction, invasive species management, precision agriculture, urban planning, etc.

This book focuses on advances in signal processing and exploitation techniques for optical remote sensing. Optical remote sensing involves acquisition and analysis of optical data—electromagnetic radiation captured by the sensing modality after reflecting off an area of interest on ground (within the sensor's field of view). Optical remote sensing has come a long way—from gray-scale photogrammetric images to hyperspectral images. The advances in imaging hardware over recent decades have enabled availability of high spatial, spectral and temporal resolution imagery to the remote sensing analyst. These advances have created unique challenges for researchers in the remote sensing community working on algorithms for representation, exploitation and analysis of such data. This book is a collection of chapters representing current state-of-the-art algorithms aimed at overcoming these

S. Prasad (✉) and L. M. Bruce
Mississippi State University, Starkville, MS, USA
e-mail: saurabh.prasad@ieee.org

J. Chanussot
Grenoble Institute of Technology, Grenoble, France

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_1,

challenges for effective processing and exploitation of remotely sensed optical data. Undergraduate students and newcomers to remote sensing have access to several textbooks on remote sensing that provide a tutorial introduction to the various remote sensing modalities and analysis techniques (e.g., [12–14]). These books are excellent resources for undergraduate and entry-level graduate students. This book is intended for a reader who has some working experience with image processing techniques for remote sensing data and wants to keep abreast with current state-of-the-art algorithms for data processing and exploitation. In particular, we believe that this book will be beneficial to graduate students and researchers who are taking advanced courses in remote sensing, image processing, target recognition and statistical pattern classification. Researchers and professionals in academia and industry working in applied areas such as electrical engineering, civil and environmental engineering, hydrology, geology, etc., who work on developing or employing algorithms for remote sensing data will also find this book useful.

1 Optical Remote Sensing: The Processing Chain

Figure 1 illustrates the processing steps in a typical optical remote sensing analysis chain. In particular, most optical remote sensing systems employ the following flow

1. *Data acquisition and processing*: this involves acquiring data from the sensing modality—handheld sensors (for on-ground data), airborne or satellite imagery (for remotely sensed data). Processing of acquired data is often necessary for mitigating affects of noise and distortion in the acquisition process, such as noise attributed to an over-heated or an improperly calibrated sensor, atmospheric distortion, luminance biases, poor contrast, etc.
2. *Data representation*: this process refers to representing data efficiently for storage, transmission or analysis. Often, optical remote sensing datasets also need to be represented efficiently due to storage and transmission limitations. Further, for effective analysis with such data (for example, for an analysis task based on statistical pattern recognition), it often becomes necessary to represent the data in a "feature" space that is amenable to the analysis task. Such a representation could be based on extracting relevant spatial statistics (e.g., texture information to exploit vicinal-pixel relationships), and spectral responses (to accurately model individual pixels and sub-pixels) from the optical imagery.
3. *Data analysis*: this process involves exploiting the data for answering the underlying remote sensing question (such as "What is the soil moisture distribution of an area", or "What is the land-cover composition of an area", or "Where are strong concentrations of invasive vegetation species for effective control") Depending upon the problem, an appropriate analysis methodology

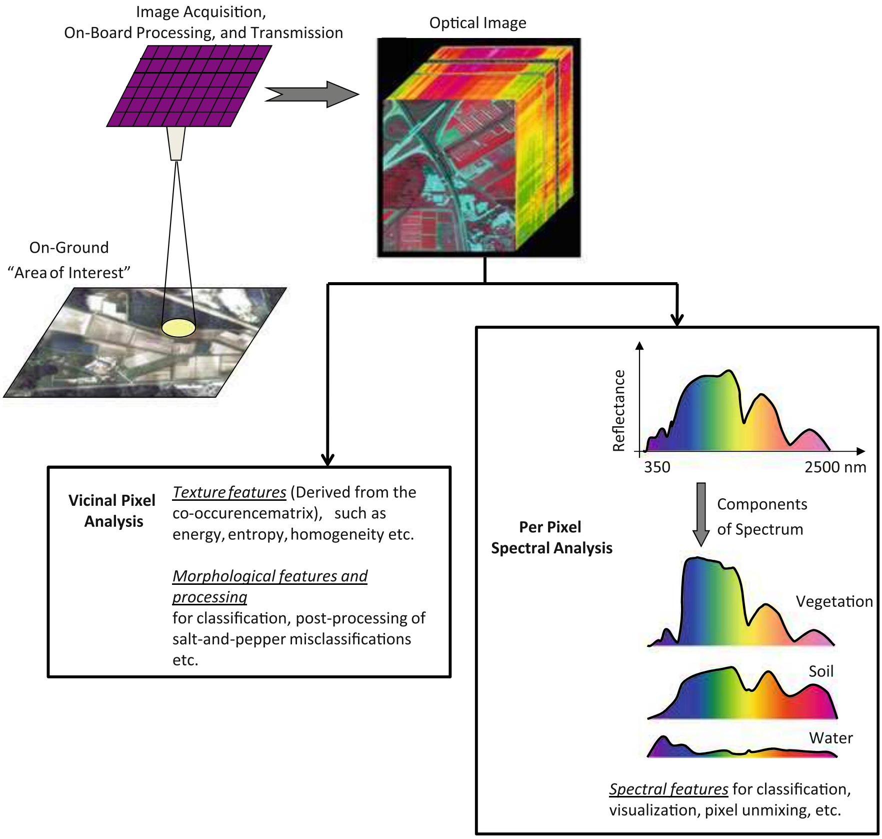


Image Acquisition,
On-Board Processing, and Transmission
Optical Image
On-Ground
"Area of Interest"
Vicinal Pixel
Analysis
Texture features (Derived from the co-occurencematrix), such as energy, entropy, homogeneity etc.
Morphological features and processing
for classification, post-processing of salt-and-pepper misclassifications etc.
Reflectance
350
2500 nm
Components
of Spectrum
Per Pixel
Spectral Analysis
Vegetation
Soil
Water
Spectral features for classification, visualization, pixel unmixing, etc.

Fig. 1 Typical flow of optical remote sensing systems

(such as statistical pattern recognition, regression analysis, unsupervised clustering, image segmentation) is invoked.

This flow results in answers to the posed remote sensing questions for a particular optical imagery. These are then interpreted for appropriate action by end-users such as scientists, government agencies, policy makers, etc.

There are three key types of optical sensing modalities: (1) handheld, (2) airborne (aerial), and, (3) spaceborne (on board a satellite). In most practical applications, spaceborne or aerial imagery is employed for analysis [2, 4, 15–18]. Trade-offs exist between spaceborne and airborne imagery, and the decision on which modality to employ for a particular application is made based on weighing in the advantages and disadvantages of each. Trade-offs include spatial and spectral resolution, ability to acquire imagery on demand, etc. versus cost, wider coverage area, repeatability, etc. Data acquired from handheld sensors is typically employed for "ground-truthing", that is, for accurately capturing spectral responses and spatial coordinates of various "classes" (objects of interest on ground), for effective training and validation of classification systems.

This book focuses on cutting-edge signal processing and exploitation techniques for addressing challenges in steps 2 and 3 of the flow outlined above. Some good references for a tutorial overview of various sensing modalities, sensor specifications, design principles, benefits and limitations of various sensors include [12–15, 19]. Kerekes et al. [20] provide an advanced overview of cutting-edge optical imaging systems, including the physics of image generation and sensing technologies, sources of noise and distortion and their impact on exploitation algorithms. Richards et al. [14] describe in detail the processing techniques employed for correcting errors due to atmospheric affects, geometric distortion, radiometric distortion, and related techniques that are carried out post-acquisition, such as georeferencing, geocoding, image registration, geometric enhancement, radiometric enhancement, etc. Examples of good tutorial introductions covering basics of image analysis and signal processing techniques for hyperspectral remotely sensed data include Landgrebe [21] and Shaw and Manolakis [22].

2 Optical Remote Sensing: Key Challenges for Signal Processing and Effective Exploitation

Early optical remote sensing systems relied on multispectral sensors, which are characterized by a small number of wide spectral bands [12, 13, 15]. Although multispectral sensors are still employed by analysts, in recent years, the remote sensing community has seen a steady shift to hyperspectral sensors, which are characterized by hundreds of fine resolution co-registered spectral bands, as the dominant technology for various tasks such as land-cover classification, environmental and ecological monitoring, etc. [2, 4, 15–17, 19–26]. Such data has the potential to reveal the underlying phenomenology as described by spectral characteristics accurately. For example, in the case of vegetation, such imagery can reveal foliar biophysical and biochemical properties, including the spectral responses at distinct wavelengths corresponding to leaf pigments, cell structure, water content, etc. [19]. This "extension" from multispectral to hyperspectral imaging does not imply that the signal processing and exploitation techniques (such as data compression, visualization and statistical pattern classification) can be simply scaled up to accommodate the extra dimensions in the data. New techniques are being developed that exploit the rich information provided by modern optical sensing modalities. In light of the above discussion, this book addresses the following key challenges:

1. *Challenges in representation and visualization of high dimensional data*: high dimensional optical data, such as hyperspectral data, is traditionally acquired in full dimensionality before being reduced in dimension prior to any processing or analysis. Hence, dataset sizes are becoming ever more voluminous, with both

spectral as well as spatial resolutions continuing to increase, resulting in extremely large quantities of data acquired in typical geospatial sensing systems, with multi-temporal data exacerbating this issue. Ramifications of this issue include: (a) it can burden transmission and storage systems, and (b) displaying the abundant information contained in this high dimensional space for effective visualization becomes challenging.

Chapters 2 through 6 will present advances in representation and visualization techniques for such datasets, including on-board compressive sensing, coded aperture imaging and visualization techniques. In Chap. 2, Christophe presents an overview of conventional and recently developed methods for compression of hyperspectral data. In Chap. 3, Fowler et al. present a review of compressive random projections for compression of hyperspectral imagery—an approach that facilitates the integration of these random projections directly into signal acquisition without incurring a significant sender side computational cost as compared to explicit dimensionality reduction. In Chap. 4, Muise et al. present an integrated sensing and processing system for hyperspectral imagery. The proposed information sensing system integrates sensing and processing, resulting in direct acquisition of data relevant to the application. In Chap. 5, Gupta et al. review various color science issues that arise in the display and representation of artificially colored remote sensing images, and analyze the current state-of-the-art solutions to these challenges. In Chap. 6, Cai et al. review several layered approaches for effective visualization of hyperspectral data. The authors propose a feature-driven multi-layer visualization technique that automatically chooses data visualization techniques based on the spatial distribution and importance of various endmembers.

2. *Challenges in statistical pattern classification and target recognition*: most image analysis techniques for exploiting optical imagery involve statistical pattern recognition or target recognition based approaches. For such analysis methods, the high dimensionality of hyperspectral data is often a double edge sword—the dense spectral sampling per pixel often provides information that can be potentially useful for target recognition and finely resolved land cover classification. This high dimensional feature space often also results in reduced generalization and statistical ill-conditioning. In many practical situations, limited training datasets for modeling class statistics further exacerbates the ill-conditioning problem.

Another issue commonly encountered when working with optical imagery is that of “mixed” pixels. Traditionally, spatial resolution is often compromised in high spectral resolution imagers. Further, in many situations, relevant features of interest may exist at sub-pixel levels. In other words, the imagery could have “mixed” pixels, representing a spectral response from a mixture of multiple objects. Hence, each pixel in such an image is typically a mixture of multiple classes/objects. However, the dense spectral sampling of hyperspectral data can help in “unmixing” (identifying the relative abundances of each class per pixel) such mixed pixels. Other issues that make this problem more challenging include affects of variations in atmospheric conditions [27], contrast and luminance variations and general variability in the spectral characteristics of the objects on ground (depending upon their interaction with their environment). Algorithms designed for the analysis of such datasets must address these issues.

Chapters 7 through 12 cover advances in statistical pattern classification and data analysis techniques, including multi-classifier systems and information fusion, morphological profiles, kernel methods, manifold learning and spectral pixel unmixing. In Chap. 7, Prasad et al. present a divide-and-conquer approach for statistical pattern classification of high dimensional hyperspectral data. In the proposed approach, a high dimensional classification task is partitioned into many independent smaller dimensional classification tasks, and a decision fusion technique is employed to merge results from this partition. In Chap. 8, Chanussot et al. study the benefits of morphological profile as a tool for analysis of remote sensing data. The chapter reviews this method based on principles of mathematical morphology and granulometry and addresses the key issues when employing this technique for multispectral and hyperspectral data. In Chap. 9, Bakos et al. present a multiple classifier, decision fusion technique for vegetation mapping and monitoring applications. The authors demonstrate the benefits of a classifier ensemble approach for vegetation mapping when employing spatial and spectral information derived from hyperspectral imagery. In Chap. 10, Camps-Valls et al. present a detailed review of applications and recent theoretical developments of kernel methods for remote sensing data analysis. In Chap. 11, Crawford et al. demonstrate the benefits of nonlinear manifold learning for dimensionality reduction and classification of hyperspectral data. In Chap. 12, Plaza et al. present a review of advances in spectral pixel unmixing and endmember extraction techniques (methods that estimate the relative abundances of various classes/endmemebers for each pixel in a mixed-pixel scenario). The chapter reviews both linear and nonlinear pixel unmixing techniques, as well as benefits of incorporating spatial information for pixel unmixing tasks

3. *Challenges in fusing multi-sensor data:* it is now possible to acquire imagery from different sensing modalities and platforms simultaneously (or nearly simultaneously) over the region of interest on ground. This implies potential availability of multiple types of optical data (e.g., high spatial resolution gray-level or multispectral imagery and high spectral resolution hyperspectral imagery), or multiple types of passive and active remotely sensed data (e.g., optical imagery and SAR or LIDAR imagery). Such multi-source data can potentially play a complimentary role—for example, (1) high spatial resolution optical imagery can provide useful vicinal-pixel and texture information, while high spectral resolution imagery can reveal valuable sub-pixel spectral characteristics, (2) optical imagery can potentially capture and help characterize surface phenomena (such as reflectance characteristics over the electromagnetic

spectrum per pixel, texture characteristics between neighboring pixels, etc.), while a ground-penetrating SAR imagery can reveal sub-surface characteristics, such as soil moisture, etc. There is hence a potential to improve analysis techniques by exploiting the diversity of information available with such multi-sensor data. In this book, we consider the following possible multi-source scenarios—optical imagery acquired from different sensors with different specifications (e.g., different spectral and spatial characteristics), or acquired from the same sensor at different times (e.g., multi-temporal imagery for change detection tasks), or a combination of optical and active remotely sensed imagery (e.g., optical and SAR imagery).

Chapters 13 through 15 cover advances in multi-sensor data fusion techniques. In Chap. 13, Bruzzone et al. study and present techniques to minimize affects of registration noise between images acquired over the same geographic area at different times on the change detection performance. Fusion of hyperspectral imagery with panchromatic or multispectral imagery for enhancing the spatial resolution of hyperspectral imagery is commonly employed by remote sensing analysts. In Chap. 14, Garzelli et al. study the effects of such spatial enhancement of hyperspectral imagery on spectral distributions. In Chap. 15, Dell'Acqua et al. demonstrate the benefits of fusion of optical and SAR data for a practical remote sensing task—seismic vulnerability mapping of buildings.

References

1. Analytical Spectral Devices FieldspecPro FR specifications. Available: http://asdi.com/productsspecifications-FSP.asp
2. Green, R., Eastwood, M., Sarture, C., Chrien, T., Aronsson, M., Chippendale, B., Faust, J., Pavri, B., Chouit, C., Solis, M., Olah, M., Williams, O.: Imaging spectroscopy and the airborne visible/infrared imaging spectromter (AVIRIS). Remote Sens. Environ. **65**, 227–248 (1998)
3. HYPERION instrument specifications, available: http://eo1.gsfc.nasa.gov/Technology/Hyperion.html
4. Pearlman, J., Segal, C., Liao, L., Carman, S., Folkman, M., Browne, B., Ong, L., Ungar, S.: Development and operations of the EO-1 hyperion imaging spectrometer. Proc. Earth Observ. Syst. V SPIE **4135**, 243–253 (2000)
5. Bruzzone, L., Marconcini, M., Wegmuller, U., Wiesmann, A.: An advanced system for the automatic classification of multitemporal SAR images. IEEE Trans. Geosci. Remote Sens. **42**, 1321–1334 (2004)
6. Chanussot, J., Mauris, G., Lambert, P.: Fuzzy fusion techniques for linear features detection in multi-temporal SAR images. IEEE Trans. Geosci. Remote Sens. **37**, 1292–1305 (1999)
7. Pellizzeri, T.M., Gamba, P., Lombardo, P., Dell'Acqua, F.: Multitemporal/multiband SAR classification of urban areas using spatial analysis: statistical versus neural kernel-based approach. IEEE Trans. Geosci. Remote Sens. **41**, 2338–2353 (2003)
8. NASA UAVSAR overview at: http://uavsar.jpl.nasa.gov/overview.html
9. CALIPSO—Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation, Available: http://www-calipso.larc.nasa.gov/
10. Winker, D.M., Pelon, J., McCormick, M.P.: The CALIPSO mission: spaceborne lidar for observation of aerosols and clouds. Proceedings of SPIE, vol. 4893 (2003)

11. Lefsky, M.A., Cohen, W.B., Parker, G.G., Harding, D.J.: Lidar remote sensing for ecosystem studies. BioScience **52**, 19–30 (2002)
12. Jensen, J.: Remote Sensing of the Environment: An Earth Resource Perspective. Prentice Hall, Englewood Cliffs (2006)
13. Campbell, J.B.: Introduction to Remote Sensing. The Guilford Press, New York (2002)
14. Richards, J.A., Jia, X.: Remote Sensing Digital Image Analysis: An Introduction. Springer, Heidelberg (2006)
15. Goetz, A., Vane, G., Solomon, J., Rock, B.: Imaging spectroscopy for earth remote sensing. Science **228**, 1147–1153 (1985)
16. Nischan, M., Kerekes, J., Baum, J., Basedow, R.: Analysis of HYDICE noise characteristics and their impact on subpixel object detection. Proc. Imaging Spect. V, SPIE **3753**, 112–123 (1999)
17. Rickard, L., Basedow, R., Zalewski, E., Silverglate, P., Landers, M.: HYDICE: an airborne system for hyperspectral imaging. Proc. Imaging Spect. Terrestr. Environ. SPIE **1937**, 173–179 (1993)
18. Schott, J.: Remote Sensing: The Image Chain Approach. Oxford University Press, New York (2006)
19. Lucas, R., Rowlands, A., Niemann, O., Merton, R.: Hyperspectral sensors and applications, Chapter 1. In: Varshney, P.K., Arora, M.K. (eds.) Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data. Springer, Heidelberg (2004)
20. Kerekes, J., Schott, J.R.: Hyperspectral imaging systems, Chapter 2. In: Chang, C.I. (ed.) Hyperspectral Data Exploitation: Theory and Applications. Wiley, New Jersey (2007)
21. Landgrebe, D.: Hyperspectral image data analysis. IEEE Signal Process. Mag. **19**, 17–28 (2002)
22. Shaw, G., Manolakis, D.: Signal processing for hyperspectral image exploitation. IEEE Signal Process. Mag. **19**, 12–16 (2002)
23. Chang, C.I.: Hyperspectral Data Exploitation: Theory and Applications. Wiley, New York (2007)
24. Kerekes, J., Baum, J.E.: Spectral imaging system analytical model for sub-pixel object detection. IEEE Trans. Geosci. Remote Sens. **40**, 1088–1101 (2002)
25. Varshney, P.K., Arora, M.K.: Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data. Springer, Heidelberg (2004)
26. Kerekes, J., Baum, J.E.: Full spectrum spectral imaging system analytical model. IEEE Trans. Geosci. Remote Sens. **43**, 571–580 (2005)
27. Kaufman, Y.: Atmospheric effect on spatial resolution of surface imagery. Appl. Opt. **23**, 3400–3408 (1984)

Hyperspectral Data Compression Tradeoff

Emmanuel Christophe

Abstract Hyperspectral data are a challenge for data compression. Several factors make the constraints particularly stringent and the challenge exciting. First is the size of the data: as a third dimension is added, the amount of data increases dramatically making the compression necessary at different steps of the processing chain. Also different properties are required at different stages of the processing chain with variable tradeoff. Second, the differences in spatial and spectral relation between values make the more traditional 3D compression algorithms obsolete. And finally, the high expectations from the scientists using hyperspectral data require the assurance that the compression will not degrade the data quality. All these aspects are investigated in the present chapter and the different possible tradeoffs are explored. In conclusion, we see that a number of challenges remain, of which the most important is to find an easier way to qualify the different algorithm proposals.

Keywords Compression · Hyperspectral · Quality evaluation · Quality comparison · Data Acquisition · Lossless compression · Lossy compression · Near-lossless compression

1 Introduction

For the past 20 years hyperspectral data are a challenge for data compression. Several factors make the constraints particularly stringent and the challenge exciting. First is the size of the data: as a third dimension is added, the amount of

E. Christophe (✉)
Centre for Remote Imaging, Sensing and Processing, National University of Singapore, Singapore, Singapore
e-mail: emmanuel.christophe@gmail.com

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_2,

data increases dramatically making the compression necessary at different steps of the processing chain. Second, the differences in spatial and spectral relation between values make the more traditional 3D compression algorithms obsolete. And finally, the high expectations from the scientists using hyperspectral data require the assurance that the compression will not degrade the data quality.

In Sect. 2, the different steps of the processing chain, where compression is required for hyperspectral data, are detailed: the specific requirements for each situation are explained and the different possible tradeoffs are explored. The following Sect. 3 goes more deeply into the exploration of some key characteristics of hyperspectral data that can be successfully used by compression algorithms. Examples are drawn from the recent published literature on the subject. Finally, in Sect. 4, requirements for an accurate assessment of image quality are explored.

2 Data Acquisition Process and Compression Properties

Compression is a way to reduce the amount of data to be transmitted or processed. The compression can be lossless without any impact on the data, or lossy, when the data values are distorted in the process and the original data cannot be retrieved in their original form. Compression is a tradeoff between processing capabilities and size (whether it is storage or transmission). Lossy compression adds a third dimension to the equation: data quality.

Before defining the compression algorithms, it is important to understand the context in which they operates and the constraints that led to their definition. There is not much in common in the requirements for compressing data onboard a satellite and compressing data on-ground. In the first case, computational power is limited and any error is unrecoverable, while in the second case, compression is used for speeding up network transfer or processing but the whole data can be transmitted if necessary.

The properties required for these algorithms will be strongly dependent on the aim.

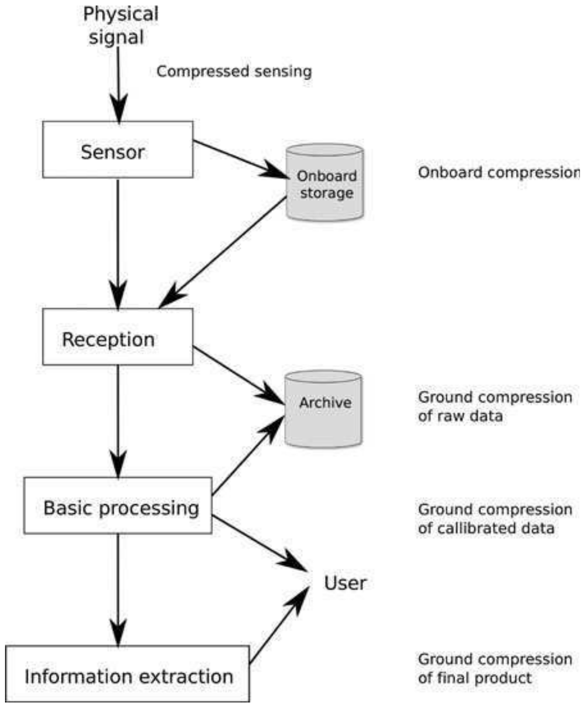
2.1 Data Acquisition Process

The first important question is to find out where the compression is going to take place. This will define the data to work on, the constraints on the algorithm and the desirable properties.

The processing chain is similar, whether the data are acquired by space-borne or air-borne sensor. Data compression usually occurs at several levels in the chain, where different tradeoffs take place. Figure 1 presents a typical processing chain.

The first place where data compression can occur in the processing chain is in the acquisition process itself. This is quite a recent paradigm, widely known as

Fig. 1 Data acquisition chain



compressed sensing but it will be treated in Reconstructions from Compressive Random Projections of Hyperspectral Imagery. We will focus in the present chapter on more traditional techniques.

After the signal acquisition, the information will be either stored onboard or directly transmitted. Direct transmission usually requires constant bitrate: this requirement can be mitigated by the use of memory buffers. The transmission which occurs in noisy environment requires redundancy coding. Both compression (source coding) and redundancy coding (channel coding) can be combined in a single operation using joint source and channel coding [1]. If the hyperspectral instrument is space-borne, all this processing is subjected to stringent requirement in term of complexity. The constraints of onboard compression are detailed in Sect. 3.

Once the signal is received on the ground, it is transmitted over cable network and stored for future use. At this part of the processing, the complexity constraint is greatly relaxed. A transcoding step can occur to keep the data in a more practical format. However, in some situation the delay between the reception and the final product must remain short. Different properties of the encoded bitstream are expected from the user. These properties are detailed in Sect. 4.

Figure 1 presents the most common processing chain where minimal operations are performed onboard. However, due to the evolution of technology, some simple operations such as calibration can be applied before compression of the signal. This will lead to different properties of the data to be compressed that should be considered during algorithm evaluation by working on the right data. These considerations are detailed in Sect. 5.

One important factor to keep in mind when designing a compression system is the end-user. Depending on who the end-user is and what information is intended to be retrieved from the data, the optimal compression solution can be very different from one case to another. The first point to consider is whether the objectives of the mission are specific or generic. A specific mission would intend to use hyperspectral data to obtain a detailed land cover map of the area for example. It could also be used to raise a warning when some anomalies are detected. In these cases, the purpose of the mission is clearly identified and the final product fully defined. On the contrary, a generic mission does not preclude any possible application. In this case, the mission has to transfer the information to the final user in a form that is as close as possible to the physical measurement.

For both these situations, specific or generic application, the compression should have no impact from the point of view of the application. No impact does not necessarily mean no differences as the error could stay within the confidence interval of the application itself. More details on the error are presented in Sect. 4.

2.2 Lossy, Lossless, Near-Lossless

Lossless compression algorithms enable the users to retrieve exactly the original data from the compressed bitstream. They are generally based on a predictor followed by an entropy coder of the residuals. The most recent publications in this domain [2–5] converge towards a compression ratio around 3:1. Such a compression ratio is insufficient to meet the constraint for onboard systems [6]. However, they are highly relevant for archiving the data and distribution to the end-user.

Lossy compression on the contrary introduces a distortion in the data and the original data cannot be retrieved in its exact form. These methods generally have parameters that can be adjusted to move along the rate–distortion curve. Reducing the bitrate increase the distortion and vice-versa.

When the distortion remains small, the algorithm can be qualified as near-lossless. Two main definitions appear in the literature for *near-lossless compression* of hyperspectral data. The first definition [6] considers that the compression is near-lossless if the noise it introduces remains below the sensor noise: the data quality remains the same. The other definition [7] considers that an algorithm is near-lossless if the distortion is bounded. We will stick to the former definition which guaranties no distortion from the application point of view: the compression remains in the noise of the sensor.

2.3 Onboard

Amazing acquisition capabilities of satellites make them the ideal candidates for regular monitoring. Many fields would benefit from regular observations. Since the launch of Hyperion on EO-1 on November 2000, the feasibility of hyperspectral space sensors has been demonstrated. Several projects in the coming years will probably increase the amount of hyperspectral data available.

For satellite sensors, the trend is towards an increase in spatial resolution, radiometric precision and possibly the number of spectral bands, leading to a dramatic increase in the amount of bits generated by such sensors. Often, continuous acquisition of data is desired, which requires scan-based compression capabilities. Scan-based compression denotes the ability to begin the compression of the image when the end of the image is still under acquisition. But due to the amount of data collected and the limited transmission capacity, there is no doubt that data has to be compressed onboard. Onboard compression presents several challenges. First, if the compression is lossy, losses are irrecoverable; this is to contrast with data compressed for transmission to the user, where the compression can be modified and data retransferred if it appears that there is a need for a higher quality. This fact makes it particularly challenging to accept onboard lossy compression even if the impact is proven to be negligible.

The second point concerns limited processing capabilities: electronics onboard a satellite need to be protected from radiation, work in a vacuum environment, have a low power consumption, limited heating, and support these conditions for several years. All these conditions cause a lag of several years in terms of processing power capabilities between consumer electronics and satellite electronics.

As onboard storage is limited, the data need to be processed on the flow as they are acquired: the start of the scene is compressed and transmitted before the end of the scene is even acquired. Satellite acquisition is done in pushbroom mode where the spectral dimension and one spatial dimension are acquired simultaneously, while the second spatial dimension is created by the satellite motion. Data ordering such as bits interleaved per pixel (BIP) or bits interleaved by line (BIL) are representative of this acquisition process.

Another consequence of this limited onboard storage is that data is often transmitted while it is acquired. One requirement to enable this transmission is a constant throughput of the compression system. This requirement can be alleviated by the use of buffers.

Desirable properties for onboard compression are summarized in Table 1.

2.4 Image Distribution

For image distribution, the challenges are very different. Data transmission is not the main problem, but the constraint is rather on processing and visualization. Due

Table 1 Desirable properties for onboard compression

Constant throughput
Low complexity
On the flow coding
Error resilient

to the huge amount of data involved, even compressed images are significant in size. In this situation, progressive data encoding enables quick browsing of the image with limited computational or network resources.

When the sensor resolution is below 1 m, images containing more than 30, 000 × 30, 000 pixels are not exceptional. In these cases, it is important to be able to decode only portions of the whole image. This feature is called random access decoding.

Resolution scalability is another feature which is appreciated within the remote sensing community. Resolution scalability enables the generation of a quicklook of the entire image using just few bits of coded data with very limited computation. It also allows the generation of low resolution images which can be used by applications that do not require fine resolution. More and more applications of remote sensing data are applied within a multiresolution framework [8, 9], often combining data from different sensors. Hyperspectral data should not be an exception to this trend. Hyperspectral data applications are still in their infancy and it is not easy to foresee what the new application requirements will be, but we can expect that these data will be combined with data from other sensors by automated algorithms.

Strong transfer constraints are ever more present in real remote sensing applications as in the case of the *International Charter: space and major disasters* [10]. Resolution scalability is necessary to dramatically reduce the bitrate and provide only the necessary information for the application.

For ground compression, error recovery is not so critical as most of the time information can be transmitted again on demand.

As the main purpose at this level is to make the image available, it is important to ensure the wide availability of the decompression algorithm. This is where the usage of an established standard is particularly relevant. Image users have a wide variety of software to analyze and process the images. This software usually implements standard formats. If the data are distributed in a specific format, transcoding into a standard format is generally required before processing. Having the data already in a standard format can save this transcoding step.

Finally, the raw pixel data is not the only product of interest for the user. First, auxiliary data are required to apply correction to the image (geometry or radiometry corrections for example), to extract geographic information, to combine with other images, etc. Going further, value added products can also be distributed directly by the data provider: classification, end-members, etc. In these cases, it is important that the format handles this information seamlessly with the image data.

Desirable properties for compression for image distribution are summarized in Table 1.

2.5 Data Availability

When designing a compression algorithm, the choice of the data on which it is going to be evaluated is important. The choice of the quality measurement is also critical and will be presented in Sect. 4. Several considerations need to be taken into account for the choice of the data (Table 2).

The first factor is availability: is there any dataset available that is representative of the mission? If not, simulations take a particularly important role. In some cases, similar data might be available from an instrument operating in different conditions (airborne sensor instead of spaceborne sensor). This is the case for example of the Aviris data sets [11]. These datasets are widely used and enable a quick and easy comparison with previous published results. Figure 2 illustrates a color composition of the four popular datasets: Moffett Field (Fig. 2a) and Jasper Ridge (Fig. 2b) represents a mix of urban area and vegetation, the two other tracks, Cuprite (Fig. 2c) and Lunar Lake (Fig. 2d) are more focused on geology application as the content is mostly minerals.

The second point is the data level to consider. If onboard compression is targeted, radiance data should probably be considered or even better, uncalibrated data if they are representative of the targeted sensor. If compression for the final user distribution is targeted, the reflectance product or even the final product can be compressed.

The third point concerns the processing required to simulate the targeted sensor. For example, if Aviris data are used to qualify a hyperspectral compression system that would be onboard of a satellite, it is unlikely that the same signal to noise ratio could be reached. In this case, additional noise should be added to the data before the compression. Some specific artifacts should also be considered. In [5], for example, it is shown that the algorithms giving the best results on unprocessed data are not the same than the best ones on calibrated data (radiance).

3 Trends in Compression Algorithms

Hyperspectral data presents an enticing challenge with an original relation between spatial and spectral information and a high information value which may rely on subtle variations of the spectrum. As a consequence, efficient compression remains an open problem. Several publications tackle the problem taking a diversity of approaches.

Table 2 Desirable properties for image distribution

Random access
Progressive decoding
Established standard
Access to value added products

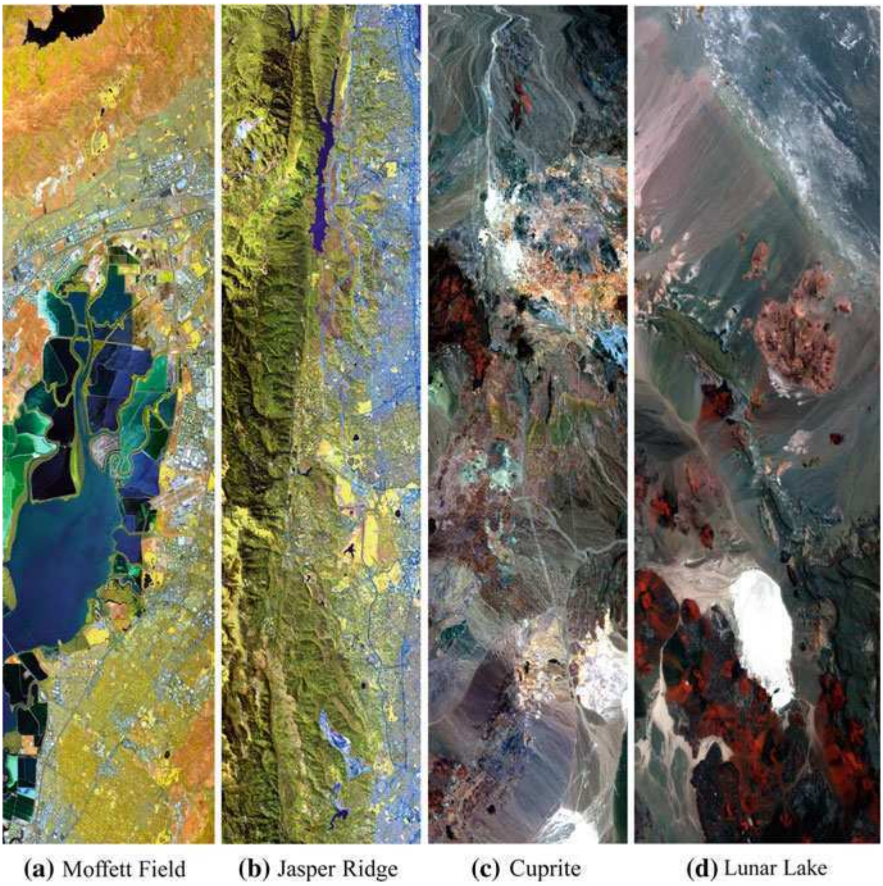

(a) Moffett Field (b) Jasper Ridge (c) Cuprite (d) Lunar Lake

Fig. 2 Classic data sets used for compression algorithms evaluation

We can separate these methods into three groups: prediction, vector quantization and transform coding. These three different approaches have been successively refined, leading to an important diversity of methods. Some of the most recent papers on the subject for prediction-based methods are [2–5, 7, 12–14], most of them in lossless compression; vector quantization recently appears in [6, 15], and transform methods in [16–28].

3.1 Prediction-Based

Directly following the main trend for lossless compression algorithms for 2D images, several adaptations for hyperspectral image compression are devised based on prediction methods. In these approaches, the data are first decorrelated by a predictor. In a second step, the prediction error is coded by an entropy coder. The predictor takes advantage of the strong correlation between spectral bands (as presented in Fig. 3). It also relies on correlation with neighboring pixel values.

As shown on Fig. 3, the correlation is not only between neighboring bands, but also between bands far apart in the spectrum. This is particularly striking for the visible part of the spectrum, which is highly correlated with the infrared (for bands 20 and 120, the correlation is above 0.6 for example), but not so much with the near-infrared (for bands 20 and 60, the correlation is around 0.3 for example). This is mainly due to the specific response of the vegetation in the near-infrared region with a strong signal due to chlorophyll. In [29] for example, it is shown that optimal reordering of the bands for Aviris can lead to a gain of 18.5% in compression performance. However, the optimal reordering might not be feasible onboard [7] and some simplifications are often used. Most of the time, only the previous band is used as a predictor. The most promising method in the domain of prediction-based compression seems to be the use of lookup tables (LUT) [2, 5, 13, 30] or the adaptation of CALIC [4, 7].

3.2 Vector Quantization

Vector quantization (VQ) of hyperspectral data is very tempting as one of the most popular application of hyperspectral data is classification. When the classification algorithm only considers pixels one by one, each pixel is assigned to the nearest class (in term of classification distance). This naturally brings the notion of codebook, each codeword being the spectrum of one material in the scene. Only the codebook (the classes) and the map (classification) have to be transmitted. This

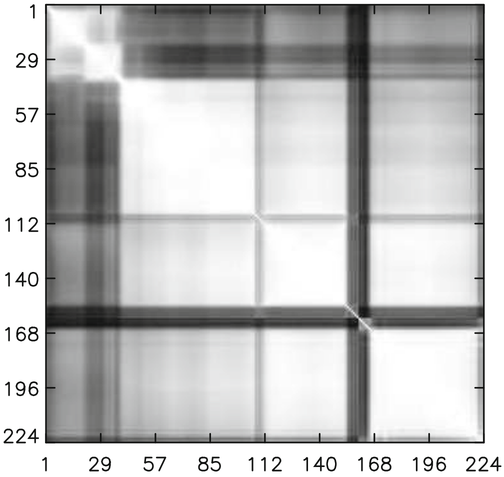


Fig. 3 Interband correlation for the Moffett hyperspectral image on a gray level scale: white corresponds to highly correlated bands while black to uncorrelated ones. Abscissa and ordinate represents the band number

is a significant reduction of the data. However, most of the time, as generic applications are targeted, the method is more complex and provides much more than a classification.

VQ compression has two separate steps: a training step, where the codebook is constructed and a coding step where each vector is assigned to a codeword. One of the common methods to generate a codebook is the Generalized Lloyd Algorithm (GLA). However, high computational costs of this algorithm presents a challenge for the compression of hyperspectral data [15]. Most of the work focuses on simplifying this step to relax the complexity constraints.

Work is going on within the Canadian Space Agency [6, 15, 31] as well as in other teams [32–34] on the use of vector quantization for the compression of hyperspectral data. In general, the targeted compression rate is high (typically 100) with a significant distortion on the image but not on classification applications where the impact is negligible. However, when the compression rate remains small, the vector quantization algorithms remains acceptable for a wider range of applications [6].

3.3 Transform Methods

Transform coding works in two steps, the first step is to transform the data in a domain where the representation of the data is more compact (energy compaction) and less correlated. The second step is to encode this information as efficiently as possible. It is during this last step, encoding, that the information loss occurs, usually through quantization. Most of the algorithms developed for hyperspectral data compression revolve around this scheme with variations on how the two steps are defined.

3.3.1 Transform

The correlation between spectral bands is important in hyperspectral data. The spectral variations are usually much slower that the spatial variations. The consequence is that hyperspectral images are more compressible than traditional images. Figure 3 presents the correlation of spectral bands with each other. The correlation coefficient is often above 0.9, even for spectral bands separated by several hundred nanometers.

From the point of view of signal theory, the most efficient transform in terms of energy compaction and decorrelation is the Karhunen–Loeve Transform (KLT) which is strongly related to the Principal Component Analysis (PCA). In [18], it is shown that using KLT transform to decorrelate the spectral bands lead to a quality gain of more than 20 dB. The main drawback is that the transform is costly in terms of computation (Table 3). The basis vectors depend on the data. In the case of an onboard compression system, a full KLT transform cannot be implemented

Table 3 Surface of silicium required to implement a KLT transform (on ASIC): without including the computation of the transform matrix according to the number of spectral bands

# Bands	Surface (mm^2)
16	6
64	99
128	400
256	1,500

The limit in 2006 was around 110 mm^2 [35]

for 200 spectral bands. Several solutions specifically target the simplification of the KLT transform.

One of the first solutions to avoid the complexity of the KLT is to precompute the transform coefficients on a set of typical images and reuse these coefficients for all images. But unfortunately, if it works for multispectral images with few bands [36], it does not for hyperspectral images as the variations in the spectra between pixels become too important to be efficiently decorrelated by an average KLT.

Some papers, such as [28, 37] design simplified versions of the KLT transform to enable its implementation onboard satellites.

The other most popular transform is the wavelet family. In [18], it is shown that using wavelet transform to decorrelate the spectral bands leads to a quality gain of more than 15 dB. The extension of the 2D wavelet transform to the 3D space led to wide range of possibilities. The first variation is on which wavelet to use. The standard 9/7 and 5/3 wavelets, which were also adopted by the JPEG 2000 standard are the most popular. Due to complexity and memory constraints, these wavelets are usually separable. The second variation is on the order in which the separable wavelets should be applied. The straightforward extention of the Mallat decomposition to 3D does not lead to the best results and another simple decomposition appears to be nearly optimal [18]. Several papers uses this decomposition which is becoming the standard [16, 19, 28, 38]. The decomposition is illustrated on Fig. 4: first the multiresolution wavelet transform is fully applied to each spectrum, then the dyadic 2D wavelet decomposition is applied on each resulting plane.

3.3.2 Coding

Once the energy is compacted to a small number of coefficients, several methods are used to code these values.

All the subtleties of the different coding methods rely on the more efficient way to order these data and/or how to predict them. The prediction methods are related to the methods presented in Sect. 3.1, but remain different as the correlation is much lower here.

Usually, the first step is quantization, which is the step where the distortion takes place. Often this occurs indirectly during a bitplane coding. Bitplane coding is a way to navigate in the binary data to be encoded starting from the one with the

Fig. 4 3D Wavelet decomposition commonly used for hyperspectral images

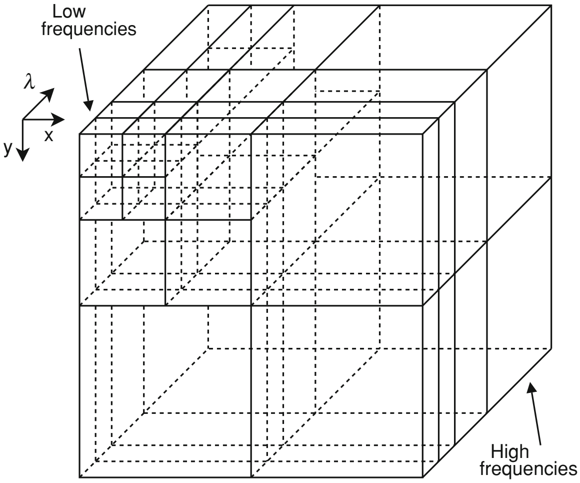


Table 4 Example of bitplane coding

Bitplane	*q*	5	63	173
7	128	0	0	1
6	64	0	0	0
5	32	0	1	1
4	16	0	1	0
3	8	0	1	1
2	4	1	1	1
1	2	0	1	0
0	1	1	1	1

This would typically lead to a bitstream 001000011010011..

greater impact first. This enables progressive decoding of the final bitstream. For example, with the example presented in Table 4, the numbers 5, 63, 173 are all going to be encoded starting by the first bitplane (quantization step $q = 128$), and then progressively refined by successive smaller quantization. This process ensures that the value with the most energy will be coded and transmitted first. As the distribution of the value to be coded, which is the result of the transform or the residual of a predictor, is close to a Laplacian distribution, for the higher bitplane, most of the values will be zero which can be very efficiently coded.

The order in which these bits are going to be visited can be further refined. As the data are sparse, there is a high number of zeros (at least in the higher bitplane). The idea is to maximize the number of zeros that can be encoded together. Some strategies exist to further increase the amount of zeros in the stream to be encoded, such as the use of signed binary digits [39]. The main strategies to take benefit of long streams of zeros are to exploit the fact that small wavelet coefficients are clustered in similar spatial areas: if the wavelet coefficient is small in the low frequency band, it is likely to be small in the high frequency band for the same spatial location. This fact is used by zerotree algorithms such as EZW, SPIHT and SPECK that have been successfully adapted to hyperspectral data [19, 40, 41]. In these algorithms, the visiting order of the value is designed to maximize the probability to code large chunks of zeros using only one symbol.

Once the data visiting order is designed, the data coding itself takes place: using the minimum number of symbols to code the stream. Arithmetic coders have been very successful, but the implementation complexity can be a deterrent. Simpler coders such as run-length, Lempel–Ziv algorithms are also used.

Most of algorithms currently used for compression are a combination of these different steps, some of them being optional: JPEG 2000 combines the wavelet transform with a contextual arithmetic coder [42]. 3D-SPIHT combines the wavelet transform with tree ordering, without the requirement of specific coding thereafter.

3.4 Lossy to Lossless

One of the current trends in the definition of new compression algorithms for hyperspectral data is to try to get the best of both worlds and provide a progressive compressed bitstream which is able to reach lossless quality. Several possibilities exist:

- use a lossless algorithm that is able to do progressive encoding;
- use an hybrid solution combining a lossy algorithm with error encoding techniques.

For the first case, JPEG2000 can be used with the 5/3 integer wavelet transform. The bitstream is progressive: decoding only the beginning of the compressed data leads to data with a lower quality but adapted to some applications. If the full stream is decoded, the data are recovered without any distortion. One main drawback of the method is the relatively low quality obtained for intermediate bitrates: the integer wavelet transform is not as efficient as the 9/7 for the decorrelation of hyperspectral images.

The second solution encodes the residual error of the lossy encoding. The residual error can be encoded using a DPCM scheme for example as in [25]. The performance of the lossy compression part is preserved and the residual error is used only if required, but this causes an increase in the complexity of the algorithm, particularly in terms of memory handling.

These methods are most likely to find an application in the ground segment for data archiving where the complexity constraints are relaxed and when no transcoding losses are tolerated.

3.5 What is in Use Now?

All these major trends have been successfully implemented and/or used in real situation. Here are some examples; note that these have been mainly used in the case of a demonstration mission to show the capabilities of hyperspectral data.

A system based on onboard classification was planned for the canceled mission of Cois on the Nemo satellite. This system, named Optical Real-time Adaptive Signature Identification System (Orasis), enables compression ratios of 30:1 while preserving good quality for classification applications [43].

On the transform side, the SPIHT algorithm is a good candidate for onboard hyperspectral data compression. A modified version of SPIHT is currently flying towards the 67P/Churyumov-Gerasimenko comet and is targeted to reach in 2014 (Rosetta mission) among other examples. This modified version of SPIHT is used to compress the hyperspectral data of the VIRTIS instrument [44]. This interest is not restricted to hyperspectral data. The current development of the Consultative Committee for Space Data Systems (CCSDS, which gathers experts from different space agencies as NASA, ESA and CNES) is oriented towards zero-trees principles [45]. The CCSDS currently has a group working on hyperspectral data compression targeting to reach a standard by 2011.

The vector quantization solution is quite advanced in terms of progress and demonstrated feasibility with hardware implementation on FPGA [6]. More importantly, this algorithm was also submitted to an extensive acceptance study by hyperspectral data users [46]. This study, using a double-blind setup, has demonstrated that compression rate of 10:1 seems acceptable for all applications and compression rates of 30:1 are for most of them. This is a gain of a factor 3 to 10 compared to lossless compression.

Of course when compression is used to distribute the data to the end user, established standards benefit from the wide availability of software able to read and process the data. The JPEG 2000 format is increasingly popular for the distribution of high resolution satellite data.

4 Ensuring Sufficient Quality

4.1 Why Bothering with Lossy Compression?

Given the fidelity requirement of the final applications whether it is target recognition (see A Divide-and-Conquer Paradigm for Hyperspectral Classification and Target Recognition), classification (see Decision Fusion of Multiple Classifiers for Vegetation Mapping and Monitoring Applications by Means of Hyperspectral Data) spectral unmixing (see Recent Developments in Endmember Extraction and Spectral Unmixing) or change detection (see Change Detection in VHR Multispectral Images: Estimation and Reduction of

Registration Noise Effects), any loss of information caused by compression is unacceptable. This is one of the main reasons why lossless compression is still so popular on hyperspectral data. However, the consideration has to be taken in a wider range than just image to image comparison. We have to look at the mission globally to find the optimal tradeoff. Compression enables gathering more data, the cost being a slight distortion on the final product. The question that needs to be answered is *does the increase in acquisition capability* (*increasing information*) *offset the quality loss* (*decreasing information*)?

For example, the MERIS sensor onboard the ENVISAT satellite acquires hyperspectral data in 520 spectral bands before averaging some of them and discarding other to produce the final selectable 15 band products [47]. To further reduce the output rate, averaging is also performed on the spatial domain.

In the setup of a specific application, the answer to this question can be validated quite easily using simulations. Using compression could enable an increase in resolution (providing more details), an increase in swath (reducing the revisit delay, thus improving multitemporal resolution), more spectral bands or an increase duty cycle (increasing the amount of images collected per orbit).

If every application would benefit from an increase in the amount of data collected, most of them would also suffer if the data quality is impacted. This is especially true when generic applications are targeted. In these conditions, it makes no sense to target compression ratio higher than 20:1 and a bitrate between 1 and 2 bit per pixel per band (bpppb) seems a reasonable target.

Lossless coding is very reassuring from the point of view of the user. This is the assurance that the compression algorithm will not change the data at all. But if it is considered in the more global situation of the mission trade-off, given the fact that sensor noise affects the data anyway, lossless compression is definitely not the optimal choice.

4.2 Quality Evaluation

With qualifying lossy compression comes the problem of quality evaluation. The important point is the impact on the end-user, but it is particularly difficult to evaluate or compare algorithms from the application point of view. The most convincing measure is to show the impact on a real application using ground truth before and after compression. However, a realistic evaluation is not often done in the literature as compression specialists are rarely also application specialists. The first shortcut which is often taken is to use a statistical distortion measure (such as SNR, PSNR or MSE). But such measures, even if widely used, have well-known drawbacks: see [48] for a review on the topic. The second widely popular shortcut is to measure how well the compression preserves the results of a benchmark application. This can be referred as Preservation of Application Results (PAR), which is a more general case of the Preservation of Classification (POC) presented in [16]. Both these cases are different than measuring the true impact of the compression on the applications.

There is currently no universal method to provide a quality evaluation. For example, if we review papers on lossy hyperspectral compression published in IEEE journals in the last three years [6, 17, 18, 20, 21–28], six papers present only statistical measurement (SNR, spectral angle) [17, 18, 20, 21, 22, 23], five present additional examples on applications comparing with the results obtained on the original image, classification [24] or anomaly detection [25–28]. Only two compare with real ground truth [28], for the classification (this paper also evaluate the anomaly detection performance, but as PAR) and only [6] provides extensive results on a wide range of real applications with the participation of several experts using the set up described in [46]. These results are not suprising, and, given the difficulty to set up a correct evaluation, such a set up cannot be expected for each paper.

Choosing a suitable quality measure is not an easy task and the amount of existing criteria to quantify the quality of compressed hyperspectral data is significant: for example see [49] for a non-exhaustive list of quality criteria for hyperspectral images.

The Preservation of Application Results (PAR) supposes that results obtained from the original data (classification, anomaly detection, …) are as close as possible to the ground truth. This only is an approximation of what we really want to measure: the classification accuracy compared to the ground truth or the real anomaly detection rate. The ideal is of course to compare the results with a ground truth, but this is not easily available.

There is a trend towards standardizing the datasets used for the evaluation of hyperspectral compression algorithms (see Sect. 5). This is already a great improvement. The trend should continue towards the availability of standard application algorithms with ground truth to make the evaluation and comparison of quality more objective. The website [50] of the Data Fusion Contest (DFC) 2008 [51] proposes the automatic evaluation of hyperspectral classification with ground truth. This system can be used to qualify the impact of hyperspectral data compression on this particular application. Another system proposing an automatic evaluation of anomaly detection would be a very valuable complement to the existing one. Anomaly detection is an important application of hyperspectral remote sensing and is neglected by most evaluations (none of the aforementioned papers compare anomaly detection with a real ground truth). A third one that would be a perfect complement would be a spectral unmixing application. With these three applications, a much better evaluation of the impact of hyperspectral compression could be done.

So we have to separate problems here: how to compare the different algorithms between each other and how to get a precise evaluation of the impact on the targeted application. Ideally, these two problems would be one, but given the number of algorithms available and number of existing applications, it is more convenient to rely on simpler measure for comparison.

4.3 Making Comparison Easier

In Sect. 5, we insisted on choosing representative data for the targeted application. But once again, as comparison is important, results should also be provided for a classic case. If the algorithm is highly specialized for one particular type of image, the results can be compared also for a standard algorithm on this case and contrasted with the ad-hoc proposed algorithm. If the algorithm is targeting the minimization of error for a particular application, once again, it can be contrasted with a standard algorithm. As it is by definition a standard, JPEG 2000, seems the ideal candidate for this task. Several implementations are freely available and easy to use. Section 5 will provide the results on the classic images.

5 Reference Results

As it is a widely available solution and standardized, the result for JPEG 2000 compression are presented for reference on some popular data sets. The user should be able to reproduce these results without trouble and compare with the implementation of its own algorithm.

However, for simplicity and because it is among the most widely used, we choose the popular SNR which can be easily converted to the ever popular PSNR or MSE (meaning that the ranking between algorithms would be the same).

When computing the SNR, one has to be careful about the variance computation which introduces an additional source of error. Hyperspectral images contains an important number of pixels on a wide range of values, computational artefacts (which becomes significant when millions of small values are added) appear in some publications. Depending on the algorithm used for computation, one has to be careful to use double precision to avoid such artifacts.

Table 5 presents the results for the popular Aviris data set for JPEG 2000 lossless compression. JPEG 2000 is used without any fancy options. The only non classic option is the use of the multicomponent decomposition (MCT) using wavelets as defined in the standard [42, 52]. Five levels of decomposition are used in the spatial and spectral directions. The decomposition is equivalent to the one illustrated in Fig. 4. The rate allocation is done considering all the wavelet subbands together (default behavior of Kakadu).

The implementation used to obtain those results is Kakadu v6.2.1 [53], most of the options are the default one apart from the MCT which requires specific parameters.

The dataset are the first scene of the three popular tracks, in radiance and reflectance. The original scenes are 614 × 512 pixels. The results are presented for the original scenes, but also for some common extracts: 512 × 512 pixels and 256 × 256 starting from the top left.

Table 6 presents the results obtained with JPEG 2000 in lossy configuration. The results are presented in terms of SNR and maximum error. These results can

Table 5 Lossless performances

Scene	Size	Rate (bpppb)	Compression ratio
Moffett Field (sc 1) Radiance	614	5.684	2.815
	512	5.654	2.830
	256	5.557	2.879
Jasper Ridge (sc 1) Radiance	614	5.598	2.858
	512	5.547	2.885
	256	5.390	2.968
Cuprite (sc 1) Radiance	614	5.291	3.024
	512	5.286	3.027
	256	5.261	3.041
Moffett Field (sc 1) Reflectance	614	6.865	2.331
	512	6.844	2.338
	256	6.767	2.365
Jasper Ridge (sc 1) Reflectance	614	6.619	2.417
	512	6.573	2.434
	256	6.428	2.489
Cuprite (sc 1) Reflectance	614	6.755	2.369
	512	6.763	2.366
	256	6.784	2.359

Table 6 Lossy performances at 1.0 bpppb

Scene	Size	SNR	MAD
Moffett Field (sc 1) radiance	614	45.233	90
	512	45.453	91
	256	45.898	87
Jasper Ridge (sc 1) radiance	614	44.605	96
	512	44.807	78
	256	45.367	60
Cuprite (sc 1) radiance	614	50.772	58
	512	50.920	54
	256	51.259	51
Moffett Field (sc 1) reflectance	614	36.110	444
	512	36.438	444
	256	37.865	260
Jasper Ridge (sc 1) reflectance	614	36.446	225
	512	36.983	201
	256	37.647	127
Cuprite (sc 1) reflectance	614	34.995	283
	512	34.952	283
	256	34.829	291

easily be obtained for other bitrates as both the images and the JPEG2000 implementation are available. Any new proposal for an hyperspectral compression algorithm can be easily compared with this reference to provide a convenient comparison point for the reader.

6 Conclusion

Despite the numerous algorithms proposed, a number of challenges remain in the area of hyperspectral image compression. One of the main challenges is the evaluation of the impact of lossy compression. The lack of confidence of the final users in these evaluations is probably the main reason for the reluctance to accept near lossless compression in spite of the significant advantage in acquisition capabilities.

An extensive study conducted in a double blind setup shows that a factor of 3 can be obtained with near lossless compression compared to lossless compression with no impact from the user point of view. This shows that near-lossless compression is the best tradeoff for onboard compression of generic missions. However the procedure to evaluate the impact of the distortion needs to be refined as it is not conceivable to conduct a new extensive study with end-users for each new algorithm proposal.

Once the procedure for impact evaluation is accepted, more advanced concepts to reduce the data volume with an acceptable complexity can be proposed. This is the case of compressed sensing which proposes a shift in the compression paradigm, shifting most of the complexity at the decoding step.

Acknowledgments The author wishes to thank Corinne Mailhes for valuable suggestions that improved this chapter and for carefully reading the early versions.

References

1. Abousleman, G., Lam, T.-T., Karam, L.: Robust hyperspectral image coding with channel-optimized trellis-coded quantization. IEEE Trans. Geosci. Remote Sens. **40**(4), 820–830 (2002)
2. Mielikainen, J., Toivanen, P.: Lossless compression of hyperspectral images using a quantized index to lookup tables. Geosci. Remote Sens. Lett. **5**(3), 474–478 (2008)
3. Huo, C., Zhang, R., Peng, T.: Lossless compression of hyperspectral images based on searching optimal multibands for prediction. Geosci. Remote Sens. Lett. **6**(2), 339–343 (2009)
4. Magli, E.: Multiband lossless compression of hyperspectral images. IEEE Trans. Geosci. Remote Sens. **47**(4), 1168–1178 (2009)
5. Kiely, A.B., Klimesh, M.A.: Exploiting calibration-induced artifacts in lossless compression of hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **47**(8), 2672–2678 (2009)
6. Qian, S.-E., Bergeron, M., Cunningham, I., Gagnon, L., Hollinger, A.: Near lossless data compression onboard a hyperspectral satellite. IEEE Trans. Aerospace Electron. Syst. **42**(3), 851–866 (2006)
7. Magli, E., Olmo, G., Quacchio, E.: Optimized onboard lossless and near-lossless compression of hyperspectral data using CALIC. IEEE Geosci. Remote Sens. Lett. **1**(1), 21–25 (2004)
8. Krishnamachari, S., Chellappa, R.: Multiresolution Gauss–Markov random field models for texture segmentation. IEEE Trans. Image Process. **39**(2), 251–267 (1997)
9. Bruce, L.M., Morgan, C., Larsen, S.: Automated detection of subpixel hyperspectral targets with continuous and discrete wavelet transforms. IEEE Trans. Geosci. Remote Sens. **39**(10), 2217–2226 (2001)

10. International charter: Space and major disasters. http://www.disasterscharter.org/main_e.html
11. Jet Propulsion Laboratory: AVIRIS free standard data product. http://aviris.jpl.nasa.gov/html/aviris.freedata.html
12. Zhang, J., Liu, G.: An efficient reordering prediction-based lossless compression algorithm for hyperspectral images. Geosci. Remote Sens. Lett. **4**(2), 283–287 (2007)
13. Aiazzi, B., Baronti, S., Alparone, L.: Lossless compression of hyperspectral images using multiband lookup tables. Geosci. Remote Sens. Lett. **16**(6), 481–484 (2009)
14. Wang, H., Babacan, S.D., Sayood, K.: Lossless hyperspectral-image compression using context-based conditional average. IEEE Trans. Geosci. Remote Sens. **45**(12), 4187–4193 (2007)
15. Qian, S.-E.: Hyperspectral data compression using a fast vector quantization algorithm. IEEE Trans. Geosci. Remote Sens. **42**(8), 1791–1798 (2004)
16. Fowler, J.E., Rucker, J.T.: 3D wavelet-based compression of hyperspectral imagery. In: Chang, C.-I. (ed.) Hyperspectral Data Exploitation: Theory and Applications, Chapter 14, pp. 379–407. Wiley, Hoboken (2007)
17. Penna, B., Tillo, T., Magli, E., Olmo, G.: Progressive 3-D coding of hyperspectral images based on JPEG 2000. IEEE Geosci. Remote Sens. Lett. **3**(1), 125–129 (2006)
18. Christophe, E., Mailhes, C., Duhamel, P.: Hyperspectral image compression: adapting SPIHT and EZW to anisotropic 3D wavelet coding. IEEE Trans. Image Process. **17**(12), 2334–2346 (2008)
19. Christophe, E., Pearlman, W.A.: Three-dimensional SPIHT coding of volume images with random access and resolution scalability. EURASIP J. Image Video Process. (2008). doi: 10.1155/2008/248905
20. Cheung, N.-M., Wang, H., Ortega, A.: Sampling-based correlation estimation for distributed source coding under rate and complexity constraints. IEEE Trans. Image Process.**17**(11), 2122–2137 (2008)
21. Wang, L., Wu, J., Jiao, L., Shi, G.: Lossy-to-lossless hyperspectral image compression based on multiplierless reversible integer TDLT/KLT. IEEE Geosci. Remote Sens. Lett. **6**(3), 587–591 (2009)
22. García-Vílchez, F., Serra-Sagristà, J.: Extending the CCSDS recommendation for image data compression for remote sensing scenarios. IEEE Trans. Geosci. Remote Sens. **47**(10), 3431–3445 (2009)
23. Du, Q., Fowler, J.E., Zhu, W.: On the impact of atmospheric correction on lossy compression of multispectral and hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **47**(1), 130–132 (2009)
24. Zhang, J., Fowler, J.E., Liu, G.: Lossy-to-lossless compression of hyperspectral imagery using three-dimensional TCE and an integer KLT. IEEE Geosci. Remote Sens. Lett. **5**(4), 814–818 (2008)
25. Carvajal, G., Penna, B., Magli, E.: Unified lossy and near-lossless hyperspectral image compression based on JPEG 2000. IEEE Geosci. Remote Sens. Lett. **5**(4), 593–597 (2008)
26. Du, Q., Zhu, W., Fowler, J.E.: Anomaly-based JPEG2000 compression of hyperspectral imagery. IEEE Geosci. Remote Sens. Lett. **5**(4), 696–700 (2008)
27. Penna, B., Tillo, T., Magli, E., Olmo, G.: Hyperspectral image compression employing a model of anomalous pixels. IEEE Geosci. Remote Sens. Lett. **4**(4), 664–668 (2007)
28. Penna, B., Tillo, T., Magli, E., Olmo, G.: Transform coding techniques for lossy hyperspectral data compression. IEEE Trans. Geosci. Remote Sens. **45**(5), 1408–1421 (2007)
29. Tate, S.R.: Band ordering in lossless compression of multispectral image. IEEE Trans. Geosci. Remote Sens. **46**(4), 477–483 (1997)
30. Mielikainen, J.: Lossless compression of hyperspectral images using lookup tables. IEEE Signal Process. Lett. **13**(3), 157–160 (2006)
31. Qian, S.-E., Hollinger, A., Williams, D., Manak, D.: Vector quantization using spectral index-based multiple subcodebooks for hyperspectral date compression. IEEE Trans. Geosci. Remote Sens. **38**(3), 1183–1190 (2000)

32. Motta, G., Rizzo, F., Storer, J.A.: Compression of hyperspectral imagery. In: Data Compression Conference, DCC, vol. 8. IEEE, Mar. 2003, pp. 333– 342
33. Ryan, M.J., Arnold, J.F.: Lossy compression of hyperspectral data using vector quantization. Remote Sens. Environ. **61**, 419–436 (1997)
34. Ryan, M., Pickering, M.: An improved M-NVQ algorithm for the compression of hyperspectral data. In: IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2000, vol. 2, pp. 600–602 (2000)
35. Implementation du décorellateur multispectral—R&T Compression. Alcatel Alenia Space, Tech. Rep. 100137101A, Nov (2006)
36. Thiebaut, C., Christophe, E., Lebedeff, D., Latry, C.: CNES studies of on-board compression for multispectral and hyperspectral images. In: SPIE, Satellite Data Compression, Communications, and Archiving III, vol. 6683. SPIE, August (2007)
37. Penna, B., Tillo, T., Magli, E., Olmo, G.: A new low complexity KLT for lossy hyperspectral data compression. In IEEE International Geoscience and Remote Sensing Symposium, IGARSS'06, August (2006), pp. 3525–3528
38. Liu, G., Zhao, F.: Efficient compression algorithm for hyperspectral images based on correlation coefficients adaptive 3D zerotree coding. IET Image Process. **2**(2), 72–82 (2008)
39. Christophe, E., Duhamel, P., Mailhes, C.: Adaptation of zerotrees using signed binary digit representations for 3 dimensional image coding. EURASIP J. Image Video Process. (2007)
40. Tang, X., Pearlman, W.A.: Scalable hyperspectral image coding. In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'05, vol. 2, pp. 401–404 (2005)
41. Cho, Y., Pearlman, W.A., Said, A.: Low complexity resolution progressive image coding algorithm: progres (progressive resolution decompression). In: IEEE International Conference on Image Processing, vol. 3, pp. 49–52 (2005)
42. Information technology—JPEG 2000 image coding system: Core coding system, ISO/IEC Std. 15 444-1 (2002)
43. Bowles, J., Gillis, D., Palmadesso, P.: New improvements in the ORASIS algorithm. Aerospace Conference Proceedings **3**, 293–298 (2000)
44. Langevin, Y., Forni, O.: Image and spectral image compression for four experiments on the ROSETTA and Mars Express missions of ESA. In Applications of Digital Image Processing XXIII, vol. 4115. SPIE, 2000, pp. 364–373.
45. Yeh, P.-S., Armbruster, P., Kiely, A., Masschelein, B., Moury, G., Schaefer, C., Thiebaut, C.: The new CCSDS image compression recommendation. In IEEE Aerospace Conference. IEEE, March (2005)
46. Qian, S.-E., Hollinger, A., Bergeron, M., Cunningham, I., Nadeau, C., Jolly, G., Zwick, H.: A multi-disciplinary user acceptability study of hyperspectral data compressed using onboard near lossless vector quantization algorithm. Int. J. Remote Sens. **26**(10), 2163–2195 (2005)
47. Rast, M., Bezy, J.L., Bruzzi, S.: The ESA medium resolution imaging spectrometer MERIS - a review of the instrument and its mission. Int. J. Remote Sens. **20**(9), 1681–1702 (1999)
48. Wang, Z., Bovik, A.C.: Mean square error: love it or leave it?. IEEE Signal Process. Mag. **26**(1), 98–117 (2009)
49. Christophe, E., Léger, D., Mailhes, C.: Quality criteria benchmark for hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **43**(09), 2103–2114 (2005)
50. DFTC fusion contest (2008). http://tlclab.unipv.it/dftc/home.do?id=3
51. Licciardi, G., Pacifici, F., Tuia, D., Prasad, S., West, T., Giacco, F., Inglada, J., Christophe, E., Chanussot, J., Gamba, P.: Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRS-S Data Fusion Contest. IEEE Trans. Geosci. Remote Sens. **47**(11), 3857–3865 (2009)
52. Information technology – JPEG 2000 image coding system: Extensions, ISO/IEC Std. 15 444-2, (2004)
53. Taubman, D.: Kakadu Software v 6.0 (2009). http://www.kakadusoftware.com/

Reconstructions from Compressive Random Projections of Hyperspectral Imagery

James E. Fowler and Qian Du

Abstract High-dimensional data such as hyperspectral imagery is traditionally acquired in full dimensionality before being reduced in dimension prior to processing. Conventional dimensionality reduction on-board remote devices is often prohibitive due to limited computational resources; on the other hand, integrating random projections directly into signal acquisition offers an alternative to explicit dimensionality reduction without incurring sender-side computational cost. Receiver-side reconstruction of hyperspectral data from such random projections in the form of compressive-projection principal component analysis (CPPCA) as well as compressed sensing (CS) is investigated. Specifically considered are single-task CS algorithms which reconstruct each hyperspectral pixel vector of a dataset independently as well as multi-task CS in which the multiple, possibly correlated hyperspectral pixel vectors are reconstructed simultaneously. These CS strategies are compared to CPPCA reconstruction which also exploits cross-vector correlations. Experimental results on popular AVIRIS datasets reveal that CPPCA outperforms various CS algorithms in terms of both squared-error as well as spectral-angle quality measures while requiring only a fraction of the computational cost.

Keywords Hyperspectral data · Principal component analysis (PCA) · Random projections · Rayleigh-Ritz theory

J. E. Fowler (✉) and Q. Du
Department of Electrical and Computer Engineering, Geosystems Research Institute,
Mississippi State University, Mississippi State, MS 39762, USA
e-mail: fowler@ece.msstate.edu

Q. Du
e-mail: du@ece.msstate.edu

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_3,

1 Introduction

In the traditional data pipeline used with high-dimensional data such as hyperspectral imagery, the data is acquired in its full dimensionality within some typically remote signal-sensing platform (e.g., a satellite), the data is downlinked to some central (i.e., earth-based) processing site, and finally the data is subjected to the desired application-specific processing. In many cases, the dimensionality of the dataset is reduced prior to the processing. For example, a variety of linear—e.g., principal component analysis (PCA) [1, 2] and independent component analysis (ICA) [3, 4]—as well as nonlinear—e.g., locally linear embedding (LLE) [5, 6]—forms of dimensionality reduction have been applied to reduce spectral dimensionality of hyperspectral imagery thereby facilitating the deployment of classifiers to detect specific endmember classes or anomalous pixels. This traditional data-flow pipeline is illustrated in Fig. 1a.

Unfortunately, there are a number of problematic issues with this traditional data-flow paradigm. Specifically, dataset sizes are becoming ever more voluminous, with both spectral as well as spatial resolutions continuing to increase, resulting in extremely large quantities of data acquired in typical geospatial sensing systems, with multi-temporal data exacerbating this issue. As a result, it would be greatly beneficial if dimensionality reduction could occur before data downlink, since many signal-acquisition platforms are severely resource-constrained (e.g., satellite- and airborne devices). On-board dimensionality reduction would dramatically cut storage and communication burdens faced by such remote sensors; however, many approaches to dimensionality reduction are data dependent and exceedingly computationally expensive so as to preclude implementation

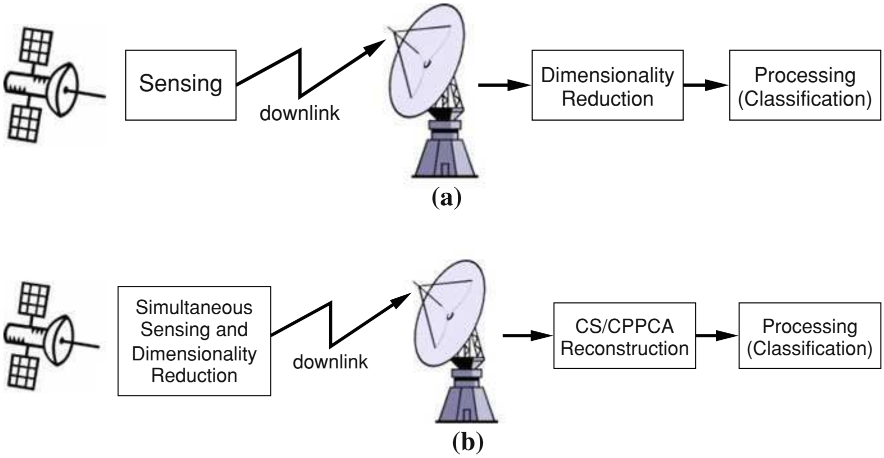

Sensing
downlink
Dimensionality Reduction
Processing (Classification)
(a)
Simultaneous Sensing and Dimensionality Reduction
downlink
CS/CPPCA Reconstruction
Processing (Classification)
(b)

Fig. 1 **a** Traditional sensing, communication, and processing data-flow pipeline. **b** Proposed data flow with simultaneous sensing and dimensionality reduction accomplished with random projections at the sender and CS or CPPCA reconstruction located at the receiving base station within resource-constrained sensing platforms. For instance, the computational complexity of nonlinear LLE restricts it to small blocks of a large image scene even on computationally powerful machines.

Recently, it has been demonstrated that projections onto randomly chosen subspaces can be a particularly powerful form of dimensionality reduction. Namely, the mathematical theory of *compressed sensing* (CS) [7–10] establishes that sparsely representable signals can be recovered exactly from data-independent random projections. Furthermore, we have recently developed *compressive-projection principal component analysis* (CPPCA) [11–13] which recovers an approximate PCA representation of the original signal from random projections. Both CS and CPPCA permit sensing platforms to enjoy the benefits of dimensionality reduction (less burdensome storage and communication requirements) without the expense of computation associated with explicit dimensionality reduction since the random projections can be accomplished simultaneously with the sensing and signal-acquisition process, while the more expensive reconstruction from the projections takes place at the receiver-side base station. Specifically, we replace the traditional data-flow pipeline of Fig. 1a with that of Fig. 1b in which random projections enable simultaneous signal-acquisition and dimensionality reduction, while CS or CPPCA reconstruction drives further processing at the receiving base station.

There have been several recent efforts (e.g., [14–16]) aimed at devising hyperspectral sensors that accomplish such simultaneous signal-acquisition and dimensionality reduction at the sender side of the system. As a consequence, we explore here options for reconstruction of hyperspectral data at the receiver side, comparing the relative merits of CPPCA and CS reconstruction. We begin by overviewing both CPPCA and CS in Sects. 2 and 3, respectively. We then present a battery of experimental results in Sect. 4 in which we observe that CPPCA usually outperforms CS in terms of both square-error and spectral-angle quality measures while requiring only a fraction of the computational cost. Finally, we make some concluding remarks in Sect. 5.

2 Compressive-Projection Principal Component Analysis (CPPCA)

In brief, CPPCA effectuates a reconstruction from random projections by recovering not only the coefficients associated with the PCA transform, but also an approximation to the PCA transform basis itself. In the next section, we briefly overview the theoretical underpinnings of CPPCA—specifically, an extension of existing Rayleigh–Ritz theory to the special case of highly eccentric distributions which permits simple approximations to orthogonal projections of eigenvectors. We then describe in Sect. 2.2 how this analytical result is used to devise the CPPCA algorithm to recover the PCA transform basis and PCA transform coefficients.

2.1 Overview of CPPCA

At the core of the CPPCA technique is a receiver-side process that produces an approximation to the PCA transform basis. Consider a dataset of M zero-mean vectors $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M]$, where each $\mathbf{x}_m \in \mathbb{R}^{\mathbb{N}}$. The covariance matrix of $\mathbf{X}$ is $\boldsymbol{\Sigma} = \mathbf{X}\mathbf{X}^T/M$, and the PCA transform matrix is the $N \times N$ matrix $\mathbf{W}$ of eigenvectors that emanates from the eigendecomposition of $\boldsymbol{\Sigma}$; i.e.,

$$\boldsymbol{\Sigma} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^T, \tag{1}$$

where $\mathbf{W}$ contains the N unit eigenvectors of $\boldsymbol{\Sigma}$ column-wise. However, central to the CPPCA paradigm is that production of the PCA transform matrix occurs at the receiver rather than at the sender as in the traditional use of PCA; that is, the CPPCA receiver cannot implement eigendecomposition (1) directly as it does not know either $\mathbf{X}$ or $\boldsymbol{\Sigma}$. Instead, the receiver knows only K-dimensional projections of $\mathbf{X}$. Specifically, suppose we have K orthonormal vectors $\mathbf{p}_k$ that form the basis of K-dimensional subspace $\mathcal{P}$ such that $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K]$ provides an orthogonal projection onto $\mathcal{P}$. Then, the orthogonal projection of $\mathbf{x}_m$ onto $\mathcal{P}$ is $\mathbf{y}_m = \mathbf{P}\mathbf{P}^T\mathbf{x}_m$; expressed with respect to the basis $\{\mathbf{p}_k\}$, we have $\widetilde{\mathbf{y}}_m = \mathbf{P}^T\mathbf{x}_m$, such that $\mathbf{y}_m = \mathbf{P}\widetilde{\mathbf{y}}_m$. The CPPCA sender produces the projected vectors $\widetilde{\mathbf{Y}} = [\widetilde{\mathbf{y}}_1 \cdots \widetilde{\mathbf{y}}_M]$, and it is from projections $\widetilde{\mathbf{Y}}$ that the CPPCA receiver approximates $\mathbf{W}$. The projected vectors have covariance

$$\widetilde{\boldsymbol{\Sigma}} = \widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T/M = \mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P}/M = \mathbf{P}^T\boldsymbol{\Sigma}\mathbf{P}, \tag{2}$$

which the CPPCA receiver calculates having received $\widetilde{\mathbf{Y}}$ from the sender.

Rayleigh–Ritz theory [17] describes the relation between the eigenvectors of $\boldsymbol{\Sigma}$ and those of $\widetilde{\boldsymbol{\Sigma}}$ as given by (2). Covariance matrix $\boldsymbol{\Sigma}$ has spectrum $\lambda(\boldsymbol{\Sigma}) = \{\lambda_1(\boldsymbol{\Sigma}), \ldots, \lambda_N(\boldsymbol{\Sigma})\}$, where the eigenvalues satisfy $\lambda_1(\boldsymbol{\Sigma}) \geq \cdots \geq \lambda_N(\boldsymbol{\Sigma})$, and the corresponding unit eigenvectors are $\mathbf{w}_n$. The eigendecomposition of $\widetilde{\boldsymbol{\Sigma}} = \mathbf{P}^T\boldsymbol{\Sigma}\mathbf{P}$ is $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{U}}^T$, where $\widetilde{\mathbf{U}} = [\widetilde{\mathbf{u}}_1 \cdots \widetilde{\mathbf{u}}_K]$, $\widetilde{\boldsymbol{\Lambda}} = \operatorname{diag}\left(\lambda_1\left(\widetilde{\boldsymbol{\Sigma}}\right), \ldots, \lambda_K\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)$, $\|\widetilde{\mathbf{u}}_k\|_2 = 1$, and $\lambda_1\left(\widetilde{\boldsymbol{\Sigma}}\right) \geq \cdots \geq \lambda_K\left(\widetilde{\boldsymbol{\Sigma}}\right)$. The K eigenvalues $\lambda_k\left(\widetilde{\boldsymbol{\Sigma}}\right)$ are called *Ritz values*; additionally, there are K vectors, known as *Ritz vectors*, defined as

$$\mathbf{u}_k = \mathbf{P}\widetilde{\mathbf{u}}_k, \quad 1 \leq k \leq K, \tag{3}$$

where $\widetilde{\mathbf{u}}_k$ are the eigenvectors of $\widetilde{\boldsymbol{\Sigma}}$. Finally, we define *normalized projection* $\mathbf{v}_n$ as the orthogonal projection of $\mathbf{w}_n$ onto $\mathcal{P}$, normalized to unit length; i.e.,

$$\mathbf{v}_n = \frac{\mathbf{P}\mathbf{P}^T\mathbf{w}_n}{\left\|\mathbf{P}\mathbf{P}^T\mathbf{w}_n\right\|_2}. \tag{4}$$

These vectors are illustrated for an example distribution in the simple case of $N = 3$ and $K = 2$ in Fig. 2.

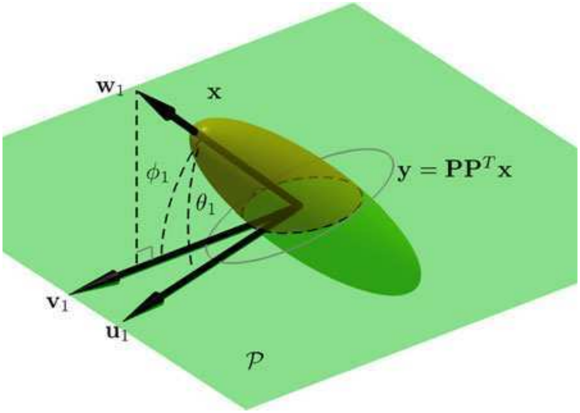


Fig. 2 Data distribution of $\mathbf{x}$ in $\mathbb{R}^3$ is projected onto 2D subspace $\mathcal{P}$ as $\mathbf{y}$; the first Ritz vector, $\mathbf{u}_1$, lies close to the normalized projection, $\mathbf{v}_1$, onto $\mathcal{P}$ of the first eigenvector, $\mathbf{w}_1$, of $\mathbf{x}$ (from [11], © 2009 IEEE)

Traditional Rayleigh–Ritz theory is rather limited in that it tells us very little about the Ritz vectors for $K < N$, giving only that the Ritz vectors do not typically align with the orthogonal projections of any of the eigenvectors [17]; i.e., $\mathbf{u}_k \neq \mathbf{v}_n$ in general. However, CPPCA is built on the central idea that, if subspace $\mathcal{P}$ is chosen randomly, and the distribution of the vectors in $\mathbf{X}$ is highly eccentric in that eigenvalue $\lambda_k(\mathbf{\Sigma})$ is sufficiently separated in value with respect to the other eigenvalues, then it is likely that its corresponding normalized projection, $\mathbf{v}_k$, will be quite close to the Ritz vector, $\mathbf{u}_k$, corresponding to the Ritz value $\lambda_k\left(\widetilde{\mathbf{\Sigma}}\right)$. Under the assumption that $\mathbf{u}_k \approx \mathbf{v}_k$, an algorithm based on projections onto convex sets (POCS) [18] was devised in [11, 13] to approximate the first L eigenvectors $\mathbf{w}_n$ from $\widetilde{\mathbf{Y}}$; this algorithm is overviewed next. Suffice it to say, however, that the entire feasibility of the CPPCA technique rests on the approximation $\mathbf{u}_k \approx \mathbf{v}_k$. However, extensive analysis in [12, 13] established the validity of this approximation.

2.2 The CPPCA Algorithm

We now use the fact, as discussed above, that we can approximate orthogonal projections of eigenvectors with Ritz vectors to enable a system that uses random projections at the sender. The corresponding receiver then implements recovery of not only the PCA coefficients for the transmitted dataset, but also an approximation to the PCA transform basis itself. In this sense, the resulting CPPCA system in effect shifts the computational complexity of PCA from the sender to the receiver.

Specifically, in the CPPCA sender, the M vectors of $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$ are merely each subjected to random projection. On the other hand, the CPPCA receiver then must recover not only the PCA transform coefficients, but also the basis vectors of the transform itself, all from the projections. We assume that the receiver knows only the projection operator and its resulting projections, but not $\mathbf{X}$ or its statistics (e.g., covariance). Below, we present an overview of the CPPCA approach which is explained in more detail in [11, 13].

2.2.1 Eigenvector Recovery

Traditional design methods for PCA produce the transform $\mathbf{W}$ via the eigendecomposition given by (1); however, in the CPPCA receiver, one has access to merely $\widetilde{\boldsymbol{\Sigma}}$ and not $\boldsymbol{\Sigma}$ as required in (1). The goal of CPPCA is thus to approximate $\mathbf{W}$ from $\widetilde{\boldsymbol{\Sigma}}$ without knowledge of $\boldsymbol{\Sigma}$, given that $\widetilde{\boldsymbol{\Sigma}}$ results from random projection. The CPPCA receiver first recovers an approximation to the PCA transform basis by recovering approximations to the first L eigenvectors of $\boldsymbol{\Sigma}$ from random projections. We observe that, if we knew the true normalized projection $\mathbf{v}$ of eigenvector $\mathbf{w}$ in subspace $\mathcal{P}$, we could form subspace $\mathcal{Q}$ as

$$\mathcal{Q} = \mathcal{P}^{\perp} \oplus \operatorname{span}\{\mathbf{v}\}, \tag{5}$$

the direct sum of the orthogonal complement of $\mathcal{P}$ with a 1D space containing $\mathbf{v}$. Clearly, $\mathbf{w}$ would lie in $\mathcal{Q}$. Suppose then that we produce J distinct random K-dimensional subspaces, $\mathcal{P}^{(1)}$ through $\mathcal{P}^{(J)}$, each containing a normalized projection, $\mathbf{v}^{(1)}$ through $\mathbf{v}^{(J)}$, respectively, produced via (4) using the corresponding projection matrices, $\mathbf{P}^{(1)}$ through $\mathbf{P}^{(J)}$. We could then form subspaces $\mathcal{Q}^{(1)}$ through $\mathcal{Q}^{(J)}$ via (5) using $\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(J)}$ and $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(J)}$. The eigenvector $\mathbf{w}$ would thus be in the intersection $\mathcal{Q}^{(1)} \cap \cdots \cap \mathcal{Q}^{(J)}$. This situation is illustrated in Fig. 3 for the case of $N = 3$, $K = 2$, $J = 2$, and the eigenvector in question being $\mathbf{w}_1$.

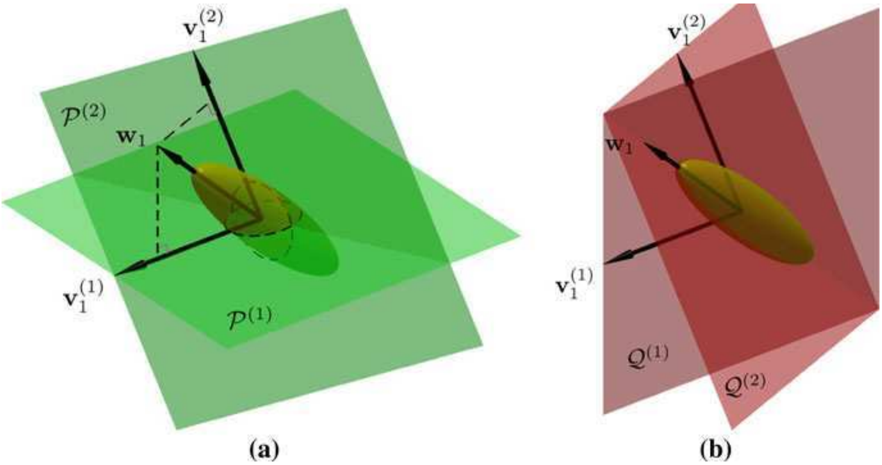


Fig. 3 **a** Two 2D subspaces $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ with corresponding normalized projections $\mathbf{v}_1^{(1)}$ and $\mathbf{v}_1^{(2)}$. **b** Subspaces $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$ whose intersection uniquely determines eigenvector $\mathbf{w}_1$ up to a sign. (from [11], © 2009 IEEE)

In the CPPCA receiver, though, we do not have access to the true normalized projections; instead, we can form Ritz vectors in each subspace $\mathcal{P}^{(j)}$ via an eigendecomposition of the corresponding projected covariance matrix $\widetilde{\mathbf{\Sigma}}^{(j)}$. Motivated by the analysis in [13], we use these Ritz vectors to approximate normalized projections; i.e., we use $\mathbf{u}_k^{(j)}$ instead of $\mathbf{v}_k^{(j)}$ to form the spaces $\mathcal{Q}^{(j)}$. Since the Ritz vectors will differ slightly from the true normalized projections, the intersection $\mathcal{Q}^{(1)} \cap \cdots \cap \mathcal{Q}^{(J)}$ is almost certain to be empty. However, since the $\mathcal{Q}^{(j)}$ are closed and convex, a parallel implementation of POCS will converge to a least-squares solution minimizing the average distance to the subspaces $\mathcal{Q}^{(j)}$ [18]; this POCS solution can then be used to approximate $\mathbf{w}$. Specifically, for iteration $i = 1, 2, \ldots,$ we form an estimate of the eigenvector as

$$\widehat{\mathbf{w}}^{(i)} = \frac{1}{J} \sum_{j=1}^{J} \mathbf{Q}^{(j)} \mathbf{Q}^{(j)^T} \widehat{\mathbf{w}}^{(i-1)}, \tag{6}$$

where projection onto $\mathcal{Q}^{(j)}$ is performed by the matrix $\mathbf{Q}^{(j)}$, and we initialize $\widehat{\mathbf{w}}^{(0)}$ to the average of the Ritz vectors. (6) will converge to $\widehat{\mathbf{w}}$; normalizing this $\widehat{\mathbf{w}}$ will approximate the desired normalized eigenvector $\mathbf{w}$ (up to sign).

In order to avoid producing multiple random projections for each vector in our dataset, the CPPCA sender splits the dataset of M vectors $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$ into J partitions $\mathbf{X}^{(j)}$, each associated with its own randomly chosen projection $\mathbf{P}^{(j)}, 1 \leq j \leq J$. It is assumed that the dataset splitting is conducted such that each $\mathbf{X}^{(j)}$ closely resembles the whole dataset $\mathbf{X}$ statistically and so has approximately the same eigendecomposition.[1] The sender transmits the projected data $\widetilde{\mathbf{Y}}^{(j)} = \mathbf{P}^{(j)} \mathbf{X}^{(j)}$ to the receiver which is assumed to know the projection operators $\mathbf{P}^{(j)}$ *a priori*. In the CPPCA receiver, $\widetilde{\mathbf{\Sigma}}^{(j)}$ is calculated from $\widetilde{\mathbf{Y}}^{(j)}$, a set of Ritz vectors $\mathbf{u}_k^{(j)}$ is produced from $\widetilde{\mathbf{\Sigma}}^{(j)}$, and then the Ritz vectors are used in place of the normalized projections to drive the POCS recovery of (6). The CPPCA receiver repeats this POCS procedure using the first L Ritz vectors to approximate the first L principal eigenvectors which are assembled into $N \times L$ matrix $\mathbf{\Psi}$, an approximation to the L-component PCA transform, $L \leq K$.

2.2.2 Coefficient Recovery

Once obtaining $\mathbf{\Psi}$, the CPPCA receiver then proceeds to recover the PCA coefficients by solving $\widetilde{\mathbf{Y}}^{(j)} = \mathbf{P}^{(j)^T} \mathbf{\Psi} \check{\mathbf{X}}^{(j)}$ for PCA coefficients $\check{\mathbf{X}}^{(j)}$ in the least-squares

[1] Dataset subsampling is commonly used to expedite covariance-matrix calculation in traditional applications of PCA, e.g., [19, 20]; we suggest modulo partitioning such as $\mathbf{X}^{(j)} = \{\mathbf{x}_m \in \mathbf{X} | (m-1) \bmod J = j-1\}$.

sense for each j. This linear reconstruction can be accomplished in several ways, for example, by using the pseudoinverse,

$$\check{\mathbf{X}}^{(j)} = \left(\mathbf{P}^{(j)^T}\boldsymbol{\Psi}\right)^{+}\widetilde{\mathbf{Y}}^{(j)}. \tag{7}$$

3 Compressed Sensing (CS)

In brief, CS (e.g., [7–10]) produces a sparse signal representation directly from a small number of projections onto another basis, recovering the sparse transform coefficients via nonlinear reconstruction. Our coverage of CS here will be brief; we refer to [10] for a more comprehensive treatment.

The main tenet of CS theory holds that, if signal $\mathbf{x} \in \mathbb{R}^{\mathbb{N}}$ can be sparsely represented (i.e., using only L nonzero coefficients) with some basis $\boldsymbol{\Psi} = [\psi_1 \cdots \psi_N]$, then we can recover $\mathbf{x}$ from K-dimensional projections $\widetilde{\mathbf{y}} = \mathbf{P}^T\mathbf{x}$ under certain conditions; here $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K]$, and $K < N$. Specifically, it is required that K must be sufficiently large with respect to the sparsity L (but still much less than N) and that $\boldsymbol{\Psi}$ and $\mathbf{P}$ be mutually *incoherent*, meaning that $\mathbf{P}$ cannot sparsely represent the ψ_n vectors. It has been shown that, if $\mathbf{P}$ is chosen randomly, then $\mathbf{P}$ and $\boldsymbol{\Psi}$ are incoherent for any arbitrary fixed $\boldsymbol{\Psi}$ with high probability [8].

The ideal recovery procedure searches for the $\check{\mathbf{x}}$ with the smallest ℓ_0 norm consistent with the observed $\widetilde{\mathbf{y}}$; i.e,

$$\check{\mathbf{x}}^* = \arg\min_{\check{\mathbf{x}}} \|\check{\mathbf{x}}\|_0, \quad \text{such that } \widetilde{\mathbf{y}} = \mathbf{P}^T\boldsymbol{\Psi}\check{\mathbf{x}}, \tag{8}$$

where the ℓ_0 norm $\|\check{\mathbf{x}}\|_0$ is the number of nonzero coefficients in $\check{\mathbf{x}}$. However, this ℓ_0 optimization being NP-complete, several alternative solution procedures have been proposed. Perhaps the most prominent of these is basis pursuit (BP) [21] which applies a convex relaxation to the ℓ_0 problem resulting in an ℓ_1 optimization:

$$\check{\mathbf{x}}^* = \arg\min_{\check{\mathbf{x}}} \|\check{\mathbf{x}}\|_1, \quad \text{such that } \widetilde{\mathbf{y}} = \mathbf{P}^T\boldsymbol{\Psi}\check{\mathbf{x}}. \tag{9}$$

Often, in practical applications that feature noisy data, or data that is only approximately sparse, BP with a quadratically relaxed constraint (e.g., [10, 22]) is employed in the form of

$$\check{\mathbf{x}}^* = \arg\min_{\check{\mathbf{x}}} \|\check{\mathbf{x}}\|_1, \quad \text{such that } \left\|\mathbf{P}^T\boldsymbol{\Psi}\check{\mathbf{x}} - \widetilde{\mathbf{y}}\right\|_2 \leq \epsilon. \tag{10}$$

BP in the form of (9) and (10) can be implemented effectively with linear programming; see, for example, ℓ_1-MAGIC [23]. However, the computational complexity of BP is often high, leading to recent interest in reduced-complexity relaxations (e.g., gradient projection for sparse reconstruction (GPSR) [24]) as well as in greedy BP variants, including matching pursuits, orthogonal matching pursuits, and sparsity adaptive matching pursuits (SAMP) [25]. Such algorithms significantly reduce computational complexity at the cost of lower reconstruction quality.

The majority of CS literature focuses on the recovery of a single vector $\mathbf{x}$ from its projection $\widetilde{\mathbf{y}} = \mathbf{P}^T\mathbf{x}$. However, for a hyperspectral image, we wish to recover a set of pixel vectors $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$ which are likely to possess a strong degree of correlation. For recovery of a set of multiple, possibly correlated vectors $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$, there have been proposals for multi-vector extensions of CS under the name of "multi-task" [26] or "distributed" CS; these, in turn, link closely to a larger body of literature on "simultaneous sparse approximation" (e.g., [27–31]). In experimental results below, we compare the performance of CPPCA to that of Multi-Task Bayesian Compressive Sensing (MT-BCS) [26] which introduces a hierarchical Bayesian framework into the multi-vector CS-recovery problem to share prior information across the multiple vectors.

We note that, on the surface, although CPPCA and MT-BCS appear somewhat similar in their functionality, there exist some crucial differences. MT-BCS, like other CS techniques, operates under an assumption of sparsity in a *known* basis $\boldsymbol{\Psi}$ but the pattern of sparsity (i.e., which L components are nonzero) is *unknown*. On the other hand, CPPCA reconstruction operates under a *known* sparsity pattern (i.e., the first L principal components), but the transform $\boldsymbol{\Psi}$ itself is *unknown*. Additionally, while MT-BCS can recover the M vectors of $\mathbf{X}$ from the same set of projections $\widetilde{\mathbf{Y}}^{(j)} = \mathbf{P}^{(j)}\mathbf{X}^{(j)}$ which drive the CPPCA recovery process, it can also function on arbitrarily small numbers of vectors, even down to $M = 1$ (in which case, MT-BCS becomes the special case of "single-task" Bayesian Compressive Sensing (ST-BCS) recovery as described in [32]). CPPCA, on the other hand, requires M to be sufficiently large to enable covariance-matrix calculation in the J subspaces.

4 Empirical Comparisons on Hyperspectral Imagery

We use hyperspectral images cropped spatially to size 100 × 100 (i.e., $M =$ 10,000); we use the popular "Cuprite," "Moffett," and "Jasper Ridge" images, AVIRIS datasets with $N = 224$ spectral bands. The mean vector has been removed from the vectors to impose a zero-mean condition.

The CPPCA and MT-BCS receivers reconstruct approximate pixel vectors $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1 \cdots \widehat{\mathbf{x}}_M]$ of the hyperspectral image from random projections. For a given vector $\widehat{\mathbf{x}}_m$, we can measure the quality of its reconstruction in several ways, for instance, by using a signal-to-noise ratio (SNR) or a spectral-angle distortion measure. In the following results, we use a vector-based SNR measured in dB; i.e.,

$$\mathrm{SNR}(\mathbf{x}_m, \widehat{\mathbf{x}}_m) = 10 \log_{10} \frac{\mathrm{var}(\mathbf{x}_\mathrm{m})}{\mathrm{MSE}(\mathbf{x}_m, \widehat{\mathbf{x}}_m)}, \qquad (11)$$

where the $\mathrm{var}(\mathbf{x}_m)$ is the variance of the components of vector $\mathbf{x}_m$, and the mean squared error (MSE) is

$$\mathrm{MSE}(\mathbf{x}_m, \widehat{\mathbf{x}}_m) = \frac{1}{N}\|\mathbf{x}_m - \widehat{\mathbf{x}}_m\|^2. \tag{12}$$

We then average the vector-based SNR over all vectors of the dataset for a measure of quality of the dataset as a whole. Alternatively, we can measure the quality of a reconstructed hyperspectral dataset using an average spectral angle, where the spectral angle in degrees between the reconstructed hyperspectral pixel vector and its corresponding original vector is averaged over the dataset; i.e., $\bar{\xi} = \mathrm{mean}(\xi_m)$ where

$$\xi_m = \angle(\mathbf{x}_m - \widehat{\mathbf{x}}_m). \tag{13}$$

4.1 Performance of Single-Task and Multi-Task CS

We first examine performance of various CS strategies on our hyperspectral datasets. As mentioned above, the most straightforward, at most common, paradigm of CS usage is to reconstruct only a single vector $\mathbf{x}$ from its projection $\widetilde{\mathbf{y}} = \mathbf{P}^T\mathbf{x}$. To apply this single-vector approach to recover a dataset $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$, of M vectors, we simply employ the single-vector reconstruction independently M times. In line with [26], we refer to this approach as “single-task” CS reconstruction.

In Figs. 4, 5, 6, 7, 8 and 9, we examine the performance of two prominent single-task CS reconstruction algorithms. Specifically, we compare GPSR[2] [24] and ST-BCS[3] [32], using the Matlab implementations available from their respective authors. For each technique, transform $\boldsymbol{\Psi}$ is the well-known length-8 Daubechies orthonormal wavelet. Clearly, the performance of the reconstructions in each case will depend on the degree of dataset reduction inherent in the projections; this quantity is characterized as a relative projection dimensionality in the form of K/N expressed as a percentage. For each value of K, we use exactly the same random projection matrix $\mathbf{P}$ for each algorithm.

In Figs. 4, 5, 6, 7, 8 and 9, we also contrast the performance of the various ST-CS reconstructions with the alternative to straightforward, single-vector processing, namely, “multi-task” CS recovery in the form of MT-BCS [26]. Again, we use the same random projections $\mathbf{P}$ and transform $\boldsymbol{\Psi}$. It is clear from Figs. 4, 5, and 6 that, for the same relative projection dimensionality, the multi-task reconstruction, which is able to exploit the substantial cross-vector correlations that exist within typical hyperspectral datasets, achieves reconstruction performance of significantly higher quality in terms of SNR. Additionally, a substantially smaller average spectral angle $\bar{\xi}$ is observed in Figs. 7, 8, and 9 for MT-BCS.

[2] http://www.lx.it.pt/mtf/GPSR/

[3] http://www.people.ee.duke.edu/lihan/cs/

Fig. 4 CS reconstruction performance for the "Cuprite" hyperspectral dataset—average SNR for varying dimensionality *K*/*N*

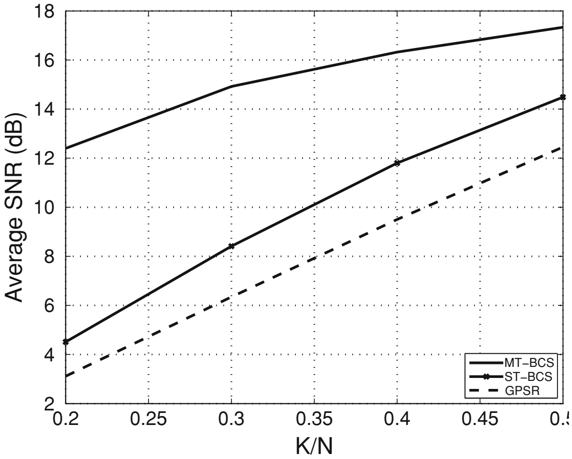


Fig. 5 CS reconstruction performance for the "Moffett" hyperspectral dataset—average SNR for varying dimensionality *K*/*N*

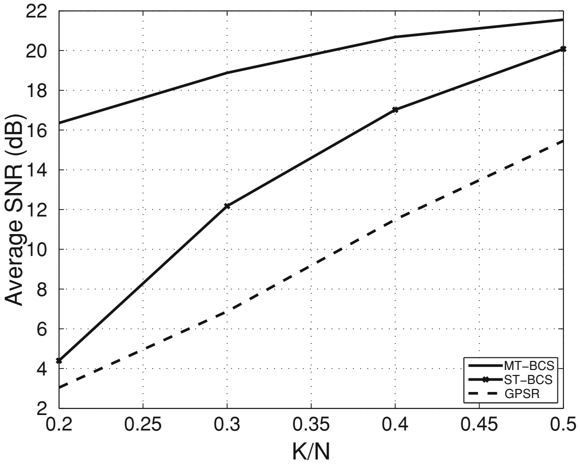
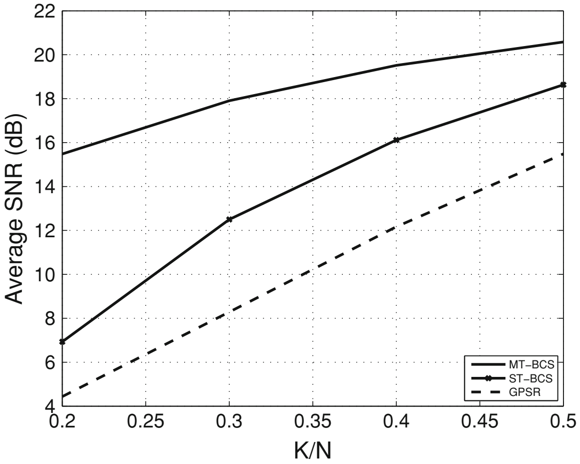


Fig. 6 CS reconstruction performance for the "Jasper Ridge" hyperspectral dataset—average SNR for varying dimensionality *K*/*N*



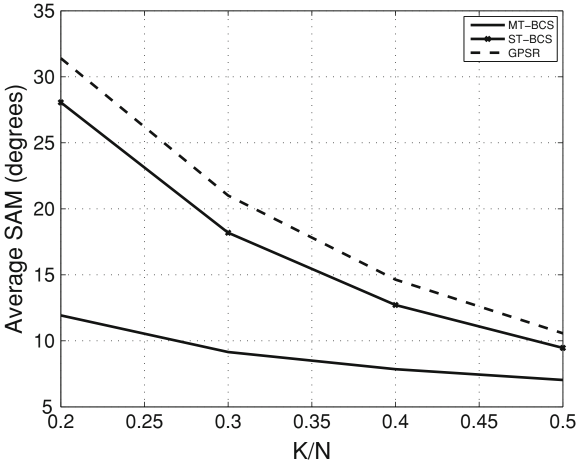


Fig. 7 CS reconstruction performance for the “Cuprite” hyperspectral dataset—average spectral angle, $\bar{\xi}$, for varying dimensionality *K*/*N*

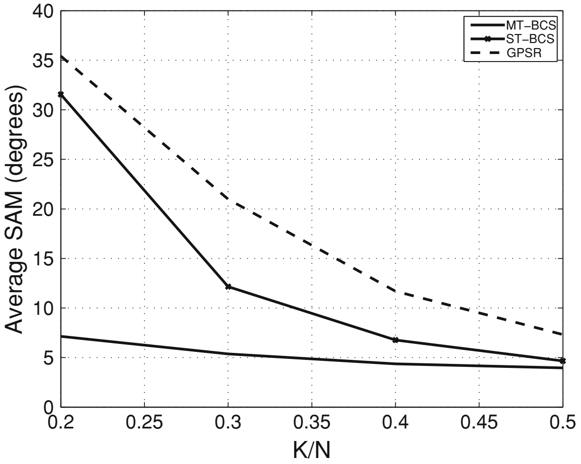


Fig. 8 CS reconstruction performance for the “Moffett” hyperspectral dataset—average spectral angle, $\bar{\xi}$, for varying dimensionality *K*/*N*

As a consequence, when we compare CS reconstruction to that of CPPCA in the next section, we limit our attention to multi-task CS.

4.2 Performance of CPPCA and CS

We now examine the performance of CPPCA reconstruction in the form of (6) and (7), comparing to MT-BCS which was seen above to be the most promising CS-based reconstruction. For CPPCA, we use $J = 20$ projection partitions while L ranges between 3 and 30, depending on the specific K used. For MT-BCS, we

Fig. 9 CS reconstruction performance for the "Jasper Ridge" hyperspectral dataset—average spectral angle, $\bar{\xi}$, for varying dimensionality *K*/*N*

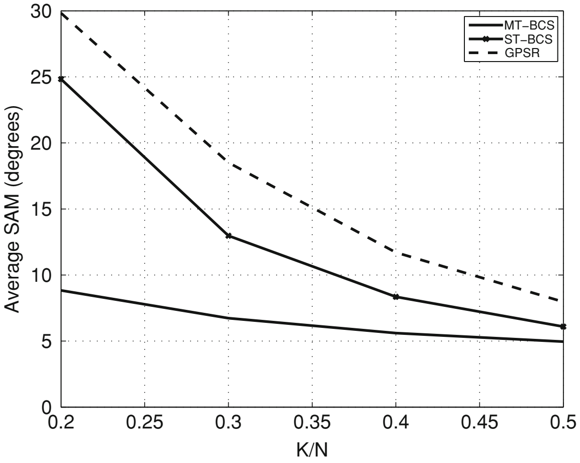


Fig. 10 CPPCA and MT-BCS reconstruction performance for the "Cuprite" hyperspectral dataset—average SNR for varying dimensionality *K*/*N*. (from [11], © 2009 IEEE)

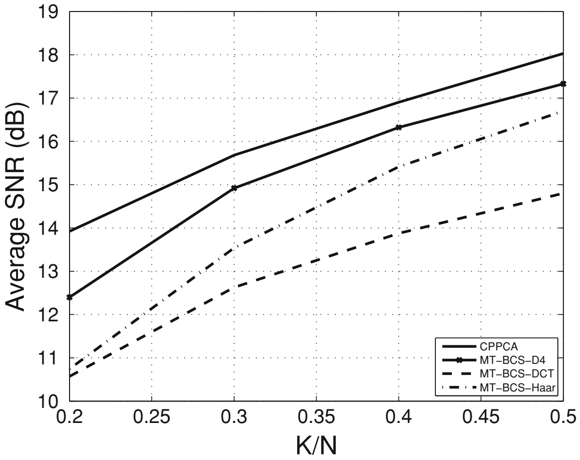
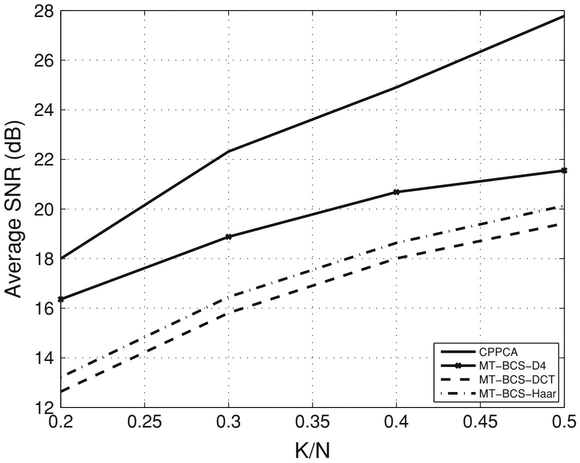


Fig. 11 CPPCA and MT-BCS reconstruction performance for the "Moffett" hyperspectral dataset—average SNR for varying dimensionality *K*/*N*



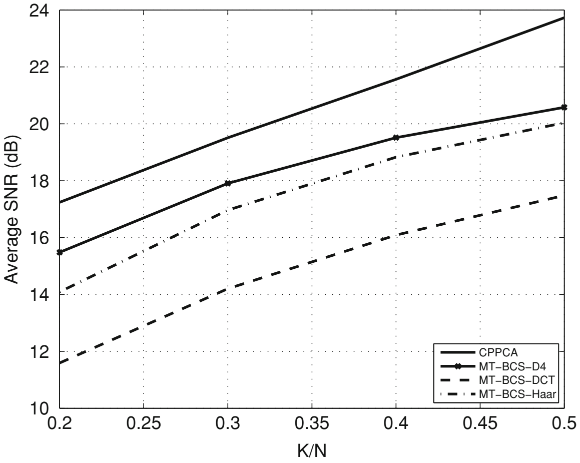


Fig. 12 CPPCA and MT-BCS reconstruction performance for the "Jasper Ridge" hyperspectral dataset—average SNR for varying dimensionality *K*/*N*. (from [11], © 2009 IEEE)

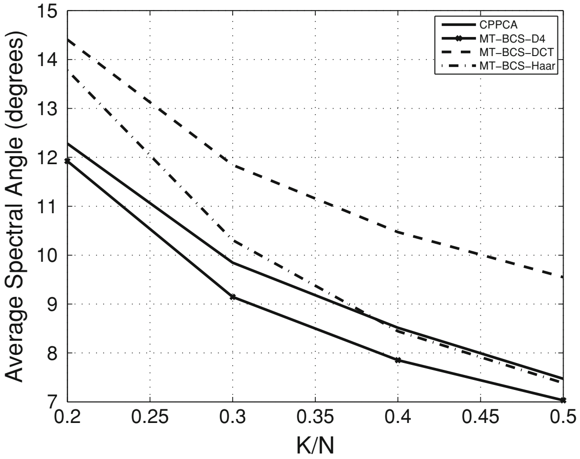


Fig. 13 CPPCA and MT-BCS reconstruction performance for the "Cuprite" hyperspectral dataset—average spectral angle, $\bar{\xi}$, for varying dimensionality *K*/*N*

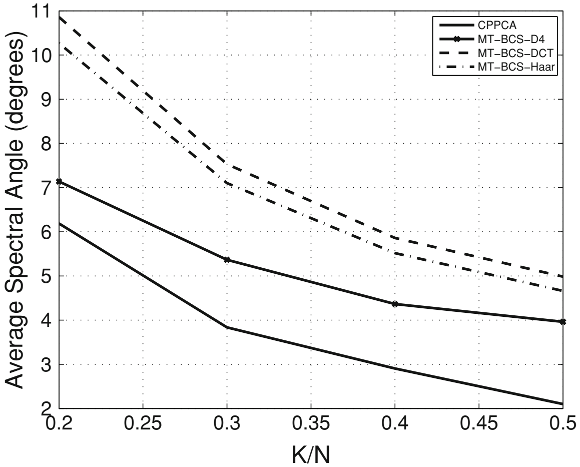


Fig. 14 CPPCA and MT-BCS reconstruction performance for the "Moffett" hyperspectral dataset—average spectral angle, $\bar{\xi}$, for varying dimensionality *K*/*N*

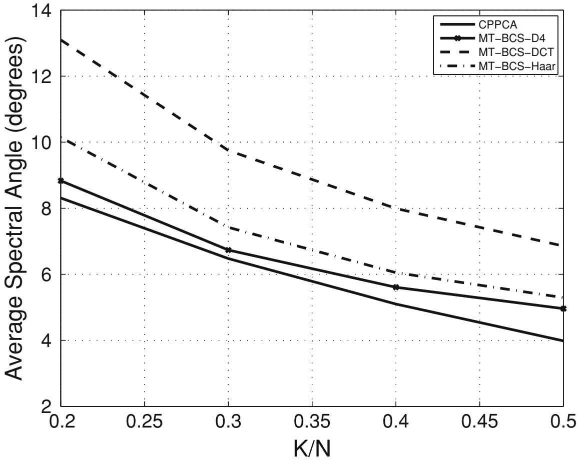


Fig. 15 CPPCA and MT-BCS reconstruction performance for the "Jasper Ridge" hyperspectral dataset—average spectral angle, $\bar{\xi}$, for varying dimensionality *K*/*N*

Table 1 Single-thread execution times in seconds for the "Cuprite" hyperspectral dataset

	K/*N*				
Algorithm	0.1	0.2	0.3	0.4	0.5
CPPCA	4	4	9	16	25
GPSR-D4	59	76	78	72	66
ST-BCS-D4	471	928	1194	1255	1294
MT-BCS-D4	1585	1507	1614	1307	1344

consider several orthonormal bases commonly used with hyperspectral data: an *N*-point discrete cosine transform (MT-BCS-DCT) as well as a discrete wavelet transform (DWT) using both the Haar basis (MT-BCS-Haar) and the length-8 Daubechies basis (MT-BCS-D4). We apply the same random projections as used for CPPCA. We see from Figs. 10, 11, and 12 that CPPCA yields average SNR substantially higher than that of the fixed-basis MT-BCS approaches over a broad range of practical *K*/*N* values. Additionally, in Figs. 13, 14, and 15, we see that CPPCA also usually produces a smaller spectral-angle distortion, the sole exception being for the "Cuprite" dataset.

4.3 Execution Times

In terms of computational complexity, none of the implementations we employ are optimized for execution speed. However, we have observed that both the POCS-based eigenvector recovery of (6) as well as the linear coefficient recovery of (7) are quite fast, yielding a relatively lightweight computational burden for the CPPCA receiver. To wit, Table 1 presents execution times for the various algorithms we have considered. In particular, we compare the execution speed of CPPCA and that of the various single-task and multi-task CS reconstructions.

The CS algorithms all use the length-8 Daubechies DWT as the sparsity transform. All implementations are run on a single core of a Sun Fire X4600 (2.6 GHz, 32 GB RAM) using Matlab R2009b running in a single thread. Table 1 presents execution times for several relative dimensionalities *K*/*N*. We find that CPPCA generally runs 3–10 times faster than GPSR, around 50–200 times faster than ST-BCS, and about 50–400 times faster than MT-BCS.

5 Conclusions

In this chapter, we have compared the performance of CPPCA to various CS reconstruction algorithms for the recovery of hyperspectral datasets subject to random-projection-based dimensionality reduction. We have seen that multi-task CS reconstruction significantly outperforms single-task CS recovery since the multi-task technique is designed specifically to exploit the significant correlation that typically exists between hyperspectral pixel vectors. On the other hand, CPPCA usually outperforms multi-task CS, consistently achieving higher SNR for all the AVIRIS datasets we consider in addition to a smaller average spectral-angle distortion for all but one of the datasets. CPPCA, which features a recovery of not only the PCA transform coefficients for the dataset in question but also an approximation to the PCA transform basis from the random projections, runs at a fraction of the computational cost as compared to the various CS approaches considered. As a consequence, we conclude that CPPCA constitutes a promising approach for computationally efficient and high-quality receiver-side reconstruction when random projection is used at the remote sender to accomplish low-cost dimensionality reduction.

References

1. Jolliffe, I.T.: Principal Component Analysis. Springer-Verlag, New York (1986)
2. Prasad, S., Bruce, L.M.: Limitations of principal component analysis for hyperspectral target recognition. IEEE Geosci. Remote Sens. Lett. **5**(4), 625–629 (2008)
3. Wang, J., Chang, C.-I.: Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. IEEE Trans. Geosci. Remote Sens. **44**(6), 1586–1600 (2006)
4. Lennon, L., Mercier, G., Mouchot, M.C., Hubert-Moy, L.: Independent component analysis as a tool for the dimensionality reduction and the representation of hyperspectral images. In: Proceedings of the International Geoscience and Remote Sensing Symposium, vol. 6, Sydney, Australia, July 2001, pp. 2893–2895
5. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
6. Mohan, A., Sapiro, G., Bosch, E.: Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images. IEEE Geosci. Remote Sens. Lett. **4**(2), 206–210 (2007)

7. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
8. Candès, E., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies?. IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
9. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
10. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)
11. Fowler, J.E.: Compressive-projection principal component analysis for the compression of hyperspectral signatures. In: Storer, J.A., Marcellin, M.W. (eds.), Proceedings of the IEEE Data Compression Conference, Snowbird, UT, March 2008, pp. 83–92
12. Fowler, J.E.: Compressive-projection principal component analysis and the first eigenvector. In: Storer, J.A., Marcellin, M.W. (eds.), Proceedings of the IEEE Data Compression Conference, Snowbird, UT, March 2009, pp. 223–232
13. Fowler, J.E.: Compressive-projection principal component analysis. IEEE Trans. Image Process. **18**(10), 2230–2242 (2009)
14. Willett, R.M., Gehm, M.E., Brady, D.J.: Multiscale reconstruction for computational spectral imaging. In: Bouman, C.A., Miller, E.L., Pollak, I. (eds.), Computational Imaging V, Proceedings of SPIE 6498, San Jose, CA, January 2007, p. 64980L
15. Pitsianis, N.P., Brady, D.J., Portnoy, A., Sun, X., Suleski, T., Fiddy, M.A., Feldman, M.R., TeKolste, R.D.: Compressive imaging sensors. In: Athale, R.A., Zolper, J.C. (eds.), Intelligent Integrated Microsystems, Proceedings of SPIE 6232, Kissimmee, FL, April 2006, p. 62320A
16. Brady, D.J.: Micro-optics and megapixels. Opt. Photon. News **17**(11), 24–29 (2006)
17. Parlett, B.N.: The Symmetric Eigenvalue Problem. Society for Industrial and Applied Mathematics, Philadelphia (1998)
18. Combettes, P.L.: The foundations of set theoretic estimation. Proc. IEEE **81**(2), 182–208 (1993)
19. Penna, B., Tillo, T., Magli, E., Olmo, G.: A new low complexity KLT for lossy hyperspectral data compression. In: Proceedings of the International Geoscience and Remote Sensing Symposium, vol. 7, Denver, CO, August 2006, pp. 3525–3528
20. Du, Q., Fowler, J.E.: Low-complexity principal component analysis for hyperspectral image compression. Int. J. High Perform. Comput. Appl. **22**(4), 438–448 (2008)
21. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
22. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
23. Candès, E., Romberg, J.: ℓ_1-MAGIC: Recovery of Sparse Signals via Convex Programming. California Institute of Technology, Pasadena (2005)
24. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Areas Commun. **1**(4), 586–597 (2007)
25. Do, T.T., Gan, L., Nguyen, N., Tran, T.D.: Sparsity adaptive matching pursuit algorithm for practical compressed sensing. In: Proceedings of the 42th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, October 2008, pp. 581–587
26. Ji, S., Dunson, D., Carin, L.: Multitask compressive sensing. IEEE Trans. Signal Process. **57**(1), 92–106 (2009)
27. Fornasier, M., Rauhut, H.: Recovery algorithms for vector-valued data with joint sparsity constraints. SIAM J. Numer. Anal. **46**(2), 577–613 (2008)
28. Mishali, M., Eldar, Y.C.: Reduce and boost: recovering arbitrary sets of jointly sparse vectors. IEEE Trans. Signal Process. **56**(10), 4692–4702 (2008)
29. Wipf, D.P., Rao, B.D.: An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. IEEE Trans. Signal Process. **55**(7), 3704–3716 (2007)

30. Tropp, J.A., Gilbert, A.C., Strauss, M.J.: Algorithms for simultaneous sparse approximation. Part I: greedy pursuit. Signal Process. **86**(3), 572–588 (2006)
31. Tropp, J.A.: Algorithms for simultaneous sparse approximation. Part II: convex relaxation. Signal Process. **86**(3), 589–602 (2006)
32. Ji, S., Xue, Y., Carin, L.: Bayesian compressive sensing. IEEE Trans. Signal Process. **56**(6), 2346–2356 (2008)

Integrated Sensing and Processing for Hyperspectral Imagery

Robert Muise and Abhijit Mahalanobis

Abstract In this chapter, we present an information sensing system which integrates sensing and processing resulting in the direct collection of data which is relevant to the application. Broadly, integrated sensing and processing (ISP) considers algorithms that are integrated with the collection of data. That is, traditional sensor development tries to come up with the "best" sensor in terms of SNR, resolution, data rates, integration time, and so forth, while traditional algorithm development tasks might wish to optimize probability of detection, false alarm rate, and class separability. For a typical automatic target recognition (ATR) problem, the goal of ISP is to field algorithms which "tell" the sensor what kind of data to collect next and the sensor alters its parameters to collect the "best" information in order that the algorithm performs optimally. We illustrate the concept of ISP using a near Infrared (NIR) hyperspectral imaging sensor. This prototype sensor incorporates a digital mirror array (DMA) device in order to realize a Hadamard multiplexed imaging system. Specific Hadamard codes can be sent to the sensor to realize inner products of the underlying scene rather than the scene itself. The developed ISP algorithms utilize these codes to overcome issues traditionally associated with hyperspectral imaging (i.e. Data Glut and SNR issues) while also performing a object detection task. The underlying integration of the sensing and processing results in algorithms which have better overall performance while collecting less data.

Keywords Hyperspectral imaging · Adaptive imaging · Compressive imaging · Hadamard multiplexing

R. Muise (✉) and A. Mahalanobis
Lockheed Martin Missiles and Fire Contro, 5600 Sand Lake Road, MP450 Orlando, FL 32819, USA
e-mail: robert.r.muise@lmco.com

A. Mahalanobis
e-mail: abhijit.mahalanobis@lmco.com

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_4,

1 Introduction

This chapter presents the development of algorithms for Integrated Sensing and Processing (ISP) utilizing a hyperspectral imaging sensor. The ISP paradigm seeks to determine the best sensing parameters for achieving the performance objectives of a given algorithm. The exploitation algorithm may also have components which adapt to the imagery being sensed. In this context, ISP is a coupling between adaptive algorithms and adaptive sensing. Considering the problem of object detection/classification in hyperspectral imagery, ISP can increase sensing and algorithm performance in several ways. Firstly, hyperspectral exploitation usually suffers from a data glut problem. That is, a hyperspectral sensor generates a cube of data where each spatial pixel is represented as a spectral vector. The first step in most exploitation algorithms is some type of data reduction, or spectral band selection. A question which should naturally arise is: Why sense particular information which is going to be immediately eliminated through a data reduction algorithm? If one can design a data collection system that integrates the sensor with the data reduction algorithm, then only information which is pertinent to the exploitation task need be sensed. Secondly, traditional hypserspectral imagers can suffer SNR degradation as compared with broadband imagers. When one is attempting high spatial resolution imaging and the sensing system separates the light into a large number of spectral components, then there is a significant loss of photons being sensed by the detector array. Thus, to get enough light to make a meaningful image, one must increase the detector integration time. If one is sensing a dynamic scene, longer integration time cannot be tolerated; which leads to significant loss of SNR in the final hyperspectral image. In Sect. 2, we show a solution to this SNR issue using spatial/spectral multiplexing.

In order to investigate algorithms for integrated sensing and processing of imagery, we use a near infrared (NIR) Hadamard multiplexing imaging sensor. This prototype sensor was developed by PlainSight Systems (PSS) and incorporates a digital mirror array (DMA) device in order to realize a Hadamard multiplexed imaging system. The known Signal-to-Noise (SNR) advantage in Hadamard spectroscopy [1] extended to imaging systems [2, 3] allows for the collection of a hyperspectral cube of data with more efficient light collection over standard "Pushbroom" hyperspectral imagers.

Fig. 1 Digital mirror array acts as an electronic shutter to select and encode spatial/spectral features in the scene

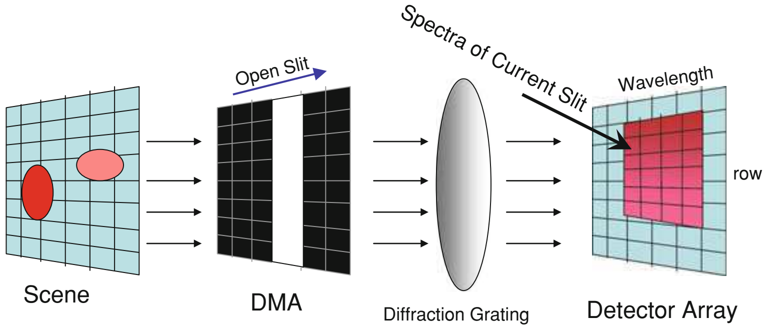


Fig. 2 Pushbroom hyperspectral imaging with a DMA device

The PlainSight NSTIS is a Spatial Light Modulator (SLM)-based multiplexing hyperspectral imaging camera, operable in the spectral range of about 900–1,700 nm, with no macro moving parts. As the SLM device, the system uses a Digital Micro-mirror Array (DMA) commercially available by Texas Instruments for projector display applications. The DMA contains 848 columns and 600 rows of mirrors and measures 10.2 mm × 13.6 mm. In Fig. 1, a DMA is shown with its glass cover removed**.**

When the scene is illuminated on the DMA device, a standard raster scan could be implemented by turning the first column of mirrors ON, sending this column to a diffraction grating which results in a spectral representation of the first spatial column of the scene being illuminated on the detector array. This process is reflected in Fig. 2.

If multiple slits (columns) in the DMA array are opened as shown in Fig. 3, the detector array will be presented with the superposition of the spectra of many columns. Such a system has the advantage of realizing optimal SNR when the sequence of open slits constitutes a Hadamard pattern [1]. Each individual frame collected at the detector array is not physically meaningful as an image, but when all the patterns of the Hadamard sequence have been recorded, the full hyperspectral data cube is recoverable by digital post-processing [2].

The PlainSight NSTIS sensor implements the process from Fig. 3 where the detector array is a standard Indigo Phoenix large-area InGaAs camera operating in the Near Infrared wavelengths.

During standard operation of the system, the sensor collects 512 raw frames of data. Each frame is 522 × 256 pixels and represents superposition of spectra vs. spatial row as shown in Fig. 3. The 512 frames are collected using the 256 Walsh (0 and 1 s) patterns that determine the columns of the DMA to be opened or closed. In other words, each column of the DMA is controlled by a bit of the Walsh code. If the bit is 0, the column is closed whereas if the bit is 1, the column is open. Since the theory of optimal SNR is based upon Hadamard (1 and -1 s) patterns, one needs to collect two Walsh patterns to generate a single Hadamard pattern. Thus, the 512 collected frames represent the required Walsh patterns to form a full set of 256 Hadamard patterns. Since each column in the DMA array

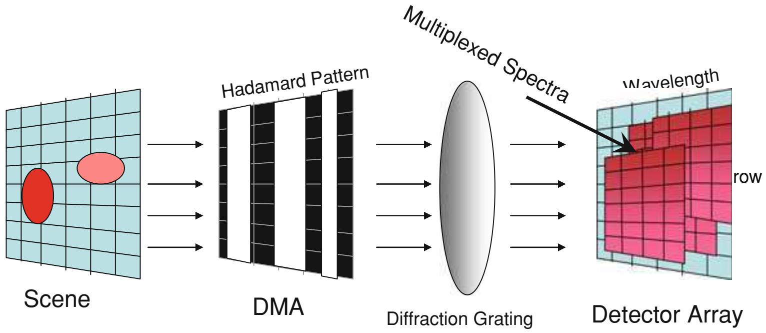


Fig. 3 Multiplexed Hadamard hyperspectral imaging

will hit the Diffraction grating at a different location, the spectra will hit the detector array at a different location per column. We refer to this as a *Skewness* in spectra which spreads the information across 522 pixels in the spectral dimension but represents only 266 actual spectral bins. Of course, this spatial/spectral mixing and skewness is invertable once all 256 Hadamard patterns have been collected. The resultant hyperspectral scene is dimension 256 × 256 with 266 spectral bands from 900 to 1,700 nm.

Given a sensor that accommodates adaptation while imaging, the ISP concepts we will discuss can be viewed as within the realm of compressive sensing (as presented by Donoho [4] and Candes et al. [5]) in that we will collect far fewer image samples than what would normally be required to exploit the entire scene of interest. Neifeld and Shankar [6] have done similar work on concepts for feature-specific imaging while Mahalanobis and Daniel [7] have looked at exploitation driven compression algorithms (another form of ISP).

The outline of this chapter is as follows. Section 2 presents an algorithm for variable resolution sensing where high resolution imagery is driven by an ATR metric. Section 3 presents the results of an experiment which demonstrates the developed algorithms implemented in a prototype ISP hyperspectral sensor, while Sect. 4 presents concluding remarks and future work.

2 Variable Resolution Hyperspectral Sensing

2.1 *Mathematical Representation*

Since the sensor encodes data identically and independently on each spatial column of the scene, we will perform the mathematical analysis given an individual, but arbitrary spatial column. Thus, for the underlying hyper-spectral scene, $S(\lambda, r, c)$, we will consider only a particular column of data $S(\lambda, r)$. We wish to establish a correspondence between the sampling of the hyperspectral row, $S(\lambda, r)$, as a digital image and a particular mirror of the DMA device. As described in the

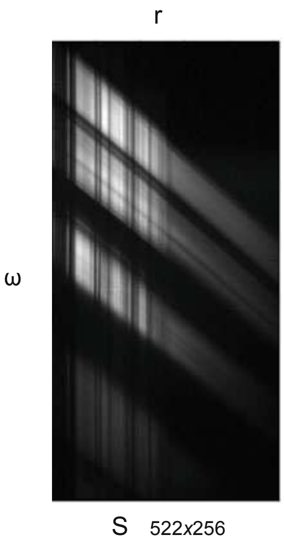


Fig. 4 Example scene matrix S representing the wavelength × spatial row information for a given spatial column. Collected from multiplexing hyperspectral imager

description of Fig. 3, each row of the scene hits the diffraction grating at a different place, and thus the entire spectrum is shifted on the focal plane as a function of the row. This is referred to as spectral "skewness". Thus, as a particular row enters the system, the underlying scene actually becomes $S(\lambda(r), r)$, where the spectrum is now a function of row. We now make the substitution $\omega = \lambda(r)$ and ignore this dependency for the moment. So we are concerned with sensing the hyperspectral row image $S(\omega, r)$. The sampling of this function brought about from the DMA generates a matrix S of dimension 522 × 256. We are thus interested in sensing this array with Hadamard vectors of length 256. An example scene matrix, S, is given in Fig. 4. Recall, this is a spectral x spatial data matrix, so there is no intrinsic interpretability.

The imaging system will sense this data from Fig. 4 with Hadamard multiplexing, thus we will measure a collection of transformations of this data rather than the data itself. Looking at the Hadamard basis, we are interested in encoding the spatial component of this data which is of dimension 256. We take a standard ordering of the Hadamard basis for $\Re^{256}$ as shown in the example below which is of dimension 8.

$$\mathrm{H}_8 = \begin{bmatrix} 1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ 1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ 1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ 1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ 1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\ 1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ 1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ 1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \end{bmatrix}.$$

It is important to note that

$$H_N = H_N^T = H_N^{-1}. \tag{1}$$

A particular single frame sensed by the camera when in multiplexed mode is resultant by a particular column of this Hadamard matrix, denoted h_i. This ith frame of collected data is the 522 × 1 vector

$$f_i = Sh_i. \tag{2}$$

Therefore, sensing with all 256 Hadamard codes yields the 522 × 256 data matrix.

$$F = SH_{256}. \tag{3}$$

This is the data which gets collected during normal operation of the sensor. It is easy to see that to exactly recover the underlying scene, **S**, from the sensed frames, **F**, we use Eq. 1 to get

$$S = FH_{256}. \tag{4}$$

This implies that if all 256 Hadamard vectors are sequentially encoded into the mirror array and sensed through the camera, then we can fully recover S from the actual collected data F. Performing this recovery on all spatial columns of this data will recover the full hyperspetral data cube.

The relationship between the underlying spectral parameter and our indexing parameter was previously given by

$$\omega = \lambda(r). \tag{5}$$

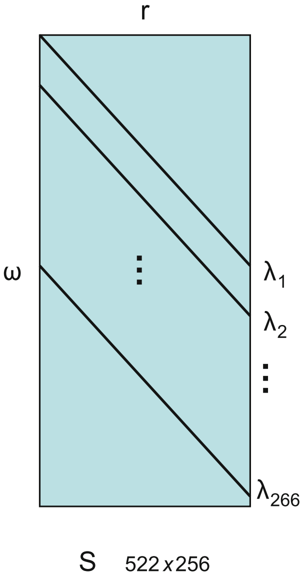


Fig. 5 Spectral skewness for the actual hyperspectral scene as sensed through the multiplexing Hadamard hyperspectral imager. The *diagonal lines* show the lines of constant wavelength

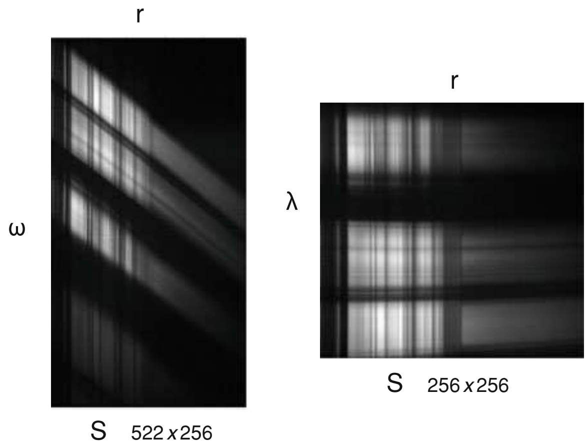


Fig. 6 Skewness correction in scene representation. *Left* is skewed data representation while *right* is the unskewed version following lines of constant wavelength from Fig. 5

Specifically, the sensing instrument being used for this discussion introduces a spectral “skewness” where the underlying spectral representation is shown in Fig. 5.

Thus, the “unskewed” scene, $S(\lambda, r)$, representation can be garnered from the recovered data, $S(\omega, r)$, by following the lines of constant wavelength from Fig. 5. This is illustrated in Fig. 6.

2.2 Reduced Resolution Imaging

Equation 4 implies that if all 256 Hadamard vectors are sequentially encoded into the mirror array and sensed through the camera, then we can fully recover the scene, S, from the data frames, F, collected by the sensor. Since we are interested in compressed sensing, we desire to know what can be recovered about S if we sense only a few of the Hadamard vectors. Consider, for example if we sense only the first 4 codes of the Hadamard matrix.

$$\mathbf{H}_{256,4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \vdots & 1 & 1 \\ 1 & \vdots & -1 & -1 \\ \vdots & 1 & -1 & -1 \\ \vdots & -1 & 1 & -1 \\ 1 & \vdots & 1 & -1 \\ 1 & \vdots & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}. \qquad (6)$$

Then, we have sensed the four vectors

$$F_{256,4} = SH_{256,4} = [Sh_1 \quad Sh_2 \quad Sh_3 \quad Sh_4]. \tag{7}$$

The dimension of F is 522 × 4. Define the approximate scene by $\hat{S}$ as follows:

$$\begin{aligned}\hat{S} = F_{256,4}H_{256,4}^T = [Sh_1 \quad Sh_2 \quad Sh_3 \quad Sh_4]H_{256,4}^T = S[h_1 \quad h_2 \quad h_3 \quad h_4]\begin{bmatrix} h_1^T \\ h_2^T \\ h_3^T \\ h_4^T \end{bmatrix} \\ = S\left[h_1h_1^T + h_2h_2^T + h_3h_3^T + h_4h_4^T\right].\end{aligned} \tag{8}$$

However, if we let 1_{64} be the 64 × 64 matrix of all 1 s, then it can be shown that

$$\left[h_1h_1^T + h_2h_2^T + h_3h_3^T + h_4h_4^T\right] = 4\begin{bmatrix} 1_{64} & 0 & \cdots & 0 \\ 0 & 1_{64} & & \vdots \\ \vdots & & 1_{64} & 0 \\ 0 & \cdots & 0 & 1_{64} \end{bmatrix}. \tag{9}$$

Thus,

$$\hat{S} \propto S\begin{bmatrix} 1_{64} & 0 & \cdots & 0 \\ 0 & 1_{64} & & \vdots \\ \vdots & & 1_{64} & 0 \\ 0 & \cdots & & 1_{64} \end{bmatrix}. \tag{10}$$

So the underlying scene is approximated by averaging. (i.e. the first 64 columns of S are averaged and establish the first 64 columns of the approximation). Again, the dimensions of the matrix $S(\omega, r)$, are wavelength, ω, and spatial row, r, implying that the spatial row information in $\hat{S}$ is the average of 64 spatial rows of the scene: A low pass filtering in the spatial row dimension. In the wavelength dimension it is somewhat more complicated. It would appear that the data in the wavelength dimension is not smoothed along the wavelength axis, but simply averaged over 64 spatial rows. However, we recall that $\omega = \lambda(r)$ is a function of the real spectral parameter which has an index which is a function of the spatial row. Thus, $\hat{S}$, the coarse scale approximation to S, smoothes S in both the wavelength and spatial row dimensions.

The lines of constant wavelength used for the coarse resolution "deskewing" are represented in Figs. 7 and 8.

At this point, one can apply a metric to the reduced resolution imagery which defines which spatial areas are to be sensed at a finer resolution. The process can then continue until the highest resolution possible is achieved over the spatial areas desired by the controlling metric.

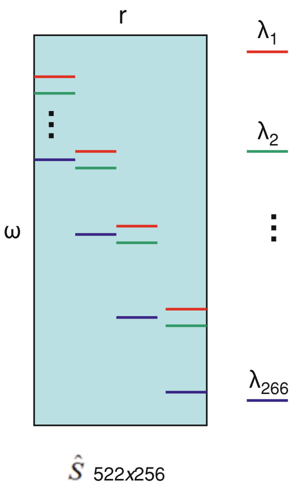


Fig. 7 Spectral skewness for the actual coarse resolution scene. The *diagonal lines* show the lines of constant wavelength

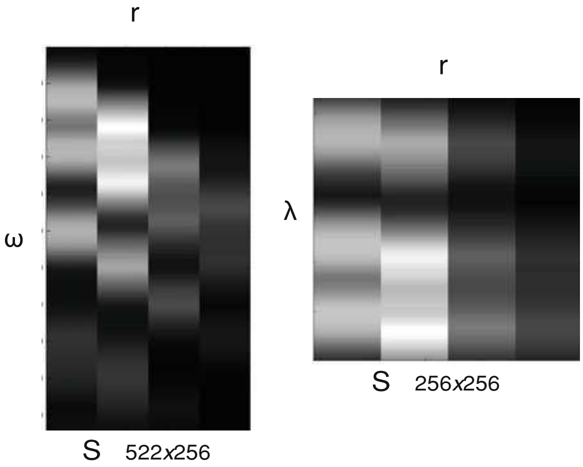


Fig. 8 Skewness correction in coarse scale scene representation. *Left* is skewed data approximation while *right* is the unskewed version following lines of constant wavelength from Fig. 7

To garner more understanding of this process, we need more detail. With the equations for a coarse scale approximation of the scene defined by sensing four frames of data,

$$\hat{S} = F_{256,4}H^{T}_{256,4}, \tag{11}$$

we are in a position to establish some type of measurable criteria as to whether any of the data matrix $\hat{S}$, needs to be approximated to a finer resolution. We will later establish an Automatic Target Recognition (ATR) criteria in more detail, but for now we will just assume that a decision is made by some criteria as to where in the array needs further detail. The extra approximation detail is collected in the same manner as previously described for the 256 row dimensional case. That is, we will consider each of the four dyadic spatial row “blocks” which have been averaged in the first approximation. The next level approximation will be made with the reduced size Hadamard basis set $H_{64,4}$. Thus, we will sense the scene as

$$F_{64,4} = SH_{64,4} = [\,Sh_1 \quad Sh_2 \quad Sh_3 \quad Sh_4\,], \tag{12}$$

where

$$\mathbf{H}_{64,4} = \left.\overbrace{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \vdots & 1 & 1 \\ 1 & \vdots & -1 & -1 \\ \vdots & 1 & -1 & -1 \\ \vdots & -1 & 1 & -1 \\ 1 & \vdots & 1 & -1 \\ 1 & \vdots & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}}^{4}\right\}64. \tag{13}$$

With this formalization,

$$\hat{S}_{64} = F_{64,4}H_{64,4}^{T}. \tag{14}$$

For the purpose of illustration, if we assume that the second dyadic block has been flagged for finer resolution approximation, then the next level approximation becomes

$$\hat{S}_{64} \propto S\begin{bmatrix} 1_{64} & 0 & \cdots & 0 \\ 0 & \begin{bmatrix} 1_{16} & 0 & \cdots & 0 \\ 0 & 1_{16} & & \vdots \\ \vdots & & 1_{16} & 0 \\ 0 & \cdots & 0 & 1_{16} \end{bmatrix} & & \vdots \\ \vdots & & 1_{64} & 0 \\ 0 & \cdots & 0 & 1_{64} \end{bmatrix}. \tag{15}$$

This procedure continues until the criteria for further resolution processing is not satisfied with any of the remaining dyadic blocks. The final approximation will be something of the form

$$\hat{S}_{\text{final}} \propto S[\text{diag}[1_{k_1}, 1_{k_2}, \ldots, 1_{k_M}]], \tag{16}$$

with the parameters $\{k_1, k_2, \ldots, k_M\}$defining the local resolution and are determined iteratively by the controlling criteria metric. For example, a spectral MACH [8] filter could be inserted at this stage as a controlling criteria for finer resolution sampling.

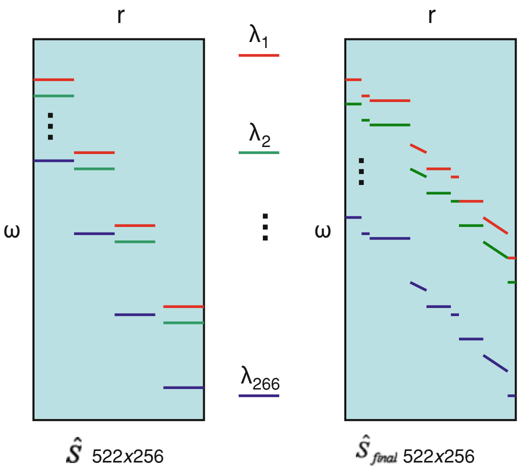


Fig. 9 Spectral skewness for multi-resolution scene. *Left* is for coarse scale representation while the *right* is for the final multi-resolution scene representation

For the final multi-resolution scene representation, the "skewness" is also multiple scales. This results in a set of piecewise linear lines of constant wavelength denoted by Fig. 9. An example on real imagery is given in Sect. 3.

3 Experimental Results

In this section, we describe experiments to demonstrate (i) the improvement in SNR by using a Hadamard encoded coded aperture, and (ii) the benefit of ISP by imaging a scene in variable resolution that dramatically reduces the amount of raw hyperspectral data which must be collected. The sensor was placed in a data collection tower and imagery was collected of a surrogate "tank" target emplaced in the grass below the tower. Figure 10 shows the target emplacement with a regular visible camera with approximately the same spatial resolution as the hyperspectral sensing system. The associated example imagery taken from the hyperspectral sensor is shown in Fig. 11, where the image is spectral band 210.

3.1 Improving SNR Using Hadamard Multiplexing

The first algorithm demonstrated was in hyperspectral imaging. The SNR gain from Hadamard multiplexing was tested by gathering a hyperspectral data cube in a standard raster scan mode. Several different cubes of the same scene were collected in this mode so that "signal" and "noise" cubes could be estimated. The "signal" cube was estimated as the average data cube and the "noise" cube was taken as the signal subtracted from each collected cube. With these estimates for signal and noise, a signal-to-noise ratio was calculated. For a 16 ms integration

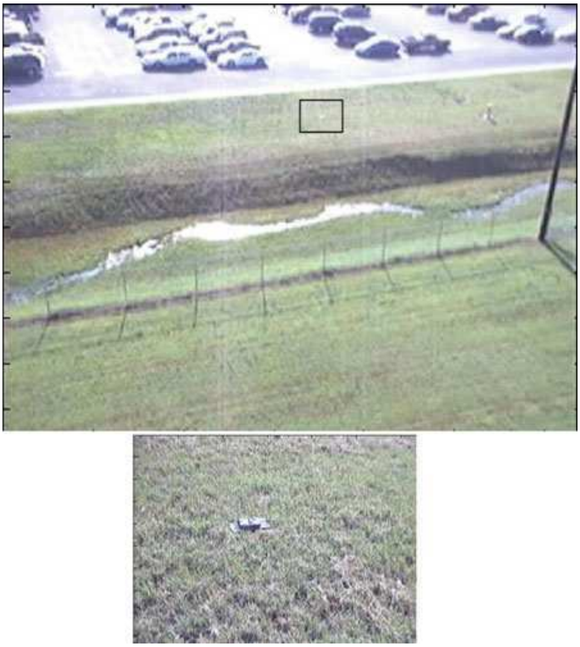


Fig. 10 *Top*: target emplacement shown with visible sensor. Tank target is inside the *box*. *Bottom*: close-up of target taken with visible camera standing directly in front of target

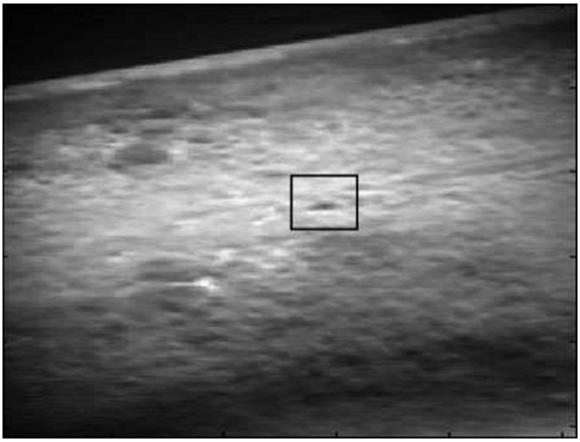


Fig. 11 Band 210 from hyperspectral sensor of target area. Tank target is inside the *box*

time per frame, the SNR in raster scan mode was calculated as 12 dB while taking imagery with a 1 ms integration time yielded an SNR of 3 dB. Samples of typical imagery collected in raster scan mode are shown in Fig. 12.

The same sensing and computations were conducted with the sensor set to Hadamard multiplexing mode. The SNR gain becomes clear as the 16 ms integration time yields an SNR of 17 dB. This reflects a 5 dB gain in SNR. For the 1 ms integration time the SNR improvements are more dramatic. The Hadamard multiplexing mode increases the SNR from 3 to 13 dB, a 10 dB improvement.

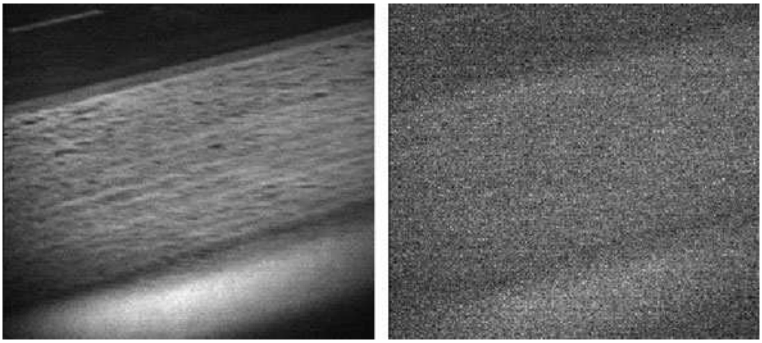


Fig. 12 Band 20 from hyperspectral sensor in raster scan mode (*left*: 16 ms integration time; *right*: 1 ms integration time)

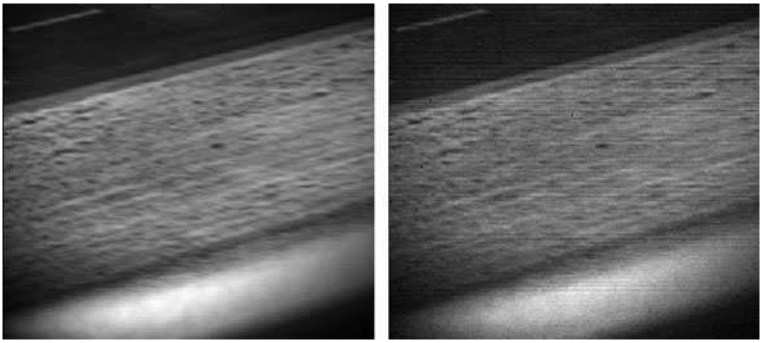


Fig. 13 Band 20 from hyperspectral sensor in Hadamard multiplexing mode (*left*: 16 ms integration time; *right*: 1 ms integration time)

Figure 13 presents typical imagery collected in Hadamard multiplexing mode. The gain in SNR becomes more pronounced as the light level is decreased. One notices that the 1 ms raster scan image contains virtually no signal information while the 1 ms Hadamard multiplexing image is comparable to the 16 ms raster scan image. The SNR gain for Hadamard multiplexing imaging is qualitatively supported by this experiment.

3.2 *Variable Resolution Hyperspectral Sensing*

This experimental setup was then used to test the variable resolution hyperspectral sensing algorithm described in Sect. 2. A training cube was collected and the average target spectral vector was calculated. This vector was then taken as the driving characteristic for what areas are identified for finer resolution sensing. For example, at each sensing level, the current approximate data cube is compared

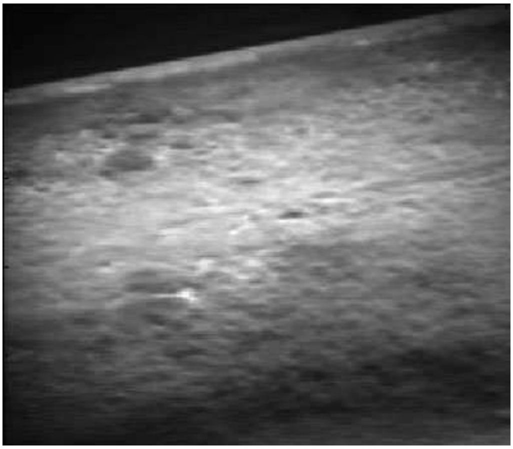


Fig. 14 Band 210 from hyperspectral sensor used for training the variable resolution imaging algorithm

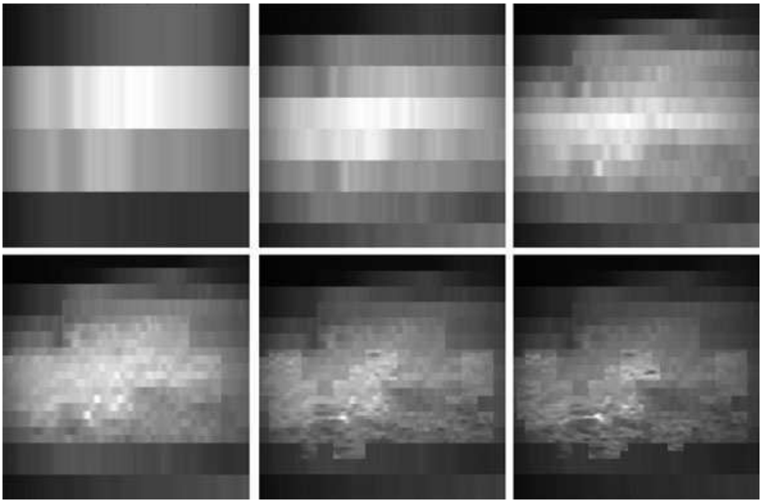


Fig. 15 Band 210 from hyperspectral sensor in variable resolution sensing mode: increased resolution progresses from the *top left* to the *bottom right*. Note the full resolution on the target and less resolution elsewhere

against the average target spectra. The L^1 norm is used for comparison and if this norm is smaller than a defined threshold, then that resolution cell is identified as requiring more resolution. The sensing continues in this manner until the highest possible resolution is attained. Figure 14 shows band 210 of the hyperspectral scene used for training. With the training signature calculated, the sequence of collected frames is shown in Fig. 15.

One notes that the final variable resolution collected by the sensor has full resolution on the target and less resolution elsewhere. Also, the sensing terminated on the parking lot area (top left of image) after the very coarsest resolution was collected. The final collected variable resolution data cube results from sensing

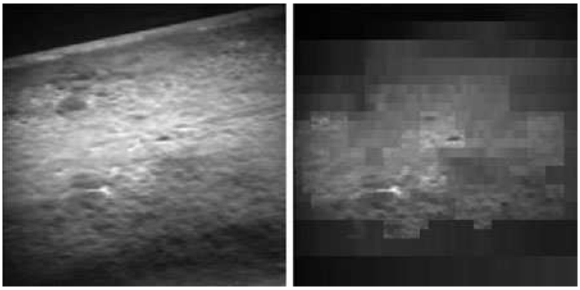


Fig. 16 Band 210 from hyperspectral sensor (*left*: full resolution mode; *right*: variable resolution mode)

only 14% of the pixels required for full resolution everywhere. This represents a substantial savings in sensing resources as well as addressing the typical data glut problem associated with hyperspectral data exploitation. The full resolution and variable resolution images are shown in Fig. 16 for comparison.

The next example in Fig. 17 shows the output of the algorithm adapted to generate fine resolution only where a certain spatial recognition criteria has been satisfied. Although any algorithm can be used, our metric is a spatial correlation filter which has been designed to detect the shape of the vehicle in the center of the scene.

Essentially, the image cube is further resolved by applying additional Hadamard vectors to sense only in regions where there is a potential match with the object of interest as determined by the response of the filter to low-resolution data. Large portions of the scene in the background and fore-ground are discontinued early in the sensing process, whereas resolution is progressively added only to the region that exhibit peaks that are potentially due to the car. This approach improves the sensing process by greatly reducing the overall volume of data and the time required to collect it. In the end, only the spatial information that is salient for the object recognition algorithm to recognize the car is gathered in detail.

4 Summary

The concept of Integrated Sensing and Processing (ISP) is a unique way to address the issue of large amounts of data associated with hyperspectral imaging.

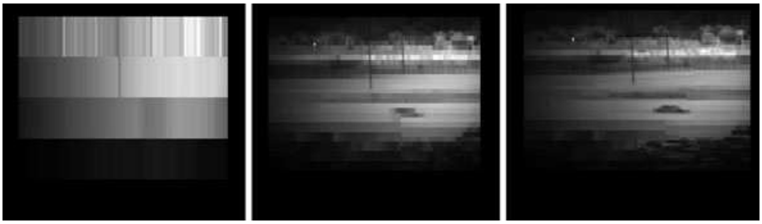


Fig. 17 The image is sensed with task-specific compressed sensing algorithm based upon a Hadamard multiplexing sensor. The resulting multi-resolution image is represented at the far right and shows fine resolution on the target and coarse resolution elsewhere

Typically, much of the data collected by a conventional sensor is often not of interest and discarded during analysis. In this chapter, we discussed a coded aperture hyperspectral imager that allows data to be collected at variable resolution by dynamically controlling the aperture. In an ISP framework, the sensor can collect relevant information only in areas where features (or objects) of interest may be present, and thereby greatly reduce the amount of raw data that needs to be sensed.

Specifically, we first described the conceptual design of the coded aperture hyperspectral imager developed by Plain Sight Systems [9]. It is noteworthy that the raw data sensed by this instrument is not a hyperspectral image, but a mix of coded spatial and spectral information which must be digitally processed to recover the hyperspectral data cube. We presented the algebraic framework for reconstructing the hyperspectral data cube using the Hadamard transform matrix, and described a method for varying resolution in the reconstructed scene.

The coded aperture imager's ability to collect less data than a conventional sensor was shown by means of illustrative examples. The essence of the experiments shows that raw data can be collected sparsely across the scene, driven by performance metrics such as pattern match criteria and therefore only a fraction of the underlying pixel need to be sensed. Fundamentally, it becomes possible to retain the salient information in the scene while avoiding the need to measure irrelevant information. This has the potential to significantly reduce the requirements for data links and on-board storage in future generation of sensors that are based on the ISP paradigm.

References

1. Harwit, M., Sloane, N.J.A.: Hadamard Transform Optics. Academic Press, New York (1979)
2. DeVerse, R.A., Hammaker, R.M., Fately, W.G.: Realization of the Hadamard multiplex advantage using a programmable optical mask in a dispersive flat-field near-infrared spectrometer. Appl. Spectrosc. **54**(12), 1751–1758 (2000)
3. Wuttig, A., Riesenberg, R.: Sensitive Hadamard transform imaging spectrometer with a simple MEMS. In: Proceedings of SPIE 4881, pp. 167–178 (2003)
4. Donoho, D.L.: Compressed sensing, Stanford University: Department of Statistics report 2004-25, October, 2004
5. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
6. Neifeld, M.A., Shankar, P.: Feature-specific imaging. Appl. Opt. **42**, 3379–3389 (2003)
7. Mahalanobis, A., Daniell, C.: Data compression and correlation filtering: a seamless approach to pattern recognition. In: Javidi, B. (ed.) Smart Imaging Systems. SPIE Press, Bellingham (2001)
8. Mahalanobis, A., Vijaya Kumar, B.V.K., Song, S., Sims, S.R.F., Epperson, J.F.: Unconstrained correlation filters. App. Opt. **33**(17), 3751–3759 (1994)
9. Fateley, W.G., Coifman, R.R., Geshwind, F., DeVerse, R.A.: System and method for encoded spatio-spectral information processing. US Patent # 6,859,275, February 2005

Color Science and Engineering for the Display of Remote Sensing Images

Maya R. Gupta and Nasiha Hrustemovic

Abstract In this chapter we discuss the color science issues that arise in the display and interpretation of artificially-colored remote-sensing images, and discuss some solutions to these challenges. The focus is on visualizing images that naturally have more than three components of information, and thus displaying them as a color image necessarily implies a reduction of information. A good understanding of display hardware and human color vision is useful in constructing and interpreting hyperspectral visualizations. After detailing key challenges, we review and propose solutions to create and refine visualizations to be more effective.

Keywords Color · Display · Basis functions · White balance · Adaptation

1 Introduction

Visualizing hyperspectral images is challenging because there is simply more information in a hyperspectral image than we can visually process at once. The best solution depends on the application. For example, if identifying and differentiating certain plants is of interest, then classifying pixels into those plants and using distinct false colors to represent them may be the most useful approach. Such labeling may be layered on top of a full visualization of the image data.

M. R. Gupta (✉) and N. Hrustemovic
Department of Electrical Engineering, University of Washington, Seattle, USA
e-mail: gupta@ee.washington.edu

N. Hrustemovic
e-mail: nasihah@u.washington.edu

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_5,

Full visualizations can be useful in orienting the viewer, analyzing image content, and understanding contextually results from classification or un-mixing algorithms.

In this chapter we discuss from a color science perspective some of the problems and issues that arise in creating and interpreting hyperspectral visualizations, and describe solutions that may be useful for a variety of visualization approaches. We hope this chapter will be useful both for those who must make sense of hyperspectral imagery, and for those who design image processing tools. First, in Sect. 2 we discuss key challenges: information loss, visual interpretation, color saturation, pre-attentive imagery, metrics, and then use principal components analysis as a case study. Then in Sect. 3 we discuss some solutions that can help address these challenges, including white balance, optimized basis functions, and adapting basis functions. In Sect. 4 we conclude and consider some of the open questions in this area.

2 Challenges

Consider a hyperspectral image H with d components, that is, each pixel is a d-dimensional vector. To visualize H, each pixel can be mapped to a three component vector that can be displayed as an RGB value on a monitor. Thus, hyperspectral visualization can be viewed as a $d \rightarrow 3$ dimensionality reduction problem, and any standard dimensionality reduction solution can be used.

However, a good visualization will present information in a way that is easy for a human to quickly and correctly interpret. In the following sections we explain some of the key issues that make achieving this ideal difficult. To dig further into the imaging and color science issues presented in this section, we refer the reader to the following references: for those new to color science, Stone's book [1] provides a friendly introduction; formulas and further details for most of the topics discussed here can be found on *Wikipedia*; most of these formulas and more details about practical digital processing for imaging can be found in the recent book by Trussell and Vrhel [2]. For those whose research involves color engineering, we recommend the compilation of surveys published as *Digital Color Imaging* [3]. Additional references for specific issues are provided where they arise in the text.

2.1 Information Loss

All (static) hyperspectral visualizations are lossy. First, one maps from d dimensions down to three displayed dimensions. Second, most displays are only 8 bits, and this often entails quantization error. Consider for example principal components analysis, which maps the pixel $H[i][j]$ to a displayable three dimensional RGB pixel as follows:

$$\begin{aligned} R[i][j] &= p_1^T H[i][j] \\ G[i][j] &= p_2^T H[i][j] \\ B[i][j] &= p_3^T H[i][j], \end{aligned} \tag{1}$$

where p_1, p_2, and p_3 are the first three principal components of the image H. Even if the original image H is 8-bit, the projections in (1) will generally be a much higher bit depth, but will have to be quantized to be displayed.

A third cause of information loss is clipping pixel values to the display range, which makes it impossible to differentiate the clipped pixels. It is common to scale hyperspectral visualizations in order to increase contrast, but this causes pixels to exceed the display value range (e.g. [0…1]), and they must be clipped. It is tempting to increase contrast with a nonlinear scaling, such as a sigmoid function, but for finite-depth displays (e.g. 8 bit displays), this just moves the information loss to middle-range pixels because the nonlinearly-scaled values must be quantized to display values.

2.2 Metrics

The right metric for a visualization depends on the application. Once a metric has been decided upon, a visualization strategy should be formed that directly optimizes for the given metric. A general metric is how well differences in the spectra correlate to perceived differences in an image. For example, Jacobson and Gupta looked at the correlation between the chrominance (a*b* distance in CIELAB space) of different pixels and the angle between the corresponding original spectra [4]. Cui et al. define a similar correlation-based metric that considers Euclidean distances of pixels in spectral space and CIELAB color space [5], as well as separation of features and aspects of interactive visualization as metrics. Visualization methods that employ PCA have used metrics such as component maximum energy/minimum correlation (MEMC) index and entropy to evaluate properties of different methods [6].

It is important to recognize that measuring the quality of a visualization by its energy or related “information measure” such as entropy can reward noise.

Ultimately, the right metric depends on the application, and should be a subjective evaluation of usefulness. However, since one cannot construct visualizations that directly optimize for subjective evaluations, we must use our knowledge of human vision and attempt to correlate subjective evaluations with optimizable objective criteria.

2.3 Visual Interpretation

How do people interpret an image? For a simple false-color visualization with only a few false colors, distinct false colors such as red, green, blue, orange, and purple can be used. Given only a few such false colors, humans will interpret the false colors as representing unrelated categories. However, sociocultural training and the natural world colors cause people to instinctively interpret some false colors in standard ways, particularly that blue means water, brown means ground, green means dense vegetation, and yellow means lighter vegetation (e.g. grass). Visualizations that agree with these intuitive mappings will be faster and easier to understand than false color maps that require the viewer to consult a legend for each color.

If more than a few false colors are displayed, then people do not see a set of unrelated colors, but instead interpret similar colors as being similar. For example, if you show a human a remote sensing image with 50 different false colors including 15 shades of green, the natural inclination is to interpret the green areas as being related. Exactly how many false colors can be used before humans start to interpret the colors as image colors depends on contextual clues (if the spatial cues suggest it is an image, people are more likely to interpret it as an image) as well as on the person. Color naming research indicates that most people perceive roughly 11 distinct color categories, and so this suggests that up to 11 false colors might be used as unrelated colors, but we hypothesize that when arranged in an image, it may take only seven colors before people intuitively interpret the colors as image colors rather than true false colors. Thus, when using a multiplicity of false colors, mapping similar concepts to similar colors will match a human's intuitive assumptions about what colors mean. This effect is of course dependent on context, for example, in a natural image we do not assume that a red fruit and a red ball are related.

Given a multiplicity of false colors, people will want to judge the similarity of the underlying spectra or generating material by how similar the colors *appear* to them. It is an ongoing quest in color science to accurately quantify how similar two colors appear. Measuring the similarity between colors using RGB descriptions is dangerous. RGB is a color description that is useful for monitors, which almost all operate by adding amounts of red (R), green (G), and blue (B) light to create an image. So an RGB color can be directly interpreted by a monitor as how much of each light the monitor will put out. However monitors differ, and thus a specific RGB value will look different on two different monitors. For this reason, RGB is called a *device-dependent* color description, and how much a difference in RGB matters depends on the display hardware.

But the problem with using RGB to measure color differences actually has more to do with the human observer. If a color distance metric is good, then any difference of 10 units between two colors should *appear* to be the same difference to a viewer. But that is simply not the case with RGB values, even if they are standardized (such as sRGB or Adobe RGB).

A solution to measuring color differences that is considered by color engineers to work tolerably in most practical situations is to describe the colors in the CIELAB colorspace, and then measure color difference as Euclidean distance in the CIELAB space (or a variant thereof, see for example ΔE_{94}). The CIELAB colorspace is a device-independent color description based on measuring the actual spectra of light representing the color. Nothing in color science is simple however, and there are a number of caveats about measuring color differences with CIELAB. One caveat is that the appearance of a color depends upon the surrounding colors (see Fig. 1 for an example), and CIELAB does not take surrounding colors into account. (More complicated color appearance models exist that do take surrounding colors into account [7].) A second caveat is that one must convert a monitor's RGB colors to CIELAB colors. Because RGB is a device-dependent colorspace and CIELAB is a device-independent colorspace, the correct way to map a monitor's RGB colors to CIELAB colors requires measuring the spectra of colors displayed by the monitor in order to fit a model of how RGB colors map to CIELAB colors for that monitor. However, most monitors have a built-in standardized setting corresponding to the device-independent sRGB space, and so it is practical to assume the monitor is sRGB and then use standard sRGB-to-CIELAB calculations.

Another concern in quantifying how humans perceive differences is that humans interpret hue differences differently than lightness or saturation differences. In the context of an image, if adjacent pixels have the same hue they are more likely to be considered part of the same object than if they have different hues. Thus it is important to consider how hue will be rendered when designing or interpreting a visualization.

2.4 Color Saturation and Neutrals

We are more sensitive to small variations in neutrals than in bright colors; there really are many shades of gray. It is easier to consistently judge differences between neutral colors than between saturated colors. For example, the perceptual difference between a pinkish-gray and a blueish gray may be the same as the perceptual difference between a bright red and a cherry red, but it is generally easier to remember, describe, and re-recognize the difference between the grays. Further, if a visualization has many strong colors they will cause *simultaneous*

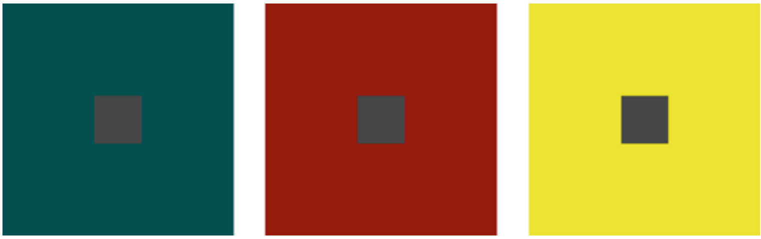


Fig. 1 Example of simultaneous contrast: the three small gray squares are exactly the same color, but the gray square on the blue background will appear pinkish compared to the center gray square, and the gray square on the yellow background will appear dark compared to the center gray square. If this figure was a visualization, a viewer would likely incorrectly judge the gray pixels to represent different spectra in the three different places

contrast, which causes the same color to look significantly different depending on the surrounding colors. An example is shown in Fig. 1. Simultaneous contrast makes it difficult to recognize the same color in different parts of an image, and to accurately judge differences between pixels.

Small regions of bright saturated colors are termed pre-attentive imagery [8], because such regions draw the viewer's attention irrespective of the user's intended focus. Thus, it is advisable to use bright saturated colors sparingly as highlights, labels, or to layer important information onto a background visualization.

2.5 Color Blindness

Roughly 5% of the population is color-blind, with the majority unable to distinguish red from green at the same luminance. Designing and modifying images to maximally inform color blind users is an area of active interest [9].

2.6 Case Study: Principal Components Analysis for Visualization

Principal components analysis (PCA) is a standard method for visualizing hyperspectral images [10]. First, the d-dimensional image is treated as a bag of d-dimensional pixels, that is, d-dimensional vectors, and orthogonal directions of maximum variance are found one after another. For visualization, usually the first three principal component directions $p_1, p_2, p_3 \in \mathcal{R}^d$ are chosen, and projecting the image onto these three principal components captures the most image variance possible with only three (linear) dimensions. This projection step maps the d-dimensional $H[i][j] \in \mathcal{R}^d$ to a new three-dimensional pixel $v[i][j] = [p_1\, p_2\, p_3]^T H[i][j]$. In order to display each projected pixel's three components as RGB values, each component is shifted up to remove negative values so that the smallest value is 0, and then scaled or clipped so that the largest value is 1. The resulting image often looks completely washed out, so the standard is to linearly stretch each of the three components so that 2% of pixels are at the minimum and 2% are at the maximum value.

The top images in Figs. 2 and 3 are example PCA visualizations. The visualizations are useful in that they highlight differences in the image. However, they are not optimized for human vision. Specifically, the colors do not have a natural or consistent meaning; rather one must learn what a color means for each image. Perceived color differences do not have a clear meaning. Saturated and bright colors are abundant in the visualizations, leaving no colors for labeling important aspects or adding a second layer of information (such as classification results).

There are many variations on PCA for visualization that can improve its performance, but which still suffer from most of the above concerns. Tyo et al. suggest

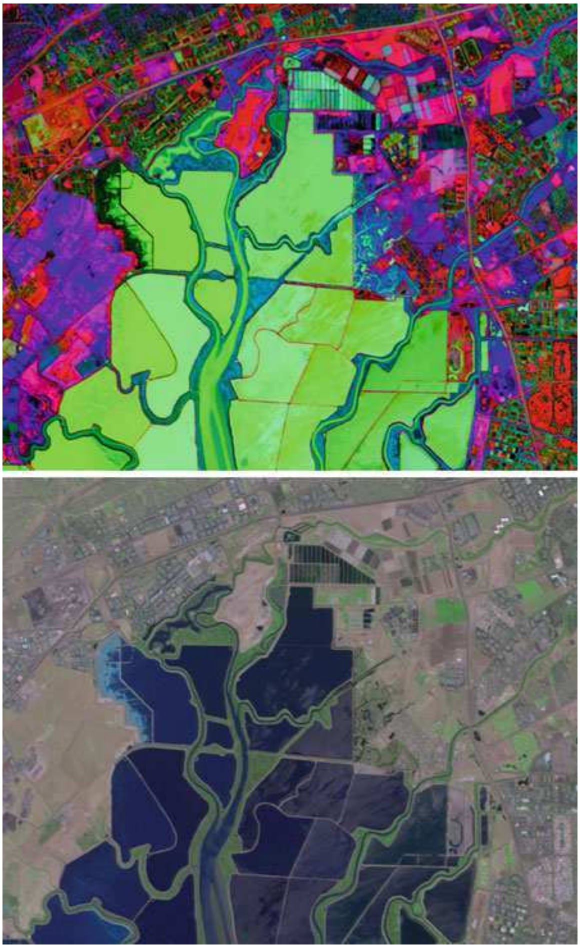


Fig. 2 Images are projections of a 224-band AVIRIS image of Moffet field [14]. *Top* Standard PCA visualization. *Bottom* Cosine-basis function visualization adapted with principal components as described in Sect. 3

treating the PCA components as YUV colors, which are more orthogonal than RGB colors, and then transforming from YUV to RGB for display [11]. ICA or other metrics can be used instead of PCA to reduce the dimensionality [12]. PCA on the image wavelet coefficients was investigated to better accentuate spatial relationships of pixels [13]. In the next section, we propose a new method to use principal components (or other such measures of relative importance of the wavelengths) to adapt basis functions optimized for human vision.

3 Some Solutions

The challenges described in the last section can be addressed in many ways; in this section we describe three approaches to these challenges that can be used separately, or combined together or with other visualization strategies. First we discuss

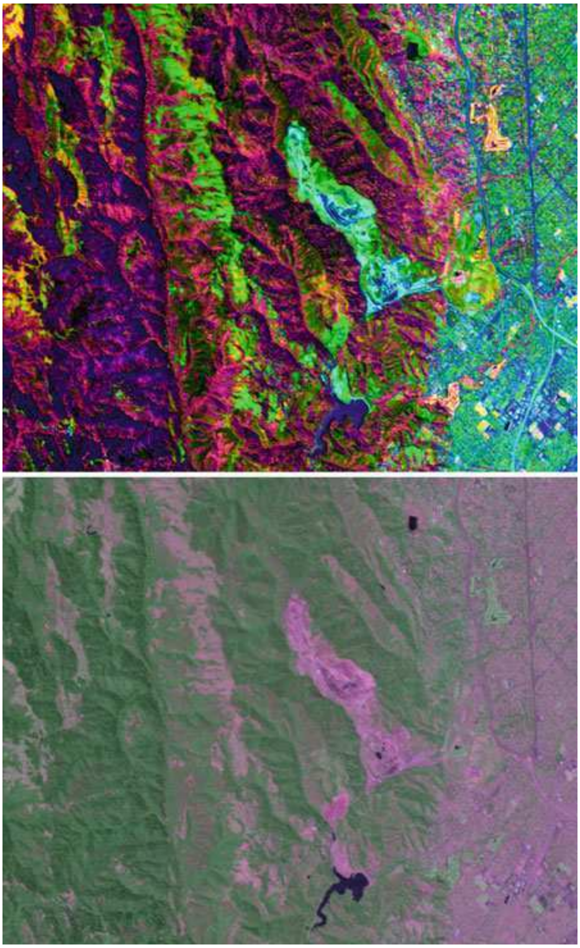


Fig. 3 Images are projections of a 224-band AVIRIS image of Jasper ridge [14]. *Top* Standard PCA visualization. *Bottom* Cosine-basis function visualization adapted with principal components as described in Sect. 3

basis functions that optimize for the peculiarities of sRGB display hardware and human vision. Then we describe two approaches to adapt any given linear projection to a function of the spectra such as variance or signal-to-noise ratio (SNR). Our last suggestion is white balance, which can be used as a post-processing step for any visualization. MATLAB code to implement these solutions is available from idl.ee.washington.edu/publications.php.

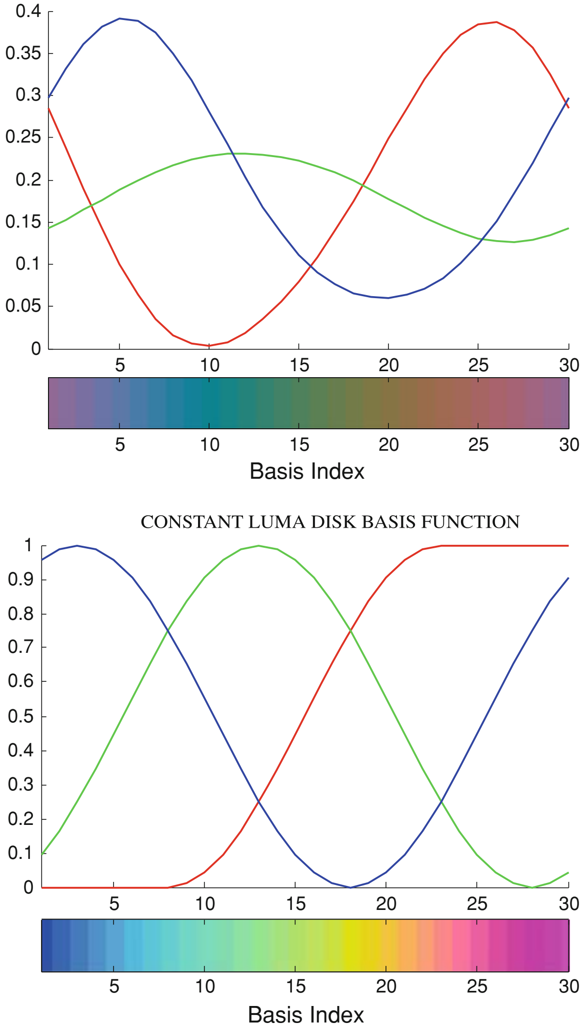
3.1 Optimized Basis Functions

Principal components form basis functions that adapt to a particular image. This has a strong advantage in highlighting the interesting image information. However, it has the disadvantages described in the previous section. Recently, Jacobson and

Gupta proposed fixed basis functions that do not adapt to image information, but do create visualizations with consistent colors and optimized human vision properties [14, 15].

Two of these basis functions are shown in Fig. 4. The top example shows the constant luma disk basis functions [15], and as shown in Fig. 5, this basis is optimal in that each component is displayed with the same brightness (luma) value, the same saturation, and the same perceptual change between components as measured in ΔE (Euclidean distance in CIELAB space). Unfortunately, in order to have those desirable properties and stay within the sRGB gamut, the constant luma disk basis function can only produce colors that are not very bright or saturated, as can be seen in the colorbar beneath the basis functions in Fig. 4 — the

Fig. 4 Two examples of basis functions designed to give equal perceptual weight to each hyperspectral component. The basis functions can be rendered for any number of components, here they are shown for $d = 30$ components



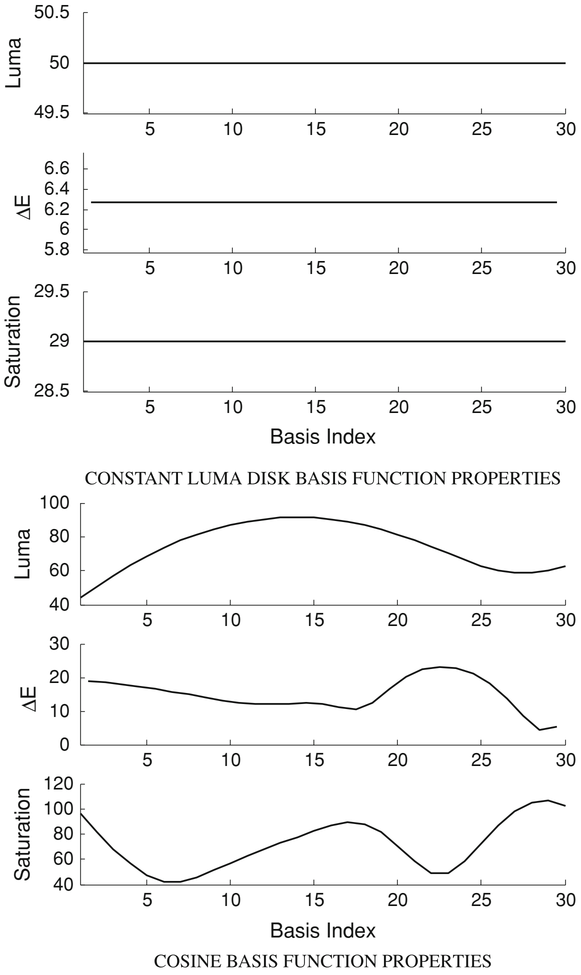
colorbar shows for each component (i.e., each wavelength) what color would be displayed if the hyperspectral image only had energy at that component.

The bottom example shows the cosine basis functions [15]. These basis functions use more of the gamut, but do not have as optimal perceptual qualities, as shown in Fig. 5.

Figure 6 (left, bottom) shows an example hyperspectral image visualized with the cosine basis function.

3.2 Adapting Basis Functions

Here we describe a new method to adapt any given set of three basis functions to take into account up to three measures of the relative importance of the wavelengths.



CONSTANT LUMA DISK BASIS FUNCTION PROPERTIES

COSINE BASIS FUNCTION PROPERTIES

Fig. 5 Color properties across wavelength are shown for two of the basis functions proposed by Jacobson et al. [15]. The basis functions can be rendered for any number of components, here they are shown for $d = 30$ components

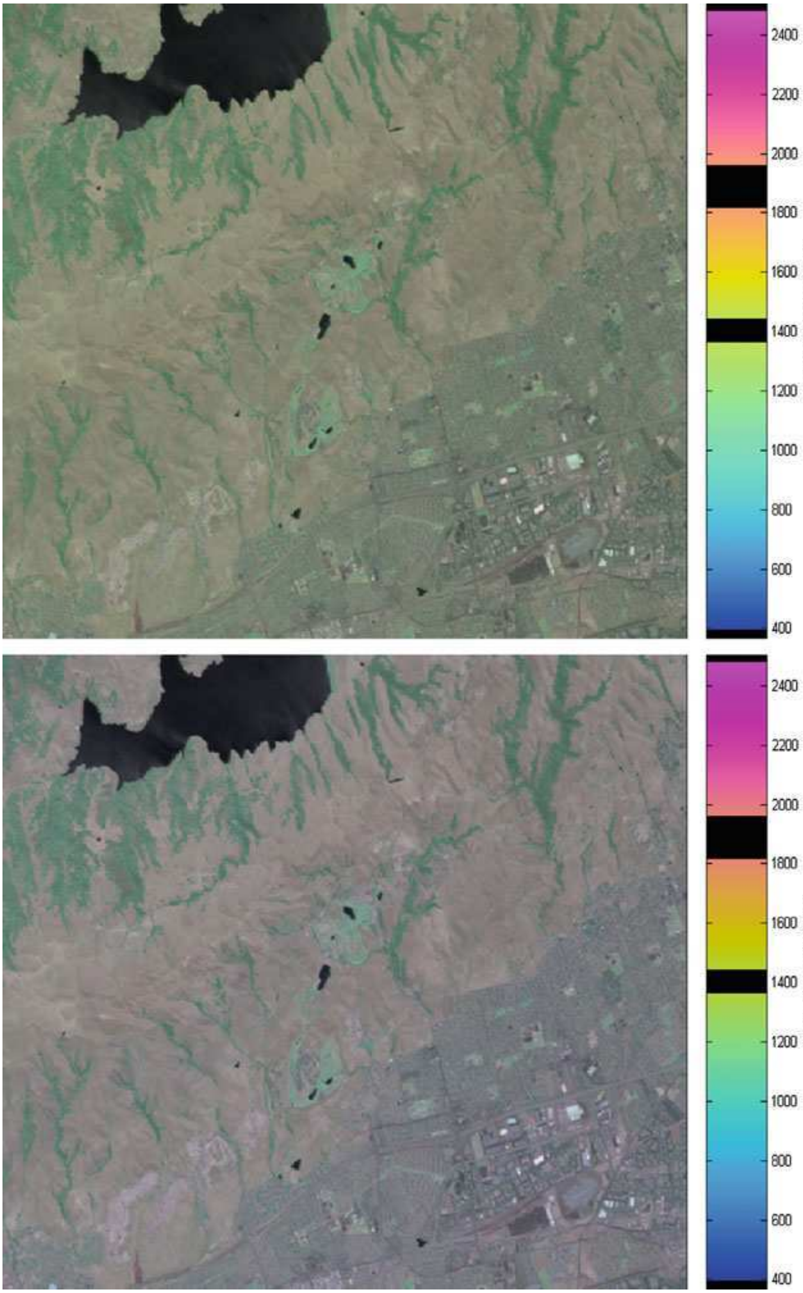


Fig. 6 *Top* The cosine basis function AVIRIS hyperspectral visualization of Jasper ridge has a slight green color-cast. *Bottom* The same visualization white-balanced

For example, the signal-to-noise ratio (SNR) is one such measure that describes how noisy each wavelength is, and the top three principal components specify three measures of relative importance of the wavelengths with respect to a specific image. The three basis functions to be adapted might be the first three principal components p_1, p_2, p_3, or the basis functions described in the last section, or the color matching basis functions to accurately reproduce what a human would see over the visible wavelengths [16], or some other set of three basis functions.

Denote the three basis functions to be adapted $r, g, b \in \mathcal{R}^d$, and the three measures of importance $f_1, f_2, f_3 \in \mathcal{R}^d$. For the case of SNR $f_1 = f_2 = f_3 = SNR$. As another example, to adapt to the top three principal component vectors, $f_1 = p_1$, $f_2 = p_2, f_3 = p_3$. In this section we discuss two ways to adapt the basis functions: a simple weighting of the basis functions, and a linear adaptation of the basis functions, which is a generalization of the SNR-adaptation proposed in [15].

A simple solution is to use the f_1, f_2, f_3 to simply weight the basis functions. Compute the normalized functions $\tilde{f}_k = f_k/(\max_\lambda f_k(\lambda))$ for $k = 1, 2, 3$. Then form new basis functions

$$\begin{aligned} r'[\lambda] &= \tilde{f}_1[\lambda] r[\lambda] \\ g'[\lambda] &= \tilde{f}_2[\lambda] g[\lambda] \\ b'[\lambda] &= \tilde{f}_3[\lambda] b[\lambda]. \end{aligned}$$

Then form normalized basis functions:

$$\begin{aligned} \tilde{r} &= \frac{r'}{\sum_\lambda r'[\lambda]} \\ \tilde{g} &= \frac{g'}{\sum_\lambda g'[\lambda]} \\ \tilde{b} &= \frac{b'}{\sum_\lambda b'[\lambda]}. \end{aligned} \tag{2}$$

Then form the linear sRGB image planes by projecting the hyperspectral image H:

$$\begin{aligned} \text{linear } R &= \tilde{r}^T H \\ \text{linear } G &= \tilde{g}^T H \\ \text{linear } B &= \tilde{b}^T H. \end{aligned} \tag{3}$$

These linear values are then gamma-corrected to form display sRGB values, usually using the standard monitor gamma of 2.2.

Weighting will cause wavelengths with high f to become brighter relative to wavelengths with high f. We recommend a slightly different solution that adapts the basis function by transferring the visualization weight of wavelengths with low f to wavelengths with high f. This re-apportions the visualization basis function, so that wavelengths with high f use up more of the visualization basis function than wavelengths with low f. For example, if f is SNR over wavelengths, and if the cosine basis function is used, then wavelengths with higher f will be both brighter and have a greater hue difference with respect to neighboring wavelengths.

For each f_k ($k = 1, 2, 3$), construct each row of the adapting matrix A_k by starting at the leftmost column of A_k which does not yet sum to 1, and add to it until the column sums to 1 or until the row sum is equal to $f_k(\lambda)$. If the $f_k(\lambda)$ for that row is not exhausted but the column already sums to 1, then add to the next column in that row until that column sums to 1 or until the row sum is equal to $f_k(\lambda)$.

Then the adapted basis functions are:

$$r' = A_1 r \quad g' = A_2 g \quad b' = A_3 b. \tag{4}$$

Example Consider a $d = 5$ component hyperspectral image H. Here we have shown how to form the adapting matrix for the basis function r:

$$f_1 = \begin{bmatrix} .2 \\ 3 \\ 1 \\ .5 \\ .3 \end{bmatrix}, \quad \text{then} \quad A_1 = \begin{bmatrix} .2 & 0 & 0 & 0 & 0 \\ .8 & 1 & 1 & .2 & 0 \\ 0 & 0 & 0 & .8 & .2 \\ 0 & 0 & 0 & 0 & .5 \\ 0 & 0 & 0 & 0 & .3 \end{bmatrix}.$$

As described above for weighting, these basis functions are then normalized to each sum to 1, the image H is projected onto them to form linear sRGB values, which are gamma-corrected to form display sRGB values.

Examples of PCA-adapted cosine basis function images are shown as the bottom images in Figs. 2 and 3. Most of the features visible in one image can be seen in the other image, but some features are easier to see in one or the other image. For example, the texture in the lakes in the Moffet field image is more clearly rendered in the PCA-adapted image, but some of the roads are better highlighted in the PCA image.

The PCA-adapted images have a number of advantages. Even though the cosine basis functions in the PCA-adapted images have been adapted independently for the Moffet field and Jasper ridge image, the color representation is similar aiding interpretation, and the colors appear somewhat natural, for example in both PCA-adapted images the vegetation displays as green. The natural palette of the cosine basis function is preserved, and in the Jasper ridge PCA-adapted image, one single bright red pixel stands out (easier to see at 200% magnification).

3.3 White Balance

When looking at a scene, we naturally adapt our vision so that the predominant illuminant appears white, and we call this adaptation white balancing. Many digital cameras also automatically white balance photos so that if you take a picture under a very yellow light, the photo does not come out looking yellowish. We suggest applying white balance to visualizations for two reasons. First, a visualization will look more like a natural scene if it is white-balanced. Second, white-balancing tends to create more grays and neutral colors, and the resulting image may appear slightly sharper and appear to have higher contrast. To actually increase contrast in an image, one can scale the pixel RGB values, however this often leads to clipping values at the extremes, which causes information loss.

A number of methods for white-balancing have been proposed [2, 17]. Typically, you must decide which original RGB value should appear as white or neutral in the white-balanced image. One standard approach is to use the maximum component values in the image to be white. A second standard approach is to make the so-called "grayworld assumption," which implies setting the average pixel value of the image to be an average gray. Both methods are effective in practice, although each one is better suited for certain types of images: average value for images rich in color, and maximum value for images with one dominant color.

Here we illustrate one method based on the grayworld assumption, as shown in Fig. 6. Note that we believe it is more justified to do white-balancing on the linear sRGB values, as we described, but white-balancing display RGB values will also be effective. Here are the steps we take:

Step 1: Let $[\bar{r}\,\bar{g}\,\bar{b}]$ be the average linear sRGB value of the image.

Step 2: Denote the linear sRGB value of the *i*th pixel as $[r_i \quad g_i \quad b_i]$. Calculate the white-balanced linear sRGB value of the *i*th pixel to be $[\tilde{r}_i\,\tilde{g}_i\,\tilde{b}_i] = [r_i/\bar{r}\,g_i/\bar{g}\,b_i/\bar{b}]$. Note that at the end of this step, the mean value of the image is $[1\,1\,1]$.

Step 3: Compute $\bar{y} = 0.2126\bar{r} + 0.7152\bar{g} + 0.0722\bar{b}$, which is the relative luminance of the average linear sRGB value of the image $[\bar{r}\,\bar{g}\,\bar{b}]$ [2].

Step 4: Calculate the normalized white-balanced linear sRGB value of the *i*th pixel to be $[\hat{r}_i\,\hat{g}_i\,\hat{b}_i] = [\tilde{r}_i\bar{y}\,\tilde{g}_i\bar{y}\,\tilde{b}_i\bar{y}]$. Note that at the end of this step, the mean value of the image is $[\bar{y}\,\bar{y}\,\bar{y}]$, so the relative luminance of the image is preserved.

Step 5: Clip values that are outside the 0 to 1 range.

Step 6: Convert the normalized white-balanced linear sRGB value $[\hat{r}_i\,\hat{g}_i\,\hat{b}_i]$ for the *i*th pixel into display sRGB values using the standard sRGB formula.

The disadvantages of white balancing a visualization are that spectra will not be rendered exactly the same in images with different white balance, and the full gamut may not be used.

4 Conclusions and Open Questions

In this chapter we have discussed the color science issues that are relevant to false-color displays of hyperspectral imagery. We hope that the challenges discussed and some of the partial solutions proposed here will spurn further thought into how to design visualizations that take into account the nonlinearities of human vision. Although theoretically the color science issues raised here are important, there is a serious lack of experimental evidence documenting what makes a hyperspectral visualization helpful or effective in practice, and this of course will depend on the exact application. This research area needs thorough and careful subjective testing that simulates as close as possible real tasks, ideally with benchmark images and standardized viewing conditions so that experimental results can be reproduced by future researchers as they compare their new ideas to the old.

References

1. Stone, M.: A Field Guide to Digital Color. AK Peters Ltd., Massachusetts (2003)
2. Trussell, J., Vrhel, M.: Fundamentals of Digital Imaging. Cambridge University Press, London (2008)
3. Sharma, G. (ed.): Digital Color Imaging. CRC Press, USA (2003)
4. Jacobson, N.P., Gupta, M.R.: Design goals and solutions for the display of hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **43**(11), 2684–2692 (2005)
5. Cui, M., Razdan, A., Hu, J., Wonka, P.: Interactive hyperspectral image visualization using convex optimization. IEEE Trans. Geosci. Remote Sens. **47**, 1673–1684 (2009)
6. Tsagaris, V., Anastassopoulos, V., Lampropoulos, G.: Fusion of hyperspectral data using segmented PCT for color representation and classification. IEEE Trans. Geosci. Remote Sens. **43**, 2365–2375 (2005)
7. Fairchild, M.: Color Appearance Models. Addison Wesley Inc., Reading, Massachusetts (2005)
8. Healey, C., Booth, K.S., Enns, J.: Visualizing real-time multivariate data using preattentive processing. ACM Trans. Model Comput. Simul. **5**(3), 190–221 (1995)
9. Jefferson, L., Harvey, R.: Accommodating color blind computer users. In: Proceedings of the International ACM SIGACCESS Conference on Computers Accessibility vol. 8, pp. 40–47, 2006
10. Ready, P.J., Wintz, P.A.: Information extraction, SNR improvement, and data compression in multispectral imagery. IEEE Trans. Commun. **21**(10), 1123–1131 (1973)
11. Tyo, J.S., Konsolakis, A., Diersen, D.I., Olsen, R.C.: Principal-components-based display strategy for spectral imagery. IEEE Trans. Geosci. Remote Sens. **41**(3), (2003)
12. Du, H., Qi, H., Wang, X., Ramanath, R., Snyder, W.E.: Band selection using independent component analysis for hyperspectral image processing. In. Proceedings of the 32nd Applied Imagery Pattern Recognition Workshop, pp. 93–98. Washington, DC, USA (2003)
13. Gupta, M.R., Jacobson, N.P.: Wavelet principal components analysis and its application to hyperspectral images. IEEE Int. Conf. Image Proc. (2006)
14. R.O.G. et al.: Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). Remote Sens. Environ. **65**, 227–248 (1998)
15. Jacobson, N.P., Gupta, M.R., Cole, J.B.: Linear fusion of image sets for display. IEEE Trans. Geosci. Remote Sens. **45**(10), 3277–3288 (2007)
16. Wyszecki, G., Stiles, W.S.: Color Science: Concepts and Methods, Quantitative Data and Formulae, 2nd edn. Wiley, New York (2000)
17. Lukac, R.: New framework for automatic white balancing of digital camera images. Signal Process. **88**, 582–592 (2008)

An Evaluation of Visualization Techniques for Remotely Sensed Hyperspectral Imagery

Shangshu Cai, Robert Moorhead and Qian Du

Abstract Displaying the abundant information contained in a remotely sensed hyperspectral image is a challenging problem. Currently no approach can satisfactorily render the desired information at arbitrary levels of detail. This chapter discusses user studies on several approaches for representing the information contained in hyperspectral information. In particular, we compared four visualization methods: grayscale side-by-side display (GRAY), hard visualization (HARD), soft visualization (SOFT), and double-layer visualization (DBLY). We designed four tasks to evaluate these techniques in their effectiveness at conveying global and local information in an effort to provide empirical guidance for better visual analysis methods. We found that HARD is less effective for global pattern display and conveying local detailed information. GRAY and SOFT are effective and comparable for showing global patterns, but are less effective for revealing local details. Finally, DBLY visualization is efficient in conveying local detailed information and is as effective as GRAY and SOFT for global pattern depiction.

Keywords Hyperspectral data visualization · Color display

1 Introduction

A hyperspectral imaging sensor collects data in hundreds of contiguous and narrow spectral bands. Comprehending such a large dataset is very challenging. Thus initially the data is usually reduced or transformed to bring out the salient aspects.

R. Moorhead (✉) and Q. Du
Mississippi State University, Mississippi State, MS, USA
e-mail: rjm@gri.msstate.edu

S. Cai
Center for Risk Studies and Safety, Santa Barbara, CA, USA

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_6,

Even then, without some visual representation it is almost impossible to understand the data.

There are several traditional ways to visualize these huge datasets. One method is grayscale side-by-side display, which visualizes the hyperspectral imagery by selecting particular bands and displaying them as grayscale side-by-side images, or displaying classification results as grayscale side-by-side images for observation. We refer to this algorithm as GRAY. Another approach is displaying multispectral/hyperspectral images as hard classification results. Generally, with hard classification, each pixel is assigned to a single class and the classified results are visualized as an image with several distinctive colors [1]; or all endmembers are displayed in one image by assigning to each pixel the color which represents the most abundant material resident in that pixel area [2]. We will call this visualization technique HARD. The third is transform-based approaches, which have been used extensively to visualize hyperspectral data recently. Principal component analysis (PCA) can condense the information in a hyperspectral data cube into several channels. It has been widely used in hyperspectral visualization [3, 4]. Jacobson et al. advocated displaying hyperspectral images as a weighted sum of signatures [5, 6]. A one-bit-transform based algorithm was introduced by Demir et al. to generate a color display [7]. A visualization technique based on convex optimization for preservation of spectral distances was proposed in [8]. A common property of these visualization techniques is that the visualized image is a color image with gradual hue transitions. We categorize these algorithms, in which a pixel represents several endmembers, as SOFT. Du et al. attests that color display using a classification approach generally produces better class separability than using a transformation-based approach [9]. Soft classification results, instead of hyperspectral images, were used to construct SOFT visualizations. A new approach was presented in [10] for visualizing mixed pixels by employing a double-layer approach (DBLY), in which one layer uses color mixing to preserve the global pattern display and the second layer, the pie-chart layer, displays the material composition information at the subpixel level. Although introduced in [10], the efficacy of the DBLY visualization algorithm was not rigorously evaluated. In this work, we present the results of a user study, evaluating the four algorithms, GRAY, HARD, SOFT, and DBLY in communicating important information in hyperspectral imagery, and provide guides for hyperspectral researchers to employ the suitable visualization methods.

User studies are broadly utilized to evaluate the effectiveness and weaknesses of visualization techniques [11]. Laidlaw et al. compared six techniques for visualizing 2D flow fields and measured user performance on three flow-related tasks for each of the six techniques [12]. Acevedo et al. investigated how the perceptual interactions among visual elements, such as brightness, icon size, etc., affect the efficiency of data exploration based on a set of 2D icon-based visualization methods [13]. With a user study, Healey built several basic rules for choosing color effectively to visualize multivariate data [14]. Kosara et al. conducted a user study to find the optimal viewing for layered texture surfaces [15]. Hagh-Shenas et al. compared two alternative algorithms for visualizing multiple discrete scalar datasets with color [16]. Ward and Theroux identified three phases of a user study: defining the goals, creating datasets, and performing studies [17]. We followed the three steps in our user study.

This study focuses on the ability of the four visualization algorithms—GRAY, SOFT, HARD, and DBLY—to convey information from both global and local aspects. Unlike Hagh-Shenas et al.'s study involving discrete variables, we investigated the effectiveness of visualization algorithms to represent the continuous datasets from hyperspectral imagery. Our experimental results are that HARD classification is less effective for global pattern display and conveying local detailed information; that GRAY and SOFT visualization are effective and comparable for showing global patterns, but are less effective for revealing local details; and that the DBLY visualization algorithm is efficient at conveying local detailed information and is as effective as the best traditional methods for global pattern depiction.

2 Image Construction

In a remotely sensed image, the reflectance of a pixel is considered a mixture of the reflectance of pure materials residing at that location. These materials are referred to as endmembers. The most commonly used model, the linear mixture model (LMM), assumes the mixture mechanism is linear [18].

Let **r** denote a pixel vector with dimensionality L, where L is the number of spectral bands. Assume the number of endmembers is p. Let M be the signature matrix of these materials denoted as $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_k, \ldots, \mathbf{m}_p]$, where $\boldsymbol{m_k}$ is the signature of the kth endmember. According to the LMM, a pixel vector r can be represented as

$$\mathbf{r} = \mathbf{M}\alpha + \mathbf{n} \tag{1}$$

where $\boldsymbol{\alpha} = \left(\alpha_1 \ldots \alpha_k \ldots \alpha_p\right)^T$ is a $p \times 1$ abundance vector, whose kth element α_k represents the proportion of the kth endmember $\mathbf{m}_k$ present in **r**. Here, **n** accommodates additive noise or sensor measurement error.

Since $\boldsymbol{\alpha}$ represents abundances, α_k for $1 \le k \le p$ should satisfy two constraints [19]: All abundances should be non-negative (the non-negativity constraint), and the sum of all abundances in a pixel should be one (the sum-to-one constraint). These two constraints can be stated as

$$\sum_{k=1}^{p} \alpha_k = 1 \quad \text{and} \quad 0 \le \alpha_k \le 1. \tag{2}$$

Fig. 1 is a 200 × 200 pixel section of a sample band from a well-analyzed dataset called Lunar Lake. The data were captured by the airborne visible/infrared imaging spectrometer (AVIRIS). Classifying the subscene with the FCLSLU algorithm [19] using prior information [20] produces six abundance images. Fig. 2 shows the spatial distribution of the six materials (Playa Lake, Rhyolite, Vegetation, Anomaly,

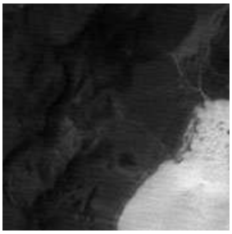


Fig. 1 Sample band of AVIRIS Lunar Lake scene cropped to a 200 × 200 pixel subscene

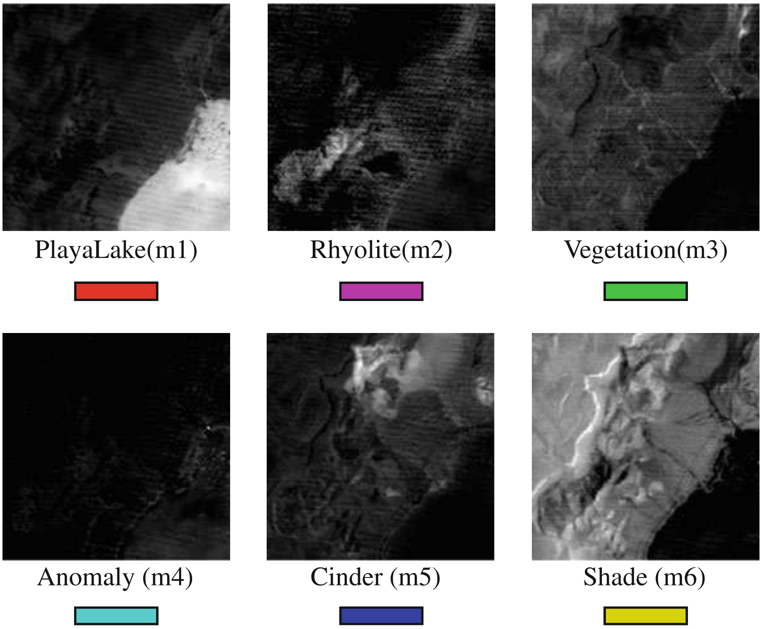


Fig. 2 The abundance images of the AVIRIS Lunar Lake scene

Cinder, Shade); highly saturated colors are used while explaining the visualization techniques. We used this dataset in part of our user study.

3 Comparative Visualization Techniques

The GRAY technique is demonstrated in Fig. 2. Normally, in a GRAY image, the material abundance from the lowest to highest is linearly mapped to the grayscale range [0, 1]. Hence, in Fig. 2, the darkest pixel represents the lowest abundance value and a white pixel represents the highest abundance value. This section briefly introduces the other three visualization techniques used in this study.

3.1 Hard Classification Visualization

In the HARD approach, a pixel is classified to contain only one material. The abundance images are converted into binary images using the following criterion:

$$\alpha_k = \begin{cases} 1, & \text{if } \alpha_k \text{ is the maximum in } \boldsymbol{\alpha} \\ 0, & \text{otherwise} \end{cases}. \tag{3}$$

The resulting HARD maps can be displayed in a single image with a color representing each material. Fig. 3a shows the resulting color image generated from the abundance images in Fig. 2. Fig. 3b shows a region of interest (ROI) that includes the anomaly marked in Fig. 3a.

3.2 Soft Classification Visualization

Unlike the HARD approach, the SOFT approach mixes the colors assigned to each abundance image to generate the final image. Assuming the color assigned to the kth endmember is $\mathbf{c}_k = (r_k, g_k, b_k)^T$, then a color matrix can be formed as

$$\mathbf{C} = \left[\mathbf{c}_1, \cdots, \mathbf{c}_k, \cdots, \mathbf{c}_p\right] = \begin{bmatrix} r_1 & \ldots & r_k & \cdots & r_p \\ g_1 & \cdots & g_k & \cdots & g_p \\ b_1 & \cdots & b_k & \cdots & b_p \end{bmatrix} \tag{4}$$

The final color $c(i, j)$ for a pixel r_{ij} with abundance vector (i, j) is

$$\mathbf{c}(i,j) = \mathbf{C}\boldsymbol{\alpha}(i,j) \tag{5}$$

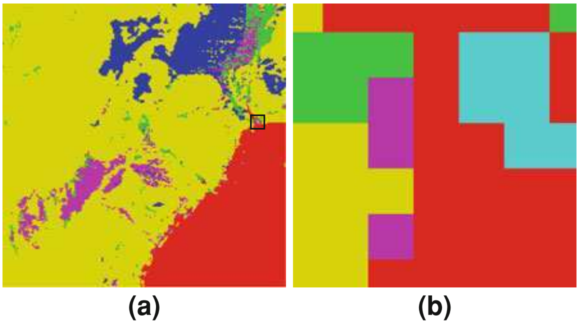


Fig. 3 HARD classification using the colors shown in Fig. 2. **a** Full image display; **b** the region of interest (ROI) highlighted in (**a**)

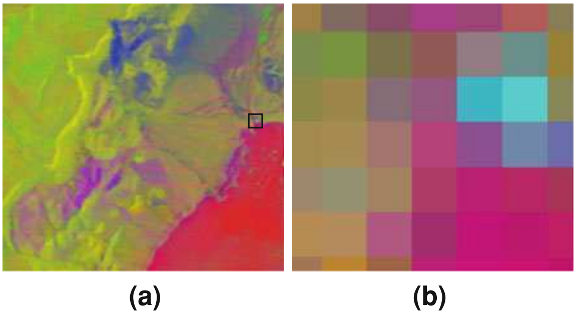


Fig. 4 SOFT classification using the colors shown in Fig. 2. **a** Full image display; **b** the ROI highlighted in (**a**)

Because the final color for each pixel is the linear combination of the colors assigned to endmembers, the final color is a function of the endmember abundances. Fig. 4 demonstrates the SOFT approach. Fig. 4a displays not only the spatial location of each endmember, but also the distribution variations.

3.3 Double Layer Visualization

The DBLY technique [10] employs the SOFT approach as a background layer and a pie-chart layer as a foreground layer. A pie-chart is formed by colored fan-shape regions which represent the different endmembers (Fig. 5). Without loss of generality, the first endmember is assigned to the first region, and so on. The area of the fan-shaped region for the *k*th endmember is proportional to the angle θ_k, which is determined by its abundance α_k, i.e.,

$$\theta_k = \alpha_k \cdot 360^{\circ} \tag{6}$$

Its starting and ending positions can be represented as

$$\beta_k^s = \sum_{j=1}^{k-1} \theta_j \quad \text{and} \quad \beta_k^e = \sum_{j=1}^{k} \theta_j \tag{7}$$

respectively. They can be related by $\theta_k = \beta_k^e - \beta_k^s$, and $\beta_1^s = 0^{\circ}$. Since the abundances of the endmembers sum to 1, a pixel is shown as a full disk, i.e., $\beta_p^e = 360^{\circ}$.

Opacity is the parameter used to control the blending of these two layers. The opacity of the pie-charts in the foreground layer is associated with a zooming parameter, which can be set automatically or manually. For the studies presented here, when the complete image is shown to display the overall distribution, the opacity of pie-charts is set to 0.2; therefore, the background layer dominates the image, as shown in Fig. 6a. If the opacity of the pie-chart layer is set to a high value when viewers zoom in for detail, then the pie-chart for each pixel pops out. Fig. 6b shows the ROI when the opacity is set to 0.8.

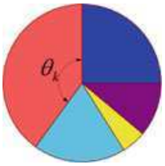


Fig. 5 A fan-shaped superpixel showing endmember percentages

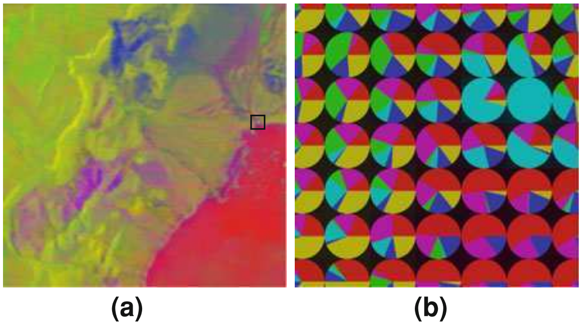


Fig. 6 Double layer visualization (DBLY). **a** Full image display; **b** the ROI highlighted in (**a**)

4 Experimental Design and Settings

Hyperspectral imagery is very useful in the discrimination of materials with similar spectral characteristics. The dominant uses of hyperspectral imagery fall into one or more of the following domain questions:

- *Perceptual Edge*: How widely distributed are the endmembers in the region? ("Where is the edge of the wheat field?")
- *Relative Position*: Where are endmembers relative to each other? ("Where is the wheat field infested with bugs?" or "Do particular materials co-exist in an area?")
- *Classification*: What and how many endmembers are present in the image scene? (A practical domain question would be "what are the different kinds of land patterns present in this image?")
- *Quantification*: How much of the endmember is in a small region or the whole area? ("How many bugs are in this area of the wheat field?" or "Where does the weed infestation exceed a certain level?")

How well a hyperspectral image is understood depends on how well these questions are answered. The goals of this study were to investigate how well these questions are answered by the four chosen methods. However, designing a user study to explicitly test these questions may not be feasible because exploring a real dataset is a complex cognitive activity. After consulting with several remote sensing experts, we were encouraged to investigate two important aspects of understanding hyperspectral images: global patterns and local information. Two tasks tested the capability of global patterns display and the other two tasks tested

Table 1 Independent and dependent variables in studies

Independent Variables			
Synthetic datasets	*Participant*	10	(random variable)
	Technique	4	GRAY (grayscale side by side) HARD (hard classification) SOFT (soft classification) DBLY (double-layer)
	Task	4	Perceptual edge detection Block value estimation Class recognition Target value estimation
Real dataset	*Participant*	15	(random variable)
	Technique	4	GRAY (grayscale side by side) HARD (hard classification) SOFT (soft classification) DBLY (double-layer)
	Task	4	Perceptual edge detection Block value estimation Class recognition Target value estimation
Dependent Variables			
Response time	Measured for each task, *seconds*		
Absolute error	$\mid user\ answer - ground\ truth \mid$		
Normalized error	$\frac{user\ answer - ground\ truth}{bar\ length} \times 100\%$ (measured for *perceptual edge detection*.)		

the ability to convey the local information. The four tasks are listed in Table 1. As described in Table 1, we designed two user studies, one based on synthetic datasets and the other based on a real dataset, namely the Lunar Lake data discussed in Sect. 2.

Our tasks had a low cognitive level and did not require a strong background in any one disciple to complete. Ten graduate students participated in the synthetic study and 15 participants in real dataset. The study was run in a conference room with a laptop computer, with its display profile set to standard RGB color space. To reduce any potential training bias, we wrote a training guide so that all participants received the same training. Testing continued for 40 min to avoid fatigue effects.

Table 1 lists the independent and dependent variables measured in the studies. In order to collect enough answers from each participant, each task was repeated several times in a test. The quantified dependent variables are response time (measured for each task), normalized error (measured for the perceptual edge detection task), and absolute error (measured for all other tasks). Standard error plots, analyses of variance (ANOVA), and post-hoc comparisons [21] were employed to analyze dependent variables. We processed outliers in the data with the procedure described by Barnett and Lewis [22]. We determined outliers on a case-by-case basis, by examining the tails of the distributions and noting values that appeared after conspicuous gaps in the histogram. Each outlier was replaced by the median of the remaining values in the experimental cell.

5 Experimental Tasks and Results

In this section, we present four tasks on the synthetic datasets and real datasets and discuss the results for each task. To limit the learning from previous responses, seven more datasets were generated by flipping and rotating the Lunar Lake dataset. A sample image is shown in Fig. 7a. Four different 20 × 20 pixel blocks (indicated by *black boxes* in Fig. 7a), were selected as ROIs. Synthetic datasets were designed for each task specifically. Fig. 7b is a sample image of synthetic data which is used in the perceptual edge detection task. The images and more details about the synthetic datasets can be found in [23].

5.1 Global Pattern Display Capability

This study investigated the perceptual edge detection and block value estimation aspects of global pattern display capability using all the image data.

5.1.1 Perceptual Edge Detection

The perceptual edge is the position where a color can no longer be perceptually distinguished from its surroundings. It is the position where the material abundance goes to zero in an image. This task was designed to test how well each visualization technique indicates the real edge, which would be important in determining where irrigation is sufficient or the extent of a covert runway.

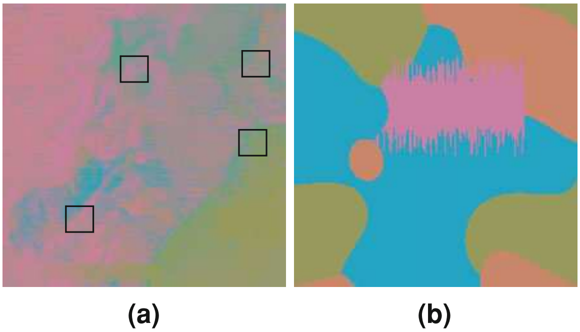


Fig. 7 Sample images used in user studies. **a** A sample image of a real dataset (SOFT visualization); **b** a sample image of a synthetic dataset (HARD visualization)

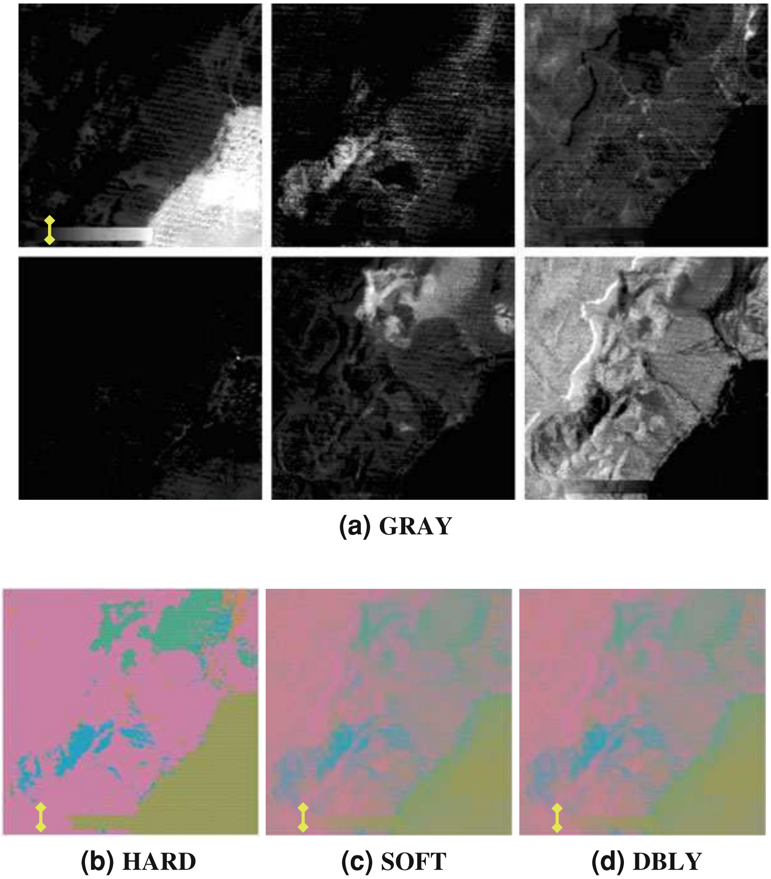


Fig. 8 An example of the perceptual edge detection task. The *yellow* lines indicate the ground-truth positions

Task

Since it is difficult to identify the precise location of the edges in a hyperspectral image, a gradient bar was embedded into the first endmember (the top left image of Fig. 8a). The gradient bar's value monotonically varied from 0.0 to 1.0. Sample images are shown in Fig. 8(b–d). For all the images, the matching pixels still satisfy the non-negative and sum-to-one constraints. For this task we asked the participants to click on the left perceptual edge of the embedded bar.

Results

We recorded the coordinate of the user's mouse click for this task. The dependent variables were *response time* and *normalized error*. We recorded a total of 351 answers for synthetic datasets and 452 answers for real datasets. We eliminated 13

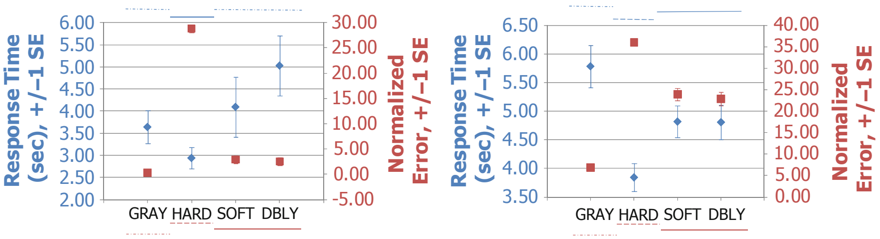


Fig. 9 Results of the perceptual edge detection task. *Left* synthetic datasets; *right* real data. For this and all figures, absent error bars indicate the standard error is smaller than the symbol size. The horizontal lines indicate the result of post-hoc comparisons (response time above the grid, error below). The response time symbol is a circle; the error is indicated by a square

and 36 *normalized error* outliers from synthetic datasets and real datasets, respectively. In addition, 27 *response time* outliers were replaced in real datasets. The means with standard error are displayed in Fig. 9. We found main effects of visualization technique on *normalized error* ($(F, p)_{\text{sd}} = (124, 0.00)$ and $(F, p)_{\text{rd}} = (166.49, 0.00)$). The subscripts “sd” and “rd” represent “synthetic datasets” and “real datasets” in this and other formulas. A post-hoc comparison indicated that GRAY provides the highest accuracy in delivering the perceptual edge information. SOFT and DBLY are in the second rank. HARD yields the biggest error, but participants took the shortest time to find the answer. A weird phenomenon we found is that the *normalized error* in SOFT and DBLY increases almost 20% from synthetic datasets to real datasets. Our explanation of this phenomenon is that the difficulty of perceptual discrimination increases as the number of endmembers increases [24]. There are four endmembers in the synthetic datasets and six in the real datasets. The *response time* does not change much from the synthetic datasets to the real datasets except for GRAY, for which the surroundings of the gradient bar have been changed. In synthetic datasets, the gradient bar is designed as an endmember; however, the gradient bar is embedded into the first endmember in real datasets.

5.1.2 Block Value Estimation

This task was designed to assess participants’ ability to read accurately the continuous values encoded by a color, a task that is known to be challenging. In each region, colors represent overlapped multiple scalars. This skill is useful in quickly accessing material quantity over an area.

Task

The task asked participants to estimate the average value of the *i*-th class within a 20 × 20 pixel block. Unlike Hagh-Shenas et al. [16], where the value in the tested region was constant, the endmember value in the block varies. A sample dataset is

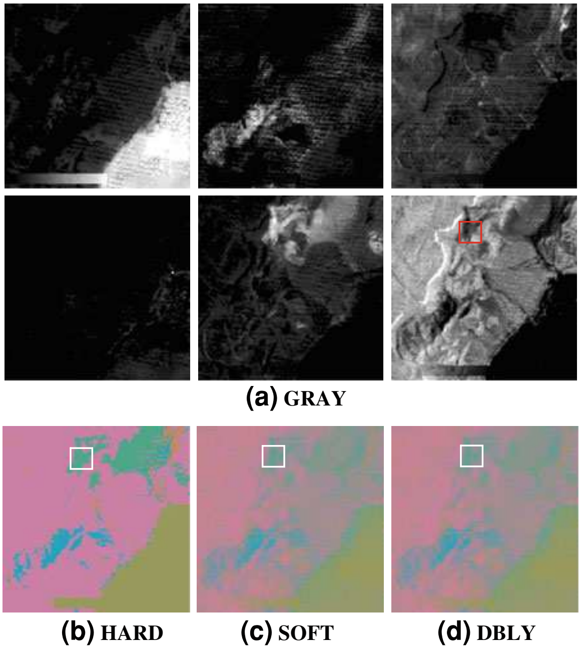


Fig. 10 A sample of the block value estimation task. The *red/white* box indicates the position of target blocks

displayed in Fig. 10. In the sample images, the participants were asked to estimate the average value of the sixth class (right-bottom in Fig. 10a).

Result

The dependent variables of *response time* and *absolute error* were measured. We recorded a total of 308 answers for the synthetic datasets and 630 answers for the real datasets. We eliminated 12 *absolute error* outliers in the synthetic datasets; a total of 52 *absolute error* outliers and 37 *response time* outliers in the real datasets were replaced. The means with error bars are plotted as Fig. 11 and show that DBLY has the best performance on both accuracy and response time. The F-value and *p*-value tests found main effects of visualization technique on *absolute error* ($(F, p)_{\mathrm{sd}} = (11.05, 0.00)$ and $(F, p)_{\mathrm{rd}} = (16.46, 0.00)$). The post-hoc analysis indicates that GRAY, SOFT, and DBLY fall in the group which has the best performance on *absolute error*.

5.2 Ability to Convey Local Information

We designed two tasks to evaluate the capability of visualization techniques to convey detailed information at the subpixel level. 20 × 20 pixel blocks were used to simulate the zooming-in operation. The blending parameter was set to 0.8 in DBLY to emphasize the pie-chart layer.

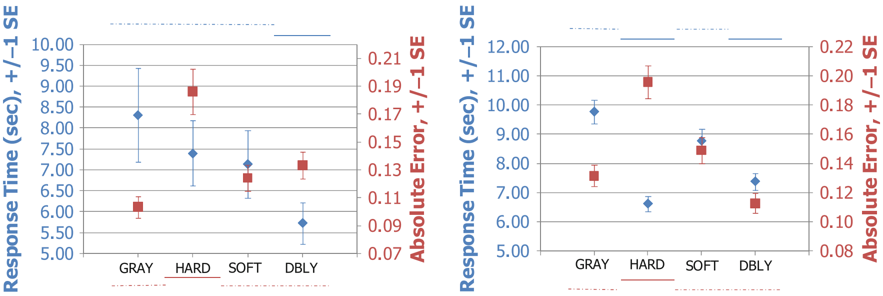


Fig. 11 The plot of means with error bars for block value estimation. *Left* synthetic datasets; *right* real data. The response time symbol is a circle; the error is indicated by a square

5.2.1 Class Recognition

The high spectral resolution of hyperspectral imagery enhances the ability to investigate detailed information in a small area, such as finding a hidden military target in the woods or the onset of a plant disease. This goal of this task was to assess participants' ability to determine the number of the endmembers present when zooming into the images.

Task

Since the size of the dataset for this task was 20 × 20, each pixel is a single color square with GRAY, HARD, and SOFT, and a single color square covered by a pie-chart in DBLY. Each pixel may contain one or more materials to simulate the real-world situation where several endmembers co-exist at the same location. In this task, we asked participants to estimate the number of classes present in the given pixel. Fig. 12 displays a sample dataset.

In the GRAY visualization (Fig. 12a), a perfectly white pixel contains 100% of that class; otherwise, other classes co-exist in that pixel. In the HARD visualization, the color of the pixel is the color of the class whose value is the maximum in the pixel. In the SOFT visualization, the color of the pixel is the mixed color of all the classes existing in the pixel. In DBLY visualization, the different colors in the fan-shape region represent the different classes and the angular extent of the wedge represents the percentage of the corresponding class in the pixel.

Results

We recorded a total of 361 answers of *response time* and *absolute error* for the synthetic datasets and 834 answers for the real datasets. The means are displayed in Fig. 13. We found main effects of visualization technique on both *absolute error* ((F, p)$_{\mathrm{sd}}$ = (43.34, 0.00) and (F, p)$_{\mathrm{rd}}$ = (91.65, 0.00)) and *response time* ((F, p)$_{\mathrm{sd}}$ = (13.10, 0.00) and (F, p)$_{\mathrm{rd}}$ = (101.93, 0.00)). The results show that DBLY

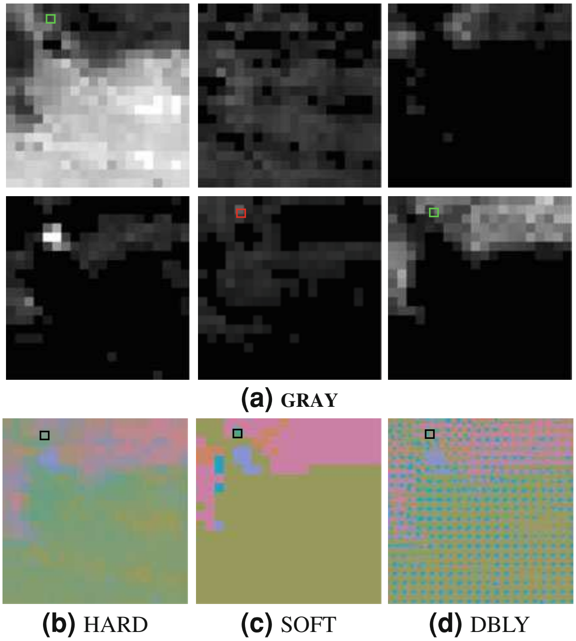


Fig. 12 A sample set of images for testing class recognition with three classes in the testing pixel. For the GRAY visualization, the testing pixel position was indicated by the *red* box in the bottom-middle gray image; the corresponding areas in other classes are marked by the *green* box. During the real test, the *green* boxes were not displayed. A *black* box marks the testing position in the other visualization techniques

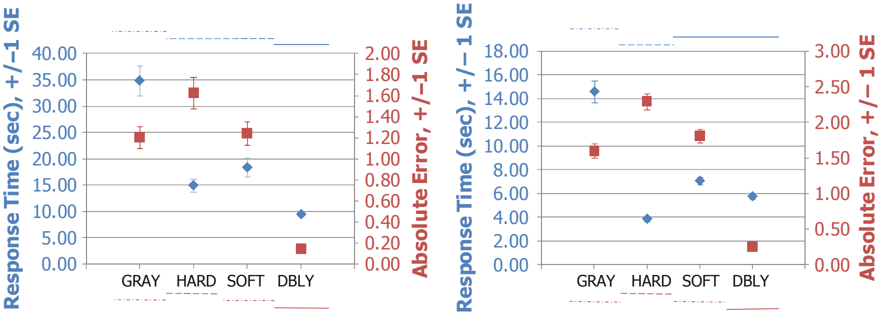
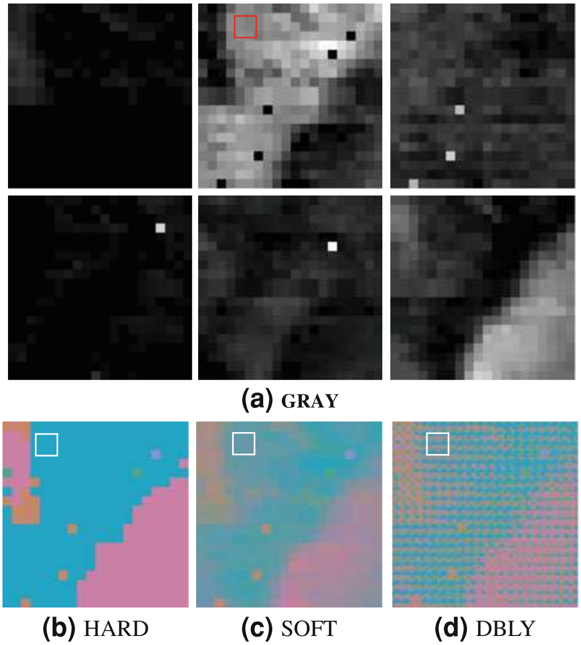


Fig. 13 The means with error bars of response time and absolute error for the class recognition task. *Left* synthetic datasets; *right* real datasets. The response time symbol is a circle; the error is indicated by a square

can achieve a very low *absolute error* (0.2) compared to the other techniques, and that participants were significantly faster with DBLY as well. With HARD the task is basically impossible. Several participants indicated that they resorted to guessing, which explains the relatively low *response time*. GRAY and SOFT provide some clues for participants to speculate on the ground truth. Even these clues do not provide the precise information; participants can determine the answer by estimating the pixel locations in other abundance images in GRAY and by considering the appearance of the mixed color in SOFT. These facts explain why GRAY and SOFT achieve better performance than HARD in *absolute error*. However,


(a) GRAY

(b) HARD (c) SOFT (d) DBLY

Fig. 14 An example of the local value estimation task. The *red/white* boxes indicate the position

participants took longer to finish the class estimation from the GRAY images. When conducting the study, we found that some participants tried to align the side-by-side displayed images by counting the pixels. This indicates that a tool that automatically aligned the pixels for side-by-side visualization would be useful.

5.2.2 Target Value Estimation

This task was designed to evaluate the ability of the four techniques to convey quantitative information.

Task

This task is very similar to the "block value estimation" task. Participants were asked to estimate the average value of a particular endmember in a 2 × 2 pixel block. Fig. 14 displays an example dataset, where the average value of Class 2 in the target block is in the range [0.4, 0.6].

Results

The measured dependent variables were *response time* and *absolute error*. We recorded a total of 317 answers for the synthetic datasets and 759 answers for the

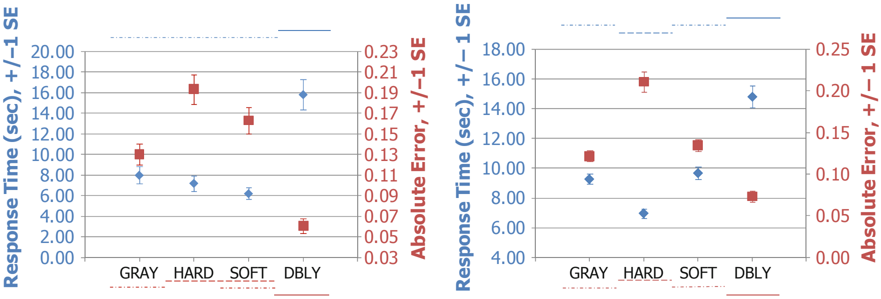


Fig. 15 The means with error bars of absolute error and response time in seconds for the target value estimation task. *Left* synthetic dataset; *right* real dataset. The response time symbol is a circle; the error is indicated by a square

real datasets. Main effects of visualization technique are found on both *absolute error* ($(F, p)_{\mathrm{sd}} = (22.54, 0.00)$ and $(F, p)_{\mathrm{rd}} = (22.54, 0.00)$) and *response time* ($(F, p)_{\mathrm{sd}} = (10.24, 0.00)$ and $(F, p)_{\mathrm{rd}} = (10.24, 0.00)$). The means with error bars of the two measured dependent variables are displayed as Fig. 15. The result shows a tradeoff between *response time* and *absolute error* (accuracy) for GRAY, SOFT, HARD, and DBLY. Participants' responses were very accurate with DBLY, but it took them longer to study the individual pie charts. GRAY and then SOFT require a mental combination of colors. The task was very difficult to perform with HARD, so participants adopted a strategy of answering quickly.

6 Discussion and Conclusions

The studies indicate that the GRAY method is effective in displaying the perceptual edge and for participants to estimate the block value, but GRAY is not sufficient to visualize the local detailed information. Moreover, the GRAY approach is space-consuming because endmembers are displayed as separated images. It is difficult to investigate the relationship among endmembers.

Since a pixel in the HARD algorithm is a pure endmember color, HARD provides a very quick impression about the information contained in a hyperspectral image. In addition, most participants thought the images of the HARD method were clearer than images of any other algorithm. That could be the reason that the HARD technique is analyzed relatively faster than other algorithms. However, the user studies illustrated that the HARD approach is less effective for perceptual edge detection, block value estimation, and local information display.

The studies attest that the SOFT approach is in the first rank for estimating block values and has good performance on perceptual edge detection. The results illustrate that the SOFT algorithm has relatively faster performance than the GRAY algorithm except on the perceptual edge detection task. However, it is less efficient in displaying local information.

The DBLY technique is verified by the studies to be the most accurate method of the four for conveying local details. Having the same advantages as the SOFT method, DBLY algorithm is effective in displaying global patterns. The user study demonstrated that adding a pie-chart layer to the SOFT approach is necessary for conveying local information while the DBLY algorithm maintains the ability to display global patterns effectively, which was exhibited by the SOFT method. However, reading the individual pie-chart results in longer response time to retrieve the detailed information.

Acknowledgments This work was supported in part by the NASA Science Mission Directorate, Earth System Division, Applied Science Program under a contract to Mississippi State University through NASA's Stennis Space Center and by the DoD High Performance Computing Visualization Initiative managed by the USACE ERDC. The authors wish to thank Dr. James Fowler for his insightful suggestions on improving some of the graphs.

References

1. Marcal, A.: Automatic color indexing of hierarchically structured classified images. In: Proceedings of IEEE Geoscience and Remote Sensing Symposium, vol. 7, pp. 4976–4979 (2005)
2. Wessels, R., Buchheit, M., Espesset, A.: The development of a high performance, high volume distributed hyperspectral processor and display system. In: Proceedings of IEEE Geoscience and Remote Sensing Symposium, vol. 4, pp. 2519–2521 (2002)
3. Tyo, J.S., Konsolakis, A., Diersen, D.I., Olsen, R.C.: Principal components-based display strategy for spectral imagery. IEEE Trans. Geosci. Remote Sens. **41**(3), 708–718 (2003)
4. Tsagaris, V., Anastassopoulos, V., Lampropoulo, G.A.: Fusion of hyperspectral data using segmented PCT for color representation and classification. IEEE Trans. Geosci. Remote Sens. **43**(10), 2365–2375 (2005)
5. Jacobson, N.P., Gupta, M.R.: Design goals and solutions for display of hyperspectral images. IEEE Trans. Geosci. Remote Sens. **43**(11), 2684–2692 (2005)
6. Jacobson, N.P., Gupta, M.R., Cole, J.B.: Linear fusion of image sets for display. IEEE Trans. Geosci. Remote Sens. **45**(10), 3277–3288 (2007)
7. Demir, B., Çelebi, A., Ertürk, S.: A low-complexity approach for the color display of hyperspectral remote-sensing images using one-bit-transform-based band selection. IEEE Trans. Geosci. Remote Sens. **47**(1), 97–105 (2009)
8. Cui, M., Razdan, A., Hu, J., Wonka, P.: Interactive hyperspectral image visualization using convex optimization. IEEE Trans. Geosci. Remote Sens. **47**(6), 1678–1684 (2009)
9. Du, Q., Raksuntorn, N., Cai, S., Moorhead, R.: Color display for hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **46**(6), 1858–1866 (2008)
10. Cai, S., Du, Q., Moorhead, R.J.: Hyperspectral imagery visualization using double layers. IEEE Trans. Geosci. Remote Sens. **45**(10), 3028–3036 (2007)
11. Kosara, R., Healey, C.G., Interrante, W., Laidlaw, D.H., Ware, C.: User studies: why, how, and when? IEEE Comput. Graphics Appl. **23**(4), 20–25 (2003)
12. Laidlaw, D.H., Kirby, R.M., Davidson, J.S., Miller, T.S., Silva, M., Warren, W.H., Tarr, M.: Quantitative comparative evaluation of 2D vector field visualization methods. In: IEEE Visualization Conference Proceedings, pp. 143–150 (2001)
13. Acevedo, D., Laidlaw, D.: Subjective quantification of perceptual interactions among some 2D scientific visualization methods. IEEE Trans. Visualization Comput. Graphics **12**(5), 1133–1140 (2006)

14. Healey, C.G.: Choosing effective colours for data visualization. In: IEEE Visualization Conference Proceedings, pp. 263–270 (1996)
15. Bair, A.S., House, D.H., Ware, C.: Texturing of layered surfaces for optimal viewing. IEEE Trans. Visualization Comput. Graphics **12**(5), 1125–1132 (2006)
16. Hagh-Shenas, H., Kim, S., Interrante, V., Healey, C.: Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. IEEE Trans. Visualization Comput. Graphics **13**(6), 1270–1277 (2007)
17. Ward, M.O., Theroux, K.J.: Perceptual benchmarking for multivariate data visualization. In: IEEE Visualization Conference Proceedings, pp. 314–321 (1997)
18. Adams, J.B., Smith, M.O., Gillespie, A.R.: Imaging spectroscopy: Interpretation based on spectral mixture analysis. In: Pieters, C.M., Englert, P. (eds.) Remote Geochemical Analysis: Elemental and Mineralogical Composition 7. Cambridge University Press, New York (1993)
19. Heinz, D.C., Chang, C.-I.: Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **39**(3), 529–545 (2001)
20. Harsanyi, J.C., Chang, C.-I.: Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. IEEE Trans. Geosci. Remote Sens. **32**(4), 779–785 (1994)
21. Howell, D.C.: Statistical Methods for Psychology, 6th edn. Wadsworth, Monterey, CA (2006)
22. Barnett, V., Lewis, T.: Outliers in Statistical Data, 3rd edn. Wiley, West Sussex, England (1994)
23. Cai, S.: Dissertation: Hyperspectral image visualization by using double and multiple layers, Mississippi State University, Dec (2008)
24. Ware, C.: Information visualization: perception for design, 2nd edn. Morgan Kaufmann, Maryland Heights, Missouri (2004)

A Divide-and-Conquer Paradigm for Hyperspectral Classification and Target Recognition

Saurabh Prasad and Lori M. Bruce

Abstract In this chapter, a multi-classifier, decision fusion framework is proposed for robust classification of high dimensional hyperspectral data in small-sample-size conditions. Such datasets present two key challenges. (1) The high dimensional feature spaces compromise the classifiers' generalization ability in that the classifier tends to over-fit decision boundaries to the training data. This phenomenon is commonly known as the Hughes phenomenon in the pattern classification community. (2) The small-sample-size of the training data results in ill-conditioned estimates of its statistics. Most classifiers rely on accurate estimation of these statistics for modeling training data and labeling test data, and hence ill-conditioned statistical estimates result in poorer classification performance. Conventional approaches, such as Stepwise Linear Discriminant Analysis (S-LDA) are sub-optimal, in that they utilize a small subset of the rich spectral information provided by hyperspectral data for classification. In contrast, the approach proposed in this chapter utilizes the entire high dimensional feature space for classification by identifying a suitable partition of this space, employing a bank-of-classifiers to perform "local" classification over this partition, and then merging these local decisions using an appropriate decision fusion mechanism. Adaptive classifier weight assignment and nonlinear pre-processing (in kernel induced spaces) are also proposed within this framework to improve its robustness over a wide range of fidelity conditions. This chapter demonstrates the efficacy of the proposed algorithms to classify remotely sensed hyperspectral data,

This chapter is based in part on Prasad and Bruce [1, 2], © 2008 and 2009 IEEE, respectively.

S. Prasad (✉) and L. M. Bruce
Geosystems Research Institute and Electrical and Computer Engineering Department,
Mississippi State University, Mississippi State, MS 39762, USA
e-mail: saurabh.prasad@ieee.org

L. M. Bruce
e-mail: bruce@bagley.msstate.edu

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_7,

since these applications naturally result in very high dimensional feature spaces and often do not have sufficiently large training datasets to support the dimensionality of the feature space. Experimental results demonstrate that the proposed framework results in significant improvements in classification accuracies over conventional approaches.

Keywords Decision Fusion · Hyperspectral Imagery · Kernel Discriminant Analysis · Multi-Classifiers · Small-Sample-Size Conditions · Statistical Pattern Recognition

1 Introduction

In the context of remote sensing applications, land cover classification and automated target recognition (ATR) systems employ statistical pattern recognition paradigms for identifying and labeling objects and features of interest in images using spatial and spectral information. Hyperspectral target recognition and classification uses the rich information available in spectral signatures of target and background pixels for identifying targets in an image. Hyperspectral imagery is a three-dimensional cube where two dimensions are spatial and one dimension is spectral. Thus, each pixel is actually a vector comprised of a hyperspectral signature containing up to hundreds or thousands of spectral bands. Recording reflectance values over a wide region of the spectrum potentially increases the class separation capacity of the data as compared to gray scale imagery (where most of the class specific information is extracted from spatial relations between pixels) or multispectral imagery (where reflectance values at a few spectral bands are recorded). Availability of this rich spectral information has made it possible to design classification systems that can perform ground cover classification and target recognition very accurately. However, this advantage of hyperspectral data is typically accompanied by the burden of requiring large amounts of training data to be available a priori for accurate representation of class conditional distributions, in order to facilitate accurate estimation of class conditional statistics of hyperspectral data and to avoid ill-conditioned formulations. This however is not guaranteed in a general remote sensing setup. In fact, in many hyperspectral applications (for example, the detection of isolated targets), the amount of ground truth pixels available to the analyst may be less than the dimensionality of the data. Another ramification of having a high dimensional feature space is over-fitting of decision boundaries by classifiers [3, 4], and consequently, poor generalization capacity. In other words, in such high dimensional spaces, it is possible that a good classifier will learn the decision boundaries based on the training data remarkably well, but may not be able to generalize well to a test set that varies slightly in its statistical structure.

As a result of the problems associated with hyperspectral data outlined above, in the absence of a large training database, it is common for researchers to either (a) limit the number of spectral bands they use for analysis, such as in best-bands selection, or, (b) perform transform based dimensionality reduction, such as with Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA), or (c) a combination of both, such as with stepwise LDA (S-LDA) prior to classification. Techniques such as PCA, LDA, best-bands selection, stepwise feature extraction (e.g., S-LDA) etc. are successful in reducing the ground truth requirement for unbiased modeling by the classifier [3, 5, 6]. However, these are not necessarily optimal from a pattern classification perspective [7–9]. For example, a PCA projection may discard useful discrimination information if it were oriented along directions of small global variance, an LDA projection will be inaccurate for multimodal class distributions, etc. Another factor that governs the efficacy of such dimensionality reduction techniques is the amount of training data required to learn the projections. For example, if the amount of training pixels is insufficient for a given feature space dimensionality, the sample scatter and covariance matrices are likely to be ill-conditioned, and transformations such as PCA and LDA may not yield optimal projections. Techniques such as best-bands selection [10] are also likely to be sub-optimal for ATR and ground cover classification tasks, considering the fact that they do not fully utilize the rich spectral information in hyperspectral (or multispectral) signatures.

The system proposed in this work employs a multi-classifier, decision fusion framework to exploit such hyperspectral data. The proposed system is capable of performing classification tasks on high dimensional data even when a relatively small amount of training data is available. Based on an intelligent partitioning scheme, the spectrum of the hyperspectral data is partitioned into smaller subspaces. After appropriate pre-processing, the data in each subspace is applied to a separate classifier (independent of other subspace classifiers). The local classifications resulting from this bank of classifiers are fused in an appropriate manner using decision fusion. This procedure partitions the single classification problem over the entire hyperspectral space into multiple classification problems, each over a subspace of a much smaller dimension. In the process, the system uses the entire available spectral information for classifying pixels, while alleviating the problems associated with high dimensional data—ill-conditioning due to small-sample-size, and, over-fitting of decision boundaries due to high dimensionality.

The outline of this chapter is as follows. Section 2 describes the functioning of the proposed Multi-Classifier and Decision Fusion framework. This includes an overview of the proposed framework, details of the procedure employed to partition the hyperspectral feature space into multiple smaller dimensional subspaces; details of two possible pre-processing steps at the subspace level; details of the classifier employed in this work and details of the decision fusion mechanism employed to fuse information from the bank of classifiers. Section 3 provides details of the experimental hyperspectral datasets (handheld and airborne) employed to demonstrate and quantify the efficacy of the proposed framework. Section 4 provides a description of the experiments employed and a discussion of the results—comparing the classification performance of the proposed framework with that of traditional single-classifier classification techniques. Section 5 concludes this chapter with a summary of benefits and limitations of the proposed system, and a discussion on potential future work in this direction.

2 The Proposed Framework

Figure 1 illustrates the proposed Multi-Classifier and Decision Fusion (MCDF) framework. The hyperspectral space is partitioned into contiguous subspaces such that the discrimination information within each subspace is maximized, and the statistical dependence between subspaces is minimized. Each subspace is then treated as a separate source in a multi-source multi-classifier setup. In doing so, we do not discard potentially useful information in the hyperspectral signatures, and also overcome the small-sample-size problem, since the number of training signatures required per subspace is substantially lower than if we directly used all the bands with a single classifier system. In fact, the minimum number of training signatures required in this scheme is governed by the size of the largest subspace formed during partitioning, which is typically much smaller than the size of the original hyperspectral space. Previous approaches to band grouping [11, 12] use a combination of correlation between variables (in this case, spectral bands) and Bhattacharya distance to partition the hyperspectral space. In this work, the

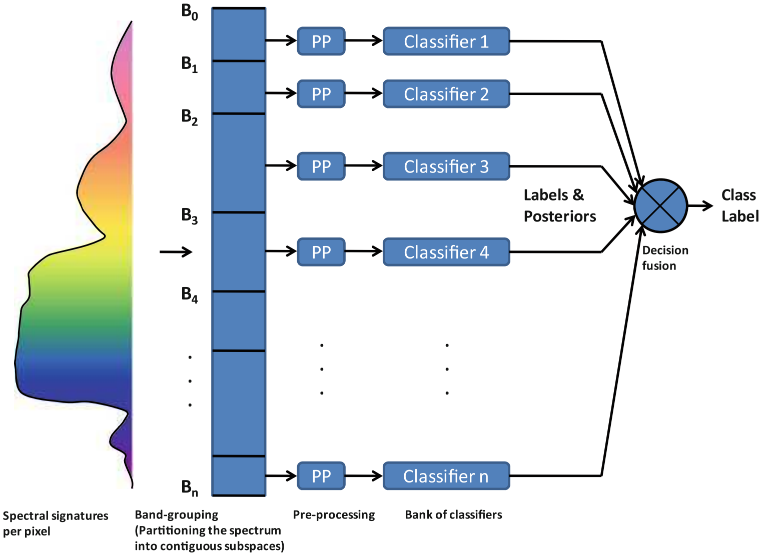


Fig. 1 The proposed divide-and-conquer paradigm for classifying high dimensional hyperspectral imagery

efficacy of higher order statistical information (using average mutual information) instead of simple correlation is studied, for a bottom-up band grouping [1, 13]. Benefits of linear (LDA) and nonlinear (KDA) pre-processing at the subspace level are also studied within the proposed framework.

2.1 Subspace Identification: Partitioning the Hyperspectral Space

Subspace identification is the first step in the proposed multi-classifier, decision fusion system. It involves intelligent partitioning of the hyperspectral feature space into contiguous subspaces such that each subspace possesses good class separation, and the statistical dependence between subspaces is minimized. A classifier is then dedicated to every subspace, and an appropriate decision fusion rule is employed to combine the local classification decisions into a final class label for every test signature. In this work, a bottom-up band grouping algorithm is proposed for subspace identification. Figure 2 depicts the application of the band grouping procedure on hyperspectral signatures. Using labeled training signatures, each subspace is grown in a bottom-up fashion (i.e., continue to add successive bands to the subspace) until the addition of bands no longer improves some performance metric. At this point, growth of the current subspace is stopped and the procedure is repeated for the next subspace. The metric employed for band grouping should be such that it simultaneously ensures good class separation within a group as well as low inter-group dependence. While good class separation per group is important for accurate decision making at the subspace level, a low inter-group dependence ensures robust decision fusion of these local decisions. A band grouping threshold (t) controls the sensitivity of partitioning to changes in the metric. This threshold is the tolerance value for the percentage change in the metric used for stopping growth of the subspace being identified. Let M_{i-1} be the

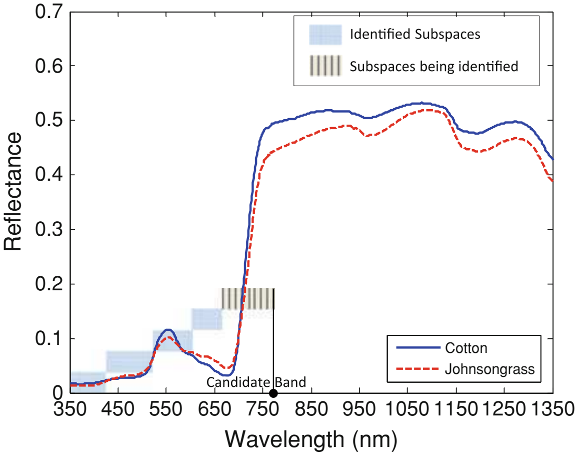


Fig. 2 Illustrating the bottom-up band grouping procedure for subspace identification. The signatures depicted are the average of all hyperspectral signatures of Cotton and Johnsongrass in the experimental hyperspectral database

performance metric of the subspace being identified without the addition of the ith band, and, let M_i be the performance metric of the subspace with the ith band included, then, the band grouping threshold, t is defined as

$$t = \frac{M_i - M_{i-1}}{M_{i-1}}. \tag{1}$$

In this work, the value of t is set to zero, that is, the growth of the subspace being identified is stopped when addition of the ith band does not change the value of the performance metric being monitored. In addition to monitoring changes in the performance metric, upper and lower bounds are imposed on the size of each subspace during the band grouping procedure. The lower bound (chosen as ten bands in this work) ensures that the number of subspaces formed does not increase unreasonably. It also ensures that subspaces are not any smaller than would be supported by the approximately block diagonal statistical structure of the correlation or mutual information matrices of hyperspectral data. The upper bound (chosen as 25 in this work) ensures that the size of each subspace is not so large that supervised dimensionality reduction and classification algorithms fail for that subspace because of ill-conditioned statistical estimates. This bound should be adjusted based on the amount of training data available for dimensionality reduction and classification.

It can be inferred from the preceding discussion that the choice of performance metric plays an important role in the performance of the proposed system. Previously [11, 12], various combinations of Bhattacharya distance and feature cross-correlation have been studied as potential performance metrics. In recent work, Tsagaris et al. [14] have suggested the use of Mutual Information for defining blocks of bands of hyperspectral data in the context of color representation. In this work, a metric using Mutual Information is proposed for band grouping.

In the subspace identification process, a good class separation in every subspace reduces the local classification errors, while statistical independence between subspaces ensures diversity in the multi-classifier setup. A multi-classifier, decision fusion system will be beneficial if there is diversity in the subspaces or in the models (e.g., classifiers). Redundancy between subspaces is not desired in a decision fusion setup since it may lead to propagation of errors (e.g., in majority vote fusion, if two different subspaces produce identical errors in classification, a single “type” of error contributes to two bad votes and so on). Instead of restricting the partitioning process to second order statistics (correlation), it is proposed that incorporating higher order statistics (as quantified by mutual information) into the metric shall generate a more meaningful partitioning of the hyperspectral space. Mutual information between two discrete valued random variables x and y is defined [15] as

$$I(x,y) = \sum_{i \in x} \sum_{j \in y} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}. \tag{2}$$

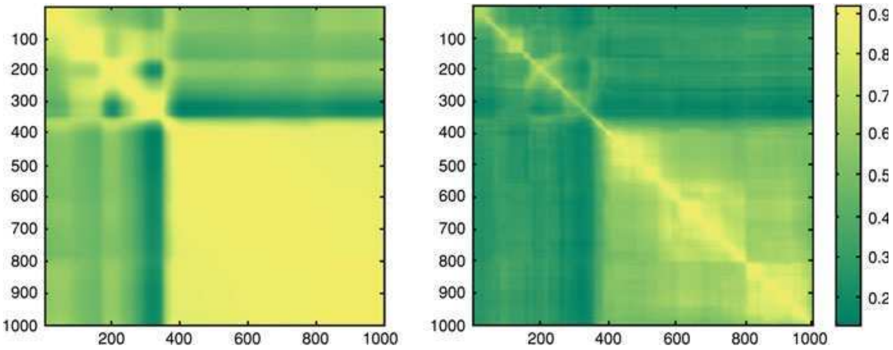


Fig. 3 Global correlation matrix (*left*) and mutual information matrix (*right*) for experimental hyperspectral data

Here, $P(i,j)$ is the joint probability distribution of x and y, and $P(i)$ and $P(j)$ are the marginal probability distributions of x and y respectively. These can be estimated using histogram approximations. In the context of hyperspectral images, x and y represent reflectance values for a pair of bands. Figure 3 shows the global correlation matrix and the global average mutual information matrix for an experimental hyperspectral dataset. Details of this dataset are provided in Sect. 3. Note that both statistical measures reveal an approximate block diagonal structure. It is this block diagonal nature of feature cross correlation (and mutual information) that allows us to partition this space into approximately independent and contiguous subspaces. Further note that the average mutual information matrix reveals a finer block diagonal structure as compared to the correlation matrix. Based on these observations, the metric employed for partitioning in this work is as follows:

$$\mathrm{JMAMI}_n = \mathrm{JM}_n \mathrm{AMI}_n, \tag{3}$$

AMI_n is the minimum average mutual information between a candidate band and the remaining bands in the current (nth) subspace, and JM_n is the between class Jeffries Matsushita (JM) distance of the current subspace, and is given by

$$\mathrm{JM} = 2(1 - e^{-BD}), \quad \text{where } \mathrm{BD} = -\ln\left(\sum_{x \in X} p(x)q(x)\right). \tag{4}$$

BD is the Bhattacharya distance; $p(x)$ and $q(x)$ are the probability distributions of the two classes between which the distance is being estimated. As will be explained later, in this chapter, both distributions are assumed to be Gaussian. JM distance is chosen to measure class separation, because unlike Bhattacharya distance it has an upper bound. This results in a normalized metric possessing lower and upper bounds. In a multi-class situation, JM_n is evaluated as the minimum pair-wise JM distance between classes in the current subspace. Previously, correlation has been employed for partitioning the space into approximately independent subspaces. The corresponding metric is similar to the one in (3) and is referred to as JMCorr, where mutual information is replaced by correlation. In recent work [1], we demonstrated that a mutual information based metric (JMAMI) for band-grouping yielded a superior partitioning of the hyperspectral space compared to the correlation based metric (JMCorr) within the MCDF framework.

2.2 Pre-processing at the Subspace Level

Since each subspace is of a much smaller dimensionality than the dimension of the original hyperspectral signature, a suitable preprocessing (such as LDA or KDA) may prove beneficial for the classification task. Note that although such preprocessing projections might have been ill-conditioned in the original high dimensional hyperspectral space, they are likely to be well-conditioned at the subspace level.

2.2.1 Linear Discriminant Analysis (LDA)

For uni-modal class conditional density functions, an LDA based dimensionality reduction is likely to preserve class separation in an even smaller dimensional projection. LDA seeks to find a linear transformation $\vec{y} = W^T\vec{x}$, where $\vec{x} \in \Re^m, \vec{y} \in \Re^n$ and $n \leq c - 1$, (c is the number of classes), such that the within-class scatter is minimized and the between-class scatter is maximized [16]. The transformation W^T is determined by maximizing Fisher's ratio,

$$J_1(W) = |W^T S_b W| / |W^T S_w W|, \tag{5}$$

which can be solved as a generalized eigenvalue problem. The solution is given by the eigenvectors of the following eigenvalue problem.

$$S_w^{-1} S_b W = \Lambda W, \tag{6}$$

where S_b is the between-class scatter matrix and S_w is the within-class scatter matrix, defined as

$$\begin{aligned} S_b &= \sum_{i=1}^{c} n_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T, \\ S_w &= \sum_{i=1}^{c} \sum_{\vec{x} \in C_i} (\vec{x} - \vec{m}_i)(\vec{x} - \vec{m}_i)^T. \end{aligned} \tag{7}$$

Recall that we impose an upper bound on the size of subspaces during the subspace identification process. One of the considerations during choosing an appropriate upper bound is for the within-class scatter matrices to be well-conditioned. Hence, LDA based dimensionality reduction at the local subspace level is going to be well-conditioned for most subspaces, as opposed to a single LDA based projection on the entire hyperspectral space, which is likely to be ill-conditioned in the absence of a lot of training data.

2.2.2 Kernel Discriminant Analysis (KDA)

Although LDA is a popular pre-processing choice in many classification tasks, in certain remote-sensing classification tasks, class-conditional distributions are multi-modal in nature (for example, due to pixel mixing—when the size of each pixel is larger than the size of the target or features on ground). In such conditions, LDA will no longer be an optimal projection. Further, if the "training" data is pure (for example, if acquired via on-ground handheld spectroradiometers), and the "testing" data comprises of mixed pixels (for example when acquired via an airborne sensor with poorer spatial resolution), decision boundaries learned from the "pure" pixels/signatures will not generalize well when classifying the test data. This mismatch can further exacerbate mixed-pixel classification. We propose that employing a nonlinear dimensionality reduction technique, such as KDA will alleviate this problem and result in a robust classification performance under severe pixel mixing conditions. In kernel methods, the key motivation behind mapping data onto a higher dimensional space is to convert nonlinear decision boundaries in the input space into linear decision boundaries in the transformed space via an appropriate nonlinear kernel function [17]. The "kernel trick" allows for computation of algorithms in a kernel mapped space without explicitly evaluating the mapping, as long as the algorithm can be expressed in terms of dot products of vectors in the input space. In its most general formulations, the kernel trick states [17] that if an algorithm can be formulated in terms of a positive definite kernel, k_1, it is possible to construct an alternate algorithm by replacing k_1 by another positive definite kernel, k_2.

In machine learning applications, the most common use of the kernel trick involves a situation where the kernel k_1 is a dot product, although, the original formulation is not limited to this case. A positive definite kernel is also endowed with a reproducing property [17]. An example usage of the kernel trick in light of this property is as follows. Assume that an algorithm in the original (input) space can be represented entirely in terms of dot products of vectors in the input space, i.e., in terms of $\langle x, x' \rangle$ where x and x' are vectors in the input space. Now consider a "kernel induced" space, created by mapping all points in the original space onto a higher (possibly infinite) dimensional space—i.e., each vector x in the original space is mapped onto $k(\cdot, x)$, a vector in the kernel induced space. The algorithm will still hold in this high dimensional kernel induced space. Further, the kernel trick and reproducing property can facilitate easy implementation of the algorithm in this space. To implement the algorithm in this kernel induced space, we need inner products of vectors in this space, $\langle k(\cdot, x), k(\cdot, x')\rangle$. Instead of performing the mapping (from the input space onto the kernel induced space) explicitly and then evaluating inner products in the kernel induced space, the reproducing property allows us to replace these inner products by the values of the kernel function evaluated using vectors in the original space, $\langle k(\cdot, x), k(\cdot, x')\rangle = k(x, x')$. For more explanation, and a more general formulation of the kernel trick and reproducing kernel Hilbert spaces, the reader is referred to [17].

Mika et al. [18] extended the conventional Fisher's LDA technique to a high dimensional, kernel induced space by employing the kernel trick. Similarly, Baudat and Anouar [19] proposed an alternative implementation to KDA, referred to as generalized discriminant analysis. In the kernel LDA setting, if Φ is a nonlinear mapping to a feature space F, the linear discriminant function that needs to be maximized is

$$J(w) = \frac{w^T S_B^{\Phi} w}{w^T S_W^{\Phi} w}, \tag{8}$$

where S_B^{Φ} and S_w^{Φ} are between-class and within-class scatter matrices [17] of the mapped training data in F, and w is a vector in F. If F is a very high dimensional space, obtaining a solution in the above formulation may become intractable. The solution proposed by Baudat and Anouar [19] is as follows:

(1) Evaluate the empirical kernel (Gram) matrix, K, as $K_{ij} = \langle \phi(x_i), \phi(x_j)\rangle = k(x_i, x_j)$, where $k(\cdot, \cdot)$ is the kernel function and $\{x_i\}$ is the set of all training data vectors.
(2) Define a block diagonal matrix, W, as $W = (W_l)_{l=1,2,\ldots,N}$, where W_l is an $(n_l \times n_l)$ matrix with all entries equal to $1/n_l$. N here is the number of classes, and n_l is the number of samples in the lth class.
(3) Perform the eigenvalue decomposition of K as $K = P\Gamma P^T$
(4) Compute the eigenvalues and eigenvectors (λ and β) of the system given by $\lambda\beta = P^T WP\beta$.
(5) Compute $\alpha = P\Gamma^{-1}\beta$.

The projection of any point (z) in the input space that maximizes (1) in the kernel space can be obtained as $w^T\phi(z) = \sum_{i=1}^{M} \alpha_i k(x_i, z)$, where $\{\alpha_i\}$ is the coefficient vector learned in the algorithm described above, M is the total number of training points $\{x_i\}$, and $k(\cdot, \cdot)$ is the kernel function. In this work, we have employed the algorithm described above to perform KDA projections on the feature space. Note that the algorithm description above is provided for completeness. The reader is referred to Baudat and Anouar [19] for a detailed proof of this algorithm (which involves reformulating the maximization problem using inner products, and then employing the kernel trick).

Such a KDA transformation provides two key advantages in pattern classification tasks: (i) the kernel mapping onto the higher dimensional space F creates a linear class separation structure, which is easier to work with and provides a better generalization ability; (ii) projection of data from the kernel space into a lower dimensional space maximizes class separation which in turn ensures good classification performance in the KDA space. In scenarios where the original (input) space contains data that is already uni-modal and linearly separable, KDA may not prove significantly beneficial over conventional LDA. However, in scenarios where the class conditional distributions in the input space are multi-modal or are not linearly separable, discriminant analysis in the kernel induced space is likely to be beneficial. With this in mind, we will study the benefits of KDA as a pre-processing transformation in the proposed MCDF framework under severe mixed pixel (targets on ground are sub-pixel) conditions.

The kernel function employed in this work is the Radial Basis Function (RBF) kernel, defined as [17]:

$$k(x_i, x_j) = \exp\left(-\left|x_i - x_j\right|^2/\sigma^2\right), \tag{9}$$

where σ is a user defined parameter of the kernel. Although the key requirement for the kernel trick to hold is for the kernel function to be positive definite, the RBF kernel has been successfully applied in machine learning applications, such as in Support Vector Machine (SVM) implementations for pattern classification tasks. In various classification applications, this kernel function has resulted in induced spaces that result in a greater degree of generalization in learning decision boundaries. Further, this kernel function results in Kernel/Gram matrices that are full ranked [17]. This is a very important advantage over other kernels, because it ensures well-conditioned formulations of kernel based algorithms.

It has been pointed out in [17] that the value of σ (width of the kernel) governs the generalization of the decision boundaries learned in the kernel induced space. The larger this value, the better that classification algorithm would generalize to arbitrary test data, and vice versa. In this chapter, classification performance will be studied over a wide range of this parameter space, in an attempt to identify appropriate parameter values for the classification task at hand.

2.3 Classifier

In this work, quadratic maximum-likelihood classifiers are employed. These classifiers assume Gaussian class distributions for the ith class, $p(x/w_i) \sim N(\mu_\mathrm{i}, \Sigma_\mathrm{i})$. Assuming equal priors, the class membership function for such a classifier is given by [3, 16]

$$M(w_i|x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i|. \tag{10}$$

Projections such as LDA and KDA tend to generate features that are approximately Gaussian distributed. In fact, in [17], the authors demonstrate that a KDA transformation followed by a maximum-likelihood classifier is as powerful as a SVM classifier. Hence, this classifier is a reasonable choice in this framework. The efficacy of decision fusion techniques (discussed in Sect. 2.4) is dependent on the accurate estimation of posterior probabilities. Although decision fusion has been used previously for remote sensing classification tasks [20–22], these methods have not been tested for alleviating the small-sample-size problem commonly encountered when classifying hyperspectral data. A typical characteristic of hyperspectral data is that adjacent bands (and hence features) are highly correlated. For normally distributed data, a high cross-feature correlation sometimes results in rank deficient covariance matrices, which makes the estimates of class membership functions or posterior probabilities unreliable. Note that this problem is not commonly encountered with multispectral data since adjacent bands of a multispectral sensor are separated by a reasonable amount in the wavelength domain. With hyperspectral data, we need to address this issue for reliable estimation of posterior probabilities or class membership functions.

It follows from the preceding discussion that for hyperspectral data, Σ_i can sometimes be rank deficient even in the presence of sufficient training data, resulting in an unstable inverse (and hence an ill-conditioned class membership function). To resolve this issue, the null space of Σ_i is discarded with the assumption that this space contains only redundant information (i.e., Σ_i is rank deficient only due to highly correlated data, not due to insufficient data). This assumption is reasonable in the proposed multi-classifier, decision fusion approach, since each classifier deals with a subspace of a much smaller dimension, and hence the small-sample-size problem is usually not encountered. Hence, to compute the inverse of Σ_i, the Singular Value Decomposition based pseudo-inverse method is used. Similarly, the determinant of Σ_i is estimated as the product of its non-zero significant singular values, in order to discard its null space. This results in stable estimates of class membership functions and posterior probabilities.

2.4 Decision Fusion

Decision fusion refers to the process of "fusing" local (at the subspace level) classification outcomes for a unified decision per pixel in the imagery. When class labels from all subspaces are employed in the fusion process (such as in a majority vote), the resulting fusion scheme is referred to as hard decision fusion. Soft decision fusion entails the use of posterior probabilities, or more generally some class membership function from every classifier for making the final decision. Unlike hard fusion techniques, soft decision fusion schemes do not rely solely on class labels from each classifier to make the final decision. A linear opinion pool [20] uses the individual posterior probabilities of each classifier ($j = 1, 2, \ldots, n$), $p_j(w_i/x)$ to estimate a global class membership function

$$C(w_i|x) = \sum_{j=1}^{n} \alpha_j p_j(w_i|x),$$
$$w = \underset{i \in \{1,2,\ldots,C\}}{\arg\max}\ C(w_i|x). \tag{11}$$

The classifier weights (α_j, $j = 1, 2, \ldots, n$) can either be uniformly distributed over all classifiers, or can be assigned based on the confidence score of each classifier. This is essentially a weighted average of posteriors across the classifier bank. In this work, we employ a linear opinion pool for decision fusion. In this chapter, a uniform weight assignment is employed, although in recent work, we have found adaptive weight assignment (where weights are estimated using an appropriate metric quantifying the strength of each local classifier) to outperform a uniform weight assignment scheme.

3 Experimental Hyperspectral Datasets

3.1 Handheld Hyperspectral Data

The handheld hyperspectral data employed for testing the proposed system was collected using an Analytical Spectral Devices (ASD) Fieldspec Pro FR handheld spectroradiometer [23]. Signatures collected from this device have 2,151 spectral bands sampled at 1 nm over the range of 350–2,500 nm with a spectral resolution ranging from 3 to 10 nm. A 25° instantaneous field of view (IFOV) foreoptic was used. The instrument was set to average ten signatures to produce each sample signature, and the sensor was held nadir at approximately four feet above the vegetation canopy. Hyperspectral signatures collected with an ASD spectroradiometer tend to have high levels of noise in the regions associated with longer wavelengths, particularly when the sensor has been in use for a longer period of time or under high temperature conditions (due to overheating of the semiconductors). Thus the signatures were truncated at 1,800 nm. Also, the reflectance values in the regions 1,350–1,430 nm were removed from all signatures to avoid noise due to atmospheric water absorption.

Signatures in the dataset (Fig. 4) form two classes: (1) an agricultural row crop, Cotton variety ST-4961, and (2) a weed that is detrimental to the crop's yield, Johnsongrass (*Sorghum halepense*). In this study, 54 signatures of Johnsongrass and 35 signatures of Cotton are used. These signatures were measured in good weather conditions in MS, USA, in 2000–2004. A target recognition scenario is created using this data treating the weed (Johnsongrass) as the target class and the crop vegetation (Cotton) as the background class, as would be the case when remote sensing is used for precision agriculture applications. Challenging target recognition tasks are created by linearly mixing target test signatures with the

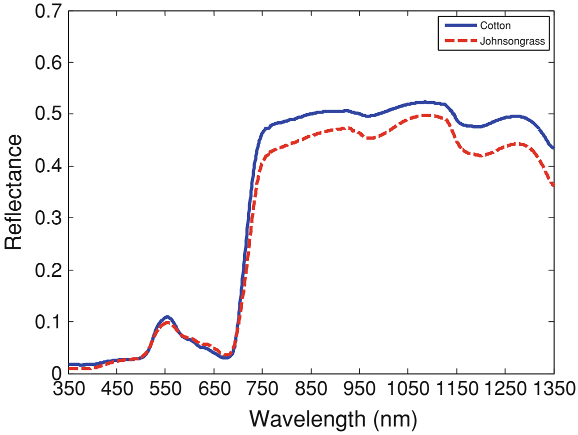


Fig. 4 Experimental hyperspectral data—hyperspectral signatures of Cotton and Johnsongrass

background at various mixing ratios (MR). All experiments reported with this dataset are performed using a leave-one-out (N-fold cross-validation) [3, 16] testing procedure. Each test target signature sequestered during the leave-one-out testing is mixed linearly with a random background signature. To ensure an unbiased setup, the background signature used in this mixing is not used for training the system. This makes it a tough and realistic ATR problem because it creates a mismatched situation where the classifiers are trained on clean target and background signatures but tested on mixed (corrupt) target signatures. The mixing ratios/MRs (background percentage to target percentage) for test target signatures reported in this work range from 10:90 (mild mixing), to 90:10 (severe mixing). With this setup, target recognition accuracies of these sub-pixel ATR tasks are estimated using the proposed MCDF system.

3.2 Airborne Hyperspectral Data

The airborne hyperspectral imagery (HSI) used in this chapter was obtained using a SpecTIR^{TM} airborne hyperspectral sensor [24]. The sensor has 128 bands, which range from 400 to 994 nm. The flight altitude was chosen to provide a 1 m spatial resolution. The image was taken on June 6, 2008. HSI acquired in this study is of a corn field in Brooksville, MS, USA. The field was sprayed with seven different concentrations of the Glufosinate herbicide diluted with water. This simulates a real-life scenario where an agricultural crop experiences stress induced by a chemical it is not resistant to [25, 26], such as when a herbicide drift event occurs between neighboring farms. The concentrations of herbicide in the solutions were 0 (no-treatment, or control), 1/32, 1/16, 1/8, 1/4, 1/2, and 1/1. Ground truth for the image was obtained using a mobile GPS unit to measure the positions of points in the field where we knew the spray concentration. Using this method, the authors

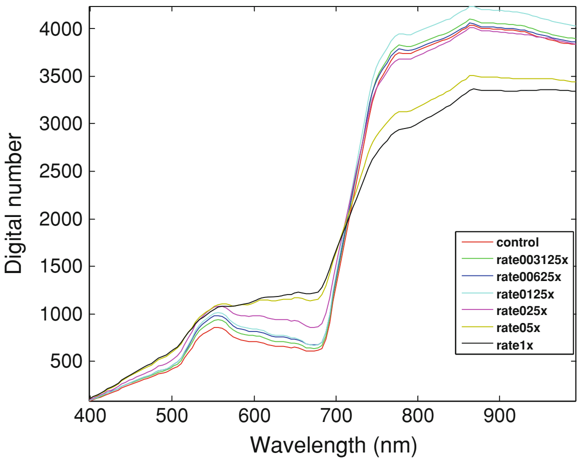


Fig. 5 Experimental hyperspectral data—hyperspectral signatures acquired from an airborne sensor (SpecTIR) for Corn crop under varying degrees of chemical stress

were able to obtain a total of 2,590 signatures for which the ground truth was known. Figure 5 illustrates how challenging this classification task is—the mean signatures of many classes are very similar. Experiments reported with this dataset are conducted using a jackknifing procedure—the available ground-truthed (labeled) data is partitioned equally into training and testing data—as before, all system parameters, supervised dimensionality reduction projections, class conditional distributions etc. are "learned" using the training data, and accuracy estimates (overall classification accuracy, target recognition accuracy, false alarm rates and other measures derived from the confusion matrix [3, 16]) are made using the testing data.

4 Experimental Setup and Results

The following sets of experiments quantify the benefits of a divide-and-conquer paradigm (MCDF) over conventional single-classifier approaches for classification of high dimensional hyperspectral data. Experiments are reported with both handheld and airborne hyperspectral data described in the previous section. In particular, the experiments are setup with the following goals: (1) To study MCDF performance using both linear and nonlinear pre-processing at the subspace level (LDA and KDA), (2) To compare performance of MCDF with conventional single-classifier approaches based on different dimensionality reduction techniques, such as PCA, S-LDA, regularized LDA, entropy based band-selection etc., (3) To study MCDF performance over a range of kernel parameter values (for a KDA based pre-processing), (4) To employ and test the efficacy of the MCDF system on a realistic operating scenario—using aerial (airborne) hyperspectral imagery to tune and train the system for land-cover classification. In all experiments, overall recognition accuracy refers to the rate of correct classification of all labeled test data relative to the number of available labeled test pixels. For the two-class target recognition task (with the handheld hyperspectral dataset), false alarm rate refers to the rate of false alarms (non-target pixels being identified as target pixels), and bars atop all bar plots indicate the 95% confidence interval in estimating these accuracies.

4.1 Experiments with Handheld HSI Data

4.1.1 Experiment 1: MCDF with LDA Based Pre-processing at the Subspace Level

In this experiment, the performance of the MCDF framework is compared with that of conventional algorithms employed by researchers for feature optimization and extraction in small-sample-size conditions. Towards this end, classification performance of the following feature extraction and classification systems is reported: (1) PCA, (2) R-LDA, (3) S-LDA, (4) BNDS, and (5) MCDF. For algorithms 1–4, a conventional single maximum-likelihood classifier is employed after each feature extraction method. These algorithms are described in Sects. 1–5 of Hyperspectral Data Compression Tradeoff. In the PCA approach, the final dimension was chosen to be equal to the number of significant eigenvalues in the spectral decomposition of the covariance matrix of the training data. In the R-LDA approach, a small constant (in this work, $1e-04$) was added to the diagonal entries of the within-class scatter matrices to avoid unstable inverses in the LDA formulation. S-LDA (also known as Discriminant Analysis Feature Extraction, or DAFE in the remote sensing community) is commonly employed by researchers in classification tasks when the training data size is small relative to the dimensionality of the data. It employs a forward selection and backward rejection algorithm to identify a smaller subset of available features (hyperspectral bands in this case) upon which a LDA transformation is applied. More details about this algorithm can be found in [27, 28]. In this work, area under Receiver Operating Characteristics (ROC) curve is employed to identify the smaller subset of hyperspectral bands upon which LDA is applied. This metric has previously shown to work well with hyperspectral data [27]. In the S-LDA algorithm, the upper limit of the intermediate feature space dimensionality in the forward selection, backward rejection procedure is set to 10, which is a reasonable value for the given amount of training data. An entropy based band-selection technique was employed in the BNDS algorithm, where, the “top” ten features were selected. For algorithm 5, the MCDF framework with *JMAMI* based band-grouping and MV based decision fusion was employed for classification, as described in Sect. 2.

Figure 6 depicts the overall recognition accuracy and false alarm rates using these algorithms, at the three mixing ratios, MR1 (30:70), MR2 (40:60) and MR3 (50:50). PCA is expected to perform poorly, and that is observed in this figure. Not only does PCA based feature extraction result in poor overall classification accuracy, the associated false-alarm rate is also very high. Regularizing the scatter

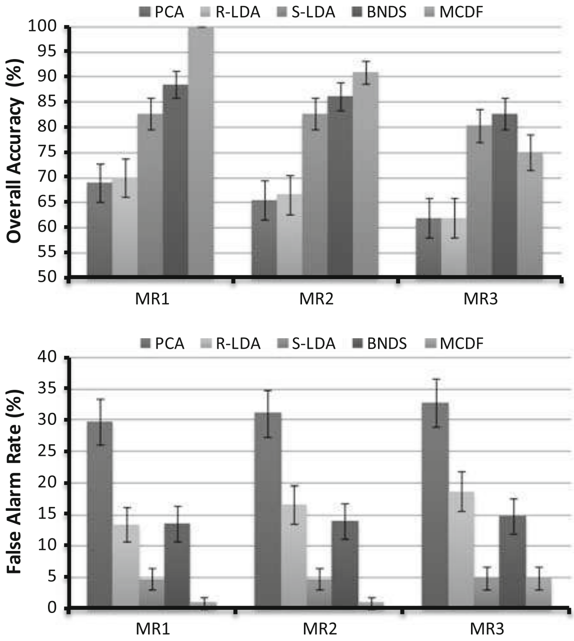


Fig. 6 Comparison of the MCDF framework with current state-of-the-art. *Error bars* atop each value indicate the 95% confidence interval for the accuracy estimates

matrices in the R-LDA approach does not yield superior classification performance either. LDA applied on a reduced subset of features based on a forward selection and backward rejection approach (S-LDA) does yield better classification performance. Entropy based band selection (BNDS) performs slightly better than S-LDA, but at the expense of a larger false-alarm rate. Finally, the proposed MCDF framework outperforms the other algorithms at most mixing ratios. It also generates the least amount of false alarms.

4.1.2 Experiment 2: MCDF with KDA Based Pre-processing at the Subspace Level

Deterioration in performance of conventional techniques and the LDA based MCDF system was observed as the pixels became more severely mixed. We propose that a nonlinear pre-processing (KDA) at the subspace level will ameliorate affects of pixel mixing and the consequent multi-modality of the class-conditional distributions. To this end, we employed a KDA projection as the pre-processing in the MCDF framework, and studied classification performance under different pixel mixing conditions. The overall implementation of this algorithm (MCDF-KDA) is similar to the description of Fig. 1, except that KDA is performed as the pre-processing.

Before comparing performance of a KDA-based MCDF system with other classification systems the generalization ability of the proposed system is studied as a function of the kernel parameter, σ. As was mentioned previously, the key

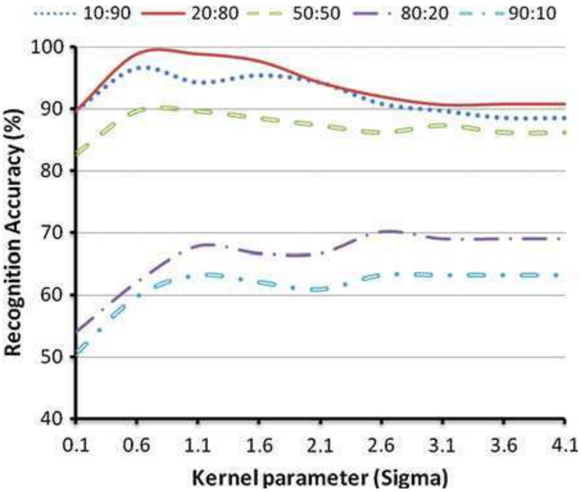


Fig. 7 Classification performance of MCDF-KDA as a function of the kernel "width", σ for the handheld hyperspectral data. Such analysis is useful in ascertaining appropriate parameters for the classification task at hand

motivation behind introducing a kernel based transformation in the MCDF framework is to improve the generalization ability of classification, that is, to ensure that the classification system is able to generalize well to arbitrary test data—even the kind that has a slightly different statistical structure as compared to the test data. This ensures a robust classification because in operational scenarios, it is rarely the case that we are able to train a classifier on data with spatial and spectral fidelity precisely similar to the actual test data.

In Fig. 7, overall classification accuracy is reported using the proposed KDA based MCDF system over a wide range of kernel parameter values, varying σ from 0.1 to 4.1. The optimal window size (maximum size of each subspace in the partition) in the partitioning process was experimentally found to be 50 for this ASD dataset [2]. Results are reported for light pixel mixing (MR 10:90 and 20:80), moderate pixel mixing (MR 50:50) and severe pixel mixing (MR 80:20 and 90:10).

As explained in [17], the value of σ, the width of the RBF kernel has an impact on the generalization ability in the kernel induced space. As σ increases, the generalization capacity of a kernel based machine typically increases. Note that for light to moderate pixel mixing conditions, the statistical structure of training and test data is very similar. This however is not the case for severe pixel mixing conditions, where not only the mismatch between training and test conditions is high, but with increased mixing, the class distributions are likely to be multi-modal in nature. This observation is reflected in the trends that can be seen in Fig. 7. For mild to moderate pixel mixing, overall accuracy increases with an increase in σ, obtaining the best classification accuracy at around $\sigma = 0.6$. However, a further increase in the parameter results in a drop in overall accuracy. For severe pixel mixing, it can again be seen that the overall accuracy increases with increasing σ. Note that under severe pixel mixing, the maximum overall accuracy is attained with a relatively wide kernel ($\sigma = 1$) as compared to the mild and moderate pixel

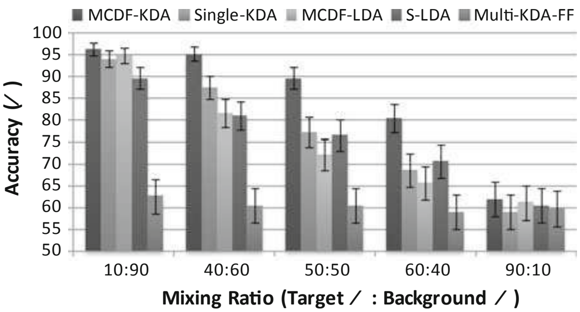


Fig. 8 Accuracy at various mixing ratios for Cotton vs. Johnsongrass. *Bars* atop each value indicate the 95% confidence interval

mixing case. This is due to the fact that under severe pixel mixing, greater generalization (obtained by a wider kernel) is needed in the classification framework to account for multi-modality of class distributions and mismatch in training and test conditions.

From this figure, it follows that without any a-prior information about the extent of pixel mixing, a value of $\sigma = 1$ appears to be a good choice for the kernel parameter, as it provides high overall accuracy over a wide range of pixel mixing conditions.

Next, the recognition performance of the proposed system (using the parameters values: window size = 50 and $\sigma = 1$) will be compared against conventional state-of-the-art approaches for hyperspectral recognition. In particular, in this experiment, overall recognition accuracy will be compared in different pixel mixing conditions using (1) MCDF-KDA (the proposed system), (2) Single-KDA (employing a single KDA transformation on the entire hyperspectral space, followed by a single maximum-likelihood classifier), (3) MCDF-LDA (The multi-classifier and decision fusion framework using LDA as the pre-processing, instead of KDA), (4) S-LDA (Stepwise LDA followed by a single maximum-likelihood classifier), (5) Multi-KDA-FF (Feature fusion of multi-KDA projections, followed by a single classifier instead of a MCDF framework). Multi-KDA-FF still employs a partitioning of the hyperspectral space, followed by a KDA transformation in each subspace of the partition. However, the outcomes of KDA transformations from each subspace are not fed into a bank-of-classifiers, and instead are fused (concatenated) into one single feature vector per hyperspectral signature. Finally, a single maximum-likelihood classifier is employed for classification. This helps in highlighting the benefits of decision fusion in the proposed MCDF-KDA system, instead of feature fusion.

Outcomes of these experiments for experimental hyperspectral datasets are depicted in Fig. 8. Note that in mild pixel mixing conditions (MR 10:90), the previously proposed MCDF-LDA system provides good classification accuracy. S-LDA and Single-KDA also perform well in these conditions. However, as pixel mixing becomes moderate (MR 40:60, 50:50) and severe (MR 60:40 and 90:10), the MCDF approach starts to break down. Performance of Single-KDA and S-LDA also starts to deteriorate. However, over this wide range of pixel mixing

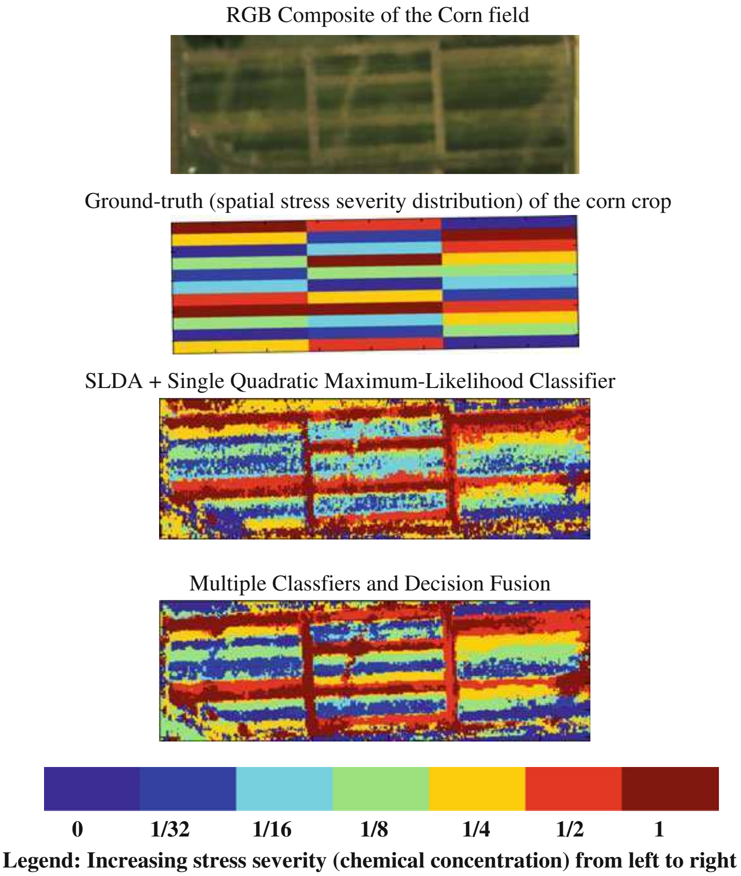

RGB Composite of the Corn field
Ground-truth (spatial stress severity distribution) of the corn crop
SLDA + Single Quadratic Maximum-Likelihood Classifier
Multiple Classfiers and Decision Fusion
0
1/32
1/16
1/8
1/4
1/2
1
Legend: Increasing stress severity (chemical concentration) from left to right

Fig. 9 Stress classification results for an aerial hyperspectral imagery of a Corn crop, with training data abundance of 3× relative to the dimensionality of the dataset—that is, the amount of training data is three times the dimensionality of the data

conditions, the proposed MCDF-KDA system outperforms other approaches (more so in moderate and severe pixel mixing conditions).

4.2 Experiments with Aerial HSI Data

Figure 9 illustrates the RGB color composite for the test field, and stress classification maps created by the S-LDA based single classifier approach, and the proposed MCDF approach. The number of labeled pixels employed for training the system was three times the dimensionality of the data. Each class, represented by a unique color in the classification maps, represents the corn crop treated by a different chemical concentration (The higher this concentration, the more "stressed" the crop). A dark-red color on this map indicates severely stressed crop, while blue color indicates healthy crop. The overall classification accuracy when using the S-LDA approach was 46.2%, and when using MCDF was 64.5%. This correlates with a reduced salt-and-pepper noise in the resulting classification map when employing MCDF. Note that this is a very challenging seven-class statistical pattern classification task. Due to a mismatch between spatial resolution (despite it being 1 m) and average size of corn canopy, pixels in this image are expected to be severely mixed across different classes.

MCDF-KDA is hence expected to outperform conventional LDA for this classification task. The KDA formulation implemented in this chapter is designed for a two-class problem. In ongoing work, the authors are working on extending it to a multi-class formulation for use in such scenarios.

5 Conclusions, Caveats and Future Work

In this chapter, we demonstrated that partitioning a high dimensional hyperspectral classification problem into multiple smaller dimensional classification tasks alleviates problems associated with over-dimensionality and limited training data. This improved performance of the multi-classifier approach was consistently observed over different sensing platforms. Although the results reported in this chapter are with handheld and airborne hyperspectral data, we obtained similar results with spaceborne HYPERION imagery [29, 30]. A data-dependent adaptation of the MCDF framework can be employed to further boost its performance. In [1], we demonstrate the benefits of an adaptive weight assignment in the decision fusion process. Such a weight assignment is expected to be beneficial when fidelity of hyperspectral signatures is non-uniform over the spectrum, or when certain classifiers in the bank of classifiers are weak. Such adaptation can be employed in a real life operating scenario by the use of development data—where the available labeled training data is partitioned further into training and testing data, and the system is optimized by maximizing classification accuracies obtained from this development dataset. In [31], we developed and demonstrated the efficacy of another possible data-dependent adaptation of the algorithms described in this chapter for improved classification performance. In this work, the feature selection process is guided by "training confusion matrices"—features that minimize confusion between the most confused classes are retained while features that most confuse such classes are pruned away.

It is important to note that although the proposed divide-and-conquer framework employs information from all features in the feature space, it is still not entirely optimal. Partitioning a high dimensional feature space into smaller dimensional subspaces can discard potentially useful cross-correlation information between features in different subspaces. Adverse affects from this issue are minimized when the feature space allows for a natural partitioning (such as with hyperspectral data). If the feature space comprises of a correlation or mutual information matrix that is full or is not block-diagonal, it would be difficult to find a good partition. Hence, the "partitioning" process when employing MCDF for a high dimensional classification task should be carefully chosen—it should minimize the loss of potentially useful cross-feature correlation information, and it should avoid a partition where the resulting classifiers in the bank result in highly correlated errors (a loss of diversity within the MCDF framework results in a reduced decision fusion performance).

In ongoing work, we are extending the two-class KDA formulation to a multi-class MCDF framework. With this, we can employ the nonlinear dimensionality reduction technique for various multi-class land-cover classification tasks involving mixed pixels and variability (mismatch) between training and test data. It is important to note that the MCDF framework can be employed on different types of feature spaces, and is not restricted to reflectance signatures. In [32, 33], this framework was successfully extended to fuse higher order spectral derivative information with reflectance information for improved classification performance. In other ongoing work, the MCDF framework is being employed on redundant wavelet transform features extracted from reflectance signatures, resulting in significant improvements in classification accuracies under very poor SNR conditions. The MCDF framework can also be extended to fuse features extracted from different remote sensing modalities. For example, vicinal pixel information, such as texture information derived from high spatial resolution imagery, can be effectively fused with sub-pixel spectral information derived from hyperspectral imagery using MCDF. In future work, we plan to test MCDF with such data fusion tasks.

Acknowledgments The authors would like to acknowledge National Aeronautical and Space Agency for partial funding of this work. The authors would also like to acknowledge Hemanth Kalluri, Terrance West, Matthew Lee, Jeff Brantley, and Jacob Bowen for help with hyperspectral data collection campaigns, and Dr. Dan Reynolds and his team for planning out and implementing the herbicide drift experiments on experimental corn fields at Mississippi State University.

References

1. Prasad, S., Bruce, L.M.: Decision fusion with confidence-based weight assignment for hyperspectral target recognition. IEEE Trans. Geosci. Remote Sens. **46**(5), 1448–1456 (2008)
2. Prasad, S., Bruce, L.M.: Information fusion in kernel-induced spaces for robust subpixel hyperspectral ATR. IEEE Geosci. Remote Sens. Lett. **6**, 572–576 (2009)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Recognition, 2nd edn. Wiley-Interscience, Hoboken (2000)
4. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 33 (2000)
5. Farrell Jr, M.D., Mersereau, R.M.: On the impact of PCA dimension reduction for hyperspectral detection of difficult targets. IEEE Geosci. Remote Sens. Lett. **2**, 192–195 (2005)

6. Swets, D.L., Weng, J.J.: Using discriminant eigenfeatures for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **18**, 831–836 (1996)
7. Prasad, S., Bruce, L.M.: Limitations of principal components analysis for hyperspectral target recognition. IEEE Geosci. Remote Sens. Lett. **5**, 625–629 (2008)
8. Prasad, S., Bruce, L.M.: Limitations of subspace LDA in hyperspectral target recognition applications. In: Proceedings of the IEEE Geoscience and Remote Sensing Symposium, pp. 4049–4052 (2007)
9. Prasad, S., Bruce, L.M.: Overcoming the small sample size problem in hyperspectral classification and detection tasks. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp. V-381–V-384 (2008)
10. Pu, R., Gong, P.: Band selection from hyperspectral data for conifer species identification. Presented at the Proceedings of the Geoinformatics'00 Conference, pp 139–146, June 2000
11. Cheriyadat, A., Bruce, L.M.: Why principal component analysis is not an appropriate feature extraction method for hyperspectral data. In: Proceedings of the IEEE Geoscience and Remote Sensing Symposium, vol. 6, pp. 3420–3422 (2003)
12. Kumar, S., Ghosh, J., Crawford, M.M.: Best-bases feature extraction algorithms for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. **39**, 1368–1379 (2001)
13. Prasad, S., Bruce, L.M.: Hyperspectral feature space partitioning via mutual information for data fusion. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp. 4846–4849 (2007)
14. Tsagaris, V., Anastassopoulos, V., Lampropoulos, G.A.: Fusion of hyperspectral data using segmented PCT for color representation and classification. IEEE Trans. Geosci. Remote Sens. **43**, 2365–2375 (2005)
15. Cover, T.: Elements of Information Theory, 2nd edn. Wiley, New York (2006)
16. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic, New York (1990)
17. Schlkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge (2001)
18. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Muller, K.-R.: Fisher discriminant analysis with kernels. In: Proceedings of IEEE Neural Networks for Signal Processing Workshop (1999)
19. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. In: Proceedings of Neural Computation (2000)
20. Benediktsson, J.A., Sveinsson, J.R.: Multisource remote sensing data classification based on consensus and pruning. IEEE Trans. Geosci. Remote Sens. **41**, 932–936 (2003)
21. Benediktsson, J.A., Swain, P.H.: Consensus theoretic classification methods. IEEE Trans. Syst. Man Cybern. **22**, 688–704 (1992)
22. Fauvel, M., Chanussot, J., Benediktsson, J.A.: Decision fusion for the classification of urban remote sensing images. IEEE Trans. Geosci. Remote Sens. **44**, 2828–2838 (2006)
23. Analytical Spectral Devices FieldspecPro FR Specifications. http://asdi.com/products specifications-FSP.asp
24. SpecTIR ProSpecTIR-VINIR Sensor Specifications. http://www.spectir.com/
25. Ellis, J.M., Griffin, J.L., Vidrine, P.R., Godley, J.L.: Corn response to simulated drift of roundup ultra and liberty and utility of drift agents. Proc. South. Weed Sci. Soc. **51**, 21 (1998)
26. Rowland, C.D.: Crop tolerance to non-target and labeled herbicide applications. M.S. Thesis, Mississippi State University, Mississippi State, MS (2000)
27. Ball, J.E.: Three stage level set segmentation of mass core, periphery, and spiculations for automated image analysis of digital mammograms. Ph.D. Dissertation, Department of Electrical Engineering, Mississippi State University, May 2007
28. Ball, J.E., West, T., Prasad, S., Bruce, L.M.: Level set hyperspectral image segmentation using spectral information divergence-based best band selection. In: Proceedings of the IEEE Geoscience and Remote Sensing Symposium, pp. 4053–4056 (2007)
29. HYPERION instrument specifications. http://eo1.gsfc.nasa.gov/Technology/Hyperion.html
30. Prasad, S.: Multi-classifiers and decision fusion for robust statistical pattern recognition with applications to hyperspectral classification. Ph.D. Dissertation, Department of Electrical and Computer Engineering, Mississippi State University (2008)

31. Kalluri, H., Prasad, S., Bruce, L.M.: Data dependant adaptation for improved classification of hyperspectral imagery. In: Proceedings of the IEEE Geoscience and Remote Sensing Symposium (IGARSS), Hawaii, USA (2010)
32. Kalluri, H., Prasad, S., Bruce, L.M.: Decision level fusion of spectral reflectance and derivative information for hyperspectral classification and target recognition. IEEE Trans. Geosci. Remote Sens. **48**(11) 4047–4058 (2010)
33. Kalluri, H., Prasad, S., Bruce, L.M.: Fusion of spectral reflectance and derivative information for robust hyperspectral land cover classification. In: Proceedings of the IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Grenoble, France (2009)

The Evolution of the Morphological Profile: from Panchromatic to Hyperspectral Images

Mauro Dalla Mura, Jon Atli Benediktsson, Jocelyn Chanussot and Lorenzo Bruzzone

Abstract Almost a decade has passed since the concept of *morphological profile* (MP) was defined for the analysis of panchromatic remote sensing images. From that time, the MP has largely proved to be a powerful tool able to model the spatial information (e.g., contextual relations) of the image by extracting structural features (e.g., size, geometry, etc.) from the objects present in the scene. The MP processes an input image with a sequence of progressively coarser filters. This leads to a stack of filtered images showing an increasing simplification of the scene. The evaluation of how the objects in the image interact with the filters gives information on the objects structural features. The great amount of contributions present in the literature that address the application of MP to many tasks (e.g., classification, object detection, segmentation, change detection, etc.) and to different types of images (e.g., panchromatic, multispectral, hyperspectral) proves how MP is still an effective and modern tool. Moreover, many variants, extensions and refinements of its definition have also appeared stating that the MP is still under continuous development. This chapter presents the MP from its early

M. Dalla Mura (✉) and L. Bruzzone
Department of Information Engineering and Computer Science, University of Trento,
Via Sommarive 14, 38123 Povo, Trento, Italy
e-mail: dallamura@disi.unitn.it

L. Bruzzone
e-mail: lorenzo.bruzzone@ing.unitn.it

M. Dalla Mura and J. A. Benediktsson
Faculty of Electrical and Computer Engineering, University of Iceland,
Hjardarhaga 2-6, 101 Reykjavik, Iceland
e-mail: benedikt@hi.is

J. Chanussot
GIPSA-Laboratory, Signal and Image Department, Grenoble Institute of Technology
(INP), France 961 rue de la Houille Blanche, 38402 Grenoble Cedex, France
e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_8,

definition to the recent advances based on morphological attribute filters. The overview of many significant contributions that have appeared in this decade allows the reader to track the evolution of the MP from the analysis of panchromatic to hyperspectral images.

Keywords Morphological profile · Extended morphological profile · Attribute profiles · Attribute filters

1 Introduction

When the geometrical resolution of remote sensing images approaches meter or even sub-meter resolution, spatial information becomes very important for the analysis of the data. It is well known that the sizes, shapes, geometries, morphologies of geospatial objects are perceptually very important features. In some cases they can provide the only discriminant feature available to distinguish the objects of interest. For example, if one aims at extracting a road network in an urban area, the spectral signature of the roads may be mixed up with the one of buildings, parking lots, etc., although their geometrical characteristics can help to discriminate them.

In the field of remote sensing, operators belonging to the mathematical morphology framework have proved to be an effective set of tools for including spatial information in the analysis [1]. Many works have been published in different application domains such as segmentation [2, 3], classification [4, 5], change detection [6], etc.

In particular, *Morphological Profiles* (MPs) are an effective tool for extracting spatial features from the image in order to describe the objects in the scene [2]. A MP performs a multiscale decomposition of an image based on a simplification of the scene through the suppression of progressively larger details. The MP is defined on the morphological operators of opening and closing by reconstruction (morphological operators particularly suitable for the analysis of high geometrical resolution images [7]) and it was first applied in 2001 on panchromatic images [2]. From its presentation, the MP was used in an increasing number of applicative domains. Remarkably, the MP definition has been generalized from the analysis of a single band image (e.g., panchromatic) to hyperspectral images made up of hundreds of spectral channels and has become one of the state of the art techniques for the analysis of such images [8].

In this chapter we give an overview of the concepts of MP and of its extension suitable for the analysis of hyperspectral images, *Extended Morphological Profile* (EMP). In addition, we present some recent advances which generalize the concepts of MP and EMP based on morphological attribute filters, which increase the capabilities of the tools in extracting structural features. Furthermore, we give an overview of the different techniques involving the MP that have appeared in the literature allowing the reader to follow the evolution of the MP over this last decade.

The chapter is organized as follows. Fundamental concepts of mathematical morphology are recalled in Sect. 2. Section 3 is devoted to the presentation of the

MP (Sect. 3.1) and its generalization based on attribute filters (Sect. 3.2) for the analysis of panchromatic images. The extension of the concepts for dealing with hyperspectral images are reported in Sect. 4. In particular, the problem of generalizing the MP from scalar to vectorial images is treated in Sect. 4.1, the definition of the EMP (Sect. 4.2) and of its extension based on attribute filters (Sect. 4.3) are given. An overview of several experiments involving MPs are presented in Sects. 3.3 and 4.4, respectively, when analyzing panchromatic and hyperspectral images.

2 Preliminaries of Mathematical Morphology

In this section, the notions about the fundamental operators in mathematical morphology necessary for the definition of the MP and its extensions are recalled.

2.1 Fundamental Properties

Let us consider a grayscale 2D image f with discrete single tone pixel values. Then, the image f can be defined as a mapping from E, the image domain (which is a subset of $\mathbb{Z}^2$) into $\mathbb{Z}$. A morphological *neighborhood transformation* transforms a pixel p of the image f according to a function ψ and a neighborhood $N(p)$ of p (set of pixels connected to p according to a connectivity rule). This can be formulated as $[\psi(f)](p) = \psi[N(p)]$ [7]. Obviously, the output of the transform depends on the function ϕ considered and on how the neighborhood N is defined. Usually the set that defines the neighborhood in such transformations is known as a *structuring element* (SE) and it is defined by a certain shape and a center. The shape is usually a discrete representation of continuous shapes (e.g., lines, rectangles, circles, etc.) on the domain lattice. The center identifies the pixel on which the SE is superposed when probing the image.

We recall below the definitions of some fundamental properties of morphological image transformations that will be useful in the following discussion.

- *Idempotence*. A transformation ψ is idempotent if the output of the transformation is independent of the number of times it is applied to the image: i.e., $\psi(\psi(f)) = \psi(f)$.
- *Increasingness*. A transformation is said to be increasing if and only if it keeps the ordering relation between images, i.e., $f \leq g \quad \Leftrightarrow \quad \psi(f) \leq \psi(g) \forall f, g$. The notation $f \leq g$ means that $f(p) \leq g(p)$ for each pixel p in the definition domain of the images.
- *Extensivity and anti-extensivity*. A transformation ψ is extensive if, for each pixel, the transformation output is greater or equal to the original image, i.e., $f \leq \psi(f)$. The correspondent property is anti-extensivity and is satisfied when $f \geq \psi(f)$ for all the pixels in the image.

- *Absorption property*. The absorption property is fullfilled when two transformations, defined by different parameters i, j, are applied to the image, and the following relation is verified: $\psi_i \psi_j = \psi_j \psi_i = \psi_{\max(i,j)}$.

Another fundamental concept is that of the so-called *connected component*. In a grayscale image a connected component (also called a "flat zone") is defined as a set of connected iso-intensity pixels. Two pixels are connected according to a connectivity rule. The connected components of a grayscale image are called flat zones. Common connectivity rules are the 4- and 8-connected, where a pixel is said to be adjacent to four or eight of its neighboring pixels, respectively. The connectivity can be extended by more general criteria defining a connectivity class [9].

2.2 Opening and Closing by Reconstruction

The two fundamental neighborhood transformations in mathematical morphology are *erosion* and *dilation*. Most morphological operations are based on a selected combination of erosion and dilation. Erosion and dilation are denoted by ε_B and δ_B, where B refers to the structuring element used in the operation. The erosion or dilation operators transform an input image by giving as output for each pixel p the *infimum* ($\wedge$) or *supremum* ($\vee$) of the intensity values of the set of pixels included by the SE when it is centered on p, respectively. It is important to note that infimum and the supremum are the minimum and maximum of an ordered set, respectively. The definition of the erosion and dilation transformation for a grayscale discrete image f is given below.

$$\varepsilon_B(f) = \bigwedge_{b \in B} f_{-b}, \tag{1}$$

$$\delta_B(f) = \bigvee_{b \in B} f_{-b}. \tag{2}$$

The sequential composition of erosion and dilation leads to the definition of the morphological *opening* and *closing* transformations. Morphological opening of an image f by a structuring element B is defined as the erosion of f by B followed by the dilation of the eroded output by $\breve{B}$, the reflected structuring element with respect to B:

$$\gamma_B(f) = \delta_{\breve{B}}[\varepsilon_B(f)]. \tag{3}$$

In contrast, a morphological closing of an image f by a structuring element B is defined as the dilation of f by B followed by the erosion of the dilated output by the reflected structuring element $\breve{B}$:

$$\phi_B(f) = \varepsilon_{\breve{B}}[\delta_B(f)]. \tag{4}$$

While the output of an erosion would have an effect on all the brighter structures independent of the size, an opening flattens bright objects that are smaller than the size of the structuring element and, because of dilation, mostly preserves the bright large areas. Similar conclusion can be drawn for darker structures when a closing is performed. The terms brighter and darker are considered with respect to the surroundings gray tones.

The morphological opening and closing operators usually lead to severe effects on the image especially when the SE is large with respect to the size of the structures in the image. Moreover, with these operators, the geometrical characteristics of the structures can be distorted or completely lost. This is obviously an undesirable effect when information on the objects of interest have to be retrieved after the filtering.

Morphological operators based on the *geodesic reconstruction* can effectively process the image by overcoming this issue. This is achieved by either completely removing or preserving the connected components in the image according to their interaction with the SE of the transformation. In greater detail, if a component in the image is larger than the SE then it will be unaffected, otherwise it will be merged to a brighter or darker adjacent region depending upon whether a closing or opening is respectively applied. An opening by reconstruction is performed in two separated phases and can be formally defined as:

$$\gamma_R^{(i)}(f) = R_f^{\delta}[\varepsilon^{(i)}(f)]. \tag{5}$$

The first transformation, $\varepsilon^i(f)$, is an erosion of the image f with an SE of size i, which defines the size of the opening. This aims at creating the so called marker image for the reconstruction operation. The second phase performs a reconstruction by dilation, $R_f^{\delta}(\cdot)$, of the marker image taking as reference mask f. This operation is an iterative procedure that applies geodesic dilation (which is defined as the infimum of the elementary dilation and the mask image) on the marker image until idempotence ($\delta_f^{(n)} = \delta_f^{(n+1)}$):

$$R_f^{\delta}(\cdot) = \delta_f^{(n)}(\cdot) = \underbrace{\delta_f^{(1)} \cdot \delta_f^{(1)} \ldots \delta_f^{(1)}(\cdot)}_{n\,\text{times}}. \tag{6}$$

The reconstruction phase permits to fully retrieve all those structures that are not completely suppressed by the erosion and it potentially needs several iterations before reaching stability.

By duality, a closing by reconstruction is defined as the reconstruction by erosion of f from the dilation of f using a structuring element of size n:

$$\phi_R^{(i)}(f) = R_f^{\varepsilon}[\delta^{(i)}(f)]. \tag{7}$$

It is important to note that the result obtained with operators by reconstruction is less dependent on the shape of the selected structuring element then in the case of morphological opening or closing. Operators by reconstruction are also less severe than the corresponding morphological ones, i.e., which can be explained by analyzing the ordering relations between the operators:

$$\gamma \leq \gamma_R \leq f \leq \phi_R \leq \phi. \tag{8}$$

2.3 Attribute Filters

Morphological attribute filters are morphological transformations that process an image according to a criterion. A generic criterion T can be defined as a mapping of the set S of values considered by T to the couple of Booleans $\{false, true\}$. The criterion is evaluated on each connected component of the image. If the criterion is verified, then the component is preserved. If it is not verified, the component is removed. The criteria are usually related to the question whether the value of an attribute α of the component C fulfills a predefined condition, e.g., $T(C) = \alpha(C) \geq \lambda$, with $\{\alpha(C), \lambda\} \in \mathbb{R}$ or $\mathbb{Z}$ for scalar attributes, where the attributes can actually be any measure computable on the image regions. This leads to great flexibility in the behavior of attribute filters, which consequently improves their capability in modeling the spatial information with respect to operators based on fixed SEs. For example, the attributes considered can be purely geometric (e.g., area, length of the perimeter, image moments, shape factors), textural (e.g., range, standard deviation, entropy), etc.

Since attribute filters can only transform an image by merging its connected components these filters belong to the family of connected filters [10]. Actually, morphological attribute filters are connected filters and the morphological operators by reconstruction are included in their definition [10].

A very important property of the criterion considered in the transformation is *increasingness*. A criterion is said to be increasing when, if it is verified for a connected component, then it will be also true for all the components nested in it. This property leads to have for example $T(C_j) = true$ when also $T(C_i) = true$ 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, non increasing attributes, such as scale invariant measures (e.g., homogeneity, shape descriptors, orientation, etc.), lead to non increasing criteria. In the following the implications of this property for attribute filters will be discussed.

Attribute openings for binary images consider an increasing criterion T. They are obtained by computing a trivial opening, Γ^T, on the output of a connected opening, Γ_F, applied to all the connected components of a binary image F. Given a pixel p in the image domain and a connected component C, the connected opening is computed as:

$$\Gamma_f(p) = \begin{cases} C & \text{if} \quad p \in C; \\ \varnothing & \text{otherwise.} \end{cases} \tag{9}$$

The trivial opening keeps the regions for which the increasing criterion T holds. This can be expressed as:

$$\Gamma_T(C) = \begin{cases} C & \text{if } T = \text{true;} \\ \varnothing & \text{otherwise.} \end{cases} \tag{10}$$

Attribute opening is then given by:

$$\Gamma^T(f) = \bigcup_{p \in F} \Gamma_T(\Gamma_F(p)). \tag{11}$$

If the criterion considered is increasing, the resulting transformation is increasing, idempotent and anti-extensive (i.e., it is an opening). In contrast, if the increasingness property is not fulfilled by the criterion, the filter remains idempotent and anti-extensive but not increasing anymore. For this reason, the transformation based on a non-increasing criterion is not an opening, but a *thinning*.

Analogous considerations can be made for the dual transformation by considering the background regions instead of the foreground ones. If the criterion is increasing, the transformation is actually a *closing* otherwise it is a *thickening*.

The extension of the operators from binary to gray-scale images is straightforward when the criterion is increasing because of the principle of threshold superposition [11]. Since a grayscale image can be expressed as the sum of all its binary thresholds, than the output image of these filterings is the sum of all the filtered input threshold images,

$$\gamma^T(f) = \sum_{k=0}^{K} \Gamma^T(F_k) \tag{12}$$

with F_a the binary threshold image f at graylevel $k \in [0,K]$ the destination domain of the grayscale values. Equation 12 can also be expressed as:

$$\gamma^T(f)(p) = \max\{k : \Gamma^T(F_k)(p) = 1\} \quad p \in E. \tag{13}$$

When the attribute criteria are not increasing, the extension to numerical functions is not straightforward anymore. For example, let us consider a numerical function f and a binary criterion T that acts on the binary sections F_k of f at successive thresholds $k_1 < k_2 < k_3$. We may have $F_{k_2} = \varnothing$, whereas $F_k \neq \varnothing$ for $k = \{k_1, k_3\}$. Thus, the results of the transformation applied to successive sections of the image do not decrease as k increases. Therefore, they cannot be considered as the stack of sections of a function. The simplest way to force the decreasingness of the sequence is to replace the image F_k by the union of all the binary thresholds from the top section, i.e., by $F'_k = \cup\{F_i(f), i \geq k\}$. This leads to the following definition of grayscale attribute thinning with a non increasing criterion $\tilde{T}$:

$$\gamma^{\tilde{T}}_{\max}(f)(p) = \max\{k : \Gamma^{\tilde{T}}(F'_k)(p) = 1\}. \tag{14}$$

This solution leads the grayscale attribute thinning (see (13)), which is referred to as *max rule* in [12]. However, other arbitrary filtering strategies can be implemented in order to achieve different output effects when extending the binary thinning and thickening to numerical images [12, 13]. For example, Urbach et al. [13] found that the so-called *subtractive rule* is particular suitable when considering shape descriptors as attributes:

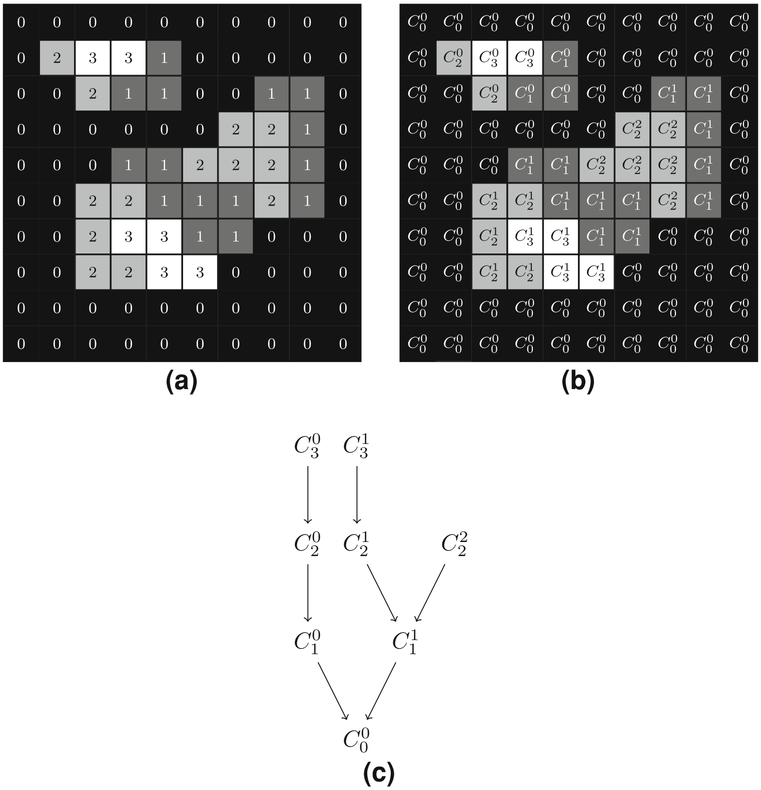
$$\gamma_{\text{sub}}^{\tilde{T}}(f)(p) = \sum_{k=0}^{K} \Gamma^{\tilde{T}}(F_k)(p). \tag{15}$$

If the criterion is increasing, then (14) and (15) are equal to (13). Similar conclusions can be drawn for attribute closing and thickening.

Attribute filters computed on gray-level images according to the definitions given in Sect. 2.3 are not efficient in terms of implementation. However, it is possible to take advantage of an efficient representation of the image called Max-tree, which represents the image as a hierarchical linked structure of its connected components [12]. An example of max-tree is reported in Fig. 1. As one can see in Fig. 1b, the image is composed by connected components of iso-intensity 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. 1c). 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 till the leaves of the tree that correspond to the regional maxima in the image. The computation of the attribute filters on the max-tree structure is composed by three steps which are detailed in the following:

1. *Max-tree creation*. This step aims at generating the tree from the image by identifying the connected components in the image and by modeling the hierarchical representations between nested nodes. This phase of the process is computationally most demanding.
2. *Evaluation of the criterion*. After the creation of the tree, the criterion is evaluated by comparing the attribute extracted from each node and the threshold value (λ) which is considered as reference and defines the degree of filtering. Then, the tree is pruned by removing those nodes that do not fulfill the criterion. If the criterion is non increasing, different filtering rules can be implemented as reported above (see (14, 15)). They correspond to different strategies in pruning the tree [12, 13].
3. *Image restitution*. The final step is the conversion of the pruned tree back to an image.

Since the max-tree is constructed by growing the tree from the lowest grayscale value to the maximum one, this structure is suitable for transformations such as opening and thinning. On the contrary, for operators of closing and thickening, the min-tree is considered. A min-tree is the representation of the image dual with respect to max-tree and can be simply computed as the max-tree of the complement of the input image.



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)

Fig. 1 Example of Max-tree. **a** Gray-scale image with intensities ranging from 0 to 3; **b** image in **a** with its connected components labelled; and **c** Max-tree of (**a**). This shows the relations between the nodes associated to the connected components in (**b**)

3 Morphological Profiles for the Analysis of Panchromatic Images

3.1 Morphological Profiles

In general, for real applications it is unlikely that filtering of an image with a single opening and closing by reconstruction completely models the spatial information in a complex scene. This behavior might limit the capability of the image for analysis. A common procedure is to filter an image with a sequence of many different SEs in order to extract more information on the scene. Granulometries and anti-granulometries are examples of this approach. A granulometry is obtained by the application of a series of opening with SEs of increasing sizes and fixed shape. An anti-granulometry is generated analogously by closing operators. By analyzing the result of a granulometry one is able to gather information on the size distribution of those objects brighter than the surrounding background. Thus, we can refer to this procedure as a multi-scale analysis. When performing such an analysis with operators based on the geodesic reconstruction, the progressive simplification of the image does not come at the detriment of the geometry of those objects that are not cancelled from the image.

The *morphological profiles* are based on these ideas. Morphological profiles were introduced by Pesaresi and Benediktsson in [2] and defined as a concatenation of an anti-granulometry followed by a granulometry performed by closing and opening by reconstruction transformations, respectively. The anti-granulometry is referred as closing profile Π_ϕ and the granulometry as opening profile Π_γ. The morphological opening n profile of an image f is an array of n openings performed on the original image using a SE of size λ, and it is defined as

$$\Pi_\gamma(f) = \Pi_{\gamma_\lambda}(f) : \Pi_{\gamma_\lambda}(f) = \gamma_R^\lambda(f)\} \quad \lambda = 0, 1, \ldots, n. \tag{16}$$

Thus by duality, the morphological closing profile composed by n levels can be denoted by

$$\Pi_\phi(f) = \{\Pi_{\phi_\lambda}(f) : \Pi_{\phi_\lambda}(f) = \phi_R^\lambda(f)\} \quad \lambda = 0, 1, \ldots, n. \tag{17}$$

Therefore, both the opening and closing profiles are generated by opening and closing by reconstruction operators with the image f taken as mask and with SEs of fixed shape and size increasing on the n levels. When a closing profile and an opening profile, both of size n, are joined a morphological profile is obtained. The MP is of size $2n - 1$, because when $\lambda = 0$ the opening and closing profiles are equal to the original image ($\Pi_{\gamma 0} = \gamma_R^0(f) = \Pi_{\Phi 0} = \Phi_R^0(f) = f$) and thus they are considered only once (see Fig. 2).

$$\mathrm{MP}(f) = \left\{ \begin{array}{lll} \Pi_{\phi_\lambda}(f), & \lambda = (n - 1 + i) & i \in [1, n]; \\ \Pi_{\gamma_\lambda}(f), & \lambda = (i - n - 1) & i \in [n + 1, 2n + 1]; \end{array} \right\}. \tag{18}$$

The derivative of a MP, denoted as *Differential Morphological Profile* (DMP) [2], is defined and can be computed as the differences between two adjacent levels of the MP,

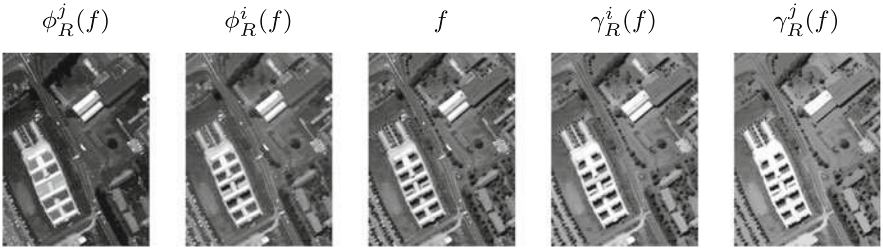


Fig. 2 Example of MP composed by five levels, obtained by two openings and two closings ($j > i$). For generating this MP a squared SE was considered, with size of 5 (i) and 9 (j) pixels

$$\mathrm{DMP}(f) = \left\{ \begin{array}{lll} \Delta_{\phi_\lambda}(f), & \lambda = (n-1+i), & i \in [1,n]; \\ \Delta_{\gamma_\lambda}(f), & \lambda = (i-n-1), & i \in [n+1,2n]; \end{array} \right\} \tag{19}$$

with the differential closing profile Δ_{ϕ_λ} and differential opening profile Δ_{γ_λ} simply defined as

$$\Delta_\gamma = \{\Delta_{\gamma_\lambda} : \Delta_{\gamma_\lambda} = \Pi_{\gamma_\lambda - 1} - \Pi_{\gamma_\lambda}\} \quad \lambda = 1, 2, \ldots, n; \tag{20}$$

$$\Delta_\phi(f) = \{\Delta_{\phi_\lambda} : \Delta_{\phi_\lambda} = \Pi_{\phi\lambda} - \Pi_{\phi\lambda} - 1\} \quad \lambda = 1, 2, \ldots, n. \tag{21}$$

The DMP stores the residuals of the sequential transformations applied to the image. This can be particular useful when the multi-scale analysis has to be visualized, since the most important components of the profiles are more evident than when the MP is considered.

Moreover, from the DMP the information on the scale of the objects in the image can be extracted. In [2], this information was used for generating from the image a multiscale segmentation map, called *morphological characteristic*. In greater detail, each pixel in the image is labelled with the index of the level in the MP in which the maximum of its derivative (i.e., DMP) occurs.

3.2 Attribute Profiles

In this section the concept of *Attribute Profile* (AP) as an extension of the morphological profile is reviewed. First the limitations of MP are reported and subsequently the AP is presented.

Although MPs proved to be an effective tool for performing a multi-scale analysis of the image, they have their main limitation in the capability to model other feature than the size of the objects. For example, if one attempts to filter the image according to different degrees of spectral homogeneity or according to different shape descriptors, the results would be rather cumbersome. This limitation is particularly important when the discriminative power of the analysis could have been increased by modeling other features rather than the size (e.g., contrast, texture, geometry, etc.).

Attribute filters can overcome this limitation of the MPs [14]. Indeed, according to the attribute considered, different information can be extracted from the image. For example, if an increasing attribute is considered (e.g., the area of the regions) the AP performs an analysis based on the scale of the structures in the scene which is analogous to a MP. Instead, if, for example, a measure of the texture, shape, morphology, or contrast (which are, usually, non increasing attributes) is considered as an attribute, it is possible to gather information on different image descriptors. Moreover, from a computational viewpoint, an AP requires much less resources than an equivalent MP. The MP always requires two complete procedures of processing the image, one performed by a closing and the other by an opening transformation for each level of the profile. In contrast, the AP builds up the trees (one max-tree for the thinnings and one min-tree for the thickening) only once and performs the sequential filtering processing as sequential prunings of the tree with different values of λ. This greatly reduces the demand of the analysis with respect to MPs.

In greater detail, attribute profiles perform a multi-level decomposition of the input image based on attribute filters [15]. The great flexibility of the attribute filters in defining the criterion which drives the filtering allow one to filter the image according to features of the image that are not only related to the size of the structures but that can be of any kind. However, this comes at a prize. While MPs are cumulative functions because they are a sequential composition of openings and closings and thus, the absorption property is always fullfilled, this characteristic might not be always verified by AP. This is an important condition because it leads to achieve a progressively increased simplification of the image when the filters values are increased and makes the computation of the derivative of the profile well defined. In greater detail, APs verify this property when the criterion considered is increasing. When it is not, a constraint on the criterion has to be applied. The family of criteria T_i considered in the profile must be formally ordered. This boils down to $\gamma^{T_i} \supseteq \gamma^{T_j}$ and $\phi^{T_i} \subseteq \phi^{T_j}$ for $i \leq j$. The imposed condition does not make the criterion increasing since this property involves two input functions and one criterion, i.e., $f \leq g \Rightarrow \gamma^T(f) \leq \gamma^T(g)$, a condition which is not fulfilled for thickening and thinning transformations.

Analogously to the MP, the AP can be defined as a concatenation of a thickening attribute profile, $\Pi_{\phi^{T'}}$, and an thinning attribute profile, $\Pi_{\gamma^{T'}}$ computed with a generic ordered criterion T':

$$AP(f) = \left\{ \begin{array}{lll} \Pi_{\phi^{T'_\lambda}}, & \lambda = (n-1+i), & \forall \lambda \in [1, n]; \\ \Pi_{\gamma^{T'_\lambda}}, & \lambda = (i-n-1), & \forall \lambda \in [n+1, 2n+1]. \end{array} \right\}. \tag{22}$$

Since $T' = \{T_1, T_2, \ldots, T_n\}$ the set of ordered criteria, for T_i, $T_j \in T'$ and $j \geq i$ the relation $T_j \supseteq T_i$ holds.

It is possible to compute the derivative of the AP analogously to the MP case. We refer the reader to [15] for further details.

3.3 Experimental Results and Discussion

Morphological profiles were first applied in [2] for segmenting two 800 × 800 pixels HR panchromatic images acquired by Indian Remote Sensing 1C (IRS-1C) with a 5-m geometric resolution on a dense urban area of Milan, Italy, and on an agricultural area of Athens, Greece. The application of operators by reconstruction to the two images showed a better representation of the geometry of the objects in the scene with respect to the processing with standard morphological operators. Moreover, the segmentation maps obtained by the morphological characteristic of

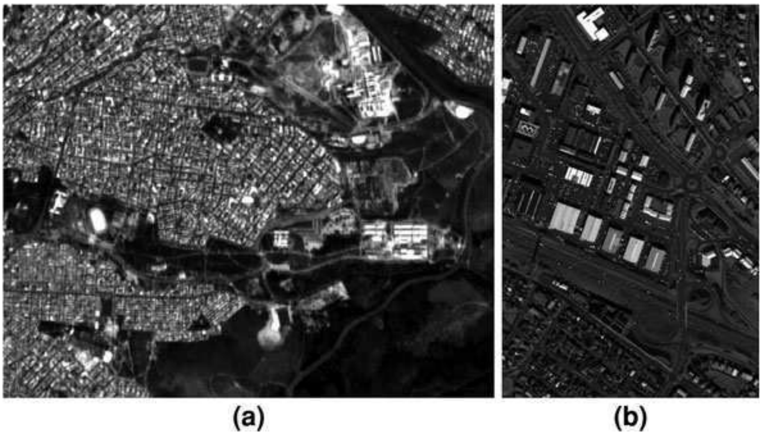


Fig. 3 Panchromatic images. **a** IRS-1C image of the city of Athens, Greece (800 × 800 pixels, 5.8 m geometrical resolution, spectral range 0.5–0.75 μm; **b** IKONOS image of the city of Reykjavik, Iceland (975 × 639 pixels, 1 m geometrical resolution, spectral range 0.53–0.93 μm

the images were not affected by the oversegmentation effect that was noticeable when a classical watershed segmentation was performed.

In [4], the MPs were applied for the first time in a classification task. An IRS-1C panchromatic image of Athen, Greece (Fig. 3a), and an IKONOS panchromatic image from Reykjavik, Iceland (Fig. 3b), were classified with a conjugate gradient neural network. In both the experiments, eight closings and eight openings were applied to the original images leading to a 17-dimensional feature vector considered as input to the neural network. In order to reduce the dimensionality of the filtered data, two feature extraction methods and a feature selection technique were investigated. The considered approaches were: (1) discriminant analysis feature extraction (DAFE) [16]; (2) decision boundary feature extraction (DBFE) [16]; and (3) a simple feature selection based on sorting the indexes of the DMP using the value of the discrete derivative. The obtained classification results showed as the use of the features extracted by the MP increased the overall accuracy from 69.4 to 70.9% of the original panchromatic image to 77.7 and 95.1% when considering the entire differential profile for the IRS-C1 and IKONOS image, respectively. Among the techniques of feature reduction, the DBFE outperformed DAFE and the feature selection technique. However, lower accuracies than those obtained by considering the whole DMP were obtained.

The morphological profile was built in [17] by applying alternating sequential filters (ASF) by reconstruction instead of the operators of opening or closing by reconstruction. Alternating sequential filters by reconstruction are iterative sequential applications of an opening and a closing by reconstruction (or vice versa) of increasing size. The MP built on ASF were applied to the IKONOS panchromatic image in Fig. 3b. The feature extracted were classified by a neural network. Although the standard MP performed better than the one with ASF on the original, the latter showed to be more robust when analyzing the image corrupted by Gaussian noise.

In [18] the DMP was interpreted as a fuzzy measure of the characteristic size and contrast of the objects in the image. The fuzzy measure extracted from the DMP was compared to predefined possibility distributions in order to derive a membership degree for the thematic classes of the samples in the image. This fuzzy measure can be compared to predefined possibility distributions to derive a membership degree for a set of given classes. The decision is taken by selecting the class with the highest membership degree. The experimental results were obtained from the analysis of the Reykjavik IKONOS image in Fig. 3b and achieved an overall accuracy of 52.1% outperforming the one obtained by a neural network of about 12%.

In order to perform a better modeling of the spatial features in the image, in [19] the computation of two MPs with SEs of different shape was proposed for classification. The authors considered in their analysis a disk-shaped SE and a linear SE with different orientations (which generate directional profiles [1]). While the MP built with the former SE is suitable to extract the smallest size of the structures, the latter allows one to infer the largest size of the objects. Moreover, an interesting variant of the geodesic reconstruction called “partial reconstruction” was presented. The proposed reconstruction procedure performs a partial geodesic reconstruction (the iterative process is converging to idempotency). This leads to reaching a trade-off between the preservation of the objects geometries and a reduction of the over segmentation effect introduced by standard reconstruction. Two study areas were considered in the analysis, an IKONOS and a Quickbird panchromatic images both acquired on the area of Ghent (Belgium). The proposed technique significantly outperformed the results obtained without considering any spatial feature in the analysis. Furthermore, an increase in the overall accuracies with respect to the case with standard reconstruction of about 2 and 7%, was achieved by considering the two MPs built with partial reconstruction for the two sites, respectively.

The APs were considered in [15] for the classification of two Quickbird panchromatic images acquired on the city of Trento (Italy). The images showed two complex urban areas of size 400 × 400 and 900 × 900 pixels. In the analysis, three attributes were selected: (1) area; (2) moment of inertia [20]; and (3) standard deviation. The area attribute was chosen for modeling the size of the structures in the image, the moment of inertia for extracting information on the shape of the regions and the standard deviation was considered as a descriptor of the spectral homogeneity of the objects. At first, each AP was considered separately and then, all the features extracted by the APs were taken into account simultaneously. The data were classified by a random forest classifier [21]. The analysis of the results was carried out by considering both thematic and the geometric errors [22] in representing some reference objects in the scene. The obtained results showed a significant increase of the accuracies, both geometric and thematic, when considering spatial features in classification with respect to the original panchromatic bands. The geometrical errors were assessed by five indeces modeling the effects of over-segmentation, under-segmentation, fragmentation, the precision in retrieving the shape and the borders of some objects in the image taken as reference. We refer the interested reader to [22] for more details on the geometrical error indeces. The thematic errors were evaluated by the overall error and the kappa error. The overall error was computed as the percentage of the number of wrongly classified patterns over the whole set of samples used for the test. The kappa error was obtained as $1 - \kappa$, being κ the kappa coefficient. The latter estimates the percentage of the agreement between the labels of the classified patterns and the correct labels which would be expected by chance and it ranges between 0 and 1 [23]. In the experiments, the classification with APs gave similar performances to the classification with the MP, but the profiles were able to extract complementary information from the scene leading to increasing accuracies when considered in classification (decreases in the kappa error were up to 38 and 17% with respect to the original panchromatic image and the MP, respectively).

4 Extended Morphological Profiles to the Analysis of Multispectral and Hyperspectral Images

4.1 Problem of Extending the Morphological Operators to Multitone Images

The extension of the concept of a morphological profile from the analysis of single-tone images to multi-tone images (e.g., multispectral and hyperspectral imagery) is certainly a non trivial task because the extension of the morphological operators for scalar to multivariate values is an ill-posed problem. In fact, the output of a generic morphological operator processesing an image, is usually the result of a function computed on an ordered set of values (e.g., the infimum for erosion, the median for the median filter, the supremum for dilation, etc.). When dealing with scalar images, the ordering of the values mapped by the image $f(p) \rightarrow k$, with $p \in E$ and $k \in \{0, \ldots, K\} \subset \mathbb{Z}$, is well defined. The scalar elements in the partially ordered set $\{0, ..., K\}$ have an unique infimum and supremum. Thus, the morphological operators are well defined. In contrast, when the image destination domain becomes a subset of a multivariate domain, e.g., $f(p) \rightarrow \mathbf{k}$, $\mathbf{k} \in \mathbb{Z}^n$ the ordering relation between the mapped vectorial values is not defined anymore. For this reason, the direct application of concepts seen in previous sections to multi- or hyperspectral images is not possible.

In order to overcome this issue several solutions have been presented in the literature. One possible approach relies on the arbitrary re-definition of the concepts of morphological filters for handling multi-valued images by forcing an ordering relation on the vectorial set of values. In in [24], Plaza et al. proposed a reduced vector ordering scheme based on the spectral purity index of the pixel vectors. The input vectors are ordered according to a spectral-based distance measure (i.e., scalar value). Three distance measures commonly used in hyperspectral analysis were considered: (1) spectral angle distance (SAD); (2) spectral information divergence (SID); and (3) hidden Markov model-based information divergence (HMMID). Having defined an ordering relation between vectorial values, the definitions of the morphological operators of opening and closing by reconstruction computed according to this ordering relation can be applied to the hyperspectral data. A reader who is interested in greater details on the extensions of the mathematical morphology concept to multi-valued images can refer to [25–28].

Another approach for extending morphological transformations to vectorial data deals at first with the reduction of the hyperspectral data to only one (or few) channels and subsequently to the application of the morphological operators to each obtained image separately.

The reduction of the dimensionality can be done by means of several techniques. The first work based on this approach considered Principal Component Analysis (PCA) as feature reduction technique [29, 30] (see Sect. 4.2 for details). In [29] Independent Component Analysis was used. Kernel PCA (KPCA) was exploited in [31, 32]. In [33] the reduction of the dimensionality was performed by PCA, KPCA, Non-parametric Weighted Feature Extraction , Decision Boundary Feature Extraction (DBFE) and Bhattacharyya Distance feature selection (BDFE) techniques [16] .

4.2 Extended Morphological Profile

The extension of the MP to hyperspectral data presented in [30], which led to the definition of the Extended Morphological Profile, is achieved through a two step procedure. At first, the multidimensional data is reduced through a PCA to few informative dimensions (i.e., the first principal components, PCs). The PCs corresponds to the eigenvectors of the estimated covariance matrix of the data and are ordered increasingly according to the values of the correspondent eigenvalues. The first PCs are meaningful for data representation since they account for most of the variance of the data in the original feature space. In general, the first considered PCs accumulate most of the total variance of the data (e.g., usually a threshold on 99% is taken). Subsequently, on each PC a full MP is computed. Thus, the EMP of the first c principal components can be formalized by

$$\mathrm{EMP}(f) = \{\mathrm{MP}(\mathrm{PC}_1), \mathrm{MP}(\mathrm{PC}_2), \ldots, \mathrm{MP}(\mathrm{PC}_c)\}. \tag{23}$$

As it is seen from (23), the EMP is the concatenation of MPs on a single stack. Since the dimensionality of the EMP can rapidly increase when increasing the number of considered PCs and the levels of the MP, in [30] the application of feature extraction techniques was proposed in order to decrease the curse of dimensionality phenomenon [34]. Feature extraction techniques for classification should be considered in order to achieve a dimensionality reduction and an

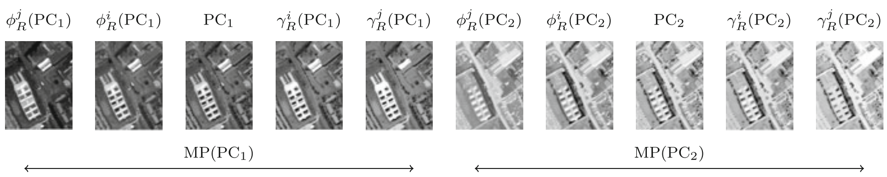

$\phi_R^j(\mathrm{PC}_1)$
$\phi_R^i(\mathrm{PC}_1)$
PC_1
$\gamma_R^i(\mathrm{PC}_1)$
$\gamma_R^j(\mathrm{PC}_1)$
$\phi_R^j(\mathrm{PC}_2)$
$\phi_R^i(\mathrm{PC}_2)$
PC_2
$\gamma_R^i(\mathrm{PC}_2)$
$\gamma_R^j(\mathrm{PC}_2)$
$\mathrm{MP}(\mathrm{PC}_1)$
$\mathrm{MP}(\mathrm{PC}_2)$

Fig. 4 Example of EMP computed on the first two PCs and composed by five levels for each MP

effective separation of the distributions of the classes in the transformed feature space [35]. An example of an EMP is reported in (Fig. 4).

4.3 *Extended Attribute Profiles*

Since the EMP is computed by morphological operators by reconstruction, similar comments can be addressed to the limited capability in modelling the spatial information as done in Sect. 3.2. Thus, it is possible to compute EMPs by composition of APs leading to the definition of *Extended Attribute Profiles* (EAPs) [36]. Again, the main advantage of using attribute filters instead of operators based on the geodesic reconstruction relies in their greater capability in modelling many image descriptors. Thus, similar to Sect. 4.2, the EAP is obtained by computing an AP on each of the c principal components extracted from the original hyperspectral data:

$$\mathrm{EAP}(f) = \{\mathrm{AP}(\mathrm{PC}_1), \mathrm{AP}(\mathrm{PC}_2), \ldots, \mathrm{AP}(\mathrm{PC}_c)\}. \tag{24}$$

It is worth noticing that the EAP includes in its definition the EMP (because the operators by reconstruction can be viewed as a particular set of morphological attribute filters) and, thus, it can be considered as its generalization. Moreover, also from a computational perspective, the generation of an EAP requires a reduced load with respect to the calculation of an EMP. In fact, taking advantage of the representation of the image as a max- and a min-tree, the demand of the filtering stage can be significantly reduced. Examples of EAPs computed with different attributes are depicted in Fig. 5

The concept of the EAP can be further extended by considering many different attributes in the analysis, by creating an EAP for each attribute considered. When the different EAPs are sequentially stacked in a single data structure, we obtain an *Extended Multi-Attribute Profile* (EMAP) [36]. An EMAP composed by m different EAPs can be easily formulated as:

$$\mathrm{EMAP}(f) = \{\mathrm{EAP}_{a_1}(f), \mathrm{EAP}'_{a_2}(f), \ldots, \mathrm{EAP}'_{a_m}(f)\} \tag{25}$$

with a_i a generic attribute and $\mathrm{EAP}' = \mathrm{EAP} \backslash \{\mathrm{PC}_1, ..., \mathrm{PC}_c\}$. The latter relation is necessary for avoiding the multiple presence of the c principal components.

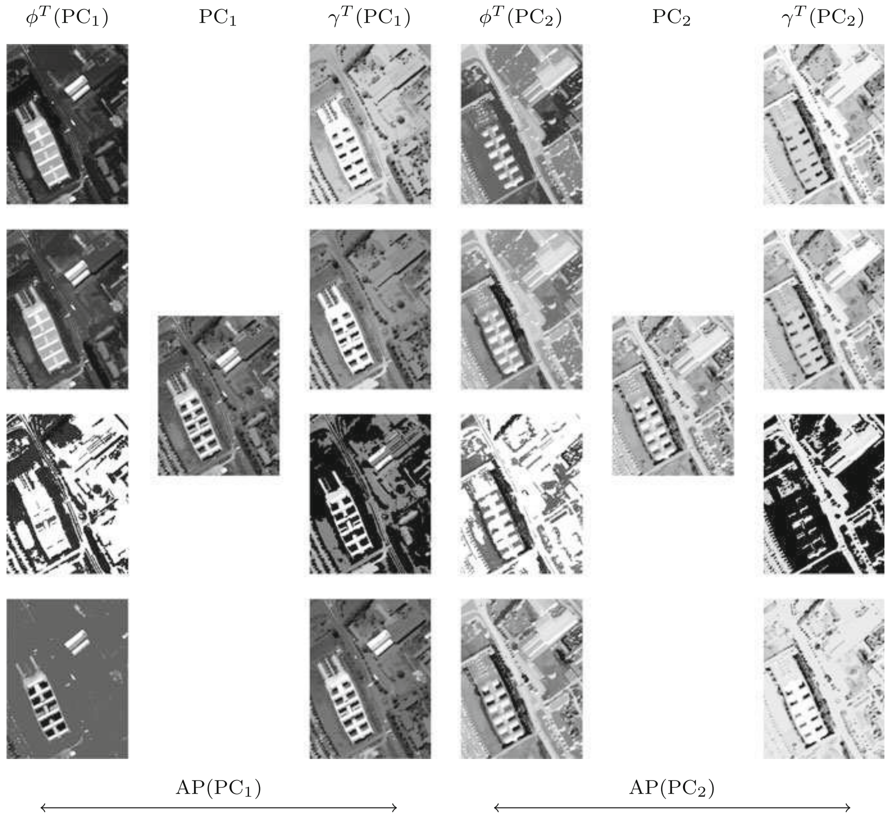

$\phi^T(\mathrm{PC}_1)$
PC_1
$\gamma^T(\mathrm{PC}_1)$
$\phi^T(\mathrm{PC}_2)$
PC_2
$\gamma^T(\mathrm{PC}_2)$
$\mathrm{AP}(\mathrm{PC}_1)$
$\mathrm{AP}(\mathrm{PC}_2)$

Fig. 5 Examples of EAPs computed on the first two PCs of a sample image. Each row shows a EAP built by different attributes. Attributes, starting from the first row are: area, length of the diagonal of the bounding box, moment of inertia and standard deviation. Each EAP is composed by the concatenation of two APs computed on PC_1 and PC_2. Each AP is composed by three levels, a thickening image ϕ^T, the original PC and a thinning image γ^T. All the thickening and thinning transformations were computed with the following attributes value, λs. Area: 5,000; Length of the diagonal: 100; Moment of inertia: 0.5; Standard deviation: 50

On the one hand, when considering an EMAP, greater capabilities in extracting spatial information are gained with respect to considering only a single EAP; on the other, this leads to an increase in the dimensionality.

4.4 Experimental Results and Discussion

The extended morphological transformations based on the ordering of the pixels multidimensional values done by considering spectral-distance metrics were applied to two hyperspectral images acquired by AVIRIS and DAIS sensors.

The AVIRIS image was acquired on Salinas Valley (CA) and is composed by 512 × 217 pixels with 192 spectral bands with 3.7 m of spatial resolution. The DAIS image showed a 400 × 400 pixels scene of the center of Pavia (Italy) with a 5-m geometrical resolution. The results obtained by the proposed techniques outperformed the classification of the original hyperspectral images up to 8 and 7% for the two images, respectively.

The EMP presented in Sect. 4.2 was applied in [30] to two hyperspectral images, one acquired by the DAIS sensor on the center of Pavia (400 × 400 pixels, 80 spectral bands, 2.4 geometrical resolution) and the other collected by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) over the Washington DC Mall area (1,280 × 307, 189 spectral bands, 2.8 m spatial resolution). The experiments were obtained by considering the first two PCs for building the EMP. When considering the features extracted by the EMP, the overall accuracy in classifying the test sites with a neural network significantly increased with respect to considering only the hyperspectral data (+45% and +12% for the two images, respectively). Moreover, DAFE, DBFE, and NWFE [16] were considered for reducing the dimensionality of the data before the classification. Although a significant reduction of the dimensionality of the data was achieved (i.e., reducing the load for the classification stage) no increase in terms of overall accuracy were accounted with respect to considering the EMP with full dimensionality. However, among the considered feature extraction techniques, in both cases NWFE performed the best among the others in terms of classification accuracies reaching values of accuracy close to those obtained by the full EMP.

In [37] the features extracted by an EMP computed on the first PC were considered by using five classifiers (Maximum Likelihood for Gaussian data, Fisher linear discriminant, the ECHO classifier, Fuzzy ARTMAP and a feed forward neural network classifier) along with two feature extraction techniques (DAFE and DBFE). The data used in the experiments involved two test sites on the urban area of Pavia, Italy, acquired with the Digital Airborne Imaging Spectrometer (DAIS). Each hyperspectral image was composed of 80 channels with a spatial resolution of 2.6 m. When considering the morphological features the overall accuracy increased by more than 27% with respect to considering the first PC alone. Moreover, the reduction of the feature size with the DBFE technique further improved the accuracy of about 2%.

In [29], ICA was considered instead of PCA for computing the EMP. ICA, in contrast to PCA, leads to a better extraction of the information sources (especially when they are non Gaussian). In experiments an hyperspectral image of the center of Pavia (Italy) acquired by the ROSIS-03 sensor (see 6(a)) was considered. The classification was performed with a maximum likelihood classifier. The overall accuracy obtained by the EMP built on the independent components outperformed by 5% the overall accuracy of the classification of the original hyperspectral data.

Kernel Principal Component Analysis instead of the conventional PCA was considered in [32] as feature reduction technique for computing the EMP. Results were obtained for three hyperspectral images, two acquired on the city of Pavia

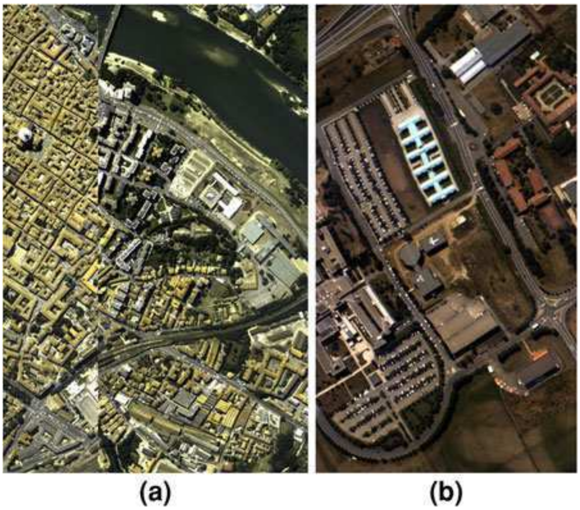


Fig. 6 Hyperspectral images acquired by ROSIS-03 sensor over the area of Pavia (Italy) with 2.6 m of spatial resolution. **a** Pavia, city center, 1,096 × 715 pixels, 102 spectral bands; **b** Pavia, University area, 610 × 340 pixels, 103 spectral bands

(Italy) (Fig. 6a, b) and one on Washington DC Mall (1,280 × 307, 189 spectral bands, 2.8 m spatial resolution). An SVM classifier with linear and Gaussian kernels was considered in the experiments. The results obtained proved that KPCA can extract more informative components with respect to PCA. In fact, the EMP computed on the KPCs increased up to +20% and +5% the overall accuracy obtained by the classification of the hyperspectral data and with the EMP with PCA, respectively.

Several feature extraction and selection methods were considered for building the EMP. The classification maps obtained with a random forest and an SVM classifier applied to the hyperspectral images reported in Fig. 6a and b, showed how the EMP with PCA is not adequate in terms of overall accuracy with respect to other techniques. In particular, NWFE and BDFS performed the best on the experiments with both the classifiers.

The work presented in [5] was devoted to the fusion of spatial features extracted through a standard EMP and the original hyperspectral data. This approach was proposed to increase the amount of spectral information considered in the classification task. The experimental analysis was carried out on two hyperspectral images of the city of Pavia (Italy) both acquired by ROSIS-03 sensor. The two original images are shown in Fig. 6a and b. Feature extraction techniques were also employed for reducing the dimensionality of the data and an SVM classifier was used for generating the classification maps. For the university site (see Fig. 6b), the overall accuracy increased from 79 to 84% without feature extraction and to 88% with feature extraction, with respect to the EMP obtained with the proposed approach.

In [38], an extension of the segmentation procedure based on the analysis of DMPs for panchromatic images [2] was proposed. The novel segmentation technique was developed for automatic object detection in high-resolution images by combining spectral and structural information. In contrast to [2], the DMPs computed on the first PCs extracted from the images were analyzed in order to extract the connected components that best represent each object in the scene. Three hyperspectral images were considered: the image of the center of Pavia (Fig. 6a), the HYDICE image acquired over Washington DC Mall (1,280 × 307, 189 spectral bands, 2.8 m spatial resolution), and a 500 × 500 pansharpened IKONOS image of Ankara (Turkey). The obtained results showed a more precise segmentation of the images and a reduced oversegmentation effect with respect to the maps obtained by the morphological characteristic of [2].

The extension of the EMP obtained by considering attribute filters (EAP, EMAP Sect. 4.3) was used for the classification of two hyperspectral images acquired on Pavia (see Fig. 6a, b) [36]. For the two data sets, the first four principal components were initially considered in the analysis in order to extract more than the 99% of the total variance of the multivariate original data. Four attributes were considered in the analysis: (1) area of the regions; (2) diagonal of the box bounding the region (as the area, this is a measure of the size of the regions); (3) first moment invariant of Hu, or moment of inertia (it measures the elongation of the regions), [20]; (4) standard deviation of the gray-level values of the pixels in the regions (index related to the homogeneity of the regions). For each attribute an EAP was computed with four thresholding values. The four EAPs were also considered together as an EMAP (see Sect. 4.3). A random forest classifier was employed for classifying the data. Again, the inclusion of the spatial information led to an increase in accuracy up to 21.9% for the university data set with respect to considering only the PCs. The EMAP performed best in terms of overall accuracy for the center image with a gain with respect to the single PCs and the EMP of about 2 and 1%, respectively. For the image of the university, the best performance was achieved by the EAP with area attribute, which showed an increase of overall accuracy of 2% and 12% over the EMAP and EMP, respectively.

5 Conclusion

In this chapter an overview of the morphological profile (MP) in remote sensing applications has been given. The MP proved to be an effective tool for the analysis of high geometrical resolution remote sensing images because it is defined as a composition of opening and closing by reconstruction transformations. Operators by reconstruction permit to filter the image by entirely preserving the geometry of those structures that are not erased from the scene. A recent generalization of the MP based on morphological attribute filters led to the definition of attribute profiles, which, in contrast to MPs, show a great flexibility in modeling many different structural features. The features of attribute filters make the attribute profiles, at the present, one of the most promising developments of the MP concept.

The problem of extending the morphological operators from scalar to multitone images was also reviewed in this chapter and the solution that led to the definition of the extended morphological profile (EMP) for multispectral and hyperspectral images was presented. The vectorial image is reduced through Principal Component Analysis for constructing the EMP to a reduced number of images, on which MPs are computed. The EMP is finally obtained as the concatenation of the single MPs. As for the MP, the extension of the EMP based on attribute filters was also investigated.

An overview of results obtained by experimental analysis of various techniques developed using the MP and EMP were reported. A significant increase in classification accuracies was observed when features extracted by MP/EMP (or its variants or extensions) were used for classification in comparison to approaches that only use spectral information. Moreover, the capabilities of modeling the structural features were further improved when considering the profiles computed with morphological attribute filters. The better characterization of the spatial information increased the classification accuracy. This analysis proves the important role of the morphological profile for classification.

Acknowledgments This work was in part supported by the Research Fund of the University of Iceland and the University of Trento.

References

1. Soille, P., Pesaresi, M.: Advances in mathematical morphology applied to geosciences and remote sensing. IEEE Trans. Geosci. Remote Sens. **40**, 2042–2055 (2002)
2. Pesaresi, M., Benediktsson, J.A.: A new approach for the morphological segmentation of high-resolution satellite imagery. IEEE Trans. Geosci. Remote Sens. **39**(2), 309–320 (2001)
3. Tarabalka, Y., Chanussot, J., Benediktsson, J.A., Angulo, J., Fauvel, M.: Segmentation and classification of hyperspectral data using watershed. In: Proceedings of the IEEE International Geoscience Remote Sensing Symposium 2008, IGARSS '08, July 7–11, **3**, pp. III–652–III–655 (2008)
4. Benediktsson, J.A., Pesaresi, M., Amason, K.: Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. IEEE Trans. Geosci. Remote Sens. **41**(9), 1940–1949 (2003)
5. Fauvel, M., Benediktsson, J.A., Chanussot, J., Sveinsson, J.R.: Spectral and spatial classification of hyperspectral data using SVMS and morphological profiles. IEEE Trans. Geosci. Remote Sens. **46**(11), 3804–3814 (2008)
6. Dalla Mura M., Benediktsson, J.A., Bovolo, F., Bruzzone, L.: An unsupervised technique based on morphological filters for change detection in very high resolution images. IEEE Geosci. Remote Sens. Lett. **5**(3), 433–437 (2008)
7. Soille, P.: Morphological Image Analysis, Principles and Applications, 2nd edn. Springer, Berlin (2003)
8. Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J.C., Trianni, G.: Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. **113**(Supp. 1), 110–122 (2009)
9. Serra, J.: Connectivity on complete lattices. J. Math. Imaging Vis. **9**(3), 231–251 (1998)

10. Breen, E.J., Jones, R.: Attribute openings, thinnings, and granulometries. Comput. Vis. Image Underst. **64**(3), 377–389 (1996)
11. Maragos, P., Ziff, R.D.: Threshold superposition in morphological image analysis systems. IEEE Trans. Pattern Anal. Mach. Intell. **12**(5), 498–504 (1990)
12. Salembier, P., Oliveras, A., Garrido, L.: Antiextensive connected operators for image and sequence processing. IEEE Trans. Image Process. **7**(4), 555–570 (1998)
13. Urbach, E.R., Roerdink, J.B.T.M., Wilkinson, M.H.F.: Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images. IEEE Trans. Image Process. **29**(2), 272–285 (2007)
14. Dalla Mura M., Benediktsson, J.A., Waske, B., Bruzzone, L.: Morphological attribute filters for the analysis of very high resolution remote sensing images. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium 2009, IGARSS '09, July **3**, pp. III–97 –III–100 (2009)
15. Dalla Mura M., Benediktsson, J.A., Waske, B., Bruzzone, L.: Morphological attribute profiles for the analysis of very high resolution images. IEEE Trans. Geosci. Remote Sens. **48**(10), 3747–3762 (2010)
16. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley-Interscience, New York (2003)
17. Chanussot, J., Benediktsson, J.A., Pesaresi, M.: On the use of morphological alternated sequential filters for the classification of remote sensing images from urban areas. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium 2003, IGARSS '03, July 21–25, **1**, pp. 473–475 (2003)
18. Chanussot, J., Benediktsson, J.A., Fauvel, M.: Classification of remote sensing images from urban areas using a fuzzy possibilistic model. IEEE Trans. Geosci. Remote Sens. **3**(1), 40–44 (2006)
19. Bellens, R., Gautama, S., Martinez-Fonte, L., Philips, W., Chan, J.C.-W., Canters, F.: Improved classification of VHR images of urban areas using directional morphological profiles. IEEE Trans. Geosci. Remote Sens. **46**(10), 2803 –2813 (2008)
20. Hu, M.: Visual pattern recognition by moment invariants. IRE Trans. Inform. Theory **8**(2), 179–187 (1962)
21. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
22. Persello, C., Bruzzone, L.: A novel protocol for accuracy assessment in classification of very high resolution images. IEEE Trans. Geosci. Remote Sens. **48**(3), 1232 –1244 (2010)
23. Congalton, R.G., Green, K.: Assessing the Accuracy of Remotely Sensed Data. CRC Press, Boca Raton (2008)
24. Plaza, A., Martinez, P., Plaza, J., Perez, R.: Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. IEEE Trans. Geosci. Remote Sens. **43**(3), 466–479 (2005)
25. Chanussot, J., Lambert, P.: Total ordering based on space filling curves for multivalued morphology. In: 4th International Symposium on Mathematical Morphology and its Applications, vol. 6, pp. 51–58 (1998)
26. Garrido, L., Salembier, P., Garcia, D.: Extensive operators in partition lattices for image sequence analysis. Signal Process. **66**(2), 157–180 (1998)
27. Lambert, P., Chanussot, J.: Extending mathematical morphology to color image processing. In: CGIP-1st Internatational Conference on Color in Graphics and Image Processing, pp. 158–163. Saint-Etienne, France (2000)
28. Aptoula, E., Lefavre, S.: A comparative study on multivariate mathematical morphology. Patt. Recog. **40**(11), 2914–2929 (2007)
29. Palmason, J.A., Benediktsson, J.A., Sveinsson, J.R.: Classification of hyperspectral ROSIS data from urban areas. In: Proceedings of the 2nd International Conference on Recent Advances in Space Technologies RAST 2005, June 9–11, pp. 63–69 (2005)
30. Benediktsson, J.A., Palmason, J.A., Sveinsson, J.R.: Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans. Geosci. Remote Sens. **43**(3), 480–491 (2005)

31. Fauvel, M., Chanussot, J., Benediktsson, J.A.: Kernel principal component analysis for feature reduction in hyperspectrale images analysis. In: Proceedings of the 7th Nordic Signal Processing Symposium NORSIG 2006, June, pp. 238–241 (2006)
32. Fauvel, M., Chanussot, J., Benediktsson, J.A.: Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas. EURASIP J. Adv. Signal Process. **2009**. (2009) doi:10.1155/2009/783194
33. Castaings, T., Waske, B., Benediktsson, J.A., Chanussot, J.: On the influence of feature reduction for the classification of hyperspectral images based on the extended morphological profile. Int. J. Remote Sens. **31**(22), 5921–5939 (2010)
34. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, USA (2000)
35. Fukunaga, K.: Introduction to Statistical Pattern Recognition. 2nd edn. Academic Press, London (1990)
36. Dalla Mura, M., Benediktsson, J.A., Waske, B., Bruzzone, L.: Extended profiles with morphological attribute filters for the analysis of hyperspectral data. Int. J. Remote Sens. **31**(22), 5975–5991 (2010)
37. Dell'Acqua, F., Gamba, P., Ferrari, A., Palmason, J.A., Benediktsson, J.A., Arnason, K.: Exploiting spectral and spatial information in hyperspectral urban data with high resolution. IEEE Trans. Geosci. Remote Sens. Lett **1**(4), 322–326 (2004)
38. Akcay, H.G., Aksoy, S.: Automatic detection of geospatial objects using multiple hierarchical segmentations. IEEE Trans. Geosci. Remote Sens. **46**(7), 2097 –2111 (2008)

Decision Fusion of Multiple Classifiers for Vegetation Mapping and Monitoring Applications by Means of Hyperspectral Data

Karoly Livius Bakos, Prashanth Reddy Marpu and Paolo Gamba

Abstract In this chapter, we introduce methodologies for fusion of multiple classifiers and a neural network architecture for mapping vegetation by means of remote sensing imagery. It is very normal that different classification schemes yield slightly different results for different classes. This effect is even more prominent in vegetation mapping applications due to the inconsistent spectral signatures of the vegetation classes. We study the possibility of combining the results of different classifiers by considering the best results for individual classes to produce an improved classification result. We propose two types of methodologies, one which uses only the classification result and the other which uses the class membership values produced by the weak classifiers. A comparison is also done with the simple majority voting scheme of the multiple classifiers. Our experiments clearly show the improvement of classification accuracy using the proposed fusion techniques.

Keywords Hyperspectral · Classification · Neural Network · Decision tree · Ensemble

K. L. Bakos (✉), P. R. Marpu and P. Gamba
Department of Electronics, Telecommunication and Remote Sensing Laboratory,
University of Pavia, Via Ferrata 1, 27100 Pavia, Italy
e-mail: karoly.bakos@unipv.it

P. R. Marpu
e-mail: prashanthmarpu@ieee.org

P. Gamba
e-mail: paolo.gamba@unipv.it

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_9,

1 Introduction

Remote sensing data interpretation techniques presently play a key role in many earth science, environmental and conservation applications. The availability of these datasets simplifies and speeds up the procedure of carrying out several tasks in those fields. Hyperspectral images are being increasingly made available in the last few years and are slowly being used in several applications. These images are characterized by their huge feature size with data recorded at a very fine spectral resolution in hundreds of narrow frequency bands. These bands provide a wealth of information regarding the physical nature of different objects in the scene imaged by the sensors. However, the high dimensionality of the data also makes it more difficult to use the data efficiently for classification. Moreover, there is a lot of redundancy in the data which has to be reduced.

In general, in hyperspectral data interpretation, multiple steps are required in order to achieve the final result which is most often a classification map of any kind. The collection of these multiple steps is often referred to as processing chain. In this study, we use the term processing chain according to [1], which consists of two levels of processing with a providers side (radiometric correction, geometric correction and atmospheric correction), and a user side (feature extraction/selection, classification, post-processing). Here, we only study the user side of the processing chain, which means that we try to find optimal solution for the data dimensionality reduction and the classification process.

In this chapter, we focus on vegetation mapping using hyperspectral data. Vegetation is generally difficult to observe using remote sensing techniques, but serves the basis of environmental and ecological applications of remotely sensed data. For a long time now, remote sensing data interpretation was used to derive land cover maps by mainly using multispectral imagery or aerial photography interpretation techniques that have the great advantage of cutting down the required field work, which is very demanding on resources. While using multispectral imagery, the information that can be derived from the data is limited, and apart from some broad approximation of a few physical properties of the observed surfaces, it is mostly limited to the identification of the general land cover types in the imagery. With hyperspectral images, it might also be possible to distinguish between the sub-classes within the general land cover types.

The inconsistent spectral signature of vegetative surfaces makes most of the existing interpretation methodologies very scene-specific. Therefore, most of them can only have some indication on how a new scene could be processed. Because of this, we suggest that only adaptive techniques can be applied in order to construct a generally applicable processing chain for vegetation mapping. Because of the nature of the problem, and the high level of complexity, it can be clearly seen that adaptive learning from the data is crucial on each scene, whereby the proposed methodologies are capable to extract useful information from the scene in a supervised manner. The learning depends on the number of training samples, the distribution of the training samples and the methodologies used to carry out data dimensionality reduction steps.

For supervised classification, different approaches and algorithms are available [2]. In most cases, as was shown in previous studies [3] single stage classification systems are not flexible enough to adapt to the complexity of the datasets. Therefore, accurate classification becomes a particularly difficult task to carry out. Although, there are both data dimensionality reduction techniques and classification algorithms that are reported to be superior than others, the different studies also shows that the performance is dependent on the class that is being detected [4]. The methodologies that will be introduced in this chapter to combine the results of multiple classifiers will address this situation by taking into account that, even on a single scene, there can be land cover classes for which different processing chains distinguish between different classes in a different way. A simple example is when we have an image on which the best performance in terms of overall accuracy can be achieved for instance using support vector machine (SVM) [5] when the first 20 components of the principal components analysis (PCA)[6] transformed image are used as inputs to the classifier. However, imagine that there are two classes that show higher separability, when the first 15 components of the minimum noise fraction (MNF) [7] transformation image are used for classification using a simple Maximum Likelihood (ML) [8] decision rule. If such multiple results can be combined to make a better decision, then more accurate classification maps can be produced.

2 Study Area

The study area we selected for the experiments is the Indian Pine AVIRIS test site located in the USA. It contains two-thirds of agriculture (some of the crops are in early stages of growth with low coverage), and one-third of forest, two highways, a rail lane and some houses. Ground truth determines 16 different classes (not mutually exclusive) shown in Table 2.

2.1 Image Data

To carry out the experiments, we used the standard dataset acquired using Airborne Visible Infrared Imaging Spectrometer (AVIRIS) on June 12, 1992. This dataset is often used for demonstration of new methodologies or making comparison among different image interpretation techniques. Water absorption bands

Table 1 Characteristics of the Indian Pines AVIRIS data set

Spectral range: 400–2,500 nm
Spectral resolution: 10 nm
Spatial resolution: 1.5 m (variable)
Swath width: 1.0 km (variable)
Sampling: scene based (145 samples, 145 lines)

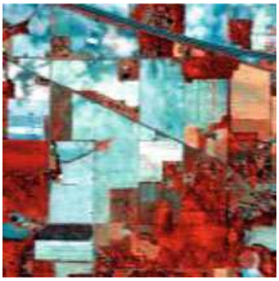


Fig. 1 The Indian Pines 1992 dataset in false colour composite

(104–108, 150–163 and 220) were removed [9], obtaining a 200 band spectrum at each pixel (Table 1).

A false colour composite of the image using bands 90, 42, 11 for red, green and blue respectively is shown in Fig. 1.

2.2 Field Data

The available field data for the Indian Pines site is shown in Fig. 2 and Table 2 showing the number of samples within each class.

The image contains 16 different land cover classes as listed in Table 2. Because of the limited number of pixels available in some of the classes, we used only a nine class subset of the original ground truth. This was done because of the fact that the classes with limited representation do not allow to appropriately train the different classifiers and to have enough samples to assess the interpretation accuracy without biasing the results. The choice of using the dataset with only the nine classes we selected is also adopted in few other studies because of the same reason.

The ground truth that was used within the framework of the study is shown in Fig. 3.

2.3 Training Data

The training set was obtained by systematic stratified random sampling of the nine class ground truth image to get 10% of the ground truth coverage. The number of training samples per class can be seen in Table 3.

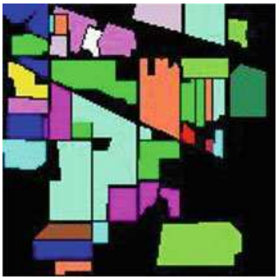


Fig. 2 The training data for the AVIRIS Indian Pines 1992 dataset

Table 2 The Indian Pines 1992 AVIRIS data land cover classes

Land cover class	Number of pixels
Alfalfa	54
Corn not tillage	1434
Corn min. tillage	834
Corn	234
Grass-pasture	497
Grass-tree	747
Grass-pasture-mowed	26
Hay windrowed	489
Oats	20
Soybean to tillage	968
Soybean min. tillage	2468
Soybean clean	614
Wheat	212
Woods	1294
Bldg.-grass-tree-drives	380
Stone-steel-towers	95

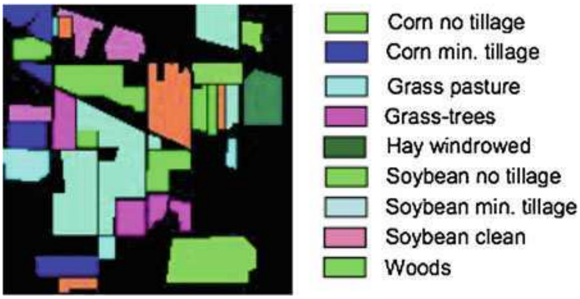


Fig. 3 The nine class ground truth map that was used in the study where the colour table is identical to the 16 class ground truth image

Table 3 The number of samples per classes within the training data

Class #	Class name	# of training samples
1	Corn not tillage	143
2	Corn min. tillage	83
3	Grass-pasture	50
4	Grass-tree	75
5	Hay windrowed	49
6	Soybean to tillage	97
7	Soybean min. tillage	247
8	Soybean clean	61
9	Woods	129

As can be observed in Table 3, there are classes that are over-represented within the training and the ground truth data. Therefore, to measure the accuracy levels we do not use the overall accuracy exclusively, but we also measure the accuracy levels on a class by class basis and identify significant changes among them as introduced in Sect. 4. In doing this, we mitigate the effect of the number of samples on the final evaluation. For instance Class 7 has significantly more samples for validation than any other class and hence very small change in the accuracy value of this class has huge effect on the final overall accuracy of the classification.

3 Methods for Fusion of Multiple Classifiers

A large number of methods have been developed for fusing the results of multiple classifiers often named as *decision fusion*, *ensemble of classifiers* or *mixture of experts* [10–16]. There are generally two ways of fusing the classifiers. We can either make a decision based on the results of individual classifiers (e.g., a simple majority voting where the pixel is assigned to the class which gets the majority votes using all the classifiers) or combine the classifiers before a decision is made by them, to generate a final result using the membership values of the classes as calculated by the individual classifiers. A very good overview of the types of multiple classifier fusion methods is given in [17–19]. In this chapter, we present methodologies of both the mentioned types.

3.1 Decision Fusion Using Hierarchical Tree Structure

Hierarchical classification and ensembles are well known in data classification methods [2]. In this part, we introduce a methodology to create an ensemble of different classifiers using a hierarchical tree structure by means of a simple learning algorithm. The learning is based on the initial analysis of the available data and it optimizes the structure of a binary decision tree (BDTC) like ensemble in terms of nodes, inputs, and decision rules to be applied at each node. This can be useful when sets of data dimensionality reduction techniques and classification algorithms are already available for the user. The aim is to combine the classification results of different processing chains using an ensemble that enables to achieve higher mapping accuracy level than any of the individual processing chains.

The proposed algorithm uses an approach where the learning uses a per class approach as opposed to traditional learning techniques where a “per sample” approach is employed. The learning mechanism starts with a series of initial pre-classifications of a limited subset of the data and class confusions are measured. Using the obtained confusion values the following information is aimed to be obtained:

- for each class the processing chain that enables the most accurate discrimination of the particular class
- a ranking of classes based on the previously identified best possible discrimination from highest to lowest accuracy level

This information then enables the generation of a hierarchical decision tree like ensemble design which we use together with a sequential data classifier. In practice, the design of the decision tree structure starts with the full set of confusion matrices obtained from the results of different classifiers, each one containing all the vegetation classes in the training data (nine classes in our test set, see Table 2). The procedure iteratively identifies the combination of a class and an input source for which the effective accuracy defined using (1) is maximal. After the class-input set has been identified, the class is removed from the full set of confusion matrices and the diagonal element of the selected class is set to 0. The procedure is then repeated until each class is selected once. Eventually, a decision tree configuration is obtained by means of the selected input and class combinations. The first chosen pair corresponds to the first node of the BDTC, while other combinations are used to build the BDTC classifier by means of a top-down approach.

The effective accuracy, A_w for the estimation procedure is calculated using the formula:

$$A_w = \frac{\sqrt{A_p^2 * A_u^2}}{2} \tag{1}$$

where A$_p$, A$_u$ are the producer and user accuracy values respectively for the class of interest. An example representation of a BDTC design can be seen in Fig. 4 and Table 4**.**

In this example, according to Table 2 the sequential classification starts with the classification of the MNF transformed data with the ML algorithm. For the next node, we mask out the pixels labelled as class A from all the available inputs. The masked pixels are labelled as belonging to class A in the final result. We then classify the PCA input using the SVM classifier and repeat the same procedure as for Class B. Iteratively, the full map is obtained.

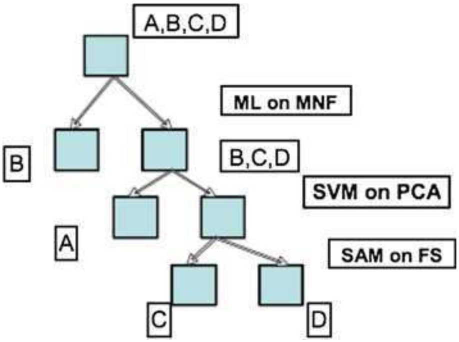


Fig. 4 Hierarchical tree structure of the ensemble classifier where at each node a different input is used and a different decision rule is applied for sequential extraction of classes

Table 4 The design for the BDTC ensemble: in Fig. 4 MNF refers to minimum noise fraction transformation, PCA referrers to principal components analysis transformation, and FS referrers to selected features

Node #	Class to label	Input	Decision rule
1	B	MNF	Maximum likelihood
2	A	PCA	Support vector machine
3	C and D	FS	Spectral Angle Mapper

3.2 Decision Fusion Using the Hierarchical Tree and Class Membership Values

The previous methodology can be improved using a simple learning mechanism that uses class membership values provided by the various classifiers. Most of the classifiers use a likelihood estimate for classification, hence for each pixel, it is possible to define a membership value to see how probable is that the pixel belongs to a particular class. We use these class membership values obtained by the weak classifiers for creating an ensemble classifier. One disadvantage of the methodology is that the different classifications must be obtained for the full scene in advance and therefore it is not computationally effective as opposed to the hierarchical tree structure estimation introduced earlier in the chapter. In order to reduce computational costs, we used the hierarchical structure and only considered those processing chains that were identified to be suitable for the hierarchical tree structure ensemble.

We carry out the classification on the full scene using therefore only the selected processing chains and store class membership values for each class. The learning algorithm then works as follows:

- at every pixel, the best three classifiers are identified by ranking the maxima of membership values calculated for each classifier
- membership values are weighted using values 3, 2 and 1 respectively and are mapped into a data cube
- the class label having the highest aggregated membership value is assigned to the pixel.

The process is shown in Fig. 5.

In the figure above it can be seen how the procedure described earlier works. The threshold unit thresholds adaptively the soft classification results and keeps only the three highest probability values. Afterward the weighting unit weights each of these triplets with 3, 2 and 1 from highest to lowest respectively and maps the values into a data cube. The data cube aggregates the weighted probability values for each class. The weights applied are only empirical and are used to emphasize the most appropriate classifications at each simple processing chain. At the last step the final map is generated by rule classification of the mapped data cube by selecting the highest possible aggregated weighted probability value and labels the pixel accordingly.

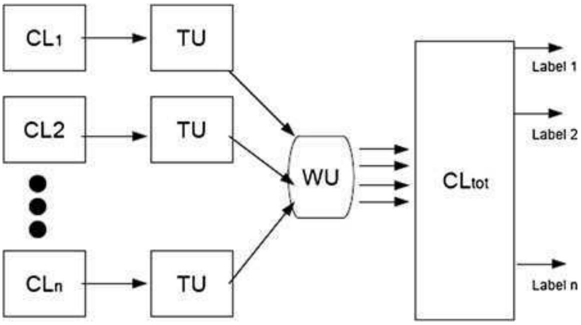


Fig. 5 The representation of class probability based ensemble classifier structure, where $\mathrm{CL_n}$ refers to different classifiers the TU is a threshold unit WU is the weighting unit and $\mathrm{CL_{tot}}$ refers to the final classifier

3.3 Class-Dependent Neural Networks Ensemble

Feed-forward neural networks are well established as one of the standard methods for classification applications [20, 21]. Class-dependent neural networks are simply the feed forward neural networks that can ideally map the input values of the class of interest to a value of 1 and the values of the rest of the classes to a value of 0 [22]. A fusion of these class-dependent networks may be done using a second level neural network which compares the pattern of the result of class-dependent networks and produces a final probability of all the classes. Fig. 6 shows the architecture of ensemble of class-dependent neural networks. $\mathrm{NN_1}$, $\mathrm{NN_2}$, …, $\mathrm{NN_K}$ are the class-dependent networks for K classes. The second level network $\mathrm{NN_{tot}}$ uses the output of the class-dependent networks and maps the patterns to final probabilities (m_1, m_2, …, m_K). To ensure a proper training of the networks, Kalman filter training is first used and the weights are refined using the scaled-conjugate gradients algorithm [23]. To avoid the problem of over fitting, only two layered neural networks are used and the number of neurons in the hidden layer of the networks depends on the number of input nodes. The number of hidden neurons is one more than the number of input nodes.

There are certain advantages using this architecture. It uses the fact that we do not require the entire feature space to distinguish one class of interest from the other classes. Every class has maximum separability with other classes in a different feature subspace. So, by dealing with the classes in a parallel manner with different inputs to each of the class-dependent networks, we may reduce the redundancy. As we are dealing with relatively smaller feature sub-spaces, it also accounts for the Hughes phenomenon which says that the classification accuracies are affected in a higher dimensional feature space when only a limited number of training samples are present [24].

While the mentioned advantages make the architecture attractive, it still has some problems when the classes are not fully separable and when there are not sufficient number of training samples. When the classes are not fully separable, the patterns produced by the class-dependent networks may vary a lot within the class and this further complicates the result of the second-level network leading to lower accuracies for the non-separable classes compared to the regular classifiers which

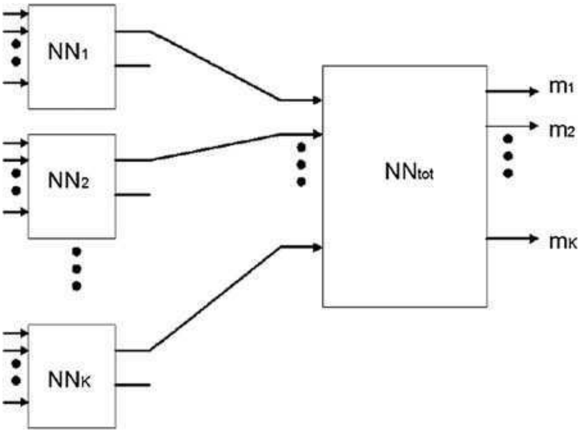


Fig. 6 The architecture of ensemble of class-dependent neural networks

deal with all the classes in the same feature space. So, it is very important that a proper feature sub-space is selected to attain an appropriate classification result.

The following options can be considered for instance, when identifying the features for every class.

- The univariate Jeffries-Matusita distance for every class combination in every band can identify the features in which the class combinations are highly separable [2, 25]. These features can be used to build the feature sub-space. In the cases where the classes are partially separable, more than one feature can be use for the classification.
- The decision boundary feature extraction (DBFE) method can be used to identify the projections in which the individual class combinations are most separable, and a feature space for every class can be built using the features extracted using DBFE over all the class combinations involving the class of interest.

Please note that although the architecture is presented as decision fusion architecture, it can be used as a regular classifier in the hierarchical decision tree classifier explained in Sect. 3.1.

4 Accuracy Assessment

For testing classification accuracy levels obtained by all the above mentioned approaches confusion matrices are used. The overall accuracy is suitable when the number of observations for each of the classes is not so different. This may not be possible in all the cases like the example used in this chapter. In this chapter, we use the accuracies of the individual classes separately to analyze which of the classes are classified well. Another reason to consider the individual classes separately is to determine the relative significance of classification accuracies of different classification results as will be discussed in Sect. 4.1.

4.1 Comparison of Classification Results

A simple comparison of the accuracies of each class when comparing the results of two different classifiers is not the right way to establish the superiority of one classifier over other. A very important consideration is the fact that accuracy assessment is a statistical method which demands a reasonable number of observations to establish the significance. The results of accuracy assessment are not reliable with a small number of observations. Following [13], this section explains how to compare two classifiers when the number of test observations is different for different classes.

The estimated variance of the misclassification rate can be derived as [13]

$$\hat{\sigma}^2 = \frac{\hat{\theta}\left(1 - \hat{\theta}\right)}{n} \tag{2}$$

where n is the number of observations and $\hat{\theta}$ is the estimated misclassification rate.

The equation of estimated variance of misclassification rate has an interesting outcome. We can determine the number of observations required to claim that two values resulted from the classifications using two different classifiers are significantly different. For example, consider a misclassification rate of $\theta = 0.1$, i.e., 90% accuracy. If we want to claim that 95 and 90% accuracies are significantly different then the standard deviations should not be greater than 0.025 i.e., 2.5%. Using Eq. 2, n is equal to 144.

It has to be noted that we need 144 samples in the above case just to say that the accuracies are significantly different but then even more samples are required to actually establish that the difference of 5% is valid. As the differences decrease, even more samples would be required.

$$\begin{aligned} Ac_1 &= \frac{(Ac + (4/n)) - \sqrt{(Ac + (4/n))^2 - ((1 + (4/n)) \cdot Ac^2)}}{(1 + (4/n))} \\ Ac_2 &= Ac + \sqrt{\frac{4}{n} \cdot Ac \cdot (1 - Ac)} \end{aligned} \tag{3}$$

On the other hand, if we have to compare two classifiers using the same number of test observations, Eq. 3 may be used to derive the lower and upper limits of range of values which are not significantly different with respect to the given classifier. Given a class with n test observations and the accuracy of the class using a classifier CL_1 as Ac, we can calculate the values Ac_1, Ac_2 as the lower and upper limits of the range of values around Ac in which the results of the other classifiers are statistically not significantly different from the result of CL_1. The result of a classifier CL_2 can be regarded as significantly different to that of the result of CL_1, if the accuracy of CL_2 is not in the range defined by $[Ac_1, Ac_2]$.

5 Results

To provide significant results, a huge number of processing chains were tested, but only the most relevant chains will be mentioned here. For the study, we used combinations of numerous feature selection/feature reduction methods and classification methods for preliminary assessment of classification performance. The applied methods are summarised in Tables 5 and 6.

We will first summarize the results of the classification algorithms and then compare them with the results of the proposed multi-classifier fusion algorithms. Even if the class dependent networks architecture is an ensemble classifier and should be seen as a fusion algorithm, it is still used as a regular classifier as it is also seen as a modification of feed forward neural networks. There are currently ongoing attempts to use a similar architecture using different types of classifiers instead of just neural networks.

5.1 Results of Various Tested Classifiers

The results of the classifiers are shown in Figs. 7, 8, 9, 10, and 11 using different inputs and will be first summarized. The results of class dependent networks architecture will be then provided with a comparison to the regular classifiers. We then provide the results of the proposed multi-classifier fusion algorithms and check if there is any significant improvement in the results after the fusion. As mentioned before, different classifiers with different inputs will produce very different classification results. The differences can be easily observed from the results provided in this section. This not only gives some idea on the differences among different processing chain performances, but also gives the opportunity for direct comparison with the introduced decision fusion methodology results.

Just by visual inspection of the different classification maps obtained by single stage classifications of the data, it can be seen that the methodologies for classification behave differently and it is also visible that the results of some of the

Table 5 The applied data dimensionality reduction techniques

1	Feature selection based on transformed divergence index (FS)
2	Principal components analysis (PCA)
3	Minimum noise fraction transformation (MNF)
4	Decision boundary feature extraction (DBFE)
5	Independent Components Analysis (ICA)

Table 6 The applied decision rules for data classification

1	Minimum distance classifier (MD)
2	Spectral Angle Mapper (SAM)
3	Mahalanobis distance classifier (MAH)
4	Maximum likelihood classifier (ML)
5	Support vector machine (radial basis function kernel) (SVM)
6	Neural network classification with back propagation algorithm (NNBP)
7	Neural Network using mixed Kalman filter and scale conjugate gradient learning (NNK+S)
8	Class dependent neural network

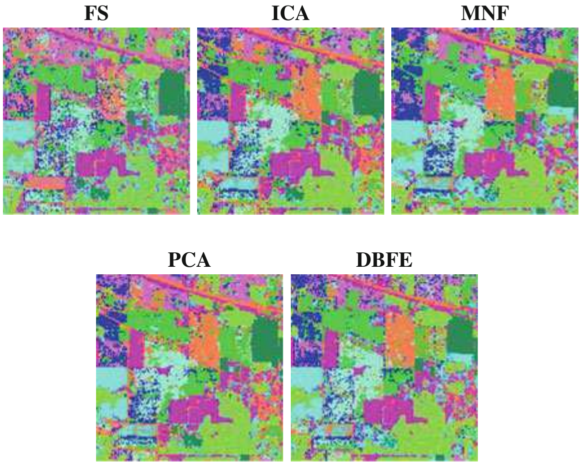


Fig. 7 Classification images obtained by Mahalanobis distance classifier using different data dimensionality reduction techniques

inputs (transformed data) have similar tendencies on the final map. For instance, it can be seen that using any of the methodologies, if the input is MNF, ICA or PCA transformed layer there are more pixels that are labelled to belong to class 2 (corn no tillage). Also, it is worth mentioning that comparing the images, only MNF and ICA images enabled any of the classifier to detect homogeneous areas of class 6 (Soybean no tillage) and most of the other data dimensionality reduction techniques resulted in only sparse individual pixels labelled as the same class. Beside the classification maps obtained, the accompanying accuracy tables are also very important in order to understand the differences among different processing chains. The accuracy tables are quite large amount of tabular data. Therefore, we do not present them entirely but we show the user and the producer accuracy values on a per class basis. The overall classification accuracy values are also given in the tables (Tables 7, 8, 9, 10, and 11) for the processing chains.

In general, it can be seen that different classifiers result different accuracy values both in terms of overall accuracy and in terms of accuracy measured on a per class basis as can be expected. The classification using the FS image did not produce better accuracies with any of the classifiers that were used, while ICA, MNF, PCA and DBFE in most cases produced satisfactory results. The more advanced

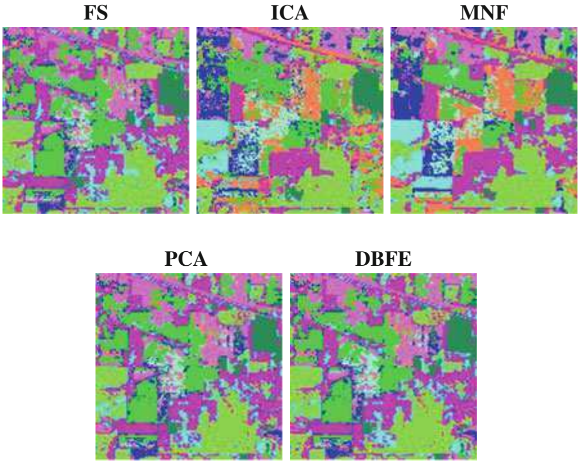


Fig. 8 Classification images obtained by minimum distance classifier using different data dimensionality reduction techniques

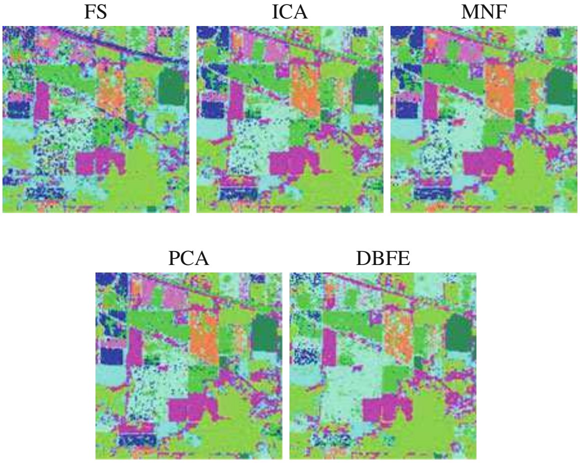


Fig. 9 Classification images obtained by maximum likelihood classifier using different data dimensionality reduction techniques

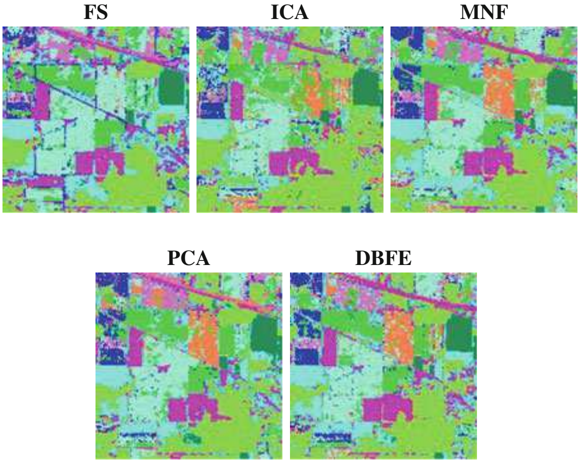


Fig. 10 Classification images obtained by SVM classifier using different data dimensionality reduction techniques

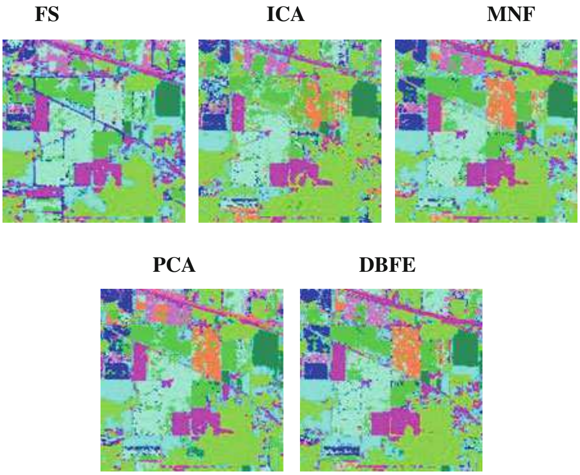


Fig. 11 Classification images obtained by neural network classifier using different data dimensionality reduction techniques

Table 7 Accuracy values for the Mahalanobis classifier

Class	FS		ICA		MNF		PCA		DBFE	
	P	U	P	U	P	U	P	U	P	U
1	66.74	56.33	77.55	70.03	80.89	71.78	77.41	67.72	74.48	68.11
2	21.58	19.89	50.72	40.52	64.39	43.69	44.96	39.39	43.53	42.81
3	69.01	60.49	66	83.25	82.29	89.69	65.79	82.78	76.46	85.01
4	93.57	83.61	94.78	82.42	94.24	89.8	94.91	83.31	95.45	89.24
5	99.59	96.63	99.59	98.58	99.59	100	99.59	98.38	99.39	99.59
6	46.8	40.77	79.55	64.33	80.37	64.56	79.13	63.57	79.44	61.67
7	41.33	61.15	50.69	79.43	49.11	80.05	48.82	77.94	54.21	80.6
8	62.7	38.85	73.45	44.43	78.18	58.54	76.87	43.78	71.34	40.78
9	80.29	97.65	91.04	99.92	95.36	100	91.42	99.92	93.66	99.84
OA (%)	59.5292		71.7817		74.9171		70.9898		72.4131	

Table 8 Accuracy values for the minimum distance classifier

Class	FS		ICA		MNF		PCA		DBFE	
	P	U	P	U	P	U	P	U	P	U
1	58.65	30.35	77.75	70.88	72.45	75.89	58.58	31.37	58.86	31.56
2	16.79	27.13	53.96	36.41	51.32	36.03	18.11	25.04	16.31	23.78
3	3.22	6.56	66	84.54	66	86.09	3.22	8.79	2.01	6.71
4	77.11	62.07	96.12	82.72	98.39	79.55	81.53	69.76	82.86	68.93
5	97.96	74.49	99.39	100	99.59	100	99.39	77.39	99.18	78.35
6	8.16	26.33	70.76	48.04	76.55	43.33	10.54	35.29	12.29	42.5
7	12.72	62.18	36.35	78.55	29.13	69.54	14.91	63.12	16.53	67.44
8	37.46	12.06	68.4	41.14	71.82	44.59	48.05	14.38	46.42	13.9
9	93.97	79.43	93.04	99.92	97.45	99.84	87.56	77.71	90.42	78.16
OA (%)	41.6372		67.4478		66.1209		42.8036		43.6169	

Table 9 Accuracy values for the maximum likelihood classifier

Class	FS		ICA		MNF		PCA		DBFE	
	P	U	P	U	P	U	P	U	P	U
1	70.92	62.13	80.68	78.49	88.15	82.51	83.33	77.2	83.54	79.5
2	58.75	51.91	79.14	74.83	83.33	72.7	79.14	75.86	38.25	89.36
3	85.71	86.06	94.57	95.53	97.18	93.79	95.37	94.42	90.34	95.74
4	96.79	97.31	98.13	96.19	98.8	97.36	97.59	96.56	97.59	96.3
5	99.39	99.79	98.98	100	99.59	100	98.98	100	99.18	100
6	61.05	55.23	77.69	78.74	85.54	80	81.3	79.1	64.88	84.98
7	55.71	68.78	79.9	81.09	76.18	87.16	75.73	83.55	90.96	62.81
8	67.92	64.95	79.8	86.27	86.48	86.76	85.67	80.18	20.68	94.07
9	98.69	96.16	99.54	99.38	99.38	99.46	99.69	99.38	99.61	97.5
OA (%)	72.7876		85.6715		87.6619		85.7571		79.9251	

Table 10 Accuracy values for the SVM classifier

Class	FS		ICA		MNF		PCA		DBFE	
	P	U	P	U	P	U	P	U	P	U
1	62.9	61.49	81.87	77.03	85.5	81.62	77.62	74.9	84.87	82.4
2	33.09	61.33	61.39	77.81	64.27	87.01	59.71	81.77	70.86	87.82
3	93.76	79.66	98.19	87.3	97.99	87.91	97.18	87.34	97.18	87.18
4	96.12	95.23	97.19	96.93	97.59	98.91	95.72	97.01	98.39	97.74
5	99.39	98.98	99.39	100	99.59	100	99.59	100	99.39	100
6	6.82	45.52	66.22	77.6	77.48	80.99	69.21	78.73	77.27	79.24
7	80.92	53.78	83.02	75.47	84.52	79.35	83.47	75.38	84.32	79.28
8	44.63	55.02	73.13	81.64	85.83	87.83	75.57	77.33	73.94	83.92
9	95.6	99.6	98.61	99.84	99.54	99.54	99.15	99.53	99.54	99.61
OA (%)	68.7212		83.4778		86.8486		83.1782		86.4955	

Table 11 Accuracies of NN classifier (using Kalman filter and scaled gradient training)

Class	FS		ICA		MNF		PCA		DBFE	
	P	U	P	U	P	U	P	U	P	U
1	64.99	64.54	79.64	78.33	88.91	83.17	78.45	72.49	88.42	83.2
2	64.39	62.08	66.43	76.52	80.1	84.77	65.47	73.39	74.7	89.51
3	90.54	83.03	94.16	87.15	93.36	87.71	91.95	87.72	93.96	88.28
4	96.65	93.89	96.52	97.3	96.12	95.23	92.9	93.91	95.85	95.21
5	98.98	98.17	99.39	99.79	99.39	98.98	100	98	99.59	98.58
6	61.26	66.04	81.3	75.31	86.16	83.99	73.97	73.29	81.61	80.04
7	73.26	70.71	82.17	82.01	83.63	85.36	80.23	79.94	86.43	84.61
8	66.61	83.47	79.97	83.93	78.5	90.26	69.87	79.15	78.34	88.42
9	96.52	97.05	98.76	98.69	99.3	98.47	97.84	97.84	98.92	98.61
OA (%)	76.88		85.13		88.56		82.42		88.23	

techniques such as SVM and NN classifiers are performing better than the other classifiers with an exception of ML classification on MNF image, which resulted in a good value in terms of accuracy. We used this classification as a reference for assessing the performance of different classifiers. The reason for this is that ML is the most commonly used methodology by users as it does not require much computational power or experience.

5.2 Results of Class Dependent Neural Networks

Based on the results of neural network classification from the previous section, only MNF, ICA and DBFE inputs are used as they give better results compared to the other inputs. When MNF and ICA subsets are used, all the available features are used at the input layer of all the class dependent networks. This is to ensure the maximum separability as it is difficult to find feature subspaces in the image data which is already a subset of MNF or ICA transformation. To use different inputs for different classes, we used the DBFE algorithm to find features for all the class combinations and those features identified for every class are used at the input layer of the corresponding class dependent network. In Tables 12 and 13, DBFE corresponds to the case when all the DBFE features are used at the input layer of the class dependent networks and DBFE-CD corresponds to using different features extracted using the DBFE algorithm for the individual class combinations.

Also, as the classes have different number of training samples, a classification is also done by making copies of the training samples of every class to have a similar number to that of the maximum available training samples for any class. This we believe removes the bias to the classes having more samples. But at the same time, as we are not adding any extra information, the overall classification is still not better when we only have limited samples to generalize the entire distribution of the class or when the classes are not completely separable. This is the case with the

Table 12 Results of class dependent networks architecture

Class	ICA		MNF		DBFE		DBFE-CD	
	P	U	P	U	P	U	P	U
1	75.03	74.57	86.05	80.87	78.87	85.1	82.15	41
2	52.52	74.87	60.79	87.72	62.35	83.6	59.71	86.16
3	91.55	77.65	85.31	95.28	84.1	88.75	83.1	91.17
4	95.58	93.21	93.04	95.86	95.05	96.47	95.58	93.95
5	99.59	99.59	99.59	93.65	99.59	99.8	98.77	97.97
6	76.45	72.91	75.1	68.78	68.7	79.07	71.69	82.13
7	76.18	70.04	79.74	74.18	85.74	68.39	84.4	71.07
8	60.26	75.05	69.22	82.52	66.12	82.69	64.66	78.93
9	98.07	98.83	99.23	96.83	97.99	99.61	99.3	97.57
OA (%)	79.5		82.95		82.63		82.88	

Table 13 Results of class dependent networks architecture after duplicating the training samples for classes having fewer samples

Class	ICA		MNF		DBFE		DBFE-CD	
	P	U	P	U	P	U	P	U
1	75.45	71.28	85.36	74.95	81.45	79.08	80.13	80.86
2	66.79	50.91	75.9	61.82	66.07	75.79	70.26	72.52
3	91.35	75.67	89.34	91.55	82.49	85.42	82.7	85.09
4	93.04	93.16	94.91	93.04	95.18	97.4	94.51	96.45
5	97.34	98.14	99.59	99.8	98.57	98.77	97.55	98.96
6	85.64	63.23	89.15	61.86	80.68	68.39	78.2	69.64
7	52.96	80.83	56.69	86.73	76.13	74.33	72.93	72.46
8	75.9	67.44	81.11	77.45	65.8	81.62	63.19	72.12
9	96.29	97.12	98.76	98.16	98.3	99.53	99.07	97.86
OA (%)	76.10		80.64		81.95		80.86	

present dataset. The class combinations 1–2, 1–6, 1–7, 2–6, 2–7, 3–4, 6–7, 7–8 are not completely separable with most of the problem with classes 2, 6, 7, 8, and there are limited samples for classes 2, 3, 5 and 8. So, we generally expect lower accuracies using the class dependent networks architecture compared to a feed forward neural network classifying all the classes at a time unlike the example in [14] where the architecture performed better due to better separability of the classes and reasonable size of the training data.

By comparing the results, it can be easily observed that the bias towards class 7 is reduced and classes 2, 6, 8 have better accuracies. The overall accuracy is misleading here because it is biased towards classes having more observations. However, as explained before, the architecture does not outperform the regular neural network classifier due to the poor separability of the classes but the results can still be used in the other decision fusion methods.

5.3 Results of Decision Fusion Using Hierarchical Tree Structure

The performance of the hierarchical ensemble was assessed by comparing the performance of classification on a per class basis with the traditional single stage classification based processing chains. Here we also use the results of the best performing single stage classifier based processing chain (ML algorithm on MNF image) to see whether there are any improvements by applying an ensemble.

As shown in Table 14, the hierarchical tree structured ensemble uses different inputs and classification rules at different nodes adapting to the data properties. In Fig. 12 the visual differences among the best single stage methodology is visualized.

Table 14 The hierarchical tree structure for the Indian Pines dataset

Node #	Class to label	Input	Decision rule
1	5	MNF	ML
2	9	DBFE	SVM
3	4	MNF	SVM
4	3	MNF	ML
5	8	MNF	SVM
6	1	MNF	NN
7	7	DBFE	CDNN
8	6/2	ICA	NN

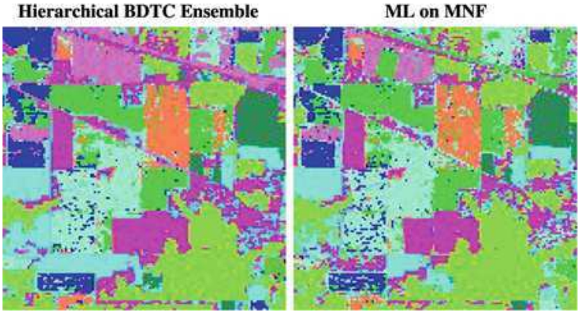


Fig. 12 Classification image of the hierarchical tree ensemble and the best single stage processing chain

For comparison we also tested the methodology against a known method of ensemble classification the majority voting approach [2]. However it is important to mention that while the majority voting approach by its nature requires the classification to be done for the full scene using all methodologies as opposed to the proposed optimization methodology for the hierarchical ensemble where only a small subset of the image must be processed. This is because majority voting approach is using the final labels of different classifications on a per pixel basis for the whole scene for labelling (Fig. 13).

As seen in the figures above, just by visual interpretation it is difficult to identify differences among the different classification results. However the accuracy levels per class given in Table 15 are better reflecting the differences among different approaches.

The hierarchical ensemble classification (see Table 15) not only outperforms the best single stage classifier based processing chain but also gains higher accuracy levels than a majority voting procedure. At the same time the approach requires less computational power than carrying out majority voting and also only the relevant inputs and methodologies are selected to be included within the classification procedure.

There are some limitations of the methodology too as mentioned earlier in this chapter such as the inability to significantly boost the accuracy level and the requirement for good quality training and validation samples for the design procedure. The former limitation is simply based on the fact that it is not a novel classification methodology but only a way to create ensemble with a reasonably

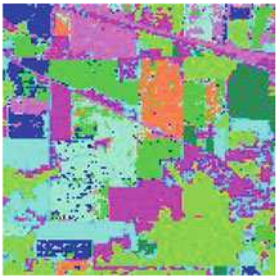


Fig. 13 Classification image obtained by the majority voting approach

Table 15 The accuracy levels per class of hierarchical ensemble, best single stage classification and majority voting procedures

Class	BDTC		ML_MNF		Majority vote	
	P	U	P	U	P	U
1	92.19	82.73	88.15	82.51	91	77.59
2	80.1	95.43	83.33	72.7	76.62	89.25
3	98.19	92.6	97.18	93.79	95.37	95.18
4	99.2	98.15	98.8	97.36	99.6	95.63
5	99.59	100	99.59	100	99.59	99.59
6	81.1	86.93	85.54	80	88.12	85.81
7	85.78	86.66	76.18	87.16	77.35	89
8	91.04	87.76	86.48	86.76	90.39	73.71
9	99.61	99.54	99.38	99.46	99.54	99.84
OA (%)	90.4839		87.6619		88.3253	

optimal design and hence the absolute performance of the system is determined by the classifiers that are used to create the ensemble structure. The latter limitation is also valid for any other classification. While using the top-down hierarchical design approach the ensemble classifier can be misled and more significant errors can be introduced than when a single stage data splitting is carried out.

5.4 Results of Hierarchical Tree Coupled with Probability Labels

The results of the approach is shown in Fig. 14 to visually compare with the best performing single stage classifier based processing chain (ML on MNF input) and also accuracy levels on a per class basis are provided in tabular format.

The visual inspection of the classification maps shows that the ensemble classifier that uses probability labels resulted in a smoother classification image containing less individual pixel errors compared to the single stage classification approach. Also the field boundaries are more recognisable without less pixel noise (Table 16).

The methodology provided the highest overall accuracy level among the different classifiers tested on the image. The classification image looks more realistic

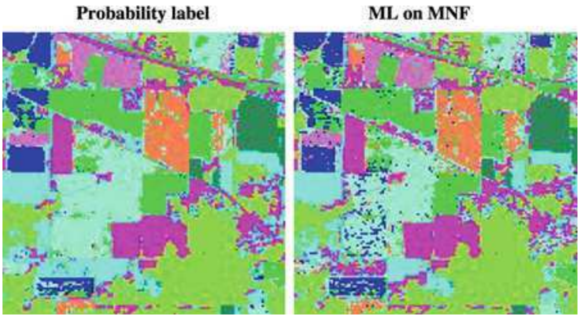


Fig. 14 The classification map obtained by using the class probability label approach compared with the best performing single stage classifier

Table 16 The accuracy levels per class of hierarchical ensemble using probability labels

Probability label		
Class	P	U
1	90.38	84.98
2	82.35	94.88
3	96.98	92.15
4	99.06	96.99
5	99.59	99.39
6	83.97	83.03
7	88.61	87.41
8	83.63	94.63
9	99.61	99.46
OA (%)	90.8964	

containing less individual pixel errors. However, as most of the classifiers are producing relatively high accuracy values, it is more difficult to improve it.

5.5 The Assessment of Significance of the Accuracy Values

The assessment was done according to the methodology described in Sect. 4 the accuracy assessment section of the chapter. The results are shown in tabular format where the sign “+” and sign “−” shows the significantly higher and the significantly lower accuracy levels against the best achieved accuracy level per classes respectively.

As can be seen in Table 17, the probability based ensemble outperformed the maximum likelihood approach on the MNF input band in case of the class 7 (Soybean no tillage) and class 1(Corn no tillage) and has similar accuracies for all the other classes. However, as mentioned before, class 7 is not separable with most of the other classes and furthermore had the highest number of samples in the scene. In comparison with the BDTC approach, the probability based ensemble performs significantly better in the case of classes 6 and 7 but has lower accuracies

Table 17 The lower and upper limit of significance calculated for the best performing classifier for producer and user accuracy levels per classes

Class	Lower limit, P	Probability, P	Upper limit, P	ML on MNF, P	CDFNN, P	BDTC, P
1	88.7	90.38	91.93	–	–	+
2	79.46	82.35	84.9		–	
3	94.79	96.98	98.36		–	
4	98.05	99.06	99.77			
5	98.49	99.59	100.17			
6	81.38	83.97	86.25		–	–
7	87.23	88.61	89.85	–	–	–
8	80	83.63	86.24		–	+
9	98.98	99.61	99.91			

in the case of classes 1 and 8. But on the whole, it is clear that the decision fusion methodologies are improving the results.

6 Conclusions

As it was shown within this chapter decision fusion is a relatively better technique for hyperspectral data processing. The class based selection of features and the actual classification that were carried out using the decision fusion methods enabled to carry out a better quality interpretation of the hyperspectral dataset as opposed to processing chains using single stage classification algorithms. The main aim of the experiments was to study the possibility of fusing decisions while classifying an image and this is successfully realized.

In this study, we introduced decision fusion methodologies that are relatively simple and are capable of improving the quality of hyperspectral data processing aimed at generic vegetation mapping applications. The challenge of taking into account the vegetation specific properties of spectral signatures were addressed by the flexible approaches that enable the class labelling procedure to be done using the most appropriate features on a per class basis. We also introduced the class dependent neural network algorithm where both the training and the actual classification are carried out using separated features that are specific for the classes that are being detected. Even if it did not provide better results for the dataset used in this study due to the issue of poor separability of classes, it can be seen as a fair methodology when the classes can be separated using a non-linear decision boundary.

A methodology was introduced that is capable of designing a hierarchical tree structured ensemble for decision fusion of multiple classifiers. A methodology was also proposed for fusing decisions by using class membership values. Although the proposed methods enable users to improve their data classification, there are certain limitations that require further research in the area. The usage of class membership values produced by different processing chains has to be further studied instead of using the current empirical weighting of the three best results. This can be very useful for end users who already have some processing chain elements implemented or as a part of a commercial software and want to combine them. Both the BDTC and the probability value based ensemble methodology are simple enough to be easily adopted. This is particularly useful in cases when there are classes present on the scene that cannot be mapped using the same processing chain because there is no processing chain that is suitable for both the classes evenly.

Also we see many areas of class-dependent neural network classifier that could be investigated to improve the classification of hyperspectral datasets even if the classes are not completely separable. This can be done for instance by using different configurations for the individual class dependent networks to better adapt to the data properties and the used inputs. Regarding the hierarchical ensemble tree structure, a simple way was introduced for selecting the optimal design for the structure. It would be worth investigating to select multiple processing chains at every node instead of basing the decision on the best processing chain at that node. Also, another significant improvement would be to use the spatial information (e.g., class labels or class membership values of the neighbouring pixels) for decision fusion of multiple classifiers.

Acknowledgments The authors would like to acknowledge that this research was carried out within the Marie-Curie founded Hyper-I-Net Research Training Network. We also would like to thank Antonio J. Plaza and Alberto Villa the help and support which they provided.

References

1. Gamba, P., Plaza, A., Benediktsson, A.J., Chanussot, J.: European perspectives in hyperspectral data analysis. In: Proceedings of 2007 IEEE Geoscience and Remote Sensing Symposium (IGARSS 2007), pp. 4794–4797 (2007)
2. Richards, J.A.: Analysis of remotely sensed data: the formative decades and the future. IEEE Trans. Geosci. Remote Sens. **43**, 422–432 (2005)
3. Bakos, K.L., Gamba, P.: Potential of hyperspectral remote sensing for vegetation mapping in high mountain ecosystems. In: Proceedings of 6th EarSEL SIG IS Workshop 2009, Unformatted CD-ROM (2009)
4. Bruzzone, L., Prieto, D.F., Serpico, S.B.: A neural statistical approach to multitemporal and multisensor and multisource remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **37**, 1350–1359 (1999)
5. Vapnik, V., Vashist, A.: A new learning paradigm: learning using privileged information. Neural Netw. **22**, 544–557 (2009)
6. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag. **2**, 559–572 (1901)
7. Green, A.A., Berman, M., Switzer, P., Craig, M.D.: A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Geosci. Remote Sens. **26**, 65–74 (1988)
8. Richards, J.A.: Remote Sensing Digital Image Analysis. Springer-Verlag, Berlin (1999)
9. Tadjudin, S., Landgrebe, D.A.: Classification of high dimensional data with limited training samples. Technical Report TR-ECE 98-8, Purdue University (1998)

10. Ho, T.H., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier system. IEEE Trans. Pattern Anal. Mach. Intell. **16**, 66–75 (1994)
11. Kuncheva, L.I., Bezdek, J.C., Sutton, M.A.: On combining multiple classifiers by fuzzy templates. In: Proceedings of NAFIPS Conference EDS, pp. 193–197 (1998)
12. Woods, K., Kegelmeyer, W.P., Bowyer, K.: Combination of multiple classifiers using local accuracy estimates. IEEE Trans. Pattern Anal. Mach. Intell. **19**, 405–410 (1997)
13. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 226–239 (1998)
14. Liu, W.G., Gopal, S., Woodcock, C.E.: Uncertainty and confidence in land cover classification using a hybrid classifier approach. Photogramm. Eng. Remote Sens. **70**, 963–971 (2004)
15. Schwenk, H., Bengio, Y.: Boosting neural networks. Neural Comput. **12**, 1869–1887 (2000)
16. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)
17. Polikar, R.: Ensemble-based systems in decision making. IEEE Circuits and Systems Magazine **6**(3), 21–45 (2006)
18. Roli, F., Giacinto, G., Vernazza, G.: Methods for designing multiple classifier systems. In: Proceedings of MCS, pp. 78–87 (2001)
19. Gabrys, B., Dymitr, R.: Genetic algorithms in classifier fusion. Appl. Soft Comput. **6**, 337–347 (2006)
20. Sinha, A., Chen, H., Danu, D.G., Kirubarajan, T., Farooq, M.: Estimation and decision fusion: a survey. Neurocomputing **71**, 2650–2656 (2008)
21. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)
22. Canty, M.J.: Image Analysis, Classification and Change Detection in Remote Sensing, with Algorithms for ENVI/IDL. CRC Press, Boca Raton (2007)
23. Marpu, P.R., Gamba, P., Niemeyer, I.: Hyperspectral data classification using an ensemble of class-dependent neural networks. In: Proceedings of IEEE GRSS Workshop on Hyperspectral Image and Signal Processing Conference (WHISPERS), Unformatted CD-ROM (2009)
24. Hughes, G.F.: On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory **14**, 55–63 (1968)
25. Lillesand, T., Kiefer, R.: Remote Sensing and Image Processing. Wiley, New York (1978)

A Review of Kernel Methods in Remote Sensing Data Analysis

Luis Gómez-Chova, Jordi Muñoz-Marí, Valero Laparra, Jesús Malo-López and Gustavo Camps-Valls

Abstract Kernel methods have proven effective in the analysis of images of the Earth acquired by airborne and satellite sensors. Kernel methods provide a consistent and well-founded theoretical framework for developing nonlinear techniques and have useful properties when dealing with low number of (potentially high dimensional) training samples, the presence of heterogenous multimodalities, and different noise sources in the data. These properties are particularly appropriate for remote sensing data analysis. In fact, kernel methods have improved results of parametric linear methods and neural networks in applications such as natural resource control, detection and monitoring of anthropic infrastructures, agriculture inventorying, disaster prevention and damage assessment, anomaly and target detection, biophysical parameter estimation, band selection, and feature extraction. This chapter provides a survey of applications and recent theoretical developments of kernel methods in the context of remote sensing data analysis. The specific methods developed in the fields of supervised classification, semisupervised classification, target detection, model inversion, and nonlinear feature extraction are revised both theoretically and through experimental (illustrative) examples. The emergent fields of transfer, active, and structured learning, along with efficient parallel implementations of kernel machines, are also revised.

Keywords Remote Sensing · Kernel Methods · Support Vector Machines (SVM) · Supervised and Semisupervised Classification · Target Detection · Biophysical ParameterEstimation · Nonlinear Feature Extraction

L. Gómez-Chova (✉), J. Muñoz-Marí, V. Laparra, J. Malo-López and G. Camps-Valls
Image Processing Laboratory, Universitat de València, Catedrático José Beltrán, 2.
Paterna, 46980, Valencia, Spain
e-mail: luis.gomez-chova@uv.es
URL: http://www.uv.es/chovago

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_10,

1 Introduction

Remotely sensed images allow Earth Observation with unprecedented accuracy. New satellite sensors acquire images with high spectral and spatial resolution, and the revisiting time is constantly reduced. Processing data is becoming more complex in such situations and many problems can be tackled with recent machine learning tools. One of the most critical applications is that of image classification, but also model inversion and feature extraction are relevant in the field. This chapter will focus on these important problems that are subsequently outlined.

1.1 Classification with Kernels

The characteristics of the acquired images allow the characterization, identification, and classification of the land covers [1]. However, traditional classifiers such as Gaussian maximum likelihood or artificial neural networks are affected by the high input sample dimension, tend to overfit data in the presence of noise, or perform poorly when a low number of training samples are available [2, 3]. In the last few years, the use of support vector machines (SVMs) [4, 5] for remote sensing image classification has been paid attention basically because the method integrates in the same classification procedure (1) a *feature extraction* step, as samples are mapped to a higher dimensional space where a simpler (linear) classification is performed, becoming nonlinear in the input space; (2) a *regularization* procedure by which model's complexity is efficiently controlled; and (3) the minimization of an upper bound of the generalization error, thus following the Structural Risk Minimization (SRM) principle. These theoretical properties, which will be reviewed in the next section, make the SVM in particular, and kernel methods in general, very attractive in the context of remote sensing image classification [6].

Another different concern is that a complete and representative training set is essential for a successful classification. In particular, it is noteworthy that few attention has been paid to the case of having an incomplete knowledge of the classes present in the investigated scene. This may be critical since, in many applications, acquiring ground truth information for all classes is very difficult, especially when complex and heterogeneous geographical areas are analyzed. In this chapter, we revise the one-class SVM for remotely-sensed image classification with incomplete training data. This method is a recent kernel-based development that only considers samples belonging to the class of interest in order to learn the underlying data class distribution. The method was originally introduced for anomaly detection [7], then analyzed for dealing with incomplete and unreliable training data [8], and recently reformulated for change detection [9].

Remote sensing image classification is hampered by both the number and quality of labeled training samples. In order to alleviate this problem, SVMs (or any other kernel-based classifier) should exploit the information contained in the abundant unlabeled samples along with the low number of labeled samples thus working under the semisupervised learning (SSL) paradigm [10]. In this chapter, we also review the SSL literature and provide some experimental evidence of the use of semisupervised approaches for classification in challenging remote sensing problems.

1.2 Model Inversion with Kernels

Remote sensing very often deals with inverting a forward model. To this aim, one has to produce an accurate and robust model able to predict physical, chemical, geological or atmospheric parameters from spectra, such as surface temperature, water vapour, ozone, etc. This has been an active research field in remote sensing for years, and kernel methods offer promising non-parametric semi-empirical solutions. Kernel developments have been published in the last years: support vector regression (SVR) methods have been used for parameter estimation [11–14], and a fully-constrained kernel least squares (FC-KLS) for abundance estimation [15]. Also, under a Bayesian perspective, other forms of kernel regression have been applied, such as the relevance vector machine (RVM) [16] or the Gaussian process (GP) regression [17, 18].

1.3 Feature Extraction with Kernels

Recently, some attention has been paid to develop kernel-based feature extraction methods for remote sensing data processing. The main interest is to extract a reduced number of (nonlinear) features with high expressive power for either classification or regression. Particular applications to remote sensing are the Kernel Principal Component Analysis (KPCA) [5] and the Kernel Partial Least Squares (KPLS) [19].

The rest of this chapter is outlined as follows. Section 2 presents a brief introduction to kernel methods, fixes notation, and reviews the basic properties. Section 3 is devoted to review the classification setting, under the paradigms of supervised, semisupervised, and one-class classification. Section 4 presents the advances in kernel methods for regression and model inversion. Section 5 reviews the field of nonlinear feature extraction with kernels. Section 6 reviews the recent developments and foresees the future trends in kernel machines for remote sensing data analysis. Section 7 concludes the chapter with some final remarks.

2 Introduction to Kernel Methods

This section includes a brief introduction to kernel methods. After setting the scenario and fixing the most common notation, we give the main properties of kernel methods. We also pay attention to kernel methods development by means of particular properties drawn from linear algebra and functional analysis [20, 21].

2.1 Measuring Similarity with Kernels

Kernel methods rely on the notion of similarity between examples. Let us define a set of empirical data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, where $\mathbf{x}_i$ are the *inputs* taken from $\mathcal{X}$ and $y_i \in \mathcal{Y}$ are called the *outputs*. Learning means using these data pairs to predict well on test examples $\mathbf{x} \in \mathcal{X}$. To develop machines that generalize well, kernel methods try to exploit the structure of the data and thus define a similarity between pairs of samples.

Since $\mathcal{X}$ may not have a proper notion of similarity, examples are mapped to a (dot product) space $\mathcal{H}$, using a mapping $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}, \mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x})$. The similarity between the elements in $\mathcal{H}$ can now be measured using its associated dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Here, we define a function that computes that similarity, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that $(\mathbf{x}, \mathbf{x}') \mapsto K(\mathbf{x}, \mathbf{x}')$. This function, called *kernel*, is required to satisfy:

$$K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathcal{H}}. \tag{1}$$

The mapping $\boldsymbol{\phi}$ is its *feature map*, and the space $\mathcal{H}$ its *feature space*.

2.2 Positive Definite Kernels

The class of kernels that can be written in the form of (1) coincides with the class of positive definite kernels.

Definition 1 *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{H}$and a feature map $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}$ such that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have $K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathcal{H}}$.*

In practice, a real symmetric $n \times n$ matrix $\mathbf{K}$, whose entries are $K(\mathbf{x}_i, \mathbf{x}_j)$ or simply K_{ij}, is called *positive definite* if for all $c_1, \ldots, c_n \in \mathbb{R}, \sum_{i,j=1}^{n} c_i c_j K_{ij} \geq 0$. Note that a positive definite kernel is equivalent to a positive definite Gram matrix in the *feature space*.

Therefore, algorithms operating on the data only in terms of dot products can be used with any positive definite kernel by simply replacing $\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathcal{H}}$ with kernel evaluations $K(\mathbf{x}, \mathbf{x}')$, a technique also known as the *kernel trick* [5]. Another direct consequence is that, for a positive definite kernel, one does not need to know the explicit form of the feature map since it is implicitly defined through the kernel.

2.3 Basic Operations with Kernels

We now review some basic properties with kernels. Note that, although the space $\mathcal{H}$ can be very high-dimensional, some basic operations can still be performed:

Translation — A translation in feature space can be written as the modified feature map $\tilde{\boldsymbol{\phi}}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}) + \Gamma$ with $\Gamma \in \mathcal{H}$. Then, the translated dot product for $\langle \tilde{\boldsymbol{\phi}}(\mathbf{x}), \tilde{\boldsymbol{\phi}}(\mathbf{x}') \rangle_{\mathcal{H}}$ can be computed if we restrict Γ to lie in the span of the functions $\{\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)\} \in \mathcal{H}$.

Centering — The previous translation allows us to center data $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$ in the *feature space*. The mean of the data in $\mathcal{H}$ is $\boldsymbol{\phi}_\mu = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\mathbf{x}_i)$ which is a linear combination of the span of functions and thus fulfills the requirement for Γ. One can center data in $\mathcal{H}$ by computing $\mathbf{K} \leftarrow \mathbf{HKH}$ where entries of $\mathbf{H}$ are $H_{ij} = \delta_{ij} - \frac{1}{n}$ and the Kronecker symbol $\delta_{i,j} = 1$ if $i = j$ and zero otherwise.

Subspace projections — Given two points Ψ and Γ in the feature space, the projection of Ψ onto the subspace spanned by Γ is $\Psi' = \frac{\langle \Gamma, \Psi \rangle_{\mathcal{H}}}{\|\Gamma\|_{\mathcal{H}}^2}\Gamma$. Therefore one can compute the projection Ψ' expressed solely in terms of kernel evaluations.

Computing distances — The kernel corresponds to a dot product in a Hilbert Space $\mathcal{H}$, and thus one can compute distances between mapped samples entirely in terms of kernel evaluations:

$$d(\mathbf{x}, \mathbf{x}') = \|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')\|_{\mathcal{H}} = \sqrt{K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}')}$$

Normalization — Exploiting the previous property, one can also normalize data in feature spaces:

$$K(\mathbf{x}, \mathbf{x}') \leftarrow \left\langle \frac{\boldsymbol{\phi}(\mathbf{x})}{\|\boldsymbol{\phi}(\mathbf{x})\|}, \frac{\boldsymbol{\phi}(\mathbf{x}')}{\|\boldsymbol{\phi}(\mathbf{x}')\|} \right\rangle = \frac{K(\mathbf{x}, \mathbf{x}')}{\sqrt{K(\mathbf{x}, \mathbf{x})K(\mathbf{x}', \mathbf{x}')}}$$

2.4 Standard Kernels

The bottleneck for any kernel method is the definition of a kernel mapping function $\boldsymbol{\phi}$ that accurately reflects the similarity among samples. However, not all kernel similarity functions are permitted. In fact, valid kernels are only those fulfilling Mercer's Theorem (roughly speaking, being positive definite similarity matrices) and the most common ones are the linear $K(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle$, the polynomial $K(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + 1)^d$, $d \in \mathbb{Z}^+$, and the Radial Basis Function (RBF) $K(\mathbf{x}, \mathbf{z}) = \exp\left(-\|\mathbf{x} - \mathbf{z}\|^2/2\sigma^2\right), \sigma \in \mathbb{R}^+$. Note that, by Taylor series expansion, the RBF kernel is a polynomial kernel with infinite degree. Thus the corresponding Hilbert space is infinite dimensional, which corresponds to a mapping into the space of smooth functions $\mathcal{C}^\infty$. The RBF kernel is also of practical convinience – stability and only one parameter to be tuned–, and it is the preferred kernel measure in standard applications.

2.5 Kernel Development

Taking advantage of some algebra and functional analysis properties [20, 21], one can derive very useful properties of kernels. Be K_1 and K_2 two positive definite kernels on $\mathcal{X} \times \mathcal{X}$, $\mathbf{A}$ a symmetric positive semidefinite matrix, $d(\cdot, \cdot)$ a metric fulfilling distance properties, f any function, and $\mu > 0$. Then, the following kernels are valid [5]:

$$K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + K_2(\mathbf{x}, \mathbf{x}') \tag{2}$$

$$K(\mathbf{x}, \mathbf{x}') = \mu K_1(\mathbf{x}, \mathbf{x}') \tag{3}$$

$$K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}') \tag{4}$$

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{A} \mathbf{x}' \tag{5}$$

$$K(\mathbf{x}, \mathbf{x}') = \exp(-d(\mathbf{x}, \mathbf{x}')) \tag{6}$$

$$K(\mathbf{x}, \mathbf{x}') = K(f(\mathbf{x}), f(\mathbf{x}')) \tag{7}$$

These basic properties give rise to the construction of refined similarity measures better fitted to the data characteristics. In remote sensing, one can sum dedicated kernels to spectral, contextual or even temporal information of pixels through (2). A scaling factor to each kernel can also be added (Eq. 3). Also, the (more appropriate) spectral angle distance between pixels is a valid kernel by (6). Recent advances for kernel development are:

Convex combinations — By exploiting (2) and (3), one can build kernels by linear combinations of kernels working on feature subsets:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} d_m K_m(\mathbf{x}, \mathbf{x}')$$

This field of research is known as multiple kernel learning (MKL) and different algorithms exist to optimize the weights and kernel parameters jointly. Note that this kernel offers some insight in the problem, since relevant features receive higher values of d_m, and the corresponding kernel parameters θ_m yield information about pairwise similarity scales.

Deforming kernels

The field of semisupervised kernel learning deals with techniques to modify the values of the training kernel including the information from the whole data distribution: K is either deformed through a graph distance matrix built with both labeled and unlabeled samples, or by means of kernels built from clustering solutions.

Generative kernels

Exploiting Eq. (7), one can construct kernels from probability distributions by defining $K(\mathbf{x},\mathbf{x}') = K(\mathbf{p},\mathbf{p}')$, where $\mathbf{p},\mathbf{p}'$ are defined on the space $\mathcal{X}$. This kind of kernels is known as *probability product kernels between distributions* and is defined as:

$$K(\mathbf{p},\mathbf{p}') = \langle \mathbf{p},\mathbf{p}' \rangle = \int_{\mathcal{X}} \mathbf{p}(\mathbf{x})\mathbf{p}'(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Joint input-output mappings

Typically, kernels are built on the input samples. Lately the framework of *structured output learning* deals with the definition of joint input-output kernels, $K((\mathbf{x},y),(\mathbf{x}',y'))$.

3 Kernel Methods in Remote Sensing Data Classification

Classification maps are the main product of remote sensing data analysis and, in the last years, kernel methods have demonstrated very good performance. The most successful kernel method are the support vector machines as extensively reported in [6]. SVMs have been applied to both multispectral [22, 23] and hyperspectral [6, 9, 24] data in a wide range of domains, including object recognition [25], land cover and multi-temporal classification [9, 26, 27], and urban monitoring [28].

3.1 Support Vector Machine

The support vector machine (SVM) is defined as follows. Notationally, given a labeled training data set $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^N$ and $y_i \in \{-1, +1\}$, and given a nonlinear mapping $\boldsymbol{\phi}(\cdot)$, the SVM method solves:

$$\min_{\mathbf{w},\xi_i,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \tag{8}$$

constrained to:

$$y_i(\langle\boldsymbol{\phi}(\mathbf{x}_i),\mathbf{w}\rangle + b) \geq 1 - \xi_i \quad \forall i = 1,\ldots,n \tag{9}$$

$$\xi_i \geq 0 \quad \forall i = 1,\ldots,n \tag{10}$$

where $\mathbf{w}$ and b define a linear classifier in the feature space, and ξ_i are positive slack variables enabling to deal with permitted errors (Fig. 1a). Appropriate choice of nonlinear mapping $\boldsymbol{\phi}$ guarantees that the transformed samples are more likely to be linearly separable in the (higher dimension) feature space. The regularization parameter C controls the generalization capability of the classifier, and it must be selected by the user. Primal problem (8) is solved using its dual problem counterpart [5], and the decision function for any test vector $\mathbf{x}_*$ is finally given by

$$f(\mathbf{x}_*) = sgn\left(\sum_{i=1}^{n} y_i\alpha_i K(\mathbf{x}_i,\mathbf{x}_*) + b\right) \tag{11}$$

where α_i are Lagrange multipliers corresponding to constraints in (9), being the support vectors (SVs) those training samples $\mathbf{x}_i$ with non-zero Lagrange multipliers $\alpha_i \neq 0; K(\mathbf{x}_i,\mathbf{x}_*)$ is an element of a kernel matrix $\mathbf{K}$ defined as in Eq. (1); and the bias term b is calculated by using the *unbounded* Lagrange multipliers as $b = 1/k\sum_{i=1}^{k}(y_i - \langle\boldsymbol{\phi}(\mathbf{x}_i),\mathbf{w}\rangle)$, where k is the number of *unbounded* Lagrange multipliers $(0\leqslant\alpha_i<C)$ and $\mathbf{w} = \sum_{i=1}^{n} y_i\alpha_i\boldsymbol{\phi}(\mathbf{x}_i)$ [5].

3.2 ν-Support Vector Machine

One interesting variation of the SVM is the ν-support vector machine (ν-SVM) introduced by Schölkopf et al. [29]. In the SVM formulation, the soft margin is controlled by parameter C, which may take any positive value. This makes difficult to adjust it when training the classifier. The idea of the ν-SVM is forcing the soft margin to lie in the range [0, 1]. This is carried out redefining the problem as

$$\min_{\mathbf{w},\xi_i,b,\rho}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i\right\} \tag{12}$$

subject to:

$$y_i(\langle\boldsymbol{\phi}(\mathbf{x}_i),\mathbf{w}\rangle + b) \geq \rho - \xi_i \quad \forall i = 1,\ldots,n \tag{13}$$

$$\rho \geq 0, \xi_i \geq 0 \quad \forall i = 1,\ldots,n \tag{14}$$

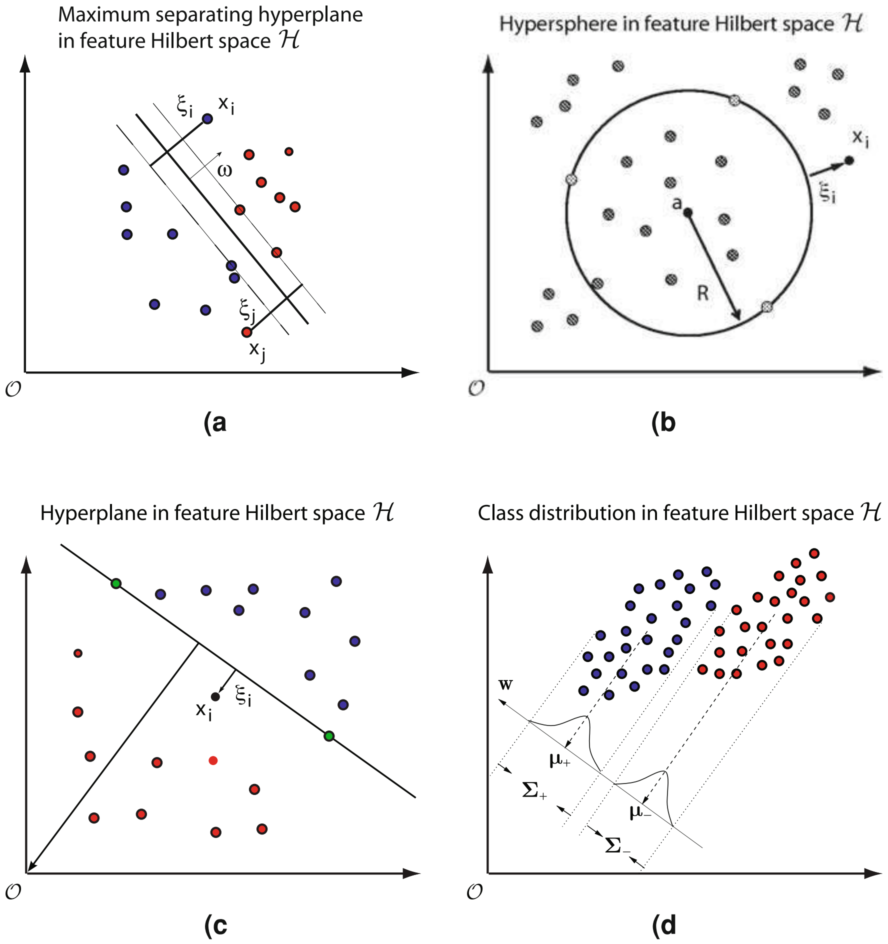

Maximum separating hyperplane in feature Hilbert space $\mathcal{H}$
ξ_i
x_i
ω
ξ_j
x_j
$\mathcal{O}$
(a
Hypersphere in feature Hilbert space $\mathcal{H}$
x_i
ξ_i
a
R
$\mathcal{O}$
(b
Hyperplane in feature Hilbert space $\mathcal{H}$
ξ_i
x_i
$\mathcal{O}$
(c
Class distribution in feature Hilbert space $\mathcal{H}$
w
μ_+
Σ_+
μ_-
Σ_-
$\mathcal{O}$
(d

Fig. 1 Illustration of kernel classifiers. **a** SVM: Linear decision hyperplanes in a nonlinearly transformed, feature space, where *slack* variables ξ_i are included to deal with errors. **b** SVDD: The hypersphere containing the target data is described by center (**a**) and radius R. Samples in the boundary and outside the ball are unbounded and bounded support vectors, respectively. **c** OC-SVM: another way of solving the data description problem, where all samples from the target class are mapped with maximum distance to the origin. **d** KFD: Kernel Fisher's Discriminant separates the classes by projecting them onto a hyperplane where the difference of the projected means (μ_1, μ_2) is large, and the variance around means σ_1 and σ_2 is small

In this new formulation, parameter C has been removed and a new variable ρ with coefficient ν has been introduced. This new variable ρ adds another degree of freedom to the margin, the size of the margin increasing linearly with ρ. The old parameter C controlled the trade off between the training error and the generalization error. In the ν-SVM formulation, this is done adjusting ν in the range [0, 1], which acts as an upper bound on the fraction of margin errors, and it is also a lower bound on the fraction of support vectors.

3.3 Support Vector Data Description

A different problem statement for classification is given by the support vector domain description (SVDD) [30]. The SVDD is a method to solve one-class problems, where one tries to describe one class of objects, distinguishing them from all other possible objects.

The problem is defined as follows. Let $\{\mathbf{x}_i\}_{i=1}^n$ be a dataset belonging to a given *class of interest*. The purpose is to find a minimum volume *hypersphere* in a high dimensional feature space $\mathcal{H}$, with radius $R > 0$ and center $\mathbf{a} \in \mathcal{H}$, which contains most of these data objects (Fig. 1b). Since the training set may contain outliers, a set of *slack variables* $\xi_i \geq 0$ is introduced, and the problem becomes

$$\min_{R,\mathbf{a}} \left\{ R^2 + C \sum_{i=1}^{n} \xi_i \right\} \tag{15}$$

constrained to

$$\|\boldsymbol{\phi}(\mathbf{x}_i) - \mathbf{a}\|^2 \leq R^2 + \xi_i \quad \forall i = 1, \ldots, n \tag{16}$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, n \tag{17}$$

where parameter C controls the trade-off between the volume of the hypersphere and the permitted errors. Parameter ν, defined as $\nu = 1/nC$, can be used as a rejection fraction parameter to be tuned as noted in [31].

The dual functional is a quadratic programming problem that yields a set of Lagrange multipliers (α_i) corresponding to constraints in (16). When the free parameter C is adjusted properly, most of the α_i are zero, giving a sparse solution. The Lagrange multipliers are also used to calculate the distance from a test point to the center $R(\mathbf{x}_*)$:

$$K(\mathbf{x}_*, \mathbf{x}_*) - 2\sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, {}_*) + \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{18}$$

which is compared with ratio R. Unbounded support vectors are those samples $\mathbf{x}_i$ satisfying $0 \leqslant \alpha_i < C$, while bounded SVs are samples whose associated $\alpha_i = C$, which are considered outliers.

3.4 One-Class Support Vector Machine

In the one-class support vector machine (OC-SVM), instead of defining a hypersphere containing all examples, a hyperplane that separates the data objects from the origin with maximum margin is defined (Fig. 1c). It can be shown that when working with normalized data and the RBF Gaussian kernel, both methods yield the same solutions [31].

In the OC-SVM, we want to find a hyperplane $\mathbf{w}$ which separates samples $\mathbf{x}_i$ from the origin with margin ρ. The problem thus becomes

$$\min_{\mathbf{w},\rho,\xi}\left\{\frac{1}{2}\|\mathbf{w}\|^2-\rho+\frac{1}{\nu n}\sum_{i=1}^{n}\xi_i\right\} \tag{19}$$

constrained to

$$\langle\boldsymbol{\phi}(\mathbf{x}_i),\mathbf{w}\rangle\geq\rho-\xi_i\quad\forall i=1,\ldots,n \tag{20}$$

$$\xi_i\geq 0\quad\forall i=1,\ldots,n \tag{21}$$

The problem is solved through its Langrangian dual introducing a set of Lagrange multipliers α_i. The margin ρ can be computed as $\rho=\langle\mathbf{w},\boldsymbol{\phi}(\mathbf{x}_i)\rangle=\sum_j\alpha_j K(\mathbf{x}_i,\mathbf{x}_j)$.

3.5 Kernel Fisher's Discriminant

Assume that, n_1 out of n training samples belong to class -1 and n_2 to class +1. Let $\boldsymbol{\mu}$ be the mean of the whole set, and $\boldsymbol{\mu}_-$ and $\boldsymbol{\mu}_+$ the means for classes -1 and +1, respectively. Analogously, let $\boldsymbol{\Sigma}$ be the covariance matrix of the whole set, and $\boldsymbol{\Sigma}_-$ and $\boldsymbol{\Sigma}_+$ the covariance matrices for the two classes. The Linear Fisher's Discriminant (LFD) seeks for projections that maximize the interclass variance and minimize the intraclass variance [32, 33]. By defining the *between class scatter matrix* $\mathbf{S}_B=(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^\top$ and the *within class scatter matrix* $\mathbf{S}_W=\boldsymbol{\Sigma}_- + \boldsymbol{\Sigma}_+$, the problem reduces to maximize

$$J(\mathbf{w})=\frac{\mathbf{w}^\top\mathbf{S}_B\mathbf{w}}{\mathbf{w}^\top\mathbf{S}_W\mathbf{w}} \tag{22}$$

The Kernel Fisher's Discriminant (KFD) is obtained by defining the LFD in a high dimensional *feature* space $\mathcal{H}$. Now, the problem reduces to maximize:

$$J(\mathbf{w})=\frac{\mathbf{w}^\top\mathbf{S}_B^{\boldsymbol{\phi}}\mathbf{w}}{\mathbf{w}^\top\mathbf{S}_W^{\boldsymbol{\phi}}\mathbf{w}} \tag{23}$$

where now $\mathbf{w},\mathbf{S}_B^{\boldsymbol{\phi}}$ and $\mathbf{S}_W^{\boldsymbol{\phi}}$ are defined in $\mathcal{H}$, $\mathbf{S}_B^{\boldsymbol{\phi}}=(\boldsymbol{\mu}_-^{\boldsymbol{\phi}}-\boldsymbol{\mu}_+^{\boldsymbol{\phi}})(\boldsymbol{\mu}_-^{\boldsymbol{\phi}}-\boldsymbol{\mu}_+^{\boldsymbol{\phi}})^\top$, and $\mathbf{S}_W^{\boldsymbol{\phi}}=\boldsymbol{\Sigma}_-^{\boldsymbol{\phi}}+\boldsymbol{\Sigma}_+^{\boldsymbol{\phi}}$.

We need to express (23) in terms of dot-products only. According to the reproducing kernel theorem [5], any solution $\mathbf{w}\in\mathcal{H}$ can be represented as a linear combination of training samples in $\mathcal{H}$. Therefore $\mathbf{w}=\sum_{i=1}^{n}\alpha_i\boldsymbol{\phi}(\mathbf{x}_i)$ and then

$$\mathbf{w}^\top\boldsymbol{\mu}_i^{\boldsymbol{\phi}}=\frac{1}{n_i}\sum_{j=1}^{n}\sum_{k=1}^{n_i}\alpha_j K(\mathbf{x}_j,\mathbf{x}_k^i)=\boldsymbol{\alpha}^\top\mathbf{M}_i \tag{24}$$

where $\mathbf{x}_k^i$ represents samples $\mathbf{x}_k$ of class i, and $(\mathbf{M}_i)_j = \frac{1}{n_i}\sum_{k=1}^{n_i} K(\mathbf{x}_j, \mathbf{x}_k^i)$. Taking the definition of $\mathbf{S}_B^{\phi}$ and (24), the numerator of (23) can be rewritten as $\mathbf{w}^{\top}\mathbf{S}_B^{\phi}\mathbf{w} = \boldsymbol{\alpha}^{\top}\mathbf{M}\boldsymbol{\alpha}$, and the denominator as $\mathbf{w}^{\top}\mathbf{S}_W^{\phi}\mathbf{w} = \boldsymbol{\alpha}^{\top}\mathbf{N}\boldsymbol{\alpha}$, where

$$\mathbf{M} = (\mathbf{M}_{-} - \mathbf{M}_{+})(\mathbf{M}_{-} - \mathbf{M}_{+})^{\top} \tag{25}$$

$$\mathbf{N} = \sum_{j=\{-1,+1\}} \mathbf{K}_j(\mathbf{I} - \mathbf{1}_{n_j})\mathbf{K}_j^{\top} \tag{26}$$

$\mathbf{K}_j$ is a $n \times n_j$ matrix with $(\mathbf{K}_j)_{nm} = K(\mathbf{x}_n, \mathbf{x}_m^j)$ (the kernel matrix for class j), $\mathbf{I}$ is the identity and $\mathbf{1}_{n_j}$ a matrix with all entries set to $1/n_j$. Finally, Fisher's linear discriminant in $\mathcal{H}$ is solved by maximizing

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^{\top}\mathbf{M}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\top}\mathbf{N}\boldsymbol{\alpha}}, \tag{27}$$

which is solved as in the linear case. The projection of a new sample $\mathbf{x}$ onto $\mathbf{w}$ can be computed through the kernel function:

$$\mathbf{w}^{\top}\boldsymbol{\phi}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \tag{28}$$

3.6 Experimental Results for Supervised Classification

Here we compare the performance of ν-SVM, OC-SVM, LFD and KFD methods in a remote sensing multisource image classication problem: the identification of classes 'urban' and 'non-urban'. For the ν-SVM, LFD and KFD the problem is binary. For OC-SVM, we take the class 'urban' as the target class. The images used are from ERS2 SAR and Landsat TM sensors acquired in 1999 over the area of Naples, Italy [34]. The dataset has seven Landsat bands, two SAR backscattering intensities (0–35 days), and the SAR interferometric coherence. Since these features come from different sensors, the first step was to perform a specific processing and conditioning of optical and SAR data, and to co-register all images. Then, all features were stacked at a pixel level. A small area of the image of 400 × 400 pixels was selected.

We used 10 randomly selected samples of each class to train the classifiers (only 10 'urban' samples for the one-class experiment). Except the LFD, the other classifiers have free parameters that must be tuned in the training process. To do this, the training set was split following a ν-fold strategy.[1] For all methods, we used

[1] In ν-fold, the training set is divided in ν subsets, then during ν times $\nu - 1$ subsets are used for training, and the remaining subset is used for validation. At the end, the parameters that have worked the best in the ν subsets are selected.

Table 1 Mean and standard deviation of estimated kappa statistic (κ), precision, recall, F-Measure and rate of support vectors for the ten realizations

Method	κ	Precicision	Recall	F-Measure	SVs (%)
ν-SVC lin	0.81 ± 0.06	0.83 ± 0.07	**0.90 ± 0.07**	0.86 ± 0.04	33 ± 0.13
ν-SVC RBF	0.80 ± 0.07	0.86 ± 0.08	0.85 ± 0.10	0.85 ± 0.05	36 ± 0.24
LFD	0.72 ± 0.06	0.76 ± 0.08	0.84 ± 0.05	0.79 ± 0.04	–
KFD	**0.82 ± 0.03**	0.87 ± 0.04	0.86 ± 0.05	**0.86 ± 0.02**	–
OC-SVM lin	0.70 ± 0.06	0.78 ± 0.11	0.79 ± 0.13	0.77 ± 0.05	15 ± 0.05
OC-SVM RBF	0.68 ± 0.16	**0.93 ± 0.06**	0.64 ± 0.21	0.74 ± 0.15	37 ± 0.12

Best results are boldfaced

the RBF kernel where σ was tuned in the range $[10^{-3}, 10^{3}]$ in logarithmic increments of 10. The ν-SVM and OC-SVM have and additional parameter to tune: ν was varied in the range [0.1, 0.5] in increments of 0.1. Experiments were repeated 10 times with different random realizations of the training sets. Averaged results are shown using four different error measures obtained from the confusion matrices: the estimated kappa statistic (κ) [35]; the *precision* (P), defined as the ratio between the number of true positives and the sum of true positives and false positives; the *recall* (R), defined as the ratio between the number of true positives and the sum of true positives and false negatives. The last one is the F-Measure (or unbiased F-Score), computed as $F = 2\frac{P \cdot R}{P+R}$, which combines both measures. Table 1 shows the mean results and the percentage of support vectors for the 10 different training sets.

3.6.1 Linear versus nonlinear

From Table 1, several conclusions can be obtained concerning the suitable kernel. In the case of ν-SVM, linear kernel yields slightly favourable results but differences to the RBF kernel are not statistically significant. On the contrary, for the case of Fisher's discriminants, KFD is better than the linear LFD. Particularly interesting is the case of the OC-SVM. Here, using the RBF Gaussian kernel has the problem of adjusting the width σ using only samples from the target class. The problem is quite difficult because, as reliable measures like the estimated kappa statistic or the F-Measure cannot be computed using only samples of the target class, σ should be adjusted by measuring only the true positive ratio and controlling model's complexity through the rate of support vectors. In those cases where a proper value for σ cannot be found, the linear kernel may perform better, as it has no free parameter to adjust.

3.6.2 ν -SVM versus OC-SVM

In terms of the estimated kappa statistic, the ν-SVM classifier generally works better than the OC-SVM in this example. This result is not surprising since this experiment is essentially a binary problem and the ν-SVM has, in the training phase, information about both classes, whereas the OC-SVM is trained using only information of the class 'urban'. Comparing the results in terms of precision, the ν-SVM performs better than OC-SVM using the linear kernel, but worse when OC-SVM uses the RBF kernel. On the other hand, the ν-SVM obtains better results in terms of recall, meaning that it has less false negatives for the target class. Evaluating the performance with the F-Measure, which takes into account both precision and recall, the ν-SVM obtains better overall results. Finally, results clearly show that sparser classifiers are obtained when using the OC-SVM with the linear kernel.

3.6.3 Support Vector versus Fisher's Discriminant

Algorithms based on support vectors using the RBF kernel have a similar (but slightly lower) performance than the KFD algorithm. This better performance may be due to the low number of training samples used (being non-sparse, KFD has a full—dense—representation of the training data) and the squared loss function used is better suited to the assumed (Gaussian) noise in the data.

3.7 Semisupervised Image Classification

Remote sensing image classification is a challenging task because only a small number of labeled pixels is typically available, and thus classifiers tend to overfit the data [2]. In this context, semisupervised learning (SSL) naturally appears as a promising tool for combining labeled and unlabeled information thus increasing the accuracy and robustness of class predictions [10, 36]. The key issue in SSL is the general assumption of *consistency*, which means that: (1) nearby points are likely to have the same label; and (2) points on the same data structure (cluster or manifold) are likely to have the same label. This argument is often called the *cluster assumption* [37, 38]. Traditional SSL methods are based on generative models, which estimate the conditional density and have been extensively applied in remote sensing image classification [39]. Recently, more attention has been paid to *discriminative* approaches, such as: (1) the Transductive SVM (TSVM) [4, 40], which maximizes the margin for labeled and unlabeled samples simultaneously; (2) Graph-based methods, in which each pixel spreads its label information to its neighbors until a global steady state is achieved on the whole image [41, 42]; and (3) the Laplacian SVM (LapSVM) [43, 44], which deforms the kernel matrix of a standard SVM with the relations found by building the graph Laplacian. Also, the design of cluster and bagged kernels [37] have been successfully presented in remote sensing [45, 46], whose essential idea is to modify the eigenspectrum of the kernel matrix that in turn implies an alteration of the distance metric. Figure 2 illustrates a typical semisupervised learning situation where distribution of unlabeled samples helps to improve the generalization of the classifier.

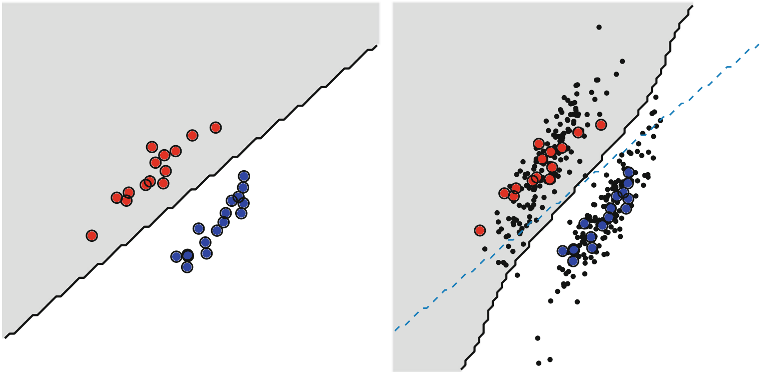


Fig. 2 *Left* classifier obtained using labeled data (*red* and *blue* denote different classes). *Right* classifier obtained using labeled data plus unlabeled data distribution (*black dots* denote unlabeled data)

3.7.1 Manifold-Based Regularization Framework

Regularization helps to produce smooth decision functions that avoid overfitting to the training data. Since the work of Tikhonov [47], many regularized algorithms have been proposed to control the capacity of the classifier [5, 48]. Regularization has been applied to both linear and nonlinear algorithms in the context of remote sensing image classification, and becomes strictly necessary when few labeled samples are available compared to the high dimensionality of the problem. In the last decade, the most paradigmatic case of regularized nonlinear algorithm is the support vector machine: in this case, maximizing the margin is equivalent to applying a kind of regularization to model weights [5, 6]. These regularization methods are especially appropriate when a low number of samples is available, but are not concerned about the geometry of the marginal data distribution. This has been recently treated within a more general regularization framework that includes Tikhonov's as a special case.

3.7.2 Semisupervised Regularization Framework

The classical regularization framework has been recently extended to the use of unlabeled samples [43] as follows. Notationally, we are given a set of l labeled samples, $\{\mathbf{x}_i\}_{i=1}^{l}$ with corresponding class labels y_i, and a set of u unlabeled samples $\{\mathbf{x}_i\}_{i=l+1}^{l+u}$. Let us now assume a general-purpose decision function f. The regularized functional to minimize is:

$$\mathcal{L} = \frac{1}{l}\sum_{i=1}^{l} V(\mathbf{x}_i, y_i, f) + \gamma_L \|f\|_{\mathcal{H}}^2 + \gamma_M \|f\|_{\mathcal{M}}^2, \tag{29}$$

where V represents a generic cost function of the committed errors on the labeled samples, γ_L controls the complexity of f in the associated Hilbert space $\mathcal{H}$, and γ_M controls its complexity in the intrinsic geometry of the data distribution. For example, if the probability distribution is supported on a low-dimensional manifold, $\|f\|_{\mathcal{M}}^2$ penalizes f along that manifold $\mathcal{M}$. Note that this semisupervised learning framework allows us to develop many different algorithms just by playing around with the loss function, V, and the regularizers, $\|f\|_{\mathcal{H}}^2$ and $\|f\|_{\mathcal{M}}^2$.

3.7.3 Laplacian Support Vector Machine

Here, we briefly review the Laplacian SVM as an instantiation of the previous framework. More details can be found in [43], and its application to remote sensing data classification in [44].

The Laplacian support vector machine (LapSVM) uses the same hinge loss function as the traditional SVM:

$$V(\mathbf{x}_i, y_i, f) = \max(0, 1 - y_i f(\mathbf{x}_i)), \tag{30}$$

where f represents the decision function implemented by the selected classifier and the predicted labels are $y_* = \operatorname{sgn}(f(\mathbf{x}_*))$. Hereafter, unlabeled or test samples are highlighted with *.

The decision function used by the LapSVM is $f(\mathbf{x}_*) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_*) \rangle + b$, where $\boldsymbol{\phi}(\cdot)$ is a nonlinear mapping to a higher dimensional Hilbert space $\mathcal{H}$, and $\mathbf{w}$ and b define a linear decision function in that space. The decision function is given by $f(\mathbf{x}_*) = \sum_{i=1}^{l+u} \alpha_i K(\mathbf{x}_i, \mathbf{x}_*) + b$. The regularization term $\|f\|_{\mathcal{H}}^2$ can be fully expressed in terms of the corresponding kernel matrix and the expansion coefficients $\boldsymbol{\alpha}$:

$$\|f\|_{\mathcal{H}}^2 = \|\mathbf{w}\|^2 = (\boldsymbol{\Phi}\boldsymbol{\alpha})^\top (\boldsymbol{\Phi}\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}. \tag{31}$$

Essentially, for manifold regularization, the LapSVM relies on the Laplacian eigenmaps (LE), which tries to map nearby inputs (pixels) to nearby outputs (corresponding class labels), thus preserving the neighborhood relations between samples.[2] Therefore, the geometry of the data is modeled with a graph in which nodes represent both labeled and unlabeled samples connected by weights W_{ij} [10]. Regularizing the graph follows from the *smoothness* (or *manifold*) assumption and intuitively is equivalent to penalize "rapid changes" of the classification function evaluated between close samples in the graph:

$$\|f\|_{\mathcal{M}}^2 = \frac{1}{(l+u)^2} \sum_{i,j=1}^{l+u} W_{ij} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 = \frac{\mathbf{f}^\top \mathbf{L} \mathbf{f}}{(l+u)^2}, \tag{32}$$

[2] In our case, nearby points are those pixels spectrally similar and thus the assumption is applied to the (high) dimensional space of image pixels.

where $\mathbf{L} = \mathbf{D}\text{-}\mathbf{W}$ is the graph Laplacian, whose entries are sample and graph-dependent; $\mathbf{D}$ is the diagonal degree matrix of $\mathbf{W}$ given by $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$ and $D_{ij} = 0$ for $i \neq j$; the normalizing coefficient $\frac{1}{(l+u)^2}$ is the natural scale factor for the empirical estimate of the Laplace operator [43]; and $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{l+u})]^\top = \mathbf{K}\boldsymbol{\alpha}$, where we deliberately dropped the bias term b.

3.7.4 Transductive SVM

The Transductive SVM (TSVM), originally proposed in [4] and further extended to deal with the peculiarities of remote sensing data in [40], aims at choosing a decision boundary that maximizes the margin on both labeled and unlabeled data. The TSVM optimizes a loss function similar to (29), but $\gamma_M \|f\|^2_{\mathcal{M}}$ is replaced by a term related to the distance of unlabeled samples to the margin. The TSVM functional to be minimized is:

$$\mathcal{L} = \frac{1}{l}\sum_{i=1}^{l} V(\mathbf{x}_i, y_i, f) + \gamma_L \|f\|^2_{\mathcal{H}} + \lambda \sum_{j=l+1}^{l+u} L^*(f(\mathbf{x}_j^*)), \tag{33}$$

where l and u are the number of labeled and unlabeled examples, λ is a free parameter that controls the relevance of unlabeled samples, and L^* is the symmetric hinge loss function:

$$L^*(f(\mathbf{x}^*)) = max(0, 1 - |f(\mathbf{x}^*)|). \tag{34}$$

The optimization of L^* can be seen as "self-learning", i.e., we use the prediction for $\mathbf{x}^*$ for training the mapping for that same example. Minimizing (34) pushes away unlabeled samples from the margin, either negative or positive, thus minimizes the absolute value.

3.8 Experimental Results for Semisupervised Classification

This section presents the experimental results of semisupervised methods in the same urban monitoring application presented in the previous section [34]. However, different sets of labeled and unlabeled training samples were used in order to test the performance of the SSL methods. Training and validation sets consisting of $l = 400$ labeled samples (200 samples *per* class) were generated, and $u = 400$ unlabeled (randomly selected) samples from the analyzed images were added to the training set for the LapSVM and TSVM. We focus on the ill-posed scenario and vary the rate of both labeled and unlabeled samples independently, i.e. $\{2, 5, 10, 20, 50, 100\}\%$ of the labeled/unlabeled samples of the training set were used to train the models in each experiment. In order to avoid skewed conclusions,

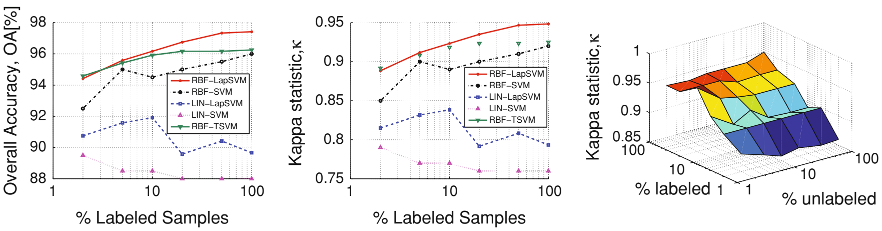


Fig. 3 Results for the urban classification. Overall Accuracy OA[%] (*left*) and Kappa statistic κ (*middle*) over the validation set as a function of the rate of labeled training samples used to build models. Kappa statistic surface (*right*) over the validation set for the best RBF-LapSVM classifier as a function of the rate of both labeled and unlabeled training samples

we run all experiments for a number of realizations where the used training samples were randomly selected.

Both linear and RBF kernels were used in the SVM, LapSVM, and TSVM. The graph Laplacian, $\mathbf{L}$, consisted of $l + u$ nodes connected using k nearest neighbors, and computed the edge weights W_{ij} using the Euclidean distance among samples. Free parameters γ_L and γ_M were varied in steps of one decade in the range $[10^{-4}, 10^4]$, the number of neighbors k used to compute the graph Laplacian was varied from 3 to 9, and the Gaussian width was tuned in the range $\sigma = \{10^{-3}, \ldots, 10\}$ for the RBF kernel. The selection of the best subset of free parameters was done by cross-validation.

Figure 3 shows the validation results for the analyzed SVM-based classifiers. Several conclusions can be obtained. First, LapSVM classifiers produce better classification results than SVM in all cases (note that SVM is a particular case of the LapSVM for $\gamma_M = 0$) for both the linear and the RBF kernels. LapSVM also produces better classification results than TSVM when the number of labeled samples is increased. Differences among methods are numerically very similar when a low number of labeled samples is available. The κ surface for the LapSVM highlights the importance of the labeled information in this problem.

4 Kernel Methods in Biophysical Parameter Estimation

Robust, fast and accurate regression tools are a critical demand in remote sensing. The estimation of physical parameters, $\mathbf{y}$, from raw measurements, $\mathbf{x}$, is of special relevance in order to better understand the environment dynamics at local and global scales [49]. The inversion of analytical models introduces a higher level of complexity, induces an important computational burden, and sensitivity to noise becomes an important issue. In the recent years, nevertheless, the use of *empirical models* adjusted to learn the relationship between the acquired spectra and actual ground measurements has become very attractive. *Parametric* models have some important drawbacks, which typically lead to poor prediction results on unseen (test) data. As a consequence, *non-parametric* and potentially *nonlinear* regression techniques have been effectively introduced [50]. Different models and architectures of neural networks have been considered for the estimation of biophysical parameters [50–52]. However, despite their potential effectiveness, neural networks present some important drawbacks: (1) design and training often results in a complex, time-consuming task; (2) following the minimization of the empirical risk (i.e. the error in the training data set), rather than the structural risk (an upper bound of the generalization error), can lead to overfit the training data; and (3) performance can be degraded when working with low-sized data sets. A promising alternative to neural networks is the use of kernel methods analyzed in this section, such as support vector regression (SVR) [11, 53], relevance vector machines (RVM) [16], and Gaussian Processes (GP) [17].

4.1 Support Vector Regression

The support vector regression (SVR) is the SVM implementation for regression and function approximation [5, 54], which has yielded good results in modeling some biophysical parameters and in alleviating the aforementioned problems of neural networks [11, 55, 56].

The standard SVR formulation uses Vapnik's ε-insensitive cost function, in which errors e_i up to ε are not penalized, and all further deviations will incur in a linear penalization. Briefly, SVR estimates weights $\mathbf{w}$ by minimizing the following regularized functional:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) \tag{35}$$

with respect to $\mathbf{w}$ and $\{\xi_i^{(*)}\}_{i=1}^n$, constrained to:

$$y_i - \mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x}_i) - b \leq \varepsilon + \xi_i \quad \forall i = 1,\ldots,n \tag{36}$$

$$\mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x}_i) + b - y_i \leq \varepsilon + \xi_i^* \quad \forall i = 1,\ldots,n \tag{37}$$

$$\xi_i, \xi_i^* \geq 0 \quad \forall i = 1,\ldots,n \tag{38}$$

where $\xi_i^{(*)}$ are positive slack variables to deal with training samples with a prediction error larger than ε ($\varepsilon > 0$), and C is the penalization parameter applied to these ones. Note that C trade-offs the minimization of errors and the regularization term, thus controling the generalization capabilities. The usual procedure for solving SVRs introduces the linear restrictions (36)–(38) into (35) using Lagrange multipliers α_i, computes the Karush-Kuhn-Tucker conditions, and solves the dual problem using QP procedures [57], which yields the final solution:

$$\hat{y}_i = \sum_{j=1}^n(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j) + b. \tag{39}$$

Again, non-zero multipliers are called SVs. Sparsity in the SVR is a direct consequence of the loss function; as the value of ε increases, the number of support vectors is reduced.

4.2 Relevance Vector Machines

Despite the good performance offered by the SVR, it has some deficiencies: (1) by assuming an explicit loss function (usually, the ε-insensitive loss function) one assumes a fixed distribution of the residuals, (2) the free parameters must be tuned usually through cross-validation methods, which result in time consuming tasks, (3) the nonlinear function used in SVR must fulfil Mercer's Theorem [58] to be valid, and (4) sparsity is not always achieved and a high number of support vectors is thus obtained.

Some of these problems of SVRs are efficiently alleviated by the relevance vector machine (RVM), which was originally introduced by Tipping in [59]. The RVM constitutes a Bayesian approximation to solve extended linear (in the parameters) models, i.e. nonlinear models. Therefore, the RVM follows a different inference principle from the one followed in SVR. In this case, a particular probability model for the support vectors is assumed and can be constrained to be sparse. In addition, it has been claimed that RVMs can produce probabilistic outputs (which theoretically permits to capture uncertainty in the predictions), RVMs are less sensitive to hyper-parameters setting than SVR, and the *kernel* function must not necessarily fulfil Mercer's conditions.

Once the kernel has been defined, and a particular Gaussian likelihood assumed for the target vector $\mathbf{y} = [y_1, \ldots, y_n]^\top$ given the weights $\mathbf{w}$, a maximum likelihood approach could be used for estimating model weights. However, a certain risk of overfitting arises and *a priori* models of weight distribution are commonly used in the Bayesian framework [60]. In the RVM learning scheme, rather than attempting to make sample-based (or point) predictions, a Gaussian *prior* distribution of zero mean and variance $\sigma^2_{w_j} \equiv \alpha_j^{-1}$ is defined over each weight:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{j=1}^{n} \mathcal{N}(w_j|0, \alpha_j^{-1}) = \prod_{j=1}^{n} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{1}{2}\alpha_j \mathbf{w}_j^2\right), \tag{40}$$

where the key to obtain sparsity is the use of n independent hyperparameters $\boldsymbol{\alpha} = (\alpha_o, \alpha_1, \ldots, \alpha_n)^\top$, one per weight (or basis function), which moderate the strength of the *prior*. After defining the *prior* over the weights, we must define the hyperpriors over $\boldsymbol{\alpha}$ and the other model parameter, the noise variance σ_n^2. These quantities were originally proposed to be Gamma distributions [59].

Now, with the *prior* (40) and the likelihood distribution, the posterior distribution over the weights is Gaussian and can be computed using Bayes' rule:

$$p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \sigma_n^2) = \frac{p(\mathbf{y}|\mathbf{w}, \sigma_n^2)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2)} \sim \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{41}$$

where the covariance and the mean are respectively given by $\boldsymbol{\Sigma} = (\sigma_n^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \mathbf{A})^{-1}$ and $\boldsymbol{\mu} = \sigma_n^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top\mathbf{y}$, with $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$. Hence, the Gaussian likelihood distribution over the training targets can be "marginalized" by integrating out the weights to obtain the *marginal likelihood* for the hyperparameters:

$$p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2) = \int p(\mathbf{y}|\mathbf{w}, \sigma_n^2)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \sim \mathcal{N}(0, \mathbf{C}) \tag{42}$$

where the covariance is given by $\mathbf{C} = \sigma_n^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\top$. For computational efficiency, the logarithm of the evidence is maximized:

$$\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2) = -\frac{1}{2}\left(n \log 2\pi + \log|\mathbf{C}| + \mathbf{y}^\top\mathbf{C}^{-1}\mathbf{y}\right), \tag{43}$$

which is commonly done using the standard *type-II maximum likelihood procedure*. However, [59] did not suggest direct minimization of the negative log evidence for training the RVM, but rather the use of an approximate Expectation-Maximization (EM) procedure [61].

In the RVM learning scheme, the estimated value of model weights is given by the mean of the posterior distribution (41), which is also the *maximum a posteriori* (MAP) estimate of the weights. The MAP estimate of the weights depends on the value of hyperparameters $\boldsymbol{\alpha}$ and the noise σ_n^2. The estimate of these two variables ($\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}_n^2$) is obtained by maximizing the marginal likelihood (42). The uncertainty about the optimal value of the weights reflected by the posterior distribution (41) is used to express uncertainty about the predictions made by the model as follows. Given a new input $\mathbf{x}_*$, the probability distribution of the corresponding output y_* is given by the (Gaussian) predictive distribution:

$$p(\mathbf{y}_*|\mathbf{x}_*, \hat{\boldsymbol{\alpha}}, \hat{\sigma}_n^2) = \int p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{w}, \hat{\sigma}_n^2)p(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}_n^2)d\mathbf{w} \sim \mathcal{N}(y_*, \sigma_*^2) \tag{44}$$

where the mean and the variance (uncertainty) of the prediction are $y_* = (\boldsymbol{\Phi})_{i,:}\boldsymbol{\mu}$ and $\sigma_*^2 = \hat{\sigma}_n^2 + (\boldsymbol{\Phi})_{i,:}\boldsymbol{\Sigma}(\boldsymbol{\Phi})_{i,:}^\top$.

In the iterative maximization of $\mathcal{L}(\boldsymbol{\alpha})$, many of the hyperparameters α_j tend to infinity, yielding *a posterior* distribution (41) of the corresponding weight w_j that tends to be a delta function centered around zero. The corresponding weight is thus deleted from the model, as well as its associated basis function, $\boldsymbol{\phi}_j(\mathbf{x})$. In the RVM framework, each basis function $\boldsymbol{\phi}_j(\mathbf{x})$ is associated to a training sample $\mathbf{x}_j$ so that $\boldsymbol{\phi}_j(\mathbf{x}) = K(\mathbf{x}_j, \mathbf{x})$. The model is built on the few training examples whose associated hyperparameters do not go to infinity during the training process, leading to a sparse solution. These examples are called the Relevance Vectors (RVs), resembling the SVs in the SVM framework.

4.3 Gaussian Processes

An important concern about the suitability of RVM Bayesian algorithms in biophysical parameter estimation was raised: oversparseness was easily obtained due to the use of an improper prior, which led to inaccurate predictions and poor predictive variance estimations outside the support. Recently, the introduction of Gaussian Processes (GPs) has alleviated the aforementioned problem at the cost of providing non-sparse models [62]. GPs are also a Bayesian approach to non-parametric kernel learning. Very good numerical performance and stability has been reported in remote sensing parameter retrieval [17, 63].

Gaussian processes for regression define a distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}$ fully described by a mean $m : \mathcal{X} \rightarrow \mathbb{R}$ and a covariance (kernel) function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and $K(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))^\top (f(\mathbf{x}') - m(\mathbf{x}'))]$. Hereafter we set m to be the zero function for the sake of simplicity. Now, given a finite labeled samples dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ we first compute its covariance matrix $\mathbf{K}$ in the same way as done for the Gram matrix in SVM. The covariance matrix defines a distribution over the vector of output values $f_\mathbf{x} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top$, such that $f_\mathbf{x} \sim \mathcal{N}(\mathbf{0}; \mathbf{K})$, which is a multivariate Gaussian distribution. Therefore the specification of the covariance function implies the form of the distribution over the functions. The role of the covariance for GPs is the same as the role of kernels in SVM, both specify the notion of similarity in the space of functions.

For training purposes, we assume that the observed variable is formed by noisy observations of the true underlying function $y = f(\mathbf{x}) + \varepsilon$. Moreover we assume the noise to be additive independently and identically Gaussian distributed with zero mean and variance σ_n^2. Let us define the stacked output values $\mathbf{y} = (y_1, \ldots, y_n)^\top$, the covariance terms of the test point $\mathbf{K}_i = [K(\mathbf{x}_i, \mathbf{x}_1), \ldots, K(\mathbf{x}_i, \mathbf{x}_n)]^\top$, and $K_{ii} = K(\mathbf{x}_i, \mathbf{x}_i)$. From the previous model assumption, the output values are distributed according to:

$$\begin{pmatrix} \mathbf{y} \\ f(\mathbf{x}_i) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_n^2 \mathbf{I} & \mathbf{K}_i \\ \mathbf{K}_i^\top & K_{ii} \end{bmatrix}\right) \tag{45}$$

For prediction purposes, the Gaussian Processes (GP) is obtained by computing the conditional distribution $f(\mathbf{x}_i) | \mathbf{y}, \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}; \mathbf{x}_i$, which can be shown to be a Gaussian distribution with predictive mean $\mathbf{K}_i^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}$ and predictive variance $K_{ii} - \mathbf{K}_i^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K}_i$. Therefore, two hyperparameters must be optimized: the kernel $\mathbf{K}$ and the noise variance σ_n^2.

Note that the GP mean predictor yields exactly the same solution that the obtained in the context of kernel ridge regression (i.e. unconstrained kernel regression with squared loss function and Tikhonov's regularization). Even more important is the fact that not only a mean prediction is obtained for each sample but a full distribution over the output values including an uncertainty of the prediction.

The optimization of GP hyperparameters can be done through standard cross-validation tecniques. However, a good property of the GP framework is the possibility to optimize all involved hyperparameters, $\boldsymbol{\theta}$, iteratively through gradient-descent. This is done by maximizing the negative log marginal likelihood, $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, and its partial derivatives w.r.t. the hyperparameters[3]:

$$\begin{aligned}\frac{\partial \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_j} &= \frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \theta_j}\mathbf{K}^{-1}\mathbf{y} \\ &\quad - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \theta_j}\right\} = \frac{1}{2}\mathrm{Tr}\left\{(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K})^{-1}\frac{\partial \mathbf{K}}{\partial \theta_j}\right\},\end{aligned} \tag{46}$$

where $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$, which is only computed once. This optimization is done by a particular gradient-based optimization, resulting in a relatively fast method that scales well for less than a few thousand training samples [62]. This technique not only avoids running heuristic cross-validation methods but also optimizing very flexible kernel functions and estimating the noise variance consistently.

4.4 Experimental Results

In this section, we evaluate the performance of SVR, RVM and GP in the estimation of oceanic chlorophyll-a concentration from measured reflectances. We compare the models in terms of accuracy, bias, and sparsity. We use the SeaBAM dataset [64], which gathers 919 *in-situ* measurements of chlorophyll concentration around the United States and Europe. The dataset contains in situ pigments and remote sensing reflectance measurements at wavelengths present in the SeaWiFS sensor.[4]

Developing a SVR requires selecting the following free parameters: σ (varied between 0.1 and 30), C (varied logarithmically between 10^{-2} and 10^5), and ε (varied logarithmically between 10^{-6} and 10^{-1}). For the case of the RVM algorithm, the σ was logarithmically varied between 0.1 and 30. For the GP, we used a scaled anisotropic RBF kernel, $K(\mathbf{x}, \mathbf{x}') = \nu \exp(-\sum_{d=1}^{D} 0.5\sigma_d^{-2}(\mathbf{x}^{(d)} - \mathbf{x}^{(d)'})^2)$ $+\sigma_n^2\delta_{\mathbf{x}\mathbf{x}'}$, where ν is a kernel scaling factor accounting for signal variance, D is the data input dimension (d indicates dimension), σ_d is a dedicated lengthscale for feature d, and σ_n is the magnitude of the independent noise component. It is worth noting that in order to obtain a good set of optimal parameters, a cross-validation methodology must be followed. The available data were randomly split into two sets: 460 samples for cross-validation and the remaining 459 samples for testing

[3] $\log p(\mathbf{y}|\mathbf{x}) \equiv \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log(det(\mathbf{K} + \sigma_n^2\mathbf{I})) - \frac{n}{2}\log(2\pi)$.

[4] More information about the data can be obtained from http://seabass.gsfc.nasa.gov/seabam/seabam.html.

Table 2 Mean error (ME), root mean-squared error (RMSE), mean absolute error (MAE), and correlation coefficient between the actual and the estimated Chl-a concentration (*r*) of models in the test set

	ME	RMSE	MAE	r	SVs/RVs (%)
Morel-1†,	−0.023	0.178	0.139	0.956	–
Ocean Chlorophyll 2, OC2	−0.031	0.169	0.133	0.960	–
NN-BP, 4 hidden nodes	−0.046	0.143	0.111	**0.971**	–
ε-SVR	−0.070	0.139	**0.105**	**0.971**	44.3
RVM	**−0.009**	0.146	0.107	0.970	**4.9**
GP	**−0.009**	**0.103**	0.107	0.961	–

Best results are boldfaced

performance. Before training, data were centered and transformed logarithmically, as in [65].

Table 2 presents results in the test set for SVR, RVM and GP models. For comparison purposes, we include results obtained with a feedforward neural network trained with back-propagation (NN-BP), which is a standard approach in biophysical parameters retrieval. Also, we include results for the model Morel-1, and the final SeaWiFS chlorophyll-a algorithm OC2 from [66]. We can observe that (1) SVR, RVM and GP show a better performance than empirical Morel and OC2 models, and also better than artificial neural networks (NN-BP); (2) the SVR and GP techniques are more accurate (RMSE, MAE); (3) RVM and GP are less biased (ME) than the rest of the models, and in the case of the RVMs, drastically much more sparse (only 4.9% of training samples were necessary to attain good generalization capabilities). Comparing SVR and RVM, we can state that RVMs provide accurate estimations (similar to SVR) with small number of relevant vectors. GP provides more accurate results than SVR and RVM.

5 Kernel Methods for Feature Extraction

The curse of dimensionality refers to the problems associated with multivariate data analysis as the dimensionality increases. This problem is specially relevant in remote sensing since, as long as new technologies improve, the number of spectral bands is continuously increasing. There are two main implications of the curse of dimensionality, which critically affect pattern recognition applications in remote sensing: there is an exponential growth in the number of examples required to maintain a given sampling density (e.g., for a density of n examples *per* bin with d dimensions, the total number of examples should be n^d); and there is an exponential growth in the complexity of the target function (e.g., a density estimate or a classifier) with increasing dimensionality. In these cases, feature extraction methods are used to create a subset of new features by combinations of the existing features. Even though the use of linear methods such as principal component analysis (PCA) or partial least squares (PLS) is quite common, recent advances to cope with nonlinearities in the data based on multivariate kernel machines have been presented [67]. In the rest of the section we will briefly review the linear and nonlinear kernel versions of PCA and PLS.

5.1 Mutivariate Analysis Methods

The family of multivariate analysis (MVA) methods comprises several algorithms for feature extraction that exploit correlations between data representation in input and output spaces, so that the extracted features can be used to predict the output variables, and viceversa.

Notationally, a set of training pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^{\mathbb{N}}, \mathbf{y}_i \in \mathbb{R}^{\mathbb{M}}$, where $\mathbf{x}$ are the observed explanatory variables in the input space (i.e. spectral channels or bands) and $\mathbf{y}$ are the target variables in the output space (e.g., class material or corresponding physical parameter), are given. This can be also expressed using matrix notation, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$, where superscript $^\top$ denotes matrix or vector transposition. $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ denote the centered versions of $\mathbf{X}$ and $\mathbf{Y}$, respectively, while $\mathbf{C}_{xx} = \frac{1}{n}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$ represents the covariance matrix of the input data, and $\mathbf{C}_{xy} = \frac{1}{n}\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}}$ the covariance between the input and output data.

Feature extraction is typically used before the application of machine learning algorithms to discard irrelevant or noisy components, and to reduce the dimensionality of the data, what helps also to prevent numerical problems (e.g., when $\mathbf{C}_{xx}$ is rank deficient). Linear feature extraction can be carried out by projecting the data into the subspaces characterized by projection matrices $\mathbf{U}$ and $\mathbf{V}$, of sizes $N \times n_p$ and $M \times n_p$, so that the n_p extracted features of the original data are given by $\tilde{\mathbf{X}}' = \tilde{\mathbf{X}}\mathbf{U}$ and $\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}}\mathbf{V}$.

5.1.1 Principal Component Analysis

Principal component analysis [68], also known as the *Hotelling transform* or the *Karhunen-Loeve transform*, projects linearly the input data onto the directions of largest input variance. To perform principal component analysis (PCA), the covariance matrix is first estimated $\mathbf{C}_{xx} = 1/n\sum_{i=1}^n \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top$. Then, the eigenvalue problem $\mathbf{C}_{xx}\mathbf{u}_i = \lambda_i\mathbf{u}_i$ is solved, which yields a set of sorted eigenvalues $\{\lambda_i\}_{i=1}^{n_p}(\lambda_i \leq \lambda_{i+1})$ and the corresponding eigenvectors $\{\mathbf{u}_i\}_{i=1}^{n_p}$. Finally, new data are projected onto the eigenvectors with largest eigenvalues $\tilde{\mathbf{X}}' = \tilde{\mathbf{X}}\mathbf{U}$.

This can also be expressed more compactly as:

$$\begin{aligned} \text{PCA:}\mathbf{U} &= \arg\max_{\mathbf{U}} \text{Tr}\{\mathbf{U}^\top\mathbf{C}_{xx}\mathbf{U}\} \\ \text{subject to:} &\quad \mathbf{U}^\top\mathbf{U} = \mathbf{I} \end{aligned} \tag{47}$$

where $\mathbf{I}$ is the identity matrix of size $n_p \times n_p$. Using Lagrange multipliers, it can be shown (see, e.g. [19]) that the solution to (47) is given by the singular value decomposition (SVD) of $\mathbf{C}_{xx}$.

The main limitation of PCA is that it does not consider class separability since it does not take into account the target variables $\mathbf{y}$ of the input vectors. PCA simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance of the original data distribution. Thus, there is no guarantee that the directions of maximum variance will contain good features for discrimination or regression.

5.1.2 Partial Least Squares

Partial least squares [69] assumes that the system of interest is driven by a few latent variables (also called factors or components), which are *linear* combinations of observed explanatory variables (spectral bands). The underlying idea of partial least squares (PLS) is to exploit not only the variance of the inputs but also their covariance with the target, which is presumably more important.

The goal of PLS is to find the directions of maximum covariance between the projected input and output data:

$$\begin{aligned} \text{PLS:}\mathbf{U},\mathbf{V} &= \underset{\mathbf{U},\mathbf{V}}{\arg\max}\text{Tr}\{\mathbf{U}^\top \mathbf{C}_{xy}\mathbf{V}\} \\ \text{subject to:} &\quad \mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I} \end{aligned} \tag{48}$$

The solution to this problem is given by the singular value decomposition of $\mathbf{C}_{xy}$.

5.2 *Kernel Multivariate Analysis*

All previous methods assume that there exists a *linear* relation between the original data matrices, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, and the extracted projections, $\tilde{\mathbf{X}}'$ and $\tilde{\mathbf{Y}}'$, respectively. However, in many situations this linearity assumption is not satisfied, and nonlinear feature extraction is needed to obtain acceptable performance. In this context, *kernel methods* are a promising approach, as they constitute an excellent framework to formulate nonlinear versions from linear algorithms [5, 19]. In this section, we describe the kernel PCA (KPCA) and kernel PLS (KPLS) implementations.

Notationally, data matrices for performing the linear feature extraction (PCA or PLS) in $\mathcal{H}$ are now given by $\mathbf{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)]^\top$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$. As before, the centered versions of these matrices are denoted by $\tilde{\mathbf{\Phi}}$ and $\tilde{\mathbf{Y}}$.

Now, the projections of the input and output data will be given by $\tilde{\mathbf{\Phi}}' = \tilde{\mathbf{\Phi}}\mathbf{U}$ and $\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}}\mathbf{V}$, respectively, where the projection matrix $\mathbf{U}$ is now of size $\dim(\mathcal{H}) \times n_p$. Note, that the input covariance matrix in $\mathcal{H}$, which is usually needed by the different MVA methods, becomes of size $\dim(\mathcal{H}) \times \dim(\mathcal{H})$ and cannot be directly computed. However, making use of the *representer's theorem* [19], we can introduce $\mathbf{U} = \tilde{\mathbf{\Phi}}^\top\mathbf{A}$ into the formulation, where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{n_p}]$ and $\boldsymbol{\alpha}_i$ is an n-length column vector containing the coefficients for the ith projection vector, and the maximization problem can be reformulated in terms of the kernel matrix.

Note that, in these kernel feature extraction methods, the projection matrix $\mathbf{U}$ in $\mathcal{H}$ might not be explicitly calculated, but the projections of the input data can be obtained. Therefore, the extracted features for a new input pattern $\mathbf{x}_*$ are given by:

$$\tilde{\boldsymbol{\phi}}'(\mathbf{x}_*) = \tilde{\boldsymbol{\phi}}(\mathbf{x}_*)\mathbf{U} = \tilde{\boldsymbol{\phi}}(\mathbf{x}_*)\tilde{\mathbf{\Phi}}^\top\mathbf{A} = \begin{bmatrix} \tilde{K}(\mathbf{x}_1, \mathbf{x}_*) \\ \vdots \\ \tilde{K}(\mathbf{x}_n, \mathbf{x}_*) \end{bmatrix}\mathbf{A} \tag{49}$$

which is expressed in terms of the inner products in the centered feature space (see Sect. 2.3).

5.2.1 Kernel Principal Component Analysis

As in the linear case, the aim of kernel principal component analysis (KPCA) is to find directions of maximum variance of the input data in $\mathcal{H}$, which can be obtained by replacing $\tilde{\mathbf{X}}$ by $\tilde{\mathbf{\Phi}}$ in (47), i.e. by replacing $\mathbf{C}_{xx}$ by $\tilde{\mathbf{\Phi}}^\top\tilde{\mathbf{\Phi}}$:

$$\begin{aligned} \text{KPCA:}\mathbf{U} &= \arg\max_{\mathbf{U}} \ \text{Tr}\ \{\mathbf{U}^\top\tilde{\mathbf{\Phi}}^\top\tilde{\mathbf{\Phi}}\mathbf{U}\} \\ \text{subject to:} &\quad \mathbf{U}^\top\mathbf{U} = \mathbf{I} \end{aligned} \tag{50}$$

Making use of the representer's theorem one can introduce $\mathbf{U} = \tilde{\mathbf{\Phi}}^\top\mathbf{A}$ into the previous formulation, and the maximization problem can be reformulated as follows:

$$\begin{aligned} \text{KPCA:}\mathbf{A} &= \arg\max_{\mathbf{A}} \ \text{Tr}\ \{\mathbf{A}^\top\tilde{\mathbf{K}}_x\tilde{\mathbf{K}}_x\mathbf{U}\} \\ \text{subject to:} &\quad \mathbf{A}^\top\tilde{\mathbf{K}}_x\mathbf{A} = \mathbf{I} \end{aligned} \tag{51}$$

where we have defined the symmetric centered kernel matrix $\tilde{\mathbf{K}}_x = \tilde{\mathbf{\Phi}}\tilde{\mathbf{\Phi}}^\top$ containing the inner products between any two points in the feature space.

The solution to the above problem can be obtained from the singular value decomposition of $\tilde{\mathbf{K}}_x\tilde{\mathbf{K}}_x$ represented by $\tilde{\mathbf{K}}_x\tilde{\mathbf{K}}_x\boldsymbol{\alpha} = \lambda\tilde{\mathbf{K}}_x\boldsymbol{\alpha}$, which has the same solution as $\tilde{\mathbf{K}}_x\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$.

5.2.2 Kernel Partial Least Squares

As in the linear case, the aim of kernel partial least squares (KPLS) is to find directions of maximum covariance between the input data in $\mathcal{H}$ and $\mathbf{Y}$, and can thus be expressed as:

$$\begin{aligned} \text{KPLS:}\mathbf{U},\mathbf{V} &= \underset{\mathbf{U},\mathbf{V}}{\arg\max}\ \text{Tr}\ \{\mathbf{U}^\top\tilde{\mathbf{\Phi}}^\top\tilde{\mathbf{Y}}\mathbf{V}\} \\ \text{subject to:} &\quad \mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I} \end{aligned} \tag{52}$$

Again, making use of the representer's theorem, one can introduce $\mathbf{U} = \tilde{\mathbf{\Phi}}^\top\mathbf{A}$ into the previous formulation, and the maximization problem can be reformulated as follows:

$$\begin{aligned} \text{KPLS:}\mathbf{A},\mathbf{V} &= \underset{\mathbf{A},\mathbf{V}}{\arg\max}\ \text{Tr}\ \{\mathbf{A}^\top\tilde{\mathbf{K}}_x\tilde{\mathbf{Y}}\mathbf{V}\} \\ \text{subject to:} &\quad \mathbf{A}^\top\tilde{\mathbf{K}}_x\mathbf{A} = \mathbf{V}^\top\mathbf{V} = \mathbf{I} \end{aligned} \tag{53}$$

The solution to the above problem can be obtained from the singular value decomposition of $\tilde{\mathbf{K}}_x\tilde{\mathbf{Y}}$.

5.3 Experimental Results

Figure 4 illustrates the performance of linear and kernel MVA feature extraction methods in a 2D toy example. KPCA and KPLS used an RBF kernel with the same sigma value fixed to the mean distance among all training samples. It can be observed that linear methods (PCA and PLS) cannot cope with the non-linearly separable problem, while kernel methods accommodate data relations in the kernel and define local boundaries. Performance of KPLS results in more accurate boundaries and perfectly separates the two classes, while KPCA fails as no class label information is used. Results in remote sensing image classification are reported in [67, 70].

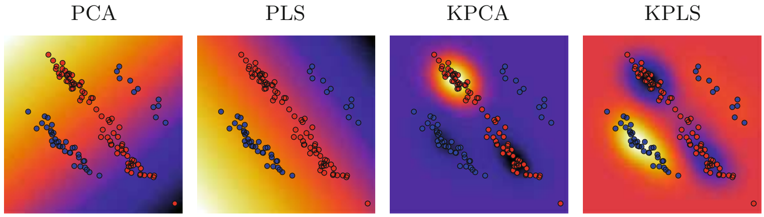


Fig. 4 First extracted component by linear (PCA, PLS) and nonlinear kernel (KPCA, KPLS) methods

6 Future Trends in Remote Sensing Kernel Learning

Even though the chapter presented an updated literature review, new kernel-based learning methodologies are being constantly explored. The special peculiarities of the acquired images lead to develop new methods. And viceversa, the new learning paradigms available offer new ways of looking at old, yet unsolved, problems in remote sensing. In what follows, we review recent research directions in the context of remote sensing kernel-based learning.

6.1 Multiple Kernel Learning

Composite kernels have been specifically designed and applied for the efficient combination of multitemporal, multisensor and multisource information [9, 71]. The previous approaches exploited some properties of kernel methods (such as the direct sum of Hilbert spaces, see Sect. 2.3) to combine kernels dedicated to process different signal sources, e.g., a kernel on spectral feature vectors can be summed up to a kernel defined over spatially-extracted feature vectors. This approach yielded very good results but it was limited to the combination of few kernels [26], as the optimization of kernel parameters was an issue. Lately, the composite framework approach has been extended to the framework of multiple kernel learning (MKL) [72]. In MKL, the SVM kernel function is defined as a weighted linear combination of kernels built using subsets of features. MKL works iteratively optimizing both the individual weights and the kernel parameters [73]. So far, the only application in remote sensing of strict MKL can be found in [74] and, taking advantage of a similar idea, spectrally weighted kernels are proposed in [75]. Not only a certain gain in accuracy is observed but also the final model yields some insight in the problem. In [46], the relevant features of remote sensing images for automatic classification are studied through this framework.

6.2 Transfer Learning

A common problem in remote sensing is that of updating land-cover maps by classifying temporal series of images when only training samples collected at one time instant are available. This is known as transfer learning or domain adaptation. This setting implies that unlabeled test examples and training examples are drawn from different domains or distributions. The problem was initially tackled with partially unsupervised classifiers, both under parametric formalisms [76] and neural networks [77]. The approach was then successfully extended to domain adaptation SVM (DASVM) [78].

A related problem is also that of classifying an image using labeled pixels from other scenes, which induces the sample selection bias problem, also known as covariance shift. Here, unlabeled test data are drawn from the same training domain, but the estimated distribution does not correctly model the true underlying distribution since the number (or the quality) of available training samples is not sufficient. These problems have been recently presented by defining mean map kernel machines that account for the dispersion of data in feature spaces [45].

6.3 Structured Learning

Most of the techniques revised so far assume a simple set of outputs. However, more complex output spaces can be imagined, e.g. predicting multiple labels (land use and land cover simultaneously), multi-temporal image sequences, or abundance fractions. Such complex output spaces are the topic of structured learning, one of the most recent developments in machine learning. Only a computer vision application [79] and the preliminary results in [80] have been presented for image processing. Certainly this field of learning joint input-output mappings will receive attention in the future.

6.4 Active Learning

When designing a supervised classifier, the performance of the model strongly depends on the quality of the labeled information available. This constraint makes the generation of an appropriate training set a difficult and expensive task requiring extensive manual human-image interaction. Therefore, in order to make the models as efficient as possible, the training set should be kept as small as possible and focused on the pixels that really help to improve the performance of the model. Active learning aims at responding to this need, by constructing effective training sets.

In remote sensing, application of active learning methods that select the most relevant samples for training is quite recent. A SVM method for object-oriented classification was proposed in [81], while maximum likelihood classifiers for pixel-based classification was presented in [82]. Recently, this approach was extended in [83] by proposing boosting to iteratively weight the selected pixels. In [84, 85] information-based active learning was proposed for target detection, and in [86], a model-independent active learning method was proposed for very-high resolution satellite images.

6.5 Parallel Implementations

Kernel methods in general, and the SVM in particular, have the problem of scaling at least quadratically with the number of training samples. With the recent explosion in the amount and complexity of hyperspectral data, and with the increasing availability of very high resolution images, the number of labeled samples to train kernel classifiers is becoming a critical problem. In this scenario, parallel processing constitutes a requirement in many remote sensing missions, especially with the advent of low-cost systems such as commodity clusters and distributed networks of computers. Several efforts are being pursued to develop parallel implementations of SVMs for remote sensing data classification: boss-worker approaches [87–89] and parallelization through decomposition of the kernel matrix have been successfully explored [90].

7 Conclusions

Kernel methods allow us to transform almost any linear method into a nonlinear one, while still operating with linear algebra. The methods essentially rely on embedding the examples into a high dimensional space where a linear method is designed and applied. Access to the mapped samples is done implicitly through kernel functions. This chapter reviewed the field of kernel machines in remote sensing data processing. The important topics of classification, model inversion, and feature extraction with kernels have been revised. The impact and development of kernel methods in this area during the last decade has been large and fruitful, overcoming some of the problems posed both by the recent satellite sensors acquired data, and the limitations of other machine learning methods. New developments are expected in the near future to encompass both remote sensing data complexity and new problem settings.

Acknowledgments This work was partially supported by projects CICYT-FEDER TEC2009-13696, AYA2008-05965-C04-03, and CSD2007-00018. Valero Laparra acknowledges the support of a Ph.D grant from the Spanish Government BES-2007-16125. The authors would like to thank a number of colleagues and collaborators in the field of kernel methods, and whose insight and points of view are at some extent reflected in this chapter: Dr. Devis Tuia from the University of Lausanne (Switzerland), Dr. Frédéric Ratle from Nuance Communications (Belgium), Dr. Jerónimo Arenas from the University Carlos III de Madrid (Spain), and Prof. Lorenzo Bruzzone from the University of Trento (Italy).

References

1. Richards, J.A., Jia, X.: Remote Sensing Digital Image Analysis. An Introduction, 3rd edn. Springer, Berlin (1999)
2. Hughes, G.F.: On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory **14**, 55–63 (1968)
3. Fukunaga, K., Hayes, R.R.: Effects of sample size in classifier design. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 873–885 (1989)
4. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)

5. Schölkopf, B., Smola, A.: Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond. MIT Press Series, Cambridge (2002)
6. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **43**, 1351–1362 (2005)
7. Mercier, G., Girard-Ardhuin, F.: Partially supervised oil-slick detection by SAR imagery using kernel expansion. IEEE Trans. Geosci. Remote Sens. **44**, 2839–2846 (2006)
8. Muñoz Marí, J., Bruzzone, L., Camps-Valls, G.: A support vector domain description approach to supervised classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. **45**, 2683–2692 (2007)
9. Camps-Valls, G., Gómez-Chova, L., Muñoz Marí, J., Martínez-Ramón, M., Rojo-Álvarez, J.L.: Kernel-based framework for multi-temporal and multi-source remote sensing data classification and change detection. IEEE Trans. Geosci. Remote Sens. **46**, 1822–1835 (2008)
10. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. 1st edn. MIT Press, Cambridge (2006)
11. Camps-Valls, G., Bruzzone, L., Rojo-Álvarez, J.L., Melgani, F.: Robust support vector regression for biophysical variable estimation from remotely sensed images. IEEE Geosci. Remote Sens. Lett. **3**, 339–343 (2006)
12. Zortea, M., De Martino, M., Moser, G., Serpico, S.B.: Land surface temperature estimation from infrared satellite data using support vector machines. In: Proceedings of the IGARSS-2006 Symposium, pp. 2109–2112, Denver, USA (2003)
13. Durbh, S.S., King, R.L., Younan, N.H.: Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Remote Sens. Environ. **107**, 348–361 (2007)
14. Yang, F., White, M., Michaelis, A., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.X., Nemani, R.: Prediction of continental-scale evapotranspiration by combining MODIS and Ameri Flux data through support vector machine. IEEE Trans. Geosci. Remote Sens. **44**, 3452–3461 (2006)
15. Broadwater, J., Chellappa, R., Banerjee, A., Burlina, P.: Kernel fully constrained least squares abundance estimates. In: Proceedings of the IGARSS-2007 Symposium, Barcelona, Spain (2007)
16. Camps-Valls, G., Gomez-Chova, L., Vila-Francés, J., Amorós-López, J., Muñoz-Marí, J., Calpe-Maravilla, J.: Retrieval of oceanic chlorophyll concentration with relevance vector machines. Remote Sens. Environ. **105**, 23–33 (2006)
17. Pasolli, L., Melgani, F., Blanzieri, E.: Estimating biophysical parameters from remotely sensed imagery with Gaussian processes. In: IEEE International Geoscience and Remote Sensing Symposium, IGARSS'08, Boston, USA (2008)
18. Camps-Valls, G., Gómez-Chova, L., Muñoz-Marí, J., Calpe-Maravilla, J.: Biophysical parameter estimation with adaptive Gaussian processes. In: IEEE International Geoscience & Remote Sensing Symposium, IGARSS'2009, Capetown, South Africa (2009)
19. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
20. Golub, G.H., Van Loan, C.F.: Matrix Computations (Johns Hopkins Studies in Mathematical Sciences). The Johns Hopkins University Press, Baltimore (1996)
21. Reed, M.C., Simon, B.: Functional Analysis. Volume I of Methods of Modern Mathematical Physics. Academic Press, New York (1980)
22. Huang, C., Davis, L., Townshend, J.: An assessment of support vector machines for land cover classification. Int. J. Remote Sens. **23**(4), 725–749 (2002)
23. Foody, G.M., Mathur, A.: Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sens. Environ. **93**, 107–117 (2004)
24. Fauvel, M., Chanussot, J., Benediktsson, J.A.: Evaluation of kernels for multiclass classification of hyperspectral remote sensing data. In: IEEE ICASSP—International conference on Acoustics, Speech and Signal Processing, pp. II-813–II-816, Toulouse, France (2006)

25. Inglada, J.: Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. ISPRS J. Photogramm. Rem. Sens. **62**, 236–248 (2007)
26. Camps-Valls, G., Gómez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., Calpe-Maravilla, J.: Composite kernels for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. **3**, 93–97 (2006)
27. Chi, M., Feng, R., Bruzzone, L.: Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem. Adv. Space Res. **41**(11), 1793–1799 (2008)
28. Fauvel, M., Benediktsson, J.A., Chanussot, J., Sveinsson, J.R.: Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. IEEE Trans. Geosci. Remote Sens. **46**(11), 3804–3814 (2008)
29. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Comput. **12**, 1207–1245 (2000)
30. Tax, D., Duin, R.P.: Support vector domain description. Pattern Recognit. Lett. **20**, 1191–1199 (1999)
31. Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J.: Support vector method for novelty detection. In: Advances in Neural Information Processing Systems 12, Denver, CO (1999)
32. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
33. Hastie, T., Tibishirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
34. Gómez-Chova, L., Fernández-Prieto, D., Calpe, J., Soria, E., Vila-Francés, J., Camps-Valls, G.: Urban monitoring using multitemporal SAR and multispectral data. Pattern Recognit. Lett. **27**, 234–243 (2006) 3rd Pattern Recognition in Remote Sensing Workshop, Kingston Upon Thames, England, Aug 27, 2004
35. Congalton, R.G., Green, K.: Assessing the Accuracy of Remotely Sensed data: Principles and Practices. Lewis Publishers, Boca Raton (1999)
36. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, USA (2005) http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf
37. Chapelle, O., Weston, J., Schölkopf, B.: Cluster kernels for semi-supervised learning. In: Becker, S., Thrun, S., Obermayer, K. (eds.) NIPS 2002, vol. 15, pp. 585–592. MIT Press, Cambridge (2003)
38. Seeger, M.: Learning with labeled and unlabeled data. Technical Report TR.2001, Institute for Adaptive and Neural Computation, University of Edinburg (2001)
39. Jackson, Q., Landgrebe, D.: An adaptive classifier design for high-dimensional data analysis with a limited training data set. IEEE Trans. Geosci. Remote Sens. **39**, 2664–2679 (2001)
40. Bruzzone, L., Chi, M., Marconcini, M.: A novel transductive SVM for the semisupervised classification of remote-sensing images. IEEE Trans. Geosci. Remote Sens. **44**, 3363–3373 (2006)
41. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, NIPS2004, vol. 16. MIT Press, Vancouver (2004)
42. Camps-Valls, G., Bandos, T., Zhou, D.: Semi-supervised graph-based hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **45**, 2044–3054 (2007)
43. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. **7**, 2399–2434 (2006)
44. Gómez-Chova, L., Camps-Valls, G., Muñoz-Marí, J., Calpe-Maravilla, J.: Semi-supervised image classification with Laplacian support vector machines. IEEE Geosci. Remote Sens. Lett. **5**, 336–340 (2008)

45. Gómez-Chova, L., Camps-Valls, G., Bruzzone, L., Calpe-Maravilla, J.: Mean map kernel methods for semisupervised cloud classification. IEEE Trans. Geosci. Remote Sens. **48**, 207–220 (2010)
46. Tuia, D., Camps-Valls, G.: Semisupervised remote sensing image classification with cluster kernels. Geosci. Remote Sens. Lett. IEEE **6**, 224–228 (2009)
47. Tikhonov, A.N.: Regularization of incorrectly posed problems. Sov. Math. Dokl. **4**, 1624–1627 (1963)
48. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**, 1–50 (2000)
49. Lillesand, T., Kiefer, R.W., Chipman, J.: Remote Sensing and Image Interpretation. 6th Edition. Wiley, New York (2008)
50. Kimes, D., Knyazikhin, Y., Privette, J., Abuelgasim, A., Gao, F.: Inversion methods for physically-based models. Remote Sens. Rev. **18**, 381–439 (2000)
51. Keiner, L.E.: Estimating oceanic chlorophyll concentrations with neural networks. Int. J. Remote Sens. **20**, 189–194 (1999)
52. Dzwonkowski, B., Yan, X.H.: Development and application of a neural network based colour algorithm in coastal waters. Int. J. Remote. Sens. **26**, 1175–1200 (2005)
53. Camps-Valls, G., Muñoz-Marí, J., Gómez-Chova, L., Richter, K., Calpe-Maravilla, J.: Biophysical parameter estimation with a semisupervised support vector machine. IEEE Geosci. Remote Sens. Lett. **6**, 248–252 (2009)
54. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**, 199–222 (2004)
55. Kwiatkowska, E., Fargion, G.: Application of machine-learning techniques toward the creation of a consistent and calibrated global chlorophyll concentration baseline dataset using remotely sensed ocean color data. IEEE Trans. Geosci. Remote Sens. **41**, 2844–2860 (2003)
56. Zhan, H., Shi, P., Chen, C.: Retrieval of oceanic chlorophyll concentration using support vector machines. IEEE Trans. Geosci. Remote Sens. **41**, 2947–2951 (2003)
57. Fletcher, R.: Practical Methods of Optimization. John Wiley & Sons, Inc. 2nd Edition (1987)
58. Courant, R., Hilbert, D.: Methods of Mathematical Physics. Interscience Publications. Wiley, New York (1953)
59. Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)
60. O'Hagan, A.: Bayesian Inference, Volume 2B of Kendall's Advanced Theory of Statistics. Arnold, London, United Kingdom (1994)
61. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Series B **39**, 1–38 (1977)
62. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, New York (2006)
63. Furfaro, R., Morris, R.D., Kottas, A., Taddy, M., Ganapol, B.D.: A Gaussian Process Approach to Quantifying the Uncertainty of Vegetation Parameters from Remote Sensing Observations. AGU Fall Meeting Abstracts A261 (1977)
64. O'Reilly, J.E., Maritorena, S., Mitchell, B.G., Siegel, D.A., Carder, K., Garver, S.A., Kahru, M., McClain, C.: Ocean color chlorophyll algorithms for SeaWiFS. J. Geophys. Res. **103**, 24937–24953 (1998)
65. Cipollini, P., Corsini, G., Diani, M., Grass, R.: Retrieval of sea water optically active parameters from hyperspectral data by means of generalized radial basis function neural networks. IEEE Trans. Geosci. Remote Sens. **39**, 1508–1524 (2001)
66. Maritorena, S., O'Reilly, J.: In: OC2v2: Update on the initial operational SeaWiFS chlorophyll a algorithm. In: Hooker, S.B., Firestone, E.R. (eds.) SeaWiFS Postlaunch Calibration and Validation Analyses, NASA Goddard Space Flight Center. Wiley, Greenbelt Part 3. NASA Tech. Memo. 2000-206892, vol. 11, pp. 3–8 (2000)
67. Camps-Valls, G., Bruzzone, L.: Kernel Methods for Remote Sensing Data Analysis. Wiley, New York (2009)
68. Jolliffe, I.T.: Principal Component Analysis. Springer, Heidelberg (1986)

69. Wold, S., Albano, C., Dunn, W.J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., Sjostrom, M.: Multivariate data analysis in chemistry. In: Kowalski, B.R. (ed.) Chemometrics, Mathematics and Statistics in Chemistry, pp. 17–95. Reidel Publishing Company, Boston (1984)
70. Arenas-García, J., Camps-Valls, G.: Efficient kernel orthonormalized PLS for remote sensing applications. IEEE Trans. Geosci. Remote Sens. **46**, 2872 –2881 (2008)
71. Tuia, D., Ratle, F., Pozdnoukhov, A., Camps-Valls, G.: Multisource composite kernels for urban image classification. IEEE Geosci. Remote Sens. Lett. **6**(2), 234–238 (2009)
72. Lancricket, G., Bie, T.D., Cristianini, N., Jordan, M., Noble, W.: A statistical framework for genomic data fusion. Bioinformatics **20**, 2626–2635 (2004)
73. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. J. Mach. Learn. Res. **7**, 1531–1565 (2006)
74. Villa, A., Fauvel, M., Chanussot, J., Gamba, P., Benediktsson, J.A.: Gradient optimization for multiple kernel parameters in support vector machines classification. In: IEEE International Geoscience and Remote Sensing Symposium, IGARSS (2008)
75. Guo, B., Gunn, S., Damper, R.I., Nelson, J.D.B.: Customizing kernel functions for SVM-based hyperspectral image classification. IEEE Trans. Image Process. **17**, 622–629 (2008)
76. Bruzzone, L., Prieto, D.: Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. IEEE Trans. Geosci. Remote Sens. **39**, 456–460 (2001)
77. Bruzzone, L., Cossu, R.: A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. IEEE Trans. Geosci. Remote Sens. **40**, 1984–1996 (2002)
78. Bruzzone, L., Marconcini, M.: Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. IEEE Trans. Geosci. Remote Sens. **47**, 1108–1122 (2009)
79. Blaschko, M., Lampert, C.: Learning to localize objects with structured output regression. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) Computer Vision: ECCV 2008, pp. 2–15. Springer, Heidelberg (2008)
80. Tuia, D., Kanevski, M., Muñoz Marí, J., Camps-Valls, G.: Structured SVM for remote sensing image classification. In: IEEE Workshop on Machine Learning for Signal Processing (MLSP09), Grenoble, France (2009)
81. Mitra, P., Uma Shankar, B., Pal, S.: Segmentation of multispectral remote sensing images using active support vector machines. Pattern Recogn. Lett. **25**, 1067–1074 (2004)
82. Rajan, S., Ghosh, J., Crawford, M.M.: An active learning approach to hyperspectral data classification. IEEE Trans. Geosci. Remote Sens. **46**(4), 1231–1242 (2008)
83. Jun, G., Ghosh, J.: An efficient active learning algorithm with knowledge transfer for hyperspectral remote sensing data. In: Proceedings of the IEEE Geoscience Remote Sensing Symposium IGARSS (2008)
84. Zhang, C., Franklin, S., Wulder, M.: Geostatistical and texture analysis of airborne-acquired images used in forest classification. Int. J. Remote Sens. **25**, 859–865 (2004)
85. Liu, Q., Liao, X., Carin, L.: Detection of unexploded ordnance via efficient semisupervised and active learning. IEEE Trans. Geosci. Remote Sens. **46**(9), 2558–2567 (2008)
86. Tuia, D., Ratle, F., Pacifici, F., Kanevski, M., Emery, W.: Active learning methods for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **47**, 2218–2232 (2009)
87. Brazile, J., Schaepman, M.E., Schläpfer, D., Kaiser, J.W., Nieke, J., Itten, K.I.: Cluster versus grid for large-volume hyperspectral image preprocessing. In: Huang, H.L.A., Bloom, H.J. (eds.) Atmospheric and Environmental Remote Sensing Data Processing and Utilization: an End-to-End System Perspective. Edited by Huang, Hung-Lung A.; Bloom, Hal J. In: Proceedings of the SPIE, vol. 5548, pp. 48–58 (2004). Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, vol. 5548, pp. 48–58. (2004)
88. Gualtieri, J.A.: A parallel processing algorithm for remote sensing classification. Technical report, Summaries of the Airborne Earth Science Workshop, Pasadena, USA (2004) http://aviris.jpl.nasa.gov/html/aviris/documents.html.

89. Plaza, A., Chang, C.I.: High Performance Computing in Remote Sensing. Chapman & Hall/ CRC Press, Boca Raton (2007)
90. Muñoz, J., Plaza, A., Gualtieri, J.A., Camps-Valls, G.: Parallel implementation of SVM in earth observation applications. In: Xhafa, F. (ed.) Parallel Programming and Applications in Grid, P2P and Networking systems, pp. 292–312. IOS Press, UK (2009)

Exploring Nonlinear Manifold Learning for Classification of Hyperspectral Data

Melba M. Crawford, Li Ma and Wonkook Kim

Abstract Increased availability of hyperspectral data and greater access to advanced computing have motivated development of more advanced methods for exploitation of nonlinear characteristics of these data. Advances in manifold learning developed within the machine learning community are now being adapted for analysis of hyperspectral data. This chapter investigates the performance of popular global (Isomap and KPCA) and local manifold nonlinear learning methods (LLE, LTSA, LE) for dimensionality reduction in the context of classification. Experiments were conducted on hyperspectral data acquired by multiple sensors at various spatial resolutions over different types of land cover. Nonlinear dimensionality reduction methods often outperformed linear extraction methods and rivaled or were superior to those obtained using the full dimensional data.

Keywords Manifold learning · Dimensionality reduction · Classification · Hyperspectral · Isometric feature mapping · Kernel principal component analysis · Locally linear embedding · Local tangent space alignment · Laplacian eigenmaps

M. M. Crawford (✉)
School of Civil Engineering and Department of Agronomy, Purdue University, West Lafayette, IN, USA
e-mail: mcrawford@purdue.edu

L. Ma
State Key Laboratory for Multi-spectral Information Processing Technologies, Huazhong University of Science and Technology, Wuhan, China
e-mail: lma@purdue.edu

W. Kim
Department of Civil Engineering, Purdue University, West Lafayette, IN, USA
e-mail: wkkim@purdue.edu

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_11,

1 Introduction

Remote sensing data from airborne and space-based hyperspectral sensors are becoming increasingly available and potentially provide greatly improved capability for discriminating, characterizing, and monitoring complex chemistry-based processes. Dimensionality reduction (DR) via feature extraction is an important preprocessing step for many approaches to analysis of hyperspectral image data, including visualization, regression, clustering, classification, and anomaly detection. While commonly used linear feature extraction methods such as principal component analysis (PCA) and linear discriminant analysis (LDA) are simple and easily implemented, the dramatic increase in spectral resolution associated with hyperspectral data allows representation of inherent nonlinearities in physical processes, motivating nonlinear dimensionality reduction, namely *manifold learning*.

The machine learning community has demonstrated the potential of manifold learning approaches for nonlinear dimensionality reduction and modeling of nonlinear structure by determining coordinate systems that lie on the nonlinear manifold represented by the data [1–5]. The potential value of manifold learning has also been demonstrated for hyperspectral remote sensing applications including feature extraction [6–10], classification [11–18], and anomaly detection [19]. However, it should be noted that while many manifold learning methods provide excellent results for synthetic data, the topology of natural data sets is often much more difficult to characterize, and linear methods often outperform nonlinear methods, particularly when applied blindly [20]. Manifold methods inherently assume smoothness in the manifold structure, but remotely sensed data acquired over disparate classes often do not vary smoothly and may contain spatially disjoint clusters in the embedded space, potentially reducing the advantages offered by these nonlinear approaches for some applications.

This study is motivated by the need to better understand characteristics of hyperspectral data in the manifold domain, with the goal of improving the development and application of these methods for analysis of hyperspectral data. In this context, we investigate nonlinear manifold learning methods for dimensionality reduction in classification of hyperspectral data. Approaches are implemented and evaluated in an empirical study involving several space-based and airborne hyperspectral data sets which are widely used by the remote sensing community to evaluate classification methods. Focusing on issues related to dimensionality reduction rather than performance related to classifiers, the k-Nearest Neighbor (k-NN) method with $k = 1$ is used as the common base classifier. The paper is organized as follows: Sect. 2 provides a summary of each of the manifold learning methods investigated in the study; the four data sets used in the study are described in Sect. 3; experimental results are contained in Sect. 4; observations are summarized in Sect. 5.

2 Nonlinear Manifold Learning for Dimensionality Reduction

Nonlinear manifold learning methods are broadly characterized as global or local approaches. Global manifold methods retain the fidelity of the overall topology of the data set, but have greater computational overhead for large data sets, while local methods preserve local geometry and are computationally efficient because they only require sparse matrix computations. Although global manifolds seek to preserve geometry across all scales of the data and have less tendency to overfit the data, which is beneficial for generalization in classification, local methods may yield good results for data sets which have significantly different sub-manifolds.

Development of representative metrics to characterize manifold topology is a current topic of significant research interest in machine learning [18, 20]. Both qualitative and quantitative approaches are being used to compare various manifold learning methods. Qualitative approaches typically involve visualization, while quantitative approaches employ metrics for reconstruction error and results of subsequent analysis such as classification, anomaly detection, or measures of intra- and inter-class distances.

The basic ideas and mathematical formulations of each of the manifold learning methods in this study are presented in the following sections. To improve comparison of the different methods, the formulation of each methodology is described using a graph embedding framework [21, 22], which provides a common formulation for both the global and the local manifold learning algorithms. A list of symbols used in this chapter is provided in Table 1.

Table 1 Symbol definitions

Symbol	Definition	Symbol	Definition
R^a	Space of real numbers (dimension a)	$\mathbf{B}$	Constraint matrix
$R^{a \times b}$	Space of real numbers ($a \times b$)	$\mathbf{C}$	Covariance matrix
Φ	Feature mapping	$\mathbf{D}$	Degree matrix
G	Graph	$\mathbf{E}$	Reconstruction error matrix
K	Kernel function	$\mathbf{F}$	Reconstruction matrix
O	Computational complexity	$\mathbf{H}$	Centering matrix
c	Number of classes	$\mathbf{I}$	Identity matrix
δ_{ij}	Kronecker delta function	$\mathbf{K}$	Gram matrix
k	Number of neighborhood samples	$\mathbf{L}$	Laplacian matrix
λ	Eigenvalue	$\mathbf{\Lambda}$	Eigenvalue matrix
m	Dimension of original data	$\mathbf{M}$	Inner product matrix
n	Number of samples	$\mathbf{\Theta}$	Local tangent space coordinates
n_1	Number of training data samples	$\mathbf{S}_{\mathrm{stp}}$	Shortest path distance matrix
n_2	Number of testing data samples	$\mathbf{V}$	Eigenvector matrix
p	Target dimension of manifold coordinates	$\mathbf{W}$	Similarity matrix
p'	Dimension of the local tangent space	$\mathbf{X}$	Data matrix
σ	Gaussian kernel parameter	$\mathbf{X}_i$	Neighborhood set of i-th sample
$\boldsymbol{\alpha}$	Expansion coefficients	$\mathbf{Y}$	Manifold coordinates
$\mathbf{e}$	Vector of ones		

2.1 Dimensionality Reduction Within a Graph Embedding Framework

Given data samples in a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, $\mathbf{x}_i \in R^m$ where n is the number of samples and m is the feature dimension, the dimensionality reduction problem seeks to find a set of manifold coordinates $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$, $\mathbf{y}_i \in R^p$, where typically, $m \gg p$, through a feature mapping Φ: $\mathbf{x} \to \mathbf{y}$, which may be analytical (explicit) or data driven (implicit), and linear or nonlinear.

For the dimensionality reduction problem, the graph embedding framework assumes an undirected weighted graph $G = \{\mathbf{X}, \mathbf{W}\}$ with data samples $\mathbf{X}$ and algorithm dependent similarity matrix $\mathbf{W}$. Once the graph is constructed, the graph Laplacian $\mathbf{L}$ plays an important role in the framework. Here, $\mathbf{L} = \mathbf{D} - \mathbf{W}$ with a diagonal degree matrix defined by $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$, $\forall\, i$.

In the one-dimensional case, where the resultant manifold coordinate for n samples is a vector $\mathbf{y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_n]$, the dimensionality reduction criterion for the methods used in this study can be represented as

$$\mathbf{y}^* = \arg \min_{\mathbf{yBy}^{\mathrm{T}}=r} \sum_{i \neq j} \left\| y_i - y_j \right\|^2 \mathbf{W}_{ij} = \arg \min_{\mathbf{yBy}^{\mathrm{T}}=r} \mathbf{yLy}^{\mathrm{T}} \tag{1}$$

where r is a constant and $\mathbf{B}$ is a constraint matrix that depends on the dimensionality reduction method. The underlying goal is for sample pairs of larger weight to have manifold coordinates that are closer to each other, under a unique data geometry characterized by the graph Laplacian $\mathbf{L}$. The solution of the optimization problem can be obtained by solving the eigen-decomposition problem $\mathbf{Ly} = \lambda\mathbf{By}$, where the one-dimensional manifold coordinates $\mathbf{y}$ are given by the eigenvector with the smallest non-zero eigenvalue. This one-dimensional case can be easily generalized to the multi-dimensional case through the following expansion

$$\mathbf{Y}^* = \arg \min_{\mathbf{YBY}^{\mathrm{T}}=\mathbf{R}} \operatorname{tr}(\mathbf{YLY}^{\mathrm{T}}) \tag{2}$$

where $\mathbf{R}$ is a diagonal matrix. Analogous to the one-dimensional case, the manifold coordinates $\mathbf{Y}$ of target dimension p can be obtained from the eigenvectors corresponding to the p smallest non-zero eigenvalues. Each of the manifold learning algorithms in this study can be described in terms of this common framework with different Laplacian matrices and constraints. In the following sections, each algorithm and the respective formulations are presented using the general framework.

2.2 Global Manifold Learning

Isometric Feature Mapping (Isomap) and Kernel PCA are the most widely utilized global manifold learning approaches for nonlinear dimensionality reduction. Basic implementations of both approaches are outlined in the following sub-sections.

2.2.1 Isometric Feature Mapping (Isomap)

The Isomap method assumes that the local feature space formed by the nearest neighbors is linear, and the global nonlinear transformation can be found by connecting these piecewise linear spaces. Isomap uses a user-defined neighborhood of size k and the shortest path algorithm to discover the manifold [1]. It first defines $\mathbf{X}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ik}]$, the set of neighborhood nodes of node $\mathbf{x}_i$, to create a distance matrix $\mathbf{S}'$. A distance to node $\mathbf{x}_j$ is computed by using the following rule. If $\mathbf{x}_j \in \mathbf{X}_i$, $S'_{ij} = s_{ij}$; otherwise $S'_{ij} = \infty$, where s_{ij} is the distance between the two nodes. Isomap then accumulates the distance beyond the set $\mathbf{X}_i$ along the shortest path to obtain a shortest path network $\mathbf{S}_{\mathrm{stp}}$. The shortest path algorithm, typically implemented in Isomap via the Dijkstra method [23], finds the paths from a root node to all other nodes to minimize the sum of the individual path lengths. The process is repeated for each sample, which in turn becomes the root node, to create $\mathbf{S}_{\mathrm{stp}}$. Dimensionality reduction is then accomplished through multidimensional scaling (MDS), a dimensionality reduction technique that places a set of samples in a meaningful dimensional space that explains the similarity between samples.

Isomap can be represented in the graph framework by defining the weight matrix $\mathbf{W}$ with the shortest path distance matrix $\mathbf{S}_{\mathrm{stp}}$. The Laplacian matrix is $\mathbf{L} = \mathbf{HTH}/2$, where $\mathbf{H} = \mathbf{I} - (1/n)\mathbf{ee}^{\mathrm{T}}$ and $\mathbf{T}_{ij} = [(\mathbf{S_{stp}})_{ij}]^2$. The matrix $\mathbf{W}$ can then be constructed by setting $\mathbf{W}_{ij} = -\mathbf{L}_{ij},\ i \neq j;$ else 0. The constraint matrix is set to the identity matrix, $\mathbf{B} = \mathbf{I}$.

The shortest path algorithm is computationally demanding. Several approaches have been proposed to mitigate this problem for large data sets. Examples include a divide and conquer method coupled with realignment of subset manifolds [6] and various implementations using landmarks (L-Isomap) in conjunction with embedding of non-landmark points via the derived embedding vectors reducing the complexity of computing the shortest path distance matrix [7, 15, 24, 25].

2.2.2 Kernel Principal Component Analysis (KPCA)

Kernel PCA is a nonlinear extension of linear PCA in a feature space induced by a kernel function [26]. The estimated covariance matrix of the sample data in the feature space is obtained from

$$\mathbf{C} := \frac{1}{n}\sum_{j=1}^{n} \Phi(\mathbf{x}_j)\Phi(\mathbf{x}_j)^{\mathrm{T}} \tag{3}$$

The principal components are then computed by solving the eigen-decomposition problem $\mathbf{CV} = \lambda\mathbf{V}$. The low dimensional embedding can be obtained from $\mathbf{y}_i = \mathbf{V}^p \cdot \Phi(\mathbf{x}_i)$, where $\mathbf{V}^p$ is a matrix of the eigenvectors corresponding to the p largest eigenvalues. However, the feature mapping Φ need not to be known explicitly if the “kernel trick” is employed, by assuming a feature space induced by a positive definite kernel function K given by $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) := (\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j))$.

Once the kernel function is defined, the feature coordinates are represented by $\mathbf{y}_i = \sum_j \alpha_j \mathrm{K}(\mathbf{x}_j, \mathbf{x}_i)$ where α denotes expansion coefficients. Using the kernel framework, the eigen-decomposition problem above is converted to $n\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}$ such that $\lambda(\boldsymbol{\alpha}^T\boldsymbol{\alpha}) = 1$, where $\mathbf{K}$ is an $n \times n$ Gram matrix whose entries are obtained from evaluation of the kernel function between the samples: $\mathbf{K}_{ij} = \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j)$. KPCA can also be formulated in the general kernel framework after a slight modification of the objective function. Given a kernel function, the optimization is over the expansion coefficients because the manifold coordinates are determined by $\boldsymbol{\alpha}$.

$$\boldsymbol{\alpha}^* = \arg \min_{\boldsymbol{\alpha}^T\mathbf{KBK}\boldsymbol{\alpha}=\mathbf{I}} \boldsymbol{\alpha}^T\mathbf{KLK}\boldsymbol{\alpha} \tag{4}$$

The similarity matrix and Laplacian matrix are given by $\mathbf{W}_{ij} = 1/n$, $\mathrm{L} = \mathbf{I}\text{-}\mathbf{ee}^{\mathrm{T}}/n$, respectively. The constraint matrix $\mathbf{B} = \mathbf{I}$.

2.3 Local Manifold Learning

Three local manifold learning methods are investigated in this study: locally linear embedding (LLE), local tangent space alignment (LTSA) and Laplacian eigenmaps (LE). All three methods are initiated by constructing a nearest neighborhood for each data point, and the local structures are then used to obtain a global manifold. According to the framework, by solving the eigenvalue problem $\mathbf{LY} = \lambda\mathbf{BY}$, the embedding $\mathbf{Y}$ is provided by the eigenvectors corresponding to the $2 \sim (p+1)$ smallest eigenvalues (the eigenvector that corresponds to the smallest zero eigenvalue is a unit vector with equal elements and is discarded).

2.3.1 Locally Linear Embedding (LLE)

In LLE [2], the local properties of each neighborhood are represented by the linear coefficients that best reconstruct each data point from its neighbors. Let $\mathbf{F} \in R^{n\times n}$ be composed of the reconstruction coefficients of all the data points, which is obtained by minimizing the reconstruction error according to

$$e(\mathbf{f}_i) = \left\| \mathbf{x}_i - \sum_i f_{ij}\mathbf{x}_{ij} \right\|^2 \quad s.t. \sum_j f_{ij} = 1 \tag{5}$$

where f_{ij} denotes the reconstruction weight of $\mathbf{x}_i$ from its j-th neighbor $\mathbf{x}_{ij}$. The embedding is then obtained by retaining these coefficients in the low dimensional space via the objective function:

$$\Phi(\mathbf{Y}) = \sum_i \left\| \mathbf{y}_i - \sum_j f_{ij}\mathbf{y}_j \right\|^2 \quad s.t. \ \ \frac{1}{n}\sum_i \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}} = \mathbf{I}, \quad \sum_i \mathbf{y}_i = 0 \tag{6}$$

which can be minimized by solving Eq. 2, where the Laplacian matrix $\mathbf{L} = (\mathbf{I} - \mathbf{F}^{\mathrm{T}})(\mathbf{I} - \mathbf{F})$, constraint matrix $\mathbf{B} = \mathbf{I}$, and the similarity matrix $\mathbf{W} = \mathbf{F} + \mathbf{F}^{\mathrm{T}} - \mathbf{F}^{\mathrm{T}}\mathbf{F}$.

2.3.2 Local Tangent Space Alignment (LTSA)

In LTSA [27], the local geometry is described by the local tangent space of each data point, and the global manifold is determined by aligning the overlapping local tangent spaces. Let $\mathbf{X}_i$ be the k nearest neighbors of point $\mathbf{x}_i$, and $\boldsymbol{\Theta}_i$ of dimensionality p' be the local tangent coordinates of $\mathbf{X}_i$. The $\boldsymbol{\Theta}_i$ relate to the global coordinates $\mathbf{Y}_i$ by an affine transformation

$$\mathbf{Y}_i\mathbf{H} = \mathbf{T}_i\Theta_i + \mathbf{E}_i \tag{7}$$

where $\mathbf{T}_i \in R^{p\times p'}$ is the transformation matrix, $\mathbf{H} \in R^{k\times k}$ is the centering matrix, and $\mathbf{E}_i \in R^{p\times k}$ is the reconstruction error matrix. $\mathbf{E}_i$ is minimized to retain the local geometry in the embedded space according to

$$\Phi(\mathbf{Y}) = \sum_{i=1}^{n} (\|\mathbf{E}_i^*\|_F^2) = \sum_{i=1}^{n} (\|\mathbf{Y}_i\mathbf{U}_i\|_F^2) \quad s.t. \ \ \mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \tag{8}$$

where $\mathbf{U}_i = \mathbf{H}(\mathbf{I} - \boldsymbol{\Theta}_i^{\mathrm{T}}(\boldsymbol{\Theta}_i\boldsymbol{\Theta}_i^{\mathrm{T}})^{-1}\boldsymbol{\Theta}_i)$. The embedding $\mathbf{Y}$ is obtained by solving Eq. 2, where $\mathbf{L}$, also referred to as the alignment matrix, is constructed with $\mathbf{L}(I_i, I_i) \leftarrow \mathbf{L}(I_i, I_i) + \mathbf{U}_i\mathbf{U}_i^{\mathrm{T}}, I_i$ is the index of $\mathbf{X}_i$, and $\mathbf{B} = \mathbf{I}$. The dimensionality of the local tangent space (p') and the dimensionality of the global manifold (p) could be different [28]; thus, they can be selected separately to provide greater flexibility in applications. This could be advantageous for large, heterogeneous data sets.

2.3.3 Laplacian Eigenmaps (LE)

In LE [29], the weighted neighborhood graph of each data point is obtained by calculating the pairwise distances between neighbors, where the distance is normally calculated using a Gaussian kernel function with parameter σ. Let $\mathbf{W} \in \mathrm{R}^{n\times n}$ be the adjacency matrix that summarizes the neighborhood relations. The embedding $\mathbf{Y}$ is obtained by minimizing the total distance between each data point and its neighbors in the low dimensional space:

$$\Phi(\mathbf{Y}) = \sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \mathbf{W}_{ij} \quad s.t. \ \mathbf{Y}\mathbf{D}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \tag{9}$$

which is analogous to Eq. 2 with $\mathbf{L} = \mathbf{D} - \mathbf{W}$, $\mathbf{B} = \mathbf{D}$. It is also equivalent (up to scaling) to the eigenvalue decomposition of the normalized graph Laplacian matrix defined by $\mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$ [30, 31].

2.4 Supervised Local Manifold Learning

Both unsupervised and supervised implementations of local manifold learning have been developed. In unsupervised local manifold learning (ULML) approaches, the k *s*pectral neighbors of a given point are searched. Supervised local manifold learning (SLML) approaches identify only the neighbors that are of the same class as the given point, making these methods more attractive for classification [18, 32, 33]. SLML maps all the training data from the same class onto a single point in the embedded space. Assuming there are c classes, the outputs of SLML are c orthogonal points $\mathbf{Y}^c = [\mathbf{y}^1,\ldots,\mathbf{y}^c] \in R^{p\times c}$. Each point $\mathbf{y}^i$ has only one nonzero element and represents all the training data from the ith class. Because the last c bands of the outputs of SLML are meaningful, and each separates one class from the others, we set $p = c$. SLML also results in an eigen-decomposition of the Laplacian matrix $\mathbf{L}$, a block diagonal matrix composed of $\mathbf{L}_i$ for each class, where $\mathbf{L}_i$ is the matrix computed over the training data from the ith class. Since each $\mathbf{L}_i$ has one zero eigenvalue, $\mathbf{L}$ has c associated zero eigenvalues.

Based on the properties of SLML and a new SLML-weighted k-NN classifier [18], we utilize SLML outputs of training data, where $\mathbf{y}^i$ is equal to 1 in the i-th coordinate and the remaining coordinates are zero. As a result, the SLML coordinates of the labeled data can be obtained without calculation. This coordinate is superior to the original SLML coordinate when used in conjunction with the k-NN classifier because it reduces the impact of imbalanced data sets [18].

2.5 Kernel-Based Out-of-Sample Extension

Traditional manifold learning methods are implemented on training data and lack generalization to new data. The kernel-based out-of-sample extension method [31, 34] mitigates this problem. From the kernel view, the manifold learning methods can be represented as kernel PCA on specially constructed kernel matrices [30]. The kernel matrix of Isomap is equal to the negative of its Laplacian matrix [31, 34]. The kernel matrix of the three local manifold learning methods can be defined as $\mathbf{K} = \mathbf{I} - \mathbf{L}$ (size of $n \times n$) on training data, and its $2 \sim (p + 1)$ major eigenvectors (the first uniform eigenvector is discarded) provide the embedding results. $\mathbf{K}$ can also be obtained by the kernel function $\mathrm{K}(\cdot,\cdot)$ computed over pairs of training data. As a result, the kernel function should be learned such that $\mathrm{K}(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{K}_{ij} = \delta_{ij} - \mathbf{L}_{ij}$, where δ is the Kronecker delta. Generally, $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j)$ is determined not only by $(\mathbf{x}_i, \mathbf{x}_j)$, but also by the other training data. Therefore, $\mathrm{K}(\cdot,\cdot)$ is a “data-dependent kernel” [31]. By defining the respective $\mathrm{K}(\cdot,\cdot)$ for each LML method, the embedding of a new data point $\mathbf{x}_0$ can be generalized via the Nyström formula

$$\mathbf{y}_0^{\mathrm{T}} = \sum_{i=1}^{n} \mathrm{K}(\mathbf{x}_0, \mathbf{x}_i)\mathbf{y}_i^{\mathrm{T}}\boldsymbol{\Lambda}^{-1} \tag{10}$$

where the $p \times p$ dimensional diagonal matrix $\boldsymbol{\Lambda}$ is composed of the largest $2 \sim (p + 1)$ eigenvalues of the kernel matrix $\mathbf{K}$. The definition of $\mathrm{K}(\cdot,\cdot)$ for LLE, LTSA and LE is introduced in Refs. [31, 34] and Ma et al. [35].

For supervised local manifold learning methods, the matrix $\mathbf{K}$ has c eigenvalues that are equal to 1 because $\mathbf{K} = \mathbf{I} - \mathbf{L}$, and $\mathbf{L}$ has c zero eigenvalues. As a result, $\boldsymbol{\Lambda} = \mathbf{I}$ in Eq. 10, and the out-of-sample extension of SLML for the testing point $\mathbf{x}_0$ becomes

$$\mathbf{y}_0 = \sum_{i=1}^{n} \mathrm{K}(\mathbf{x}_0, \mathbf{x}_i)\mathbf{y}_i \tag{11}$$

In classification of hyperspectral data, two strategies are used to obtain the manifold coordinates for (unsupervised) manifold learning methods. One applies manifold learning to both training and testing data; the other employs manifold learning on training data and the kernel-based out-of-sample extension methods on testing data. The former can obtain more accurate manifold coordinates, while the latter is suitable when there are large quantities of testing data. Because the data sets in this study are small enough for manifold learning methods to handle directly, we employed the first strategy. However, for supervised methods, we must use the out-of-sample extension method for testing data.

3 Remotely Sensed Data for Comparative Experiments

Four hyperspectral remotely sensed data sets which are commonly used to evaluate classification methods were analyzed in this comparative study. The data were acquired by sensors covering approximately the same range of wavelengths in the visible and short wave infrared portions of the spectrum in 10-nm spectral bands at spatial resolutions from 2 to 30 m. Spectral signatures of classes are complex and often overlapping, and spatial patterns include natural vegetation and agricultural fields in both fragmented and regular geometric patterns. Site and class related information is listed in Table 2. Important characteristics of each data set are summarized in the remainder of this section.

3.1 Botswana Hyperion Data (BOT)

Hyperion data with 9 identified classes of complex natural vegetation were acquired over the Okavango Delta, Botswana, in May 2001. The general class

Table 2 Class labels and number of labeled samples

BOT		KSC		ACRE	
ID	Class name	ID	Class name	ID	Class name
1	Water (158)	1	Scrub (761)	1	Corn—heavy till A (116)
2	Floodplain (228)	2	Willow swamp (243)	2	Corn—heavy till B (116)
3	Riparian (237)	3	Cabbage palm hammock (256)	3	Soybeans—Med till A (116)
4	Firescar (178)	4	Cabbage palm/oak (252)	4	Soybeans—Med till B1 (116)
5	Island interior (183)	5	Slash pine (161)	5	Corn—no till A (116)
6	Woodlands (199)	6	Oak/broadleaf hammock (229)	6	Corn—no till B (116)
7	Savanna (162)	7	Hardwood swamp (105)	7	Soybeans—no till (116)
8	Short mopane (124)	8	Graminoid marsh (431)	8	Grass (116)
9	Exposed soils (111)	9	Spartina marsh (520)	9	Soybeans—combined (116)
		10	Cattail marsh (404)	10	Soybeans—Med till B2 (116)
		11	Salt marsh (419)		
		12	Mud flats (503)		
		13	Water (927)		

IND PINE	
ID	Class name
3	Corn—min till (834)
12	Soybeans—heavy till (614)

groupings include seasonal swamps, occasional swamps, and woodlands. Signatures of several classes are spectrally overlapped, typically resulting in poor classification accuracies. After removing water absorption, noisy, and overlapping spectral bands, 145 bands were used for classification experiments. Classification results are reported for all 9 classes and separately for Class 3 (riparian) and Class 6 (woodlands), which are particularly difficult to discriminate, with the goal of illustrating both the capability of the dimensionality reduction methods in a general *c*-class setting and for specific classes of interest.

3.2 Kennedy Space Center AVIRIS Data (KSC)

Airborne hyperspectral data were acquired by the NASA AVIRIS sensor at 18-m spatial resolution over Kennedy Space Center during March 1996. Noisy and water absorption bands were removed, leaving 176 features for 13 wetland and upland classes of interest. Cabbage Palm Hammock (Class 3) and Broad Leaf/Oak Hammock (Class 6) are upland trees; Willow Swamp (Class 2), Hardwood Swamp

(Class 7), Graminoid Marsh (Class 8) and Spartina Marsh (Class 9) are trees and grasses in wetlands. Their spectral signatures are mixed and often exhibit only subtle differences. Results for all 13 classes and for these "difficult" classes are reported for the manifold learning experiments.

3.3 Indian Pine AVIRIS Data (IND PINE)

Experiments included the Indiana Indian Pine 16-class data set acquired by the NASA AVIRIS sensor in June 1992 at 20-m spatial resolution. After removing water absorption bands, 200 bands were available for analysis. The scene is primarily comprised of agricultural fields with regular geometry, providing an opportunity to evaluate the impact of within-class variability at medium spatial resolution. The corn and soybean fields, which had been recently planted, exhibit differences related to tillage practices and soils (including soil moisture). Selected results are included for Class 3 (corn, min tillage) and Class 12 (soybeans, high tillage) which are difficult to discriminate during the early part of the growing season.

3.4 ACRE ProspectTIR Data (ACRE)

Airborne hyperspectral data were collected by a ProspecTIR instrument at 2-m spatial resolution in November 2008 over the Agronomy Center for Research and Education (ACRE) farm operated by Purdue University. The agricultural research plots have 10 classes, which include corn, soybeans, and sorghum which had been harvested and tilled using various practices for research on crop yield, erosion, water quality, and carbon sequestration. Crop rows are visible in many of the fields, soil moisture varies within fields, and signatures represent mixtures of soil and remaining residue. Classification results obtained from 178 bands are reported for 10 classes, with class dependent results provided for class pair (3, 9). These results demonstrate the impact of intra-class spectral variability at very high spatial resolution on dimensionality reduction.

4 Experimental Results

All labeled data sets were randomly sampled to provide 50% training and 50% testing samples, with 20 replications of each experiment. The same data points for each of the four data sets were used for each dimensionality reduction method to provide consistent comparisons. A grid search was used to select parameters for the classification process. Extensive experiments were performed, and selected results are reported here to illustrate trends and demonstrate important characteristics of the manifold learning methods. Classification results are provided for the two global manifold learning methods (Isomap and KPCA) and the three local manifold learning methods (unsupervised and supervised LLE, LTSA, and LE), and compared to two linear dimensionality reduction methods (PCA and LDA). The base classifier was k-NN with $k = 1$ in all experiments. Results are also included for the original data, where similarities are based Euclidean distance and spectral angle [36, 37]. In the implementation of the manifold learning methods, dimensionality reduction was performed using spectral angle similarity in the nearest neighbor search on the original data, and Euclidean distance was employed thereafter in the transformed manifold space, as the physical relationship with spectral angle is lost in the transformation.

The parameters and computational complexity of the 10 dimensionality reduction methods are listed in Table 3. We consider the complexity of the most demanding step, which is eigen-decomposition of the $m \times m$ matrix for PCA and LDA, eigen-decompostion of $n \times n$ matrix for Isomap and KPCA, and identification of the k nearest neighbors for the local methods (unsupervised and supervised). It should be noted that for SLML, the manifold coordinates of the training data are obtained without calculation. Therefore, searching neighbors for the testing data is the most computationally demanding step, which requires mn_1n_2, where n_1 and n_2 are the number of training data and testing data, respectively.

Experiments were conducted across all methods to investigate performance related to the dimensionality of the resulting data. Classification accuracies are provided in Sect. 4.1 and 4.2 as a function of dimension for several c-class experiments and for some example classes which are difficult to discriminate. Additionally, plots of transformed data are provided in Sect. 4.3 to illustrate the impact of linear and nonlinear dimensionality reduction methods on class separation in various band combinations.

4.1 Performance of Dimensionality Reduction Methods (DR) for BOT Hyperion Data

Experimental results for Botswana data are included to illustrate performance of the various dimensionality reduction methods for all 9 classes and for Classes 3

Table 3 Classifier parameter and computational complexity

Method	Parameter	Computational complexity
PCA	p	$O(m^3)$
LDA	None	$O(m^3)$
Isomap	k; p	$O(n^3)$
KPCA	p	$O(n^3)$
LLE	k; p	$O(mn^2)$
LTSA	k; p; p'	$O(mn^2)$
LE	k; p; σ	$O(mn^2)$
SLLE	k	$O(mn_1n_2)$
SLTSA	k; p'	$O(mn_1n_2)$
SLE	k; σ	$O(mn_1n_2)$

(riparian) and 6 (woodlands). Figures 1–6 contain plots of experimental results obtained from the 1-NN classifier, over a range of the parameter values for the respective methods. The mean values of the Kappa statistics are plotted as a function of dimension for each nonlinear learning method and for PCA, and compared to those obtained using the original full dimensional data with both Euclidean (EU) and Spectral Angle Mapper (SAM). Figures 1–5 show results for unsupervised global and local manifold learning methods where dimensionality is variable, and Fig. 6 shows results of supervised local manifold learning as a function of k, as dimensionality is fixed $p = 5 \sim 100$. For KPCA, we use the radial basis function (RBF) kernel. For global methods dimension $p = 1 \sim 20$, while for local methods. Several trends are suggested by these results. Similar results were observed in experiments using the KSC, IND PINE, and ACRE data sets.

- Figure 1 indicates that KPCA and PCA have very similar performance for the BOT data. While neither dimensionality reduction method is able to achieve the accuracy of the 1-NN classifier, obtained using the full 145 band data set, the differences are larger for the 2-class problem than for the full 9-class experiments. Results are insensitive to the parameter σ, except for $\sigma = 1$. High dimensional inputs are needed to compensate for the small value of σ. Negligible improvement in accuracy is achieved with more than 5 bands in the Class (3, 6) experiments or more than 7 bands for the 9-class studies.
- Figure 2 illustrates results for Isomap, where it outperforms both PCA and 1-NN (EU) for the 2-class problem, but is somewhat worse for the full 9-class experiments. The 1-NN classifier with SAM achieves the highest accuracies for both sets of experiments. Isomap is generally able to achieve higher accuracies at lower dimensions than PCA. With proper values of k, Isomap achieves higher accuracies than PCA for the 2-class problem and similar results for the 9-class problem. Results are affected by the value of k at low values of p, but asymptotic accuracies are insensitive, except when k is extremely large in the 9-class problem.

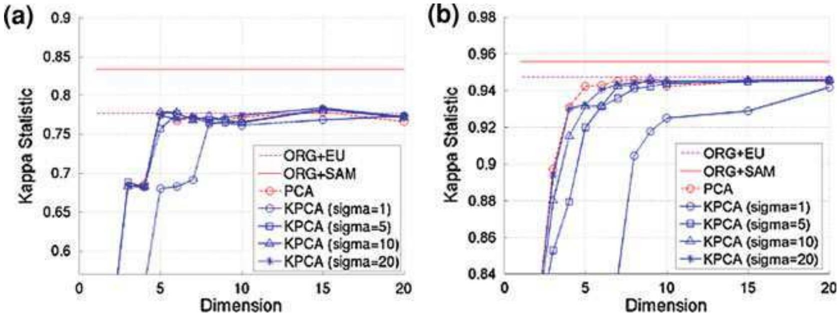


Fig. 1 Classification results using 1-NN with KPCA DR data **a** BOT Classes 3 and 6, **b** BOT Classes 1–9

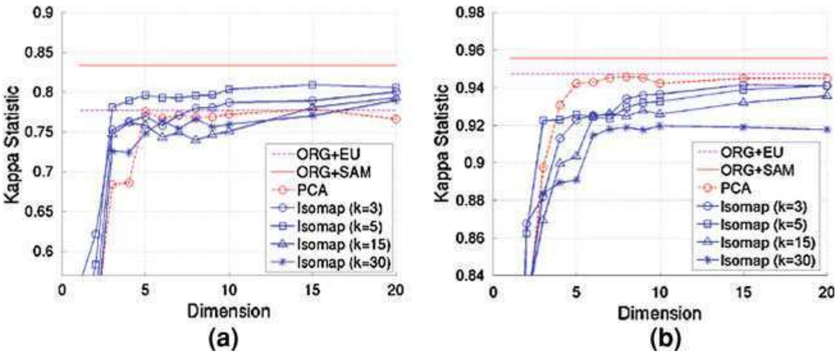


Fig. 2 Classification results using 1-NN with ISOMAP DR data **a** BOT Classes 3 and 6, **b** BOT Classes 1–9

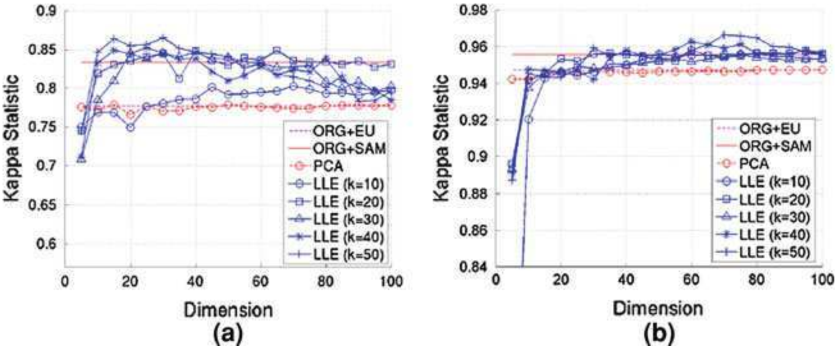


Fig. 3 Classification results using 1-NN with LLE DR data **a** BOT Classes 3 and 6, **b** BOT Classes 1–9

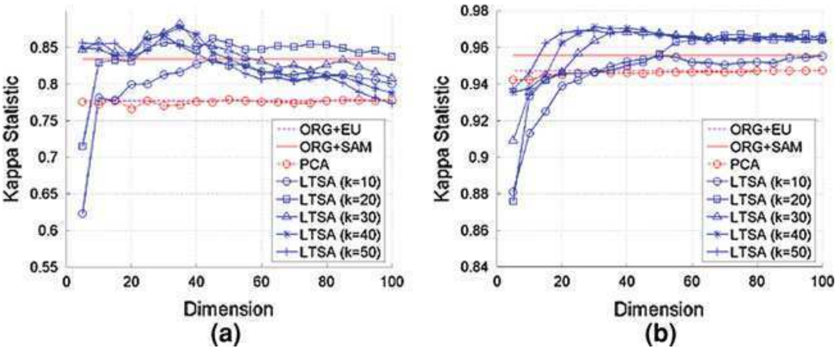


Fig. 4 Classification results using 1-NN with LTSA DR data **a** BOT Classes 3 and 6, **b** BOT Classes 1–9

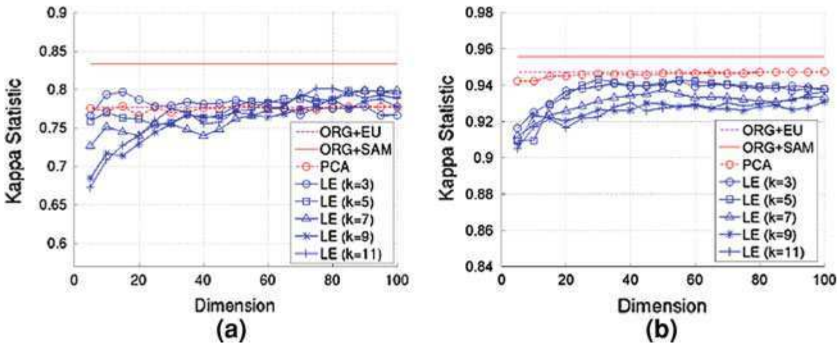


Fig. 5 Classification results using 1-NN with LE DR data **a** BOT Classes 3 and 6, **b** BOT Classes 1–9

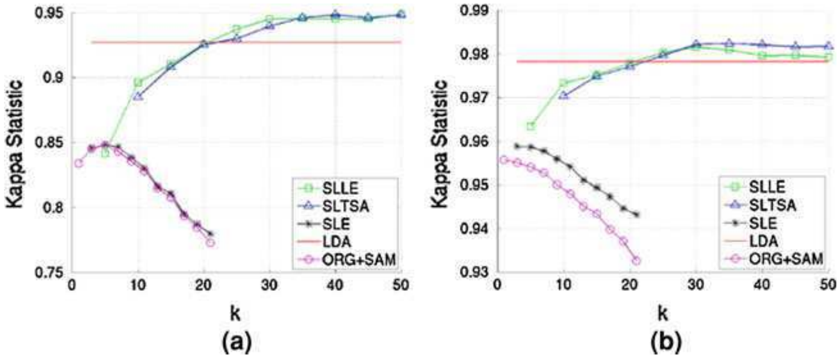


Fig. 6 Classification results using 1-NN with SLML DR data **a** BOT Classes 3 and 6, **b** BOT Classes 1–9

- Figure 3 indicates that LLE outperforms PCA and 1-NN (EU and SAM) for both sets of BOT data. For the 2-class problem, the highest accuracy is achieved for $p = 15$, while larger values are required for the more complex 9-class problem. LLE achieves higher accuracies for the Class (3, 6) experiments than both global methods, and marginally better results than global methods for the 9-class problem, but requires significantly larger $p(> 20)$ to obtain these results. For the 2-class problem LLE performance degrades when p is very large (> 40) for some values of k.
- Performance of LTSA shown in Fig. 4 is better than either PCA or 1-NN. For the 2-class problem, good performance can be achieved for $p = 5$ with some large values of k. Similarly for the 9-class problem, LTSA achieves higher accuracies than PCA and the 1-NN methods, but requires more bands. Larger values of dimension p are required to compensate for smaller numbers of local neighbors k, although results are relatively insensitive to $p(> 20)$.

- LE outperforms PCA and 1-NN (EU) for the 2-category classification, but has lower accuracies for all 9-class experiments. The best results are obtained for $k = 3$, with small values of p being adequate for the 2-class problem, and larger $p(> 20)$ being required for the 9-class experiments.
- The performance of the supervised learning methods, which have fixed dimension, is shown in Fig. 6, where the Kappa statistic is plotted versus the parameter k. SLLE and SLTSA have better performance than LDA for $k > 25$. SLE and k-NN (applied to the original data with spectral angle similarity) have lower accuracies for both the 2-class and 9-class experiments, with the best performance for $k < 5$, after which it steadily degrades for both classifiers. The optimal $''k''$ is much smaller for SLE than the other two methods, SLLE and SLTSA. Note: Recall that $''k''$in ORG + SAM is the number of neighbors for the k-NN classifier, whereas $''k''$ in the other supervised methods (SLE, SLLE, SLTSA) relates to the number of neighbors searched during manifold learning, and does not relate to k-NN.

4.2 Comparison of DR Methods for BOT, KSC, IND PINE, and ACRE Sites

Results obtained by each of the methods were also evaluated across sites. Accuracies from all methods are shown in Fig. 7, where optimal parameters were used for each method, and results are shown as function of dimension, where $p \leq 40$. Table 4 contains a summary of mean classification accuracies and associated standard deviations obtained from Fig. 7 with the best value of p. The following were observed:

- With an adequate number of features, accuracies of the best DR methods are always at least as high as 1-NN applied to the original data (EU and SAM).
- When applied to the original data, the SAM-based similarity measure consistently yields higher accuracies than Euclidean distance for the k-NN classifier.
- Results obtained from PCA and KPCA are very similar for all data sets.
- Isomap generally yields better results than both KPCA and LE, achieving good results using a small number of features. The only exceptions are the full class experiments for BOT and ACRE, where Isomap has worse results for $p > 3$.
- For a large number of features, the LTSA and LLE local manifold learning methods are consistently better than both global DR methods, while LE is not.
- Accuracies for difficult classes exhibit more variability than results from the full set of classes at each site. Supervised manifold learning methods mitigate this effect, yielding higher accuracies with smaller standard deviations (Table 4).
- In general, both local and global manifold learning methods are better able to capture the nonlinear structure of the associated manifolds and achieve higher

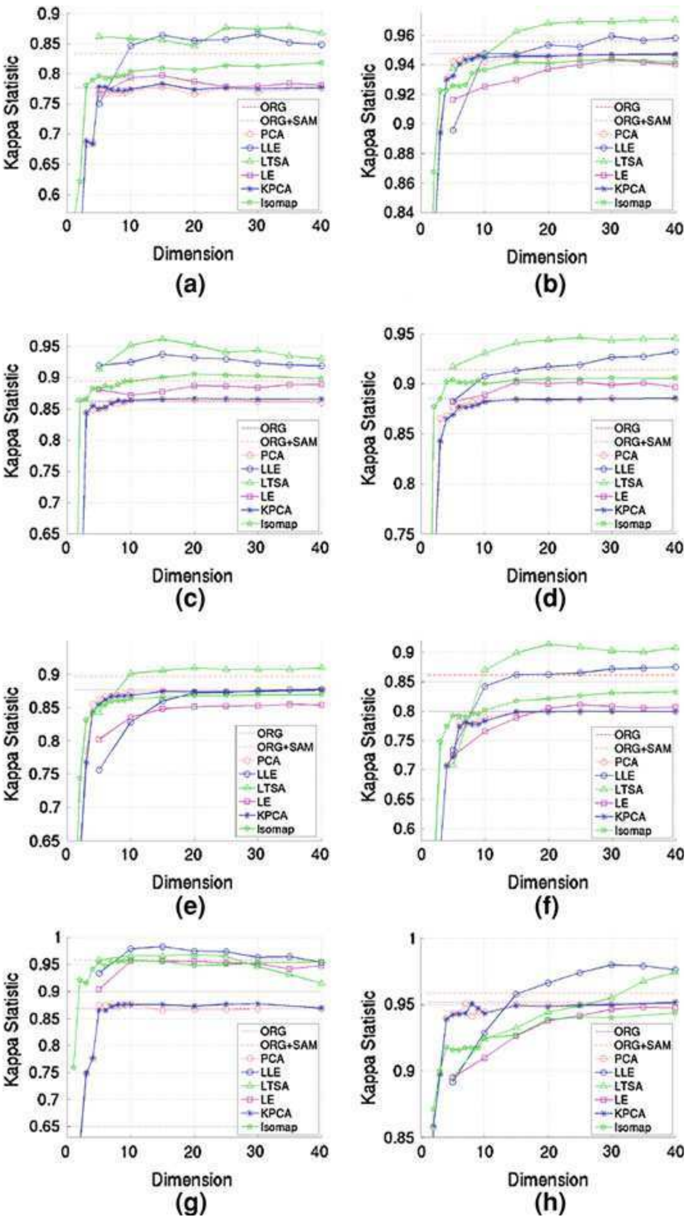


Fig. 7 Comparison of manifold learning methods **a** BOT Classes 3 and 6, **b** BOT Classes 1–9, **c** KSC Classes 3 and 6, **d** KSC Classes 2, 7, 8, and 9, **e** KSC Classes 1–13, **f** IND PINE Classes 3 and 12, **g** ACRE Classes 3 and 9, and **h** ACRE Classes 1–10

Table 4 Overall classification accuracies [Kappa statistic (SD)]

Classes	BOT				KSC				ACRE				IND PINE			
	3, 6		1 ~ 9		3, 6		2, 7, 8, 9		1 ~ 13		3, 9		1 ~ 10		3, 12	
ORG-EU	77.7	(3.70)	94.7	(0.70)	86.2	(3.00)	88.5	(1.50)	87.7	(0.70)	86.8	(4.50)	95.2	(0.80)	79.9	(1.90)
ORG-SAM	83.4	(3.50)	95.6	(0.60)	89.5	(2.30)	91.4	(3.20)	89.6	(0.60)	95.9	(2.50)	95.9	(0.90)	86.2	(1.90)
PCA	77.9	(3.90)	94.6	(0.80)	86.6	(3.00)	88.6	(2.70)	87.6	(0.70)	87.8	(4.30)	95.2	(0.80)	80.1	(1.70)
KPCA	78.4	(3.90)	94.7	(0.70)	86.7	(2.90)	88.6	(1.50)	87.6	(0.70)	87.8	(4.10)	95.2	(0.80)	80.0.	(1.70)
Isomap	81.8	(3.00)	94.4	(0.60)	90.5	(2.50)	90.6	(1.20)	86.9	(0.40)	95.8	(2.60)	94.4	(0.80)	83.2	(1.70)
LLE	86.5	(2.10)	95.9	(0.60)	93.7	(1.80)	93.2	(2.50)	87.8	(0.60)	98.4	(2.40)	98	(0.70)	87.5	(1.20)
LTSA	87.7	(3.70)	97.1	(0.60)	96.1	(1.70)	94.7	(1.10)	91	(0.60)	96.9	(2.80)	97.4	(0.80)	91.4	(1.30)
LE	79.7	(3.80)	94.3	(0.70)	88.9	(2.60)	90.2	(1.00)	85.5	(0.70)	95.7	(4.50)	94.8	(0.70)	81.1	(1.80)
LDA	92.7	(1.90)	97.8	(0.50)	98.5	(1.20)	94	(1.10)	93.5	(0.50)	99.9	(0.40)	98.7	(0.50)	91.7	(1.40)
SLLE	94. 9	(2.20)	98.2	(0.60)	96.9	(1.90)	94.5	(1.90)	92.6	(0.50)	99.5	(1.10)	99	(0.40)	95.3	(1.00)
SLTSA	94.8	(2.10)	98.3	(0.60)	99.3	(0.90)	94.9	(1.90)	93.3	(0.40)	100	0.00	99.2	(0.40)	95.5	(1.00)
SLE	84.8	(3.70)	95.9	(0.70)	90.9	(2.90)	92.2	(1.20)	90.4	(0.50)	93.5	(2.40)	96.7	(0.70)	85.5	(2.00)

accuracies for small numbers of classes than for problems with large numbers of categories. This is due to the complexity of the manifolds represented in problems with many disparate classes. The data for similar classes actually reside on sub-manifolds which are more easily characterized with fewer features than the full *c*-class problems.

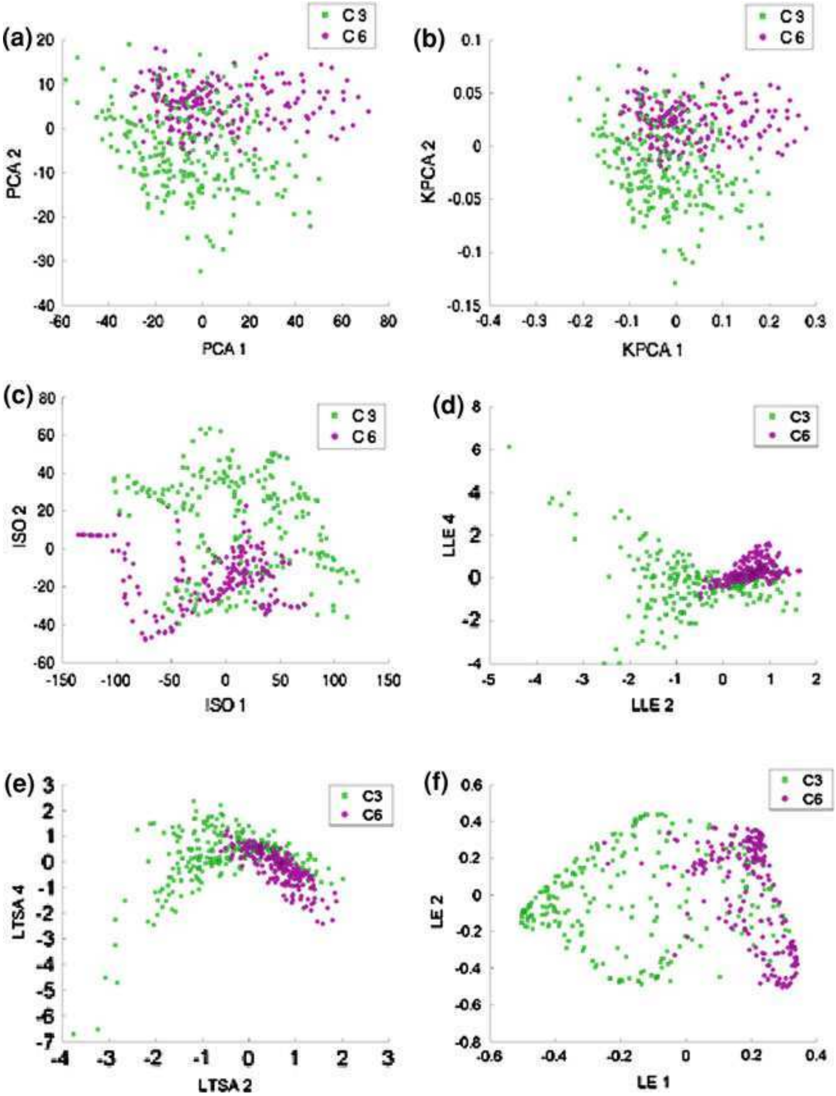


Fig. 8 Manifold coordinates of BOT Classes 3 and 6 **a** PCA, **b** KPCA, **c** Isomap, **d** LLE, **e** LTSA, **f** LE, **g** SLLE, **h** SLTSA, **i** SLE, **j** LDA

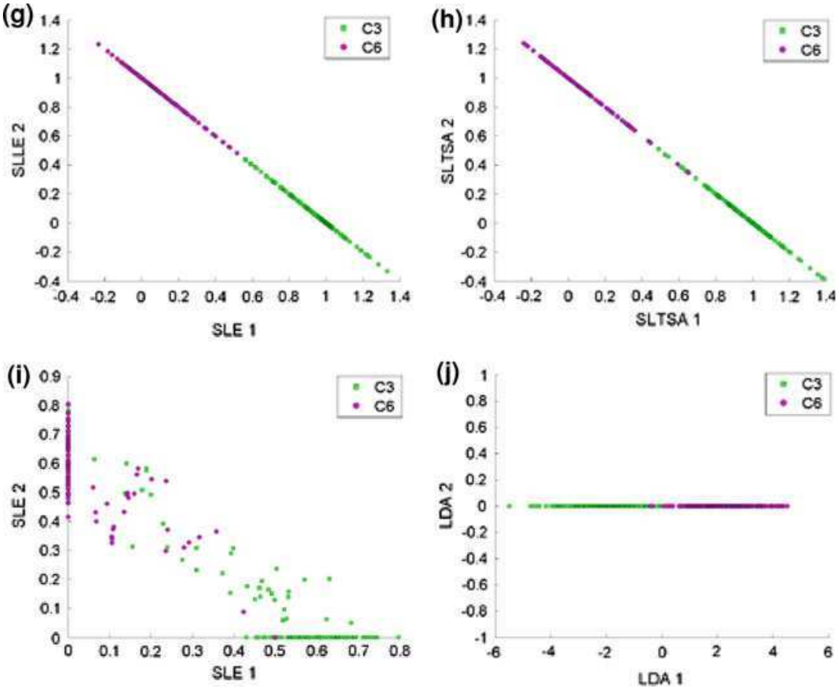


Fig. 8 (continued)

4.3 Manifold Coordinates for DR Methods

To better understand the performance of the DR methods, plots of manifold coordinates are provided for the 9-class BOT data, with optimal parameters for each set of results. The following discussion summarizes observations. Plots of selected coordinates for each DR method are shown in Fig. 8 for the BOT class (3, 6) data. Figures 9–16 contain plots of the first eight coordinates (in pairs) for each method applied to the BOT 9-class data. The similarity between PCA and KPCA is consistent with the classification accuracies. Nonlinear structures are clearly exhibited, and are quite different for the various methods. The superior performance of LDA and the supervised local manifold learning methods for the 2-class problem is clear.

- In Fig. 9a–d the first eight bands of PCA demonstrate that the information contributing to classification is primarily contained in the first four bands. Figure 10a–d shows that the KPCA transformed data have similar characteristics to that of PCA.
- Both the nonlinearity and the potential contribution of larger numbers of bands to classification are exhibited in the Isomap plots shown in Fig. 11.

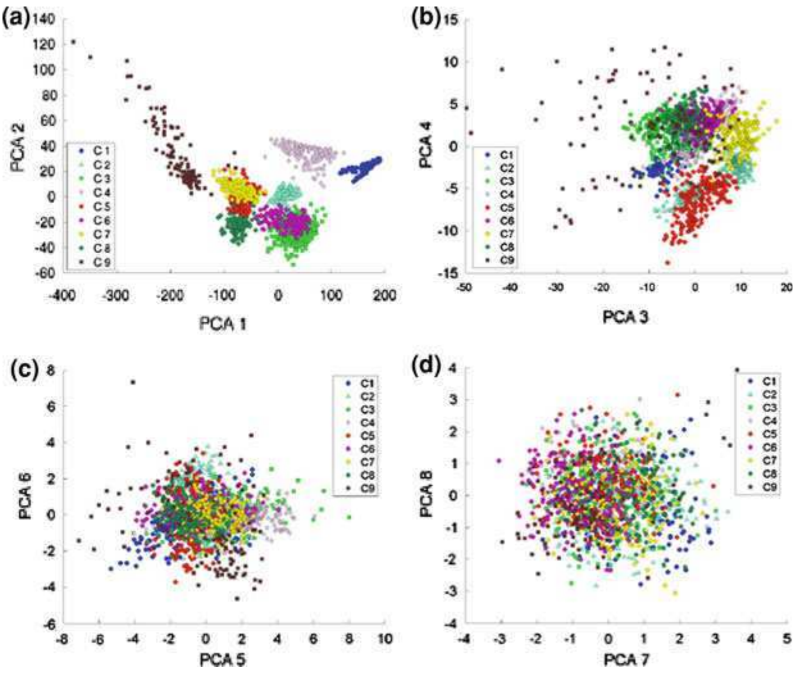


Fig. 9 PCA DR results of BOT Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

- Figure 12 indicates that many bands of LLE transformed data are meaningful for classification. For example, class 3 (riparian) and class 6 (woodlands), the most difficult class pair, can be well separated using bands 7 and 8 (Fig. 12d). Class 5 (island interior) and class 7 (savannah) can be distinguished by band 4 (Fig. 12b).
- The first eight bands for the LTSA transformation also illustrate the contribution of larger numbers of bands for improved classification, with band 7 being meaningful for classifying class pair (5, 7) and band 8 providing good separation for class pair (3, 6) (Fig. 13d) .
- Structures provided by LE (Fig. 14) are significantly different from those of LLE and LTSA. Similar to other nonlinear methods, difficult classes are not well separated by low order bands. Class pairs (3, 6) and (5, 7) are well distinguished using bands 7 and 8 (Fig. 14d).
- Results for the supervised local manifold learning methods are similar, so plots are provided only for SLLE. Here, each band clearly separates one class from the rest. For example, in band 1 (Fig. 15a), only points from class 1 have large values while points from all other classes have small values. Every band is thus equally useful for classification.

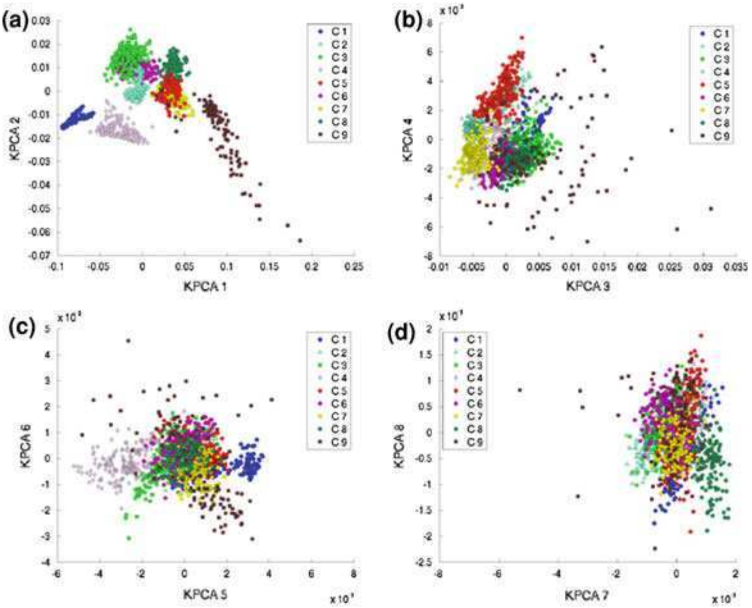


Fig. 10 KPCA DR results of BOT Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

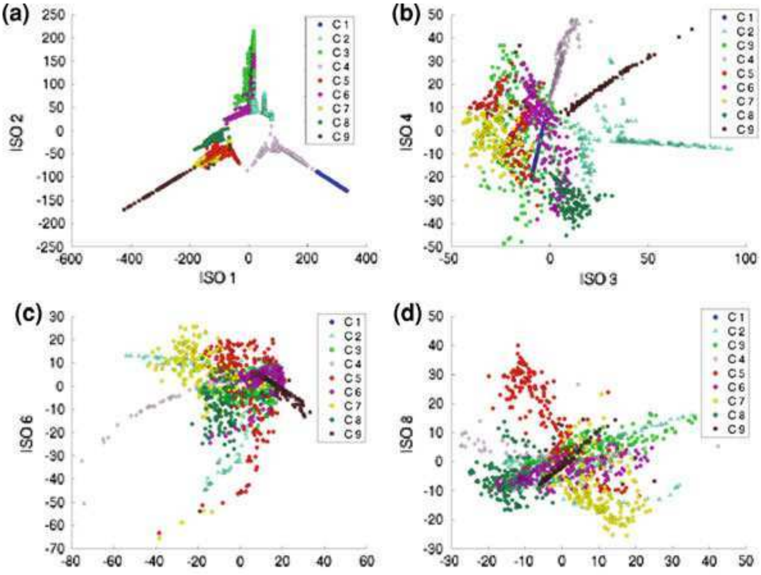


Fig. 11 Isomap DR results of BOT Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

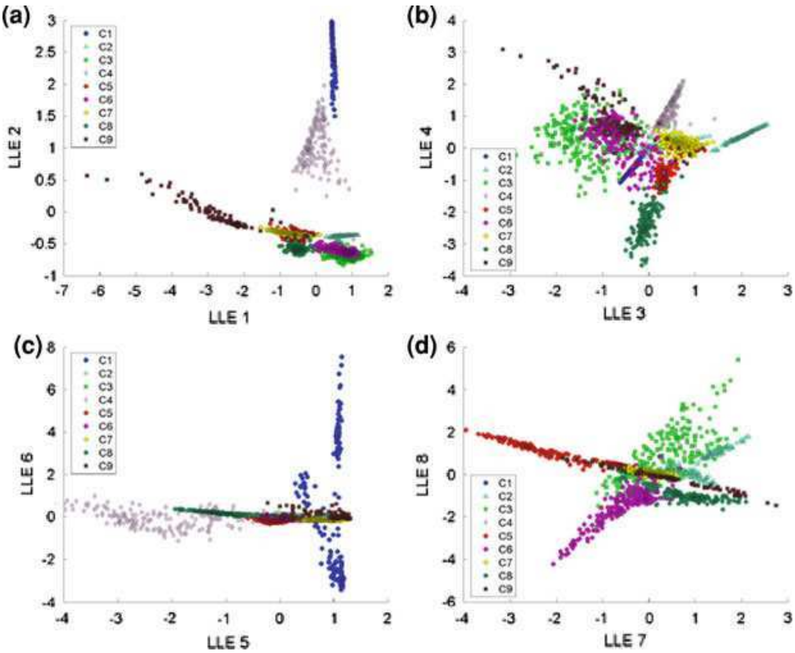


Fig. 12 LLE DR results of BOT Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

- Results of LDA (Fig. 16), the supervised linear transformation included for comparison, indicate that the first 4 bands are primarily useful for discrimination of the BOT data.

5 Summary and Conclusions

The goal of this study was to investigate the characteristics of nonlinear manifold learning methods for dimensionality reduction in classification of hyperspectral data. The investigation sought to better understand the performance of nonlinear dimensionality reduction on real world data with characteristics commonly observed in classification of multi-class land cover problems. The study was also motivated by the greater complexity of hyperspectral data relative to example data sets typically used machine learning examples, which could impact the value of these approaches for analysis of remote sensing data. While achieving the best possible classification accuracies was not a focus of the study, results should inform other investigations for which that may be the goal.

An extensive empirical study was conducted, where the most common global and local manifold learning methods were implemented in conjunction with a nearest neighbor classifier for a range of parameter settings. While results are specific to the data sets used in this study, several trends emerged which may generalize to other multi-class hyperspectral data sets.

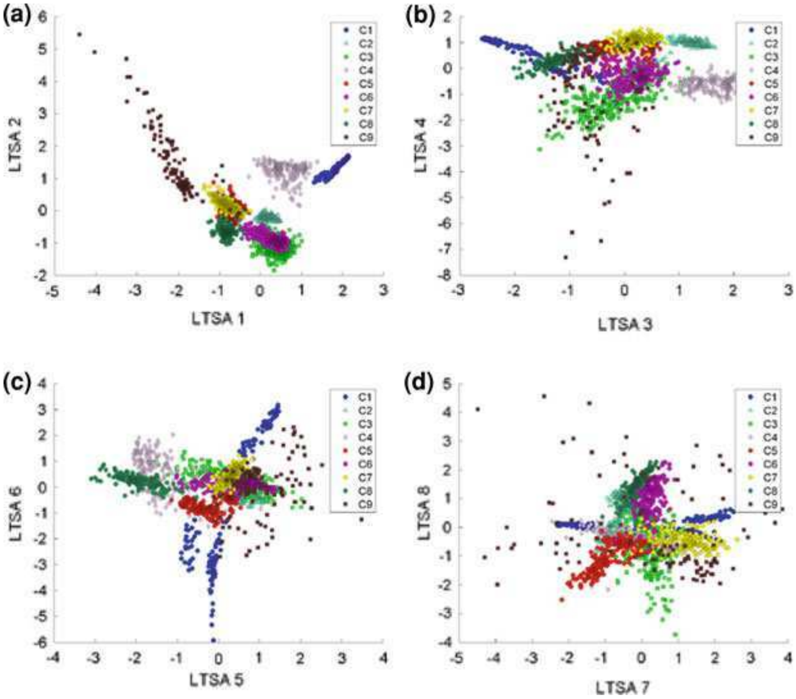


Fig. 13 LTSA DR results of BOT Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

- Global manifold learning methods capture the overall structure of these data with fewer bands than local manifold learning methods.
- For complex geometry, local manifold methods represent the data better than the global methods evaluated herein, but require more features.
- Isomap generally outperforms KPCA with the RBF kernel as a global learning method, and for these data sets KPCA offers little advantage relative to PCA.
- Among the local manifold learning methods, LLE and LTSA consistently yield better results than LE. Isomap also outperforms LE.
- When implemented with the best parameters, LLE and LTSA have the best overall performance, with higher mean accuracies and greatly reduced standard deviations for difficult classes. They also have good capability over a wide range of parameter values for both two-category data and problems with many classes.
- Supervised local manifold learning methods are advantageous in the classification framework, if training data are reliable.
- Both global and local methods have better performance for small numbers of classes than for large *c*-class problems. This is indicative of the difficulty of recovering structure for complex data, and possibly indicates that hierarchical or pairwise approaches may be advantageous for dimensionality reduction within a classification framework.

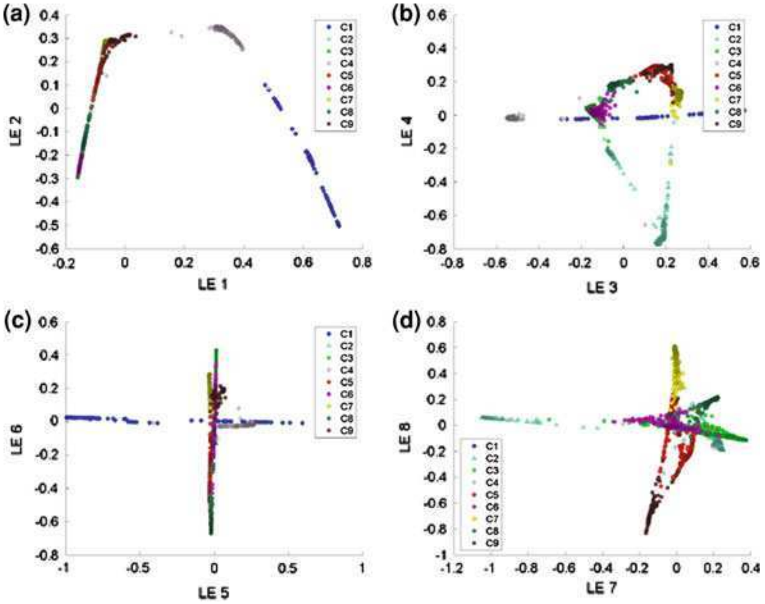


Fig. 14 LE DR results of BOT Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

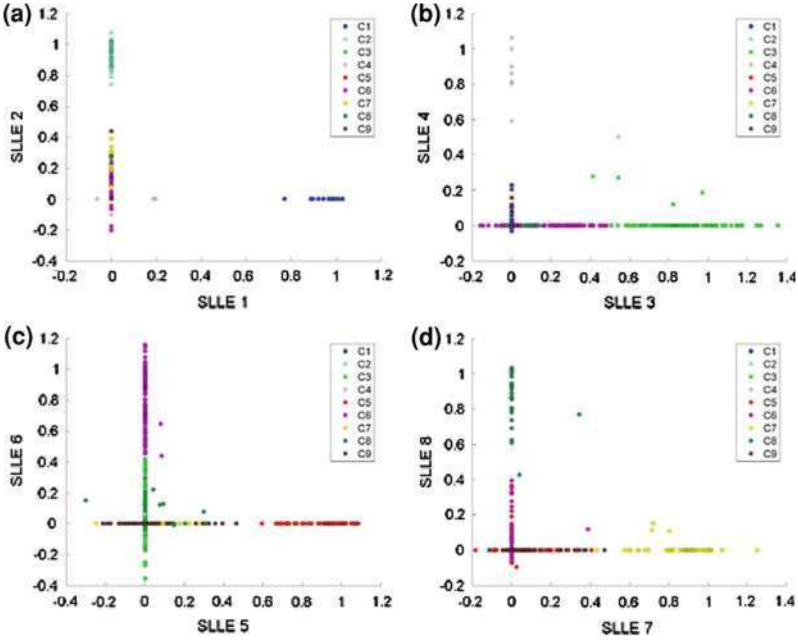


Fig. 15 SLLE DR results of BOT May 2001 Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

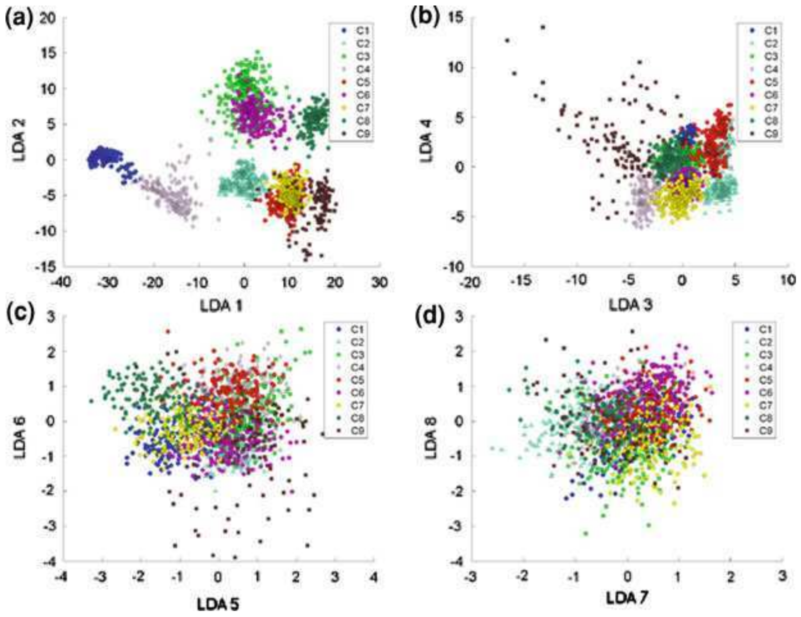


Fig. 16 LDA DR results of BOT Classes 1–9 **a** bands 1 and 2, **b** bands 3 and 4, **c** bands 5 and 6, **d** bands 7 and 8

Overall, nonlinear manifold learning methods are promising as dimensionality reduction methods. However, PCA can outperform manifold methods if methods are implemented without regard to parameter settings and dimensionality selection. Computational complexity of nonlinear methods suggests that these approaches should be used judiciously and that their greatest advantage is realized for discriminating difficult classes. Further, investigation of multi-manifold representations may have merit for supervised classification problems. The interaction of dimensionality reduction and classifier design should be explored, particularly with respect to generalization. Further investigation with additional hyperspectral data sets is also warranted.

Acknowledgment This work was conducted at the Purdue Laboratory for Applications of Remote Sensing (LARS) and supported by the National Science Foundation Grant (III-CXT: 0705836)

References

1. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)

2. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by local linear embedding. Science **290**(5500), 2323–2326 (2000)
3. Donoho, D.L., Grimes, C.: Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. USA **100**(10), 5591–5596 (2003)
4. Saul, L.K., Roweis, S.T.: Think globally, fit locally: unsupervised learning of low dimensional manifolds. J. Mach. Learn Res. **4**, 119–155 (2003)
5. Agrafiotis, D.K.: Stochastic proximity embedding. J. Comput. Chem. **24**(10), 1215–1221 (2003)
6. Bachmann, C.M., Ainsworth, T.L., Fusina, R.A.: Exploiting manifold geometry in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **43**(3), 441–454 (2005)
7. Bachmann, C.M., Ainsworth, T.L., Fusina, R.A.: Improved manifold coordinate representations of large-scale hyperspectral scenes. IEEE Trans. Geosci. Remote Sens. **44**(10), 2786–2803 (2006)
8. Bachmann, C.M., Ainsworth, T.L., Fusina, R.A., Montes, M.J., Bowles, J.H., Korwan, D.R., et al.: Bathymetric retrieval from hyperspectral imagery using manifold coordinate representations. IEEE Trans. Geosci. Remote Sens. **47**(3), 884–897 (2009)
9. Mohan, A., Sapiro, G., Bosch, E.: Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images. IEEE Geosci. Remote Sens. Lett. **4**(2), 206–210 (2007)
10. Han, T., Goodenough, D.G.: Investigation of nonlinearity in hyperspectral imagery using surrogate data methods. IEEE Trans. Geosci. Remote Sens. **46**(10), 2840–2847 (2008)
11. Chen, Y., Crawford, M.M., Ghosh, J.: Improved nonlinear manifold learning for land cover classification via intelligent landmark selection. In: IEEE Int. Geosci. Remote Sens. Symp., Denver, Colorado, USA, pp. 545–548, July 2006
12. Kim, W., Chen, Y., Crawford, M.M., Tilton, J.C., Ghosh, J.: Multiresolution manifold learning for classification of hyperspectral data. In: IEEE Int. Geosci. Remote Sens. Symp., Barcelona, Spain, pp. 3785–3788, July 2007
13. Kim, W., Crawford, M.M., Ghosh, J.: Spatially adapted manifold learning for classification of hyperspectral imagery with insufficient labeled data. In: IEEE Int. Geosci. Remote Sens. Symp., Boston, Massachusetts, USA, vol. 1, pp. I213–I216, July 2008
14. Kim, W., Crawford, M.M.: A novel adaptive classification method for hyperspectral data using manifold regularization kernel machines. In: First Workshop Hyperspectral Image Signal Process Evol. Remote Sens., Grenoble, France, pp. 1–4, August 2009
15. Crawford, M.M., Kim, W.: Manifold learning for multi-classifier systems via ensembles. Mult. Classif. Syst. **5519**, 519–528 (2009)
16. He, J., Zhang, L., Wang, Q., Li, Z.: Using diffusion geometric coordinates for hyperspectral imagery representation. IEEE Geosci. Remote Sens. Lett. **6**(4), 767–771 (2009)
17. Fauvel, M., Chanussot, J., Benediktsson, J.A.: Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas. EURASIP J. Adv. Signal Process (2009). doi:10.1155/2009/783194
18. Ma, L., Crawford, M.M., Tian, J.W.: Local manifold learning based *k*-nearest-neighbor for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens, **48**(11), 4099–4109 (2010)
19. Ma, L., Crawford, M.M., Tian, J.W.: Anomaly detection for hyperspectral images based on robust locally linear embedding. J. Infrared Millimeter Terahertz Waves **31**(6), 753–762 (2010)
20. van der Maaten, L.J.P., Postma, E., Herik, H.J.: Dimensionality reduction: a comparative review. Tech. Rep. http://homepage.tudelft.nl/19j49/Publications.html (2009). Accessed Oct 2009
21. Yan, S., Dong, X., Zhang, B., Zhang, H.J.: Graph embedding: a general framework for dimensionality reduction. In: IEEE Comput. Soc. Conf. Comp. Vis. Pattern Recognit. (CVPR 2005), San Diego, CA, USA, vol. 2, pp. 800–837, June 2005
22. Yan, S., Dong, X., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 40–51 (2007)

23. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numer. Math. **1**(1), 269–271 (1959)
24. de Silva, V., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In: Adv. Neural Inf. Process Syst., Hyatt Regency, Vancouver, BC, Canada, vol. 15, pp. 713–720, December 2002
25. Chen, Y., Crawford, M.M., Ghosh, J.: Applying nonlinear manifold learning to hyperspectral data for land cover classification. In: IEEE Int. Geosci. Remote Sens. Symp., Seoul, Korea, pp. 4311–4314, June 2005
26. Schölkopf, B., Smola, A.J., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**(5), 583–588 (1998)
27. Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J. Sci. Comput. **26**(1), 313–338 (2004)
28. Wang, J.: Improve local tangent space alignment using various dimensional local coordinates. Neurocomputing **71**(16–18), 3575–3581 (2008)
29. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
30. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A kernel view of the dimensionality reduction of manifolds. In: Int. Conf. Mach. Learn., ACM, Banff, Alberta, Canada, vol. 69, pp. 369–376, August 2004
31. Bengio, Y., Delalleau, O., Roux, N.L., Paiement, J.F., Vincent, P., Ouimet, M.: Learning eigenfunctions links spectral embedding and kernel PCA. Neural Comput. **16**(10), 2197–2219 (2004)
32. de Ridder, D., Duin, R.P.W.: Locally linear embedding for classification. Tech. Rep. (PH-2002-01), Delft University of Technology, Delft, The Netherlands, 2002
33. Li, H.Y., Teng, L., Chen, W.B., Shen, I.F.: Supervised learning on local tangent space. Lect. Notes Comput. Sci. **3496**, 546–551 (2005)
34. Bengio, Y., Paiement, J.F., Vincent, P.: Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. In: Adv. Neural Inf. Process Syst., 16, Cambridge, MA, MIT Press, pp. 177–184, July 2003
35. Ma, L., Crawford, M.M., Tian, J.W.: Generalized supervised local tangent space alignment for hyperspectral image classification. Electron. Lett, **46**(7):497–498 (2010)
36. Yuhas, R.H., Goetz, A.F.H., Boardman, J.W.: Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In: Annu. JPL Airborne Geosci. Workshop, Pasadena, CA, vol. 1, pp. 147–149, June 1992
37. Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J., et al.: The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. **44**(2–3), 145–163 (1993)

Recent Developments in Endmember Extraction and Spectral Unmixing

Antonio Plaza, Gabriel Martín, Javier Plaza, Maciel Zortea and Sergio Sánchez

Abstract Spectral unmixing is an important task for remotely sensed hyperspectral data exploitation. The spectral signatures collected in natural environments are invariably a mixture of the pure signatures of the various materials found within the spatial extent of the ground instantaneous field view of the imaging instrument. Spectral unmixing aims at inferring such pure spectral signatures, called *endmembers*, and the material fractions, called *fractional abundances*, at each pixel of the scene. In this chapter, we provide an overview of existing techniques for spectral unmixing and endmember extraction, with particular attention paid to recent advances in the field such as the incorporation of spatial information into the endmember searching process, or the use of nonlinear mixture models for fractional abundance characterization. In order to substantiate the methods presented throughout the chapter, highly representative hyperspectral scenes obtained by different imaging spectrometers are used to provide a quantitative and comparative algorithm assessment. To address the computational requirements introduced by hyperspectral imaging algorithms, the chapter also

A. Plaza (✉), G. Martín, J. Plaza and S. Sánchez
Department of Technology of Computers and Communications, University of Extremadura, Avda. de la Universidad s/n, 10071 Caceres, Spain
e-mail: aplaza@unex.es

G. Martín
e-mail: gamahefpi@unex.es

J. Plaza
e-mail: jplaza@unex.es

S. Sánchez
e-mail: sersanmar@unex.es

M. Zortea
Department of Mathematics and Statistics, University of Tromso, 9037 Tromso, Norway
e-mail: maciel.zortea@hyperinet.eu

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_12,

includes a parallel processing example in which the performance of a spectral unmixing chain (made up of spatial–spectral endmember extraction followed by linear spectral unmixing) is accelerated by taking advantage of a low-cost commodity graphics co-processor (GPU). Combined, these parts are intended to provide a snapshot of recent developments in endmember extraction and spectral unmixing, and also to offer a thoughtful perspective on future potentials and emerging challenges in designing and implementing efficient hyperspectral imaging algorithms.

Keywords Hyperspectral imaging · Spectral unmixing · Endmember extraction · Neural networks · Intelligent training · Parallel processing · GPUs

1 Introduction

Spectral mixture analysis (also called *spectral unmixing*) has been an alluring exploitation goal from the earliest days of hyperspectral imaging [1] to our days [2, 3]. No matter the spatial resolution, the spectral signatures collected in natural environments are invariably a mixture of the signatures of the various materials found within the spatial extent of the ground instantaneous field view of the imaging instrument [4]. For instance, it is likely that the pixel collected over a vegetation area in Fig. 1 actually comprises a mixture of vegetation and soil. In this case, the measured spectrum may be decomposed into a combination of pure spectral signatures of soil and vegetation, weighted by areal coefficients that indicate the proportion of each *macroscopically* pure signature in the mixed pixel [5].

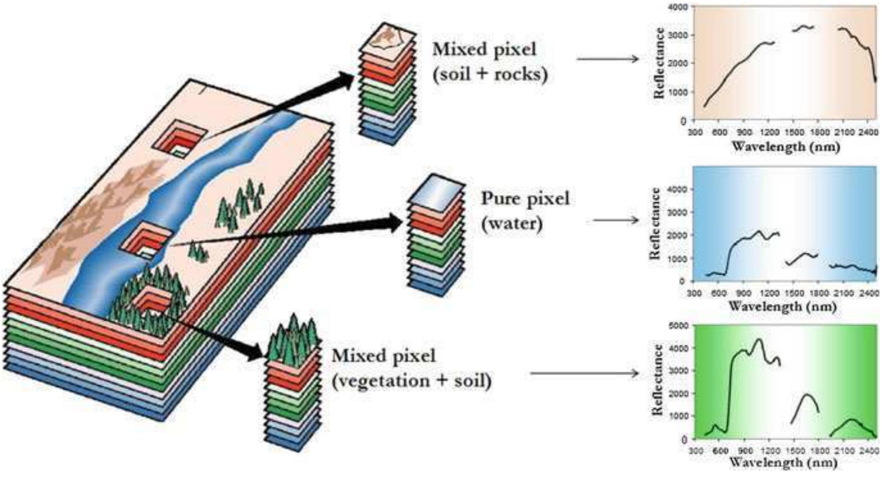


Fig. 1 The mixture problem in remotely sensed hyperspectral data analysis

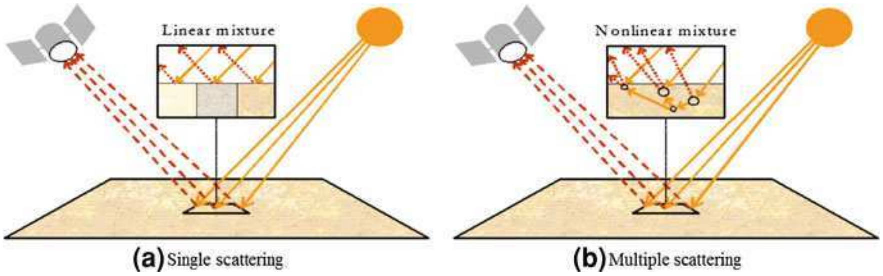


Fig. 2 Linear versus nonlinear mixture models: single versus multiple scattering

The availability of hyperspectral imagers with a number of spectral bands that exceeds the number of spectral mixture components [6] has allowed to cast the unmixing problem in terms of an over-determined system of equations in which, given a set of pure spectral signatures (called *endmembers*) the actual unmixing to determine apparent pixel *abundance fractions* can be defined in terms of a numerical inversion process [7].

A standard technique for spectral mixture analysis is *linear* spectral unmixing [8, 9], which assumes that the collected spectra at the spectrometer can be expressed in the form of a linear combination of endmembers weighted by their corresponding abundances. It should be noted that the linear mixture model assumes minimal secondary reflections and/or multiple scattering effects in the data collection procedure, and hence the measured spectra can be expressed as a linear combination of the spectral signatures of materials present in the mixed pixel (see Fig. 2a). Although the linear model has practical advantages such as ease of implementation and flexibility in different applications [10], *nonlinear* spectral unmixing may best characterize the resultant mixed spectra for certain endmember distributions, such as those in which the endmember components are randomly distributed throughout the field of view of the instrument [11, 12]. In those cases, the mixed spectra collected at the imaging instrument is better described by assuming that part of the source radiation is multiply scattered before being collected at the sensor (see Fig. 2b).

In this chapter, we provide an overview of existing techniques for spectral unmixing and endmember extraction, covering advances in both the linear and nonlinear mixture model, and with particular attention paid to recent advances in the field. The chapter is organized as follows. In Sect. 2, the chapter first focuses on the linear mixture model, introducing the formulation of the mixture problem under this model and further describing several classic approaches for endmember extraction (using different concepts) and linear spectral unmixing models (unconstrained, partially constrained and fully constrained). This section also covers recent developments in the linear mixture model by means of the incorporation of spatial information into the process of automatically extracting spectral endmembers from the image data, and further analyzes the impact of spatial information in the subsequent unmixing process. Section 3 addresses the nonlinear mixture model by means of neural network-based techniques aimed at learning the complexity of nonlinear mixtures by means of automatic training sample selection algorithms which are used in the framework of a supervised learning procedure to characterize other mixed signatures in the input data. Section 4 presents a quantitative and comparative assessment of the different techniques for spectral unmixing presented in this chapter (linear and nonlinear), using hyperspectral data sets obtained by different instruments, such as the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS), operated by NASA/JPL, and the Digital Airborne (DAIS 7915) and Reflective Optics System (ROSIS) imaging spectrometers, operated by DLR in Germany. Section 5 presents an implementation case study in which a spectral unmixing chain made up of a spatial–spectral endmember extraction algorithm followed by a linear (unconstrained) fractional abundance estimation technique are implemented in parallel using commodity graphics processing units (GPU). Finally, Sect. 6 concludes with some remarks and hints at plausible future research avenues.

2 Linear Spectral Unmixing

2.1 Problem Formulation

Let us assume that a remotely sensed hyperspectral scene with n bands is denoted by $\mathbf{I}$, in which the pixel at the discrete spatial coordinates (i, j) of the scene is represented by a vector $\mathbf{X}(i,j) = [x_1(i,j), x_2(i,j), \ldots, x_n(i,j)] \in \Re^n$, where $\Re$ denotes the set of real numbers in which the pixel's spectral response $x_k(i, j)$ at sensor channels $k = 1, \ldots, n$ is included. Under the linear mixture model assumption, each pixel vector in the original scene can be modeled using the following expression:

$$\mathbf{X}(i,j) = \sum_{z=1}^{p} \Phi_z(i,j) \cdot \mathbf{E}_z + \mathbf{n}(i,j), \tag{1}$$

where $\mathbf{E}_z$ denotes the spectral response of endmember z, $\Phi_z(i, j)$ is a scalar value designating the fractional abundance of the endmember z at the pixel $\mathbf{X}(i, j)$, p is the total number of endmembers, and $\mathbf{n}(i, j)$ is a noise vector. Two physical constrains are generally imposed into the model described in (1), these are the abundance non-negativity constraint (ANC), i.e., $\Phi_z(i,j) \geq 0$, and the abundance sum-to-one constraint (ASC), i.e., $\sum_{z=1}^{p} \Phi_z(i,j) = 1$ [8]. The solution of the fully constrained linear spectral mixture problem described in (1) relies on two major requirements:

1. A successful estimation of how many endmembers, p, are present in the input hyperspectral scene $\mathbf{I}$, and

2. The correct determination of a set $\mathbf{E} = \{\mathbf{E}_z\}_{z=1}^{p}$ of endmembers and their correspondent abundance fractions $\Phi(i,j) = \{\Phi_z(i,j)\}_{z=1}^{p}$ at each pixel $\mathbf{X}(i,j)$.

In order to address the first requirement, a successful technique in the literature has been the virtual dimensionality (VD) [13]. The VD concept formulates the issue of whether a distinct signature is present or not in each of the spectral bands as a binary hypothesis testing problem, where a so-called Neyman-Pearson detector is generated to serve as a decision-maker based on a prescribed P_F (i.e., false alarm probability). In light of this interpretation, the issue of determining an appropriate value for p is further simplified and reduced to setting a specific value of P_F. As will be shown in experiments, a suitable empirical choice is $P_F = 10^{-3}$ or $P_F = 10^{-4}$, where the method used in this work to estimate the VD is the one developed by Harsanyi, Farrand and Chang [13] (referred to as HFC method) later modified by including a noise whitening process as preprocessing to remove the second-order statistical correlation. The purpose is that signal sources can be decorrelated from the noise to achieve better signal detection. The resulting method will be referred to as the noise-whitened HFC (NWHFC). The second requirement for successful implementation of the linear mixture model (availability of endmember extraction and abundance estimation techniques) will be addressed in the following subsections.

2.2 Endmember Extraction

Over the last decade, several algorithms have been developed for automatic or semi-automatic extraction of spectral endmembers [9]. Classic techniques include the pixel purity index (PPI) [14], N-FINDR [15–17], iterative error analysis (IEA) [18], optical real-time adaptive spectral identification system (ORASIS) [19], convex cone analysis (CCA) [20], vertex component analysis (VCA) [21], and an orthogonal subspace projection (OSP) technique in [22]. Other advanced techniques for endmember extraction have been recently proposed [23–29], but none of them considers spatial adjacency. However, one of the distinguishing properties of hyperspectral data is the multivariate information coupled with a two-dimensional (pictorial) representation amenable to image interpretation. Subsequently, most endmember extraction algorithms listed above could benefit from an integrated framework in which both the spectral information and the spatial arrangement of pixel vectors are taken into account. An example of this situation is given in Fig. 3, in which a hyperspectral data cube collected over an urban area (high spatial correlation) is modified by randomly permuting the spatial coordinates (i,j) of the pixel vectors, thus removing the spatial correlation. In both scenes, the application of a spectral-based endmember extraction method would yield the same analysis results while it is clear that a spatial–spectral technique could incorporate the spatial information present in the original scene into the endmember searching process.

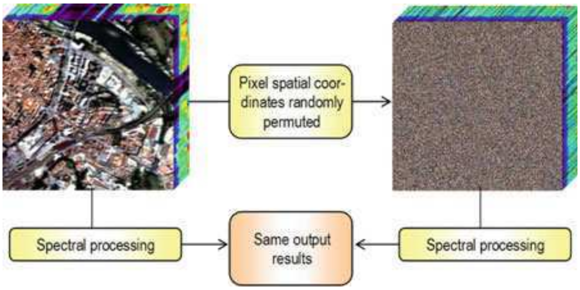


Fig. 3 Example illustrating the importance of spatial information in hyperspectral analysis

To the best of our knowledge, only a few attempts exist in the literature aimed at including the spatial information in the process of extracting spectral endmembers. Extended morphological operations [30] have been used as a baseline to develop an automatic morphological endmember extraction (AMEE) algorithm [31] for spatial–spectral endmember extraction. Also, spatial averaging of spectrally similar endmember candidates found via singular value decomposition (SVD) was used in the development of the spatial spectral endmember extraction (SSEE) algorithm [32]. Recently, a spatial preprocessing (SPP) algorithm [33] has been proposed which estimates, for each pixel vector in the scene, a spatially derived factor that is used to weight the importance of the spectral information associated to each pixel in terms of its spatial context. The SPP is intended as a preprocessing module that can be used in combination with an existing spectral-based endmember extraction algorithm.

In the following, we describe in more detail three selected spectral-based algorithms (N-FINDR, OSP, VCA) and three spatial–spectral endmember extraction algorithms (AMEE, SSEE, SPP) that will be used in our comparison in this chapter. The reasons for our selection are: (1) these algorithms are representative of the class of convex geometry-based and spatial processing-based techniques which have been successful in endmember extraction; (2) they are fully automated; (3) they always produce the same final results for the same input parameters (for the N-FINDR, there is a random initialization step that also conditions the final output); and (4) the number of endmembers to be extracted, p, is an input parameter for all algorithms, while the AMEE, SSE and SPP have additional input parameters related with the definition of spatial context around each pixel in the scene.

2.2.1 N-FINDR

This algorithm looks for the set of pixels with the largest possible volume by *inflating* a simplex inside the data. The procedure begins with a random initial selection of pixels (see Fig. 4a). Every pixel in the image must be evaluated in order to refine the estimate of endmembers, looking for the set of pixels that maximizes the volume of the simplex defined by selected endmembers.

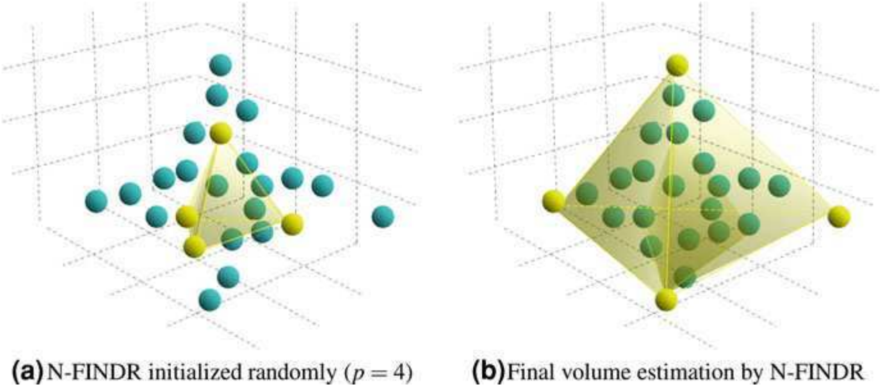

(a) N-FINDR initialized randomly ($p = 4$)
(b) Final volume estimation by N-FINDR

Fig. 4 Graphical interpretation of the N-FINDR algorithm in a three-dimensional space. **a** N-FINDR initialized randomly (p=4); **b** final volume estimation by N-FINDR

The corresponding volume is calculated for every pixel in each endmember position by replacing that endmember and finding the resulting volume. If the replacement results in a an increase of volume, the pixel replaces the endmember. This procedure is repeated until there are no more endmember replacements (see Fig. 4b). The mathematical definition of the volume of a simplex formed by a set of endmember candidates is proportional to the determinant of the set augmented by a row of ones. The determinant is only defined in the case where the number of features is $p - 1$, p being the number of desired endmembers [34]. Since in hyperspectral data typically $n \gg p$, a transformation that reduces the dimensionality of the input data, is required. In this study, the principal component transform (PCT) has been used [35, 36], although another widely used alternative that decorrelates the noise in the data is the maximum noise fraction (MNF) [37]. As a final comment, it has been observed that different random initializations of N-FINDR may produce different final solutions. Thus, our N-FINDR algorithm was implemented in iterative fashion, so that each sequential run was initialized with the previous algorithm solution, until the algorithm converges to a simplex volume that cannot be further maximized. Our experiments show that, in practice, this approach allows the algorithm to converge in a few iterations only.

2.2.2 Orthogonal Subspace Projection (OSP)

This algorithm starts by selecting the pixel vector with maximum length in the scene as the first endmember. Then, it looks for the pixel vector with the maximum absolute projection in the space orthogonal to the space linearly spanned by the initial pixel, and labels that pixel as the second endmember. A third endmember is found by applying an orthogonal subspace projector to the original image [22], where the signature that has the maximum orthogonal projection in the space orthogonal to the space linearly spanned by the first two endmembers. This procedure is repeated until the desired number of endmembers, *p*, is found [38].

2.2.3 Vertex Component Analysis (VCA)

This algorithm also makes use of the concept of orthogonal subspace projections. However, as opposed to the OSP algorithm described above, the VCA exploits the fact that the endmembers are the vertices of a simplex, and that the affine transformation of a simplex is also a simplex [21]. As a result, VCA models the data using a positive cone, whose projection onto a properly chosen hyperplane is another simplex whose vertices are the final endmembers. After projecting the data onto the selected hyperplane, the VCA projects all image pixels to a random direction and uses the pixel with the largest projection as the first endmember. The other endmembers are identified in sequence by iteratively projecting the data onto a direction orthogonal to the subspace spanned by the endmembers already determined. The new endmember is then selected as the pixel corresponding to the extreme projection, and the procedure is repeated until a set of *p* endmembers is found [21]. In our experiments with VCA, we select the corresponding pixel original spectra as the VCA solution, not the noise-smoothed solution produced by the original algorithm. In practice, our approach is expected to slightly reduce the performance of VCA for low signal-to-noise (SNR) ratios, but we also believe that this decision allows a fair comparison of VCA to N-FINDR and OSP, which do not incorporate such noise reduction stage.

2.2.4 Automatic Morphological Endmember Extraction (AMEE)

The automatic morphological endmember extraction (AMEE) [31] algorithm runs on the full data cube with no dimensional reduction, and begins by searching spatial neighborhoods around each pixel vector $\mathbf{X}(i, j)$ in the image for the most spectrally pure and mostly highly mixed pixel. This task is performed by using extended mathematical morphology operators [30] of dilation and erosion, which are graphically illustrated on Fig. 5. Here, dilation selects the most spectrally pure pixel in a local neighborhood around each pixel vector $\mathbf{X}(i, j)$, while erosion selects the most highly mixed pixel in the same neighborhood. Each spectrally pure pixel is then assigned an *eccentricity* value, which is calculated as the spectral angle distance (SAD) [5, 10] between the most spectrally pure and mostly highly mixed pixel for each given spatial neighborhood. This process is repeated iteratively for larger spatial neighborhoods up to a maximum size that is predetermined. At each iteration the eccentricity values of the selected pixels are updated. The final endmember set is obtained by applying a threshold to the resulting greyscale eccentricity image, which results in a large set of endmember candidates. The final endmembers are extracted after applying the OSP method to

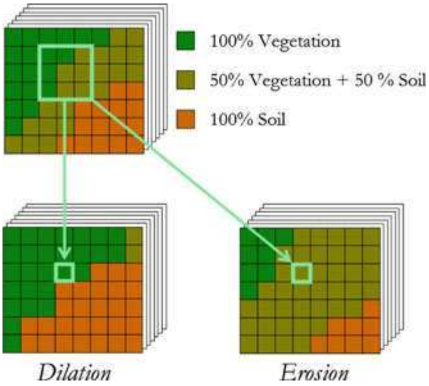


Fig. 5 Toy example illustrating extended morphological operations of dilation and erosion

the set of candidates in order to derive a final set of spectrally distinct endmembers $\{\ \mathbf{E}_z\ \}_{z=1}^{p}$, where p is an input parameter to the OSP algorithm.

2.2.5 Spatial Spectral Endmember Extraction (SSEE)

The spatial–spectral endmember extraction tool (SSEE) uses spatial constraints to improve the relative spectral contrast of endmember spectra that have minimal unique spectral information, thus improving the potential for these subtle, yet potentially important endmembers, to be selected. With SSEE, the spatial characteristics of image pixels are used to increase the relative spectral contrast between spectrally similar, but spatially independent endmembers. The SSEE algorithm searches an image with a local search window centered around each pixel vector $\mathbf{X}(i, j)$ and comprises four steps [32]. First, the singular value decomposition (SVD) transform is applied to determine a set of eigenvectors that describe most of the spectral variance in the window or partition (see Fig. 6). Second, the entire image data are projected onto the previously extracted eigenvectors to determine a set of candidate endmember pixels (see Fig. 7).

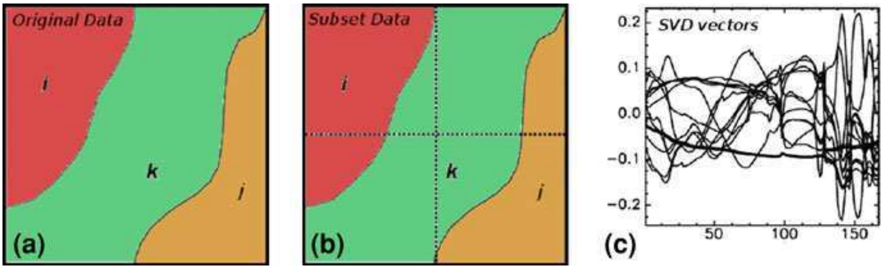


Fig. 6 First step of the SSEE algorithm. **a** Original data. **b** Subset data after spatial partitioning. **c** Set of representative SVD vectors used to describe spectral variance. This scene is reproduced from the one originally presented in [32]

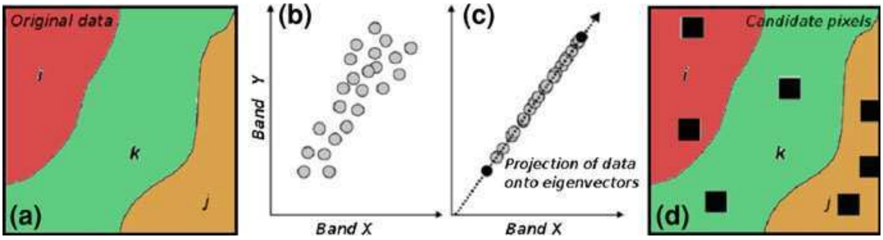


Fig. 7 Second step of the SSEE algorithm. **a** Original data. **b** Spectral distribution in two-dimensional space. **c** Projection of data onto eigenvectors. **d** Set of candidate pixels. This scene is reproduced from the one originally presented in [32]

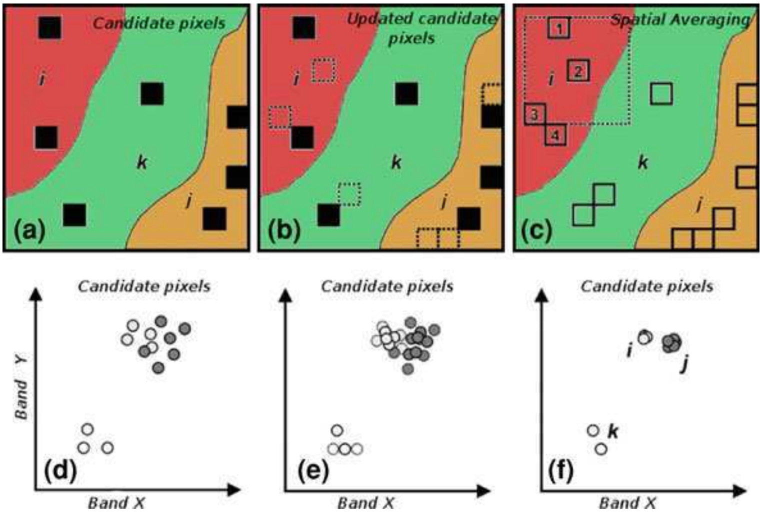


Fig. 8 Third step of the SSEE algorithm. **a** Set of candidate pixels. **b** Updated candidate pixels after including pixels which are spectrally similar to those in the original set. **c** Spatial averaging process of candidate endmember pixels using a sliding window centered on each candidate. **d** First iteration of spatial–spectral averaging. Averaged pixels shown as thick lines, with original pixels shown as thinner lines. **e** Second iteration of spatial–spectral averaging. **f** Continued iterations compress endmembers into clusters with negligible variance. This scene is reproduced from the one originally presented in [32]

Then, spatial constraints are used to combine and average spectrally similar candidate endmember pixels by testing, for each candidate pixel vector, which other pixel vectors are sufficiently similar in spectral sense (see Fig. 8). Instead of using a manual procedure as recommended by the authors in [32], we have used the OSP technique in order to derive a final set of spectrally distinct endmembers $\{\mathbf{E}_z\}_{z=1}^{p}$, where p is an input parameter to the OSP algorithm.

2.2.6 Spatial Pre-Processing (SPP)

The SPP [33] serves as a preprocessing module which can be combined with existing spectral-based algorithms such as the N-FINDR, OSP and VCA. The method estimates, for each input pixel vector, a scalar factor which is intimately related to the spatial similarity between the pixel and its spatial neighbors, and then uses this scalar factor to spatially weight the spectral information associated to the pixel. A simple geometric interpretation of the scalar factor is illustrated in Fig. 9, given as a toy example in which only two spectral bands of an input hyperspectral scene are represented against each other for visualization purposes. The idea behind the SPP is to center each spectral feature in the data cloud around its mean value, and then shift each feature straight towards the centroid of the data cloud (denoted by O′ in Fig. 9). The shift is proportional to a similarity measure calculated using both the spatial neighborhood around the pixel under consideration and the spectral information associated to the pixel, but without averaging the spectral signature of the pixel. The correction is performed so that pixels located in spatially homogenous areas (such as the pixel vector labeled as ‘1’ in Fig. 9) are expected to have a smaller displacement with regards to their original location in the data cloud than pure pixels surrounded by spectrally distinct substances (e.g., the pixel vectors labeled as ‘2’ and ‘3’ in Fig. 9).

Resulting from the above operation, a modified simplex is formed, using not only spectral but also spatial information. It should be noted that the vertices of the modified simplex are more likely to be pure pixels located in spatially homogenous areas. Although the proposed method is expected to privilege homogeneous areas for the selection of endmembers, no pixel is excluded from the competitive

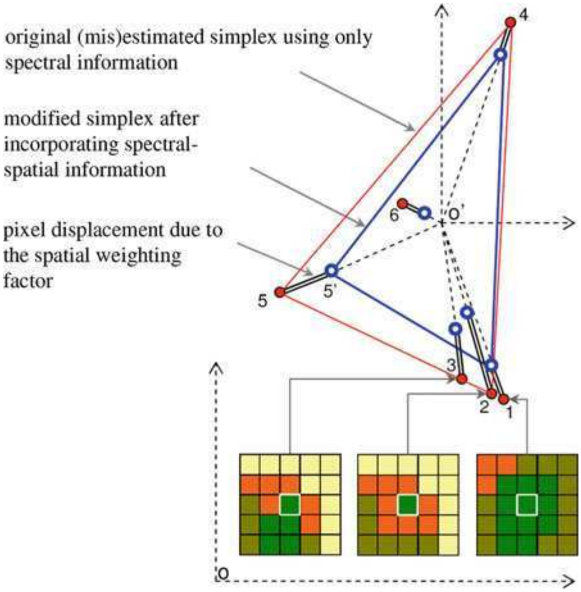

original (mis)estimated simplex using only spectral information
modified simplex after incorporating spectral-spatial information
pixel displacement due to the spatial weighting factor
4
6
O'
5
5'
3
2
1
O

Fig. 9 Geometric interpretation of the SPP method for spatial preprocessing prior to endmember extraction. This scene is reproduced from the one originally presented in [33]

endmember extraction process that follows the preprocessing. As it can be inferred from Fig. 9, the proposed method is also expected to be robust in the presence of outliers. It is important to notice that the modified simplex in Fig. 9 is mainly intended to serve as a guide for a subsequent competitive endmember extraction process, conducted using a user-defined algorithm. However, such modified simplex is not intended to replace the simplex in the input hyperspectral scene. To achieve this, the spatial coordinates of the endmembers extracted from the preprocessed image are retained, but the spectral signatures associated to those spatial coordinates are obtained from the original hyperspectral scene.

2.3 Unconstrained Versus Constrained Linear Spectral Unmixing

Once a set of endmembers $\mathbf{E} = \{\mathbf{E}_z\}_{z=1}^{p}$ have been extracted, their correspondent abundance fractions $\Phi(i, j) = \{\Phi_z(i, j)\}_{z=1}^{p}$ in a specific pixel vector $\mathbf{X}(i, j)$ of the scene can be estimated (in least squares sense) by the following unconstrained expression [10]:

$$\hat{\Phi}_{\mathrm{UC}}(i,j) = \left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{X}(i,j). \tag{2}$$

However, it should be noted that the fractional abundance estimations obtained by means of Eq. 2 do not satisfy the ASC and ANC constraints. Imposing the ASC constraint results in the following optimization problem:

$$\begin{gathered}\min_{\Phi(i,j)\in\Delta}\left\{(\mathbf{X}(i,j) - \Phi(i,j)\cdot\mathbf{E})^T(\mathbf{X}(i,j) - \Phi(i,j)\cdot\mathbf{E})\right\},\\ \text{subject to: } \Delta = \left\{\Phi(i,j)\,{}_{|}\sum_{z=1}^{p}\Phi_z(i,j) = 1\right\}.\end{gathered} \tag{3}$$

Similarly, imposing the ANC constraint results in the following optimization problem:

$$\begin{gathered}\min_{\Phi(i,j)\in\Delta}\left\{(\mathbf{X}(i,j) - \Phi(i,j)\cdot\mathbf{E})^T(\mathbf{X}(i,j) - \Phi(i,j)\cdot\mathbf{E})\right\},\\ \text{subject to: } \Delta = \left\{\Phi(i,j)\,{}_{|}\,\Phi_z(i,j) \geq 0 \text{ for all } 1 \leq z \leq p\right\}.\end{gathered} \tag{4}$$

As indicated in [13], a non-negative constrained least squares (NCLS) algorithm can be used to obtain a solution to the ANC-constrained problem described in Eq. 4 in iterative fashion [39]. In order to take care of the ASC constraint, a new endmember signature matrix, denoted by $\mathbf{E}'$, and a modified version of the pixel vector $\mathbf{X}(i, j)$, denoted by $\mathbf{X}'(i, j)$, are introduced as follows:

$$\mathbf{E}' = \begin{bmatrix}\delta\mathbf{M}\\ \mathbf{1}^T\end{bmatrix}, \Phi'(i,j) = \begin{bmatrix}\delta\Phi(i,j)\\ 1\end{bmatrix}, \tag{5}$$

where $\mathbf{1} = (\underbrace{1, 1, \ldots, 1}_{p})^T$ and δ controls the impact of the ASC constraint. Using the two expressions in (5), a fully constrained estimate can be directly obtained from the NCLS algorithm by replacing $\mathbf{E}$ and $\Phi(i, j)$ used in the NCLS algorithm with $\mathbf{E}'$ and $\Phi'(i, j)$. Hereinafter, we will refer to the fully constrained (i.e. ASC-constrained and ANC-constrained) linear spectral unmixing model by the acronym FCLSU.

3 Nonlinear Spectral Unmixing

3.1 Problem Formulation

Under the nonlinear mixture model assumption, each pixel vector in the original scene can be modeled using the following expression:

$$\mathbf{X}(i,j) = f(\mathbf{E}, \Phi(i,j)) + \mathbf{n}(i,j), \tag{6}$$

where f is an unknown nonlinear function that defines the interaction between $\mathbf{E}$ and $\Phi(i, j)$. Various learning-from-data techniques have been proposed in the literature to estimate f. In particular, artificial neural networks have demonstrated great potential to decompose mixed pixels due to their inherent capacity to approximate complex functions [40]. Although many neural network architectures exist, for decomposition of mixed pixels in terms of nonlinear relationships mostly feed-forward networks of various layers, such as the multi-layer perceptron (MLP), have been used [12, 41, 42]. It has been shown in the literature that MLP-based neural models, when trained accordingly, generally outperform other non-linear models such as regression trees or fuzzy classifiers [43].

A variety of issues have been investigated in order to evaluate the impact of training in mixed pixel classification accuracy, including the size and location of training sites, and the composition of training sets, but most of the attention has been paid to the issue of training set size, i.e., the number of training samples required for the learning stage [44]. Sometimes the smallness of a training set represents a major problem [45]. This is especially apparent for analyses using hyperspectral sensor data, where the requirement of large volumes of training sites is a serious limitation [46]. Even if the endmembers participating in mixtures in a certain area are known, proportions of these endmembers on a per-pixel basis are difficult to be estimated a priori. Therefore, one of the most challenging aspects in the design of neural network-based techniques for spectral mixture analysis is to reduce the need for very large training sets. Studies have investigated a range of issues [47], including the use of feature selection and feature extraction methods to reduce the dimensionality of the input data [48], the use of unlabeled and semi-labeled samples [46], the accommodation of spatial dependence in the data to define an efficient sampling design [32], or the use of statistics derived on other locations [49]. Our speculation is that the problem of mixed pixel interpretation demands intelligent training sample selection algorithms, able to seek for the most informative training samples, thus optimizing the compromise between estimation accuracy (to be maximized) and ground-truth knowledge (to be minimized).

A second issue that has not received attention in neural network-based mixed pixel analysis has to do with initial model conditions. For instance, the MLP neural network is typically trained using the error back-propagation algorithm [40]. It is a supervised technique of training with three phases. In the first one, an initial vector is presented to the network, which leads to the activation of the network as a whole. The second phase computes an error between the output vector and a vector of desired values for each output unit, and propagates it successively back through the network. The last phase computes the changes for the connection weights, which are randomly generated in the beginning. According to algorithm design, a good and effective learning algorithm should not depend on initial conditions, which can only affect the algorithm convergence rate, but should not alter the final results. The matter of fact is that this is generally not true in learning algorithms used for neural networks, where the choice of initial weights determines which minimum the algorithm will converge to [11]. In order for a mixture model to be effective, initial values must be representative and cannot be arbitrary.

In this section, we develop a combined linear/nonlinear mixture model which assumes that most of the mixed spectra in the data can be modeled via a combination of single and multiple scattering effects. The abundance fractions of endmember substances are first estimated via a linear mixture model and used to establish the initial condition, including the initial weight matrix. Such an initial estimation is then refined using an MLP neural network, coupled with unsupervised algorithms for intelligent selection of training samples from the available data. One of our main reasons to select an MLP neural network for demonstration is that this architecture has been often claimed to be sensitive to network architecture parameters, such as the arrangement and number of neurons in the different layers [41]. In our experience, however, a very simple MLP network configuration can produce stable results when initialized and trained accordingly, a fact that leads us to believe that both initialization and training can indeed be more important than the choice of a specific network architecture in mixture analysis applications.

3.2 Neural Network-Based Spectral Unmixing

Figure 10 shows a schematic block diagram of the proposed neural network-based unmixing architecture. The first step consists of an estimation of the number of endmembers, p, in the input data. For this purpose, in this work we use the VD concept [13]. Then, the model is initialized via a fully constrained linear mixture model based on automatic endmember extraction. Finally, the model is refined by

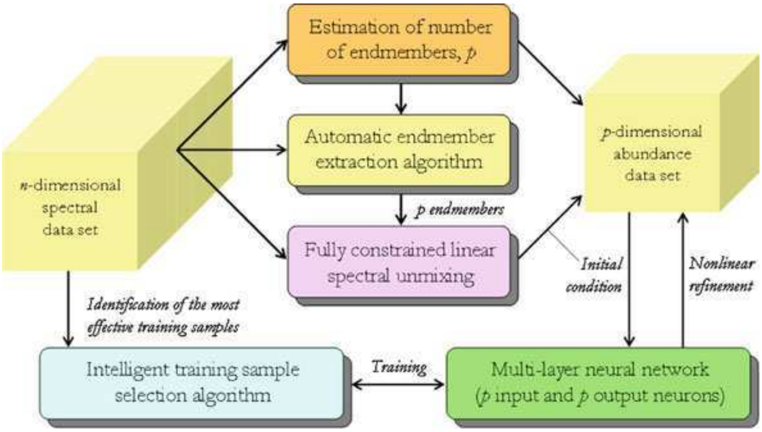

Estimation of number of endmembers, p
Automatic endmember extraction algorithm
p endmembers
Fully constrained linear spectral unmixing
n-dimensional spectral data set
p-dimensional abundance data set
Initial condition
Nonlinear refinement
Identification of the most effective training samples
Intelligent training sample selection algorithm
Training
Multi-layer neural network (p input and p output neurons)

Fig. 10 Neural network-based spectral unmixing architecture

a supervised MLP neural network. The latter step is supported by an unsupervised algorithm for intelligent selection of training samples (both pure and mixed) from the data in order to estimate the final endmember fractional abundances. The number of neurons at the input layer of the MLP architecture equals the number of spectral endmembers found in the initialization stage. The input patterns to the input layer are vectors of endmember fractional abundances for each sample vector $\mathbf{X}(i, j)$, first estimated by FCLSU. The second layer is the hidden layer, and the third layer is the output layer. The number of neurons at the output layer, p, equals the number of estimated endmembers. It should be noted that the number of hidden neurons in the MLP architecture can be fine-tuned depending on the problem under consideration [40]. However, in this work we are mainly interested in exploring training mechanisms and their implications, without particular emphasis on careful adjustment of neural network configuration parameters. Subsequently, finding optimal parameters for the hidden layer is beyond our scope. Based on previous results in the literature and our own experimentation, we have considered one hidden layer only, with the number of neurons empirically set to the square root of the product of the number of input features and information classes, a configuration that has been shown to be successful for MLP-based mixed pixel characterization in previous work [43].

At this point, it is worth noting that most available neural models for multi-dimensional data analysis in the literature assume that the neuron count at the input layer equals the dimension of the input vectors, i.e., each neuron in the input layer is associated with one of the n spectral bands in which a pixel's reflectance spectrum is measured. However, the above configuration may easily suffer from limited training samples in hyperspectral analyses, where training data are often of limited quantity relative to input space dimensionality [36]. This leads any induced classifier to potentially feature a poor generalization capability, an effect known as the Hughes effect or *curse of the dimensionality*. Numerous analyses have been undertaken founded on the desire to reduce the dimensionality of the input data prior to the analysis. In order to overcome the limitations above, in this work we adopt a simple, yet natural approach to represent an n-dimensional pixel vector as a p-dimensional vector of endmember fractional abundances at the pixel. This strategy allows for a reduction in the number of network connections without losing the information that is crucial for spectral unmixing applications. It should be noted that the issue of how to select the most informative training data (in terms of mixing knowledge) is of great importance for the success of the nonlinear learning stage. In the following subsection, we develop an unsupervised algorithm which selects training samples based on the mixture information they contain, thus allowing us to accommodate the information provided by mixed pixels into the learning process.

3.3 Automatic Selection and Labeling of Training Samples

The quality of training has a significant effect on mixed pixel characterization using neural networks [44]. Conventional approaches for selection of training samples often perform this task randomly, or by choosing the samples located in exemplar regions of each class only, while atypical cases are often removed or down-weighted in training set refinement operations. Such exemplar training patterns are located near the central *core* of the class in feature space. However, a key concern in the context of mixed pixel interpretation is how to identify and characterize the response of sites that lie away from the class core, and near to the decision boundaries commonly used in conventional, *pure* pixel classification. Therefore, *border* [47] (or, equivalently, *mixed*) training samples may be useful to refine a set of fractional abundance estimations obtained by using only spectrally pure training samples.

In this section, we describe a new technique for automatic selection and labeling of training samples from the input hyperspectral data. The proposed technique, called mixed training algorithm (MTA), first uses Winter's N-FINDR algorithm [15] as an approach to automatically label spectrally pure training samples (endmembers) without prior knowledge. Then, it iteratively seeks for the most highly mixed pixels in the input data set by following a procedure which behaves in an opposite way as N-FINDR and other convex geometry-based endmember extraction methods [9], i.e. it automatically selects and labels highly mixed training samples. Different sets of training samples, obtained by the MTA discussed in this section, will be used in the following section to investigate the impact of the composition of the training set on the characterization of mixed pixels. The MTA can be summarized by the following steps:

1. Compute $\mathbf{C}_p = (1/p)\sum_{z=1}^{p}\mathbf{E}_z$, i.e., the centroid of the simplex defined by the set of spectral endmembers $\mathbf{E} = \{\mathbf{E}_z\}^{p}{}_{z=1}$ produced for the input hyperspectral scene by an endmember extraction algorithm such as N-FINDR.

2. At iteration $j \geq 1$, calculate a point-wise spectral *distance* between each pixel vector $\mathbf{X}(i, j)$ in the input hyperspectral data and $\mathbf{C}_p$, and mark the pixel vector which provides the lowest *distance* value (i.e., the most spectrally similar to $\mathbf{C}_p$) as a new training sample $\mathbf{T}_j$.
3. Remove the pixel previously selected as a training sample from the input hyperspectral scene and apply a spectral screening algorithm to identify the pixel vectors with associated spectral signatures within a small spectral angle θ from any of the previously selected training samples, removing those samples from the input data as well.
4. Repeat from step 2 until a final set of k mixed labeled training samples $\{\mathbf{T}_j\}_{j=1}^{k}$ is generated from the input hyperspectral scene.

It should be noted that the MTA algorithm was implemented using various spectral similarity measures [5, 10], such as the SAD or the spectral information divergence (SID). In all cases, the results obtained were very similar. As a result, this paper only reports experiments based on using SAD for demonstration purposes.

4 Experimental Results

In this section we present two experiments focused on evaluating the endmember extraction and spectral unmixing techniques discussed throughout the chapter. In our first experiment, we focus on a mineral mapping application and further discuss the role of endmember extraction and the use of spatial information for linear spectral unmixing purposes, using AVIRIS image data collected over the Cuprite mining district in Nevada. In our second experiment, we provide a comparison of linear versus nonlinear spectral unmixing techniques in the context of a real agriculture and farming application in the region of Extremadura, Spain, using hyperspectral data sets collected by the DAIS 7915 and the ROSIS imaging spectrometers operating simultaneously at multiple resolutions.

4.1 First Experiment: AVIRIS Hyperspectral Data

In this experiment we use the well-known AVIRIS Cuprite data set, available online in reflectance units[1] after atmospheric correction. This scene has been widely used to validate the performance of endmember extraction algorithms. The portion used in experiments corresponds to a 350 × 350-pixel subset of the sector labeled as f970619t01p02_r02_sc03.a.rfl in the online data. The scene (displayed in Fig. 11a) comprises 224 spectral bands between 0.4 and 2.5 μm, with full width

[1] http://aviris.jpl.nasa.gov/html/aviris.freedata.html

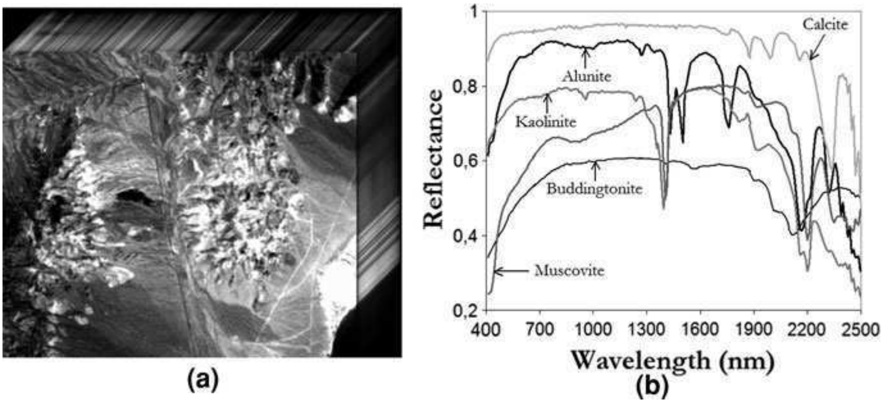


Fig. 11 **a** AVIRIS Cuprite data cube. **b** USGS spectral signatures of five representative minerals in the Cuprite mining district

at half maximum of 10 nm and spatial resolution of 20 m per pixel. Prior to the analysis, several bands were removed due to water absorption and low SNR in those bands, leaving a total of 192 reflectance channels to be used in the experiments. The Cuprite site is well understood mineralogically [50, 51], and has several exposed minerals of interest included in a spectral library compiled by the U.S. Geological Survey (USGS).[2] A few selected spectra from the USGS library, corresponding to several highly representative minerals in the Cuprite mining district (see Fig. 11b), are used in this work to substantiate endmember signature purity.

Two different metrics have been used to compare the performance of endmember extraction and spectral unmixing algorithms in the AVIRIS Cuprite scene. The first metric is the SAD between each extracted endmember and the set of available USGS ground-truth spectral signatures. For the sake of clarity, we remind that the SAD between two pixel vectors $\mathbf{X}(i, j)$ and $\mathbf{X}(r, s)$ can be simply defined as follows:

$$\mathrm{SAD}(\mathbf{X}(i,j), \mathbf{X}(r,s)) = \cos^{-1} \frac{\mathbf{X}(i,j) \cdot \mathbf{X}(r,s)}{\|\mathbf{X}(i,j)\| \|\mathbf{X}(r,s)\|}. \qquad (7)$$

It should be noted that SAD is given by the spectral angle formed by n-dimensional vectors (in radians). As a result, low SAD scores mean high spectral similarity between the compared vectors. This spectral similarity measure is invariant in the multiplication of $\mathbf{X}(i, j)$ and $\mathbf{X}(r, s)$ by constants and, consequently, is invariant before unknown multiplicative scalings that may arise due to differences in illumination and angular orientation [5]. The SAD metric allows us to identify the USGS signature which is most similar to each endmember automatically

[2] http://speclab.cr.usgs.gov/spectral-lib.htm

extracted from the scene by observing the minimum SAD distance reported for such endmember across the entire set of USGS signatures. The second metric is based on the assumption that a set of high-quality endmembers (and their corresponding FCLSU-estimated abundance fractions) may allow reconstruction of the original hyperspectral scene (by means of Eq. 1) with higher precision than a set of low-quality endmembers.

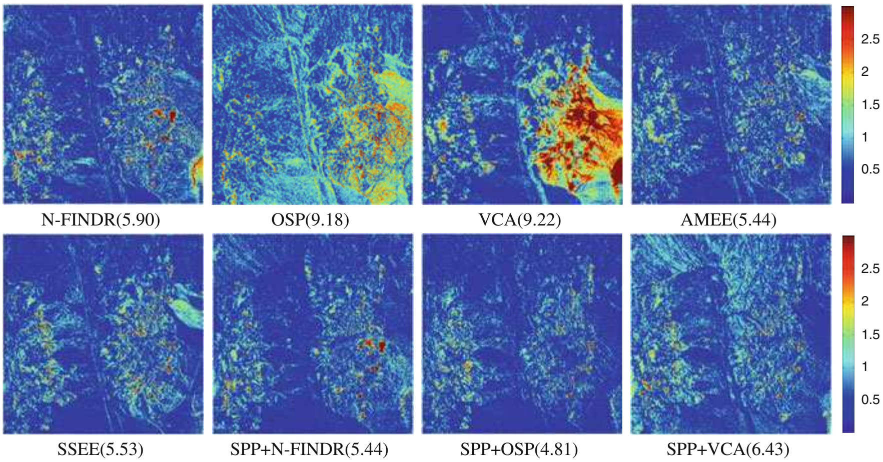
A second metric employed to evaluate the goodness of the reconstruction is the RMSE between the original and the reconstructed hyperspectral scene, which can be defined as follows. Let us assume that $\mathbf{I}^{(\mathrm{O})}$ is the original hyperspectral scene, and that $\mathbf{I}^{(\mathrm{R})}$ is a reconstructed version of $\mathbf{I}^{(\mathrm{O})}$, obtained using Eq. 1 with a set of endmembers, automatically derived by a certain algorithm from the original scene, and their corresponding FCLSU-estimated fractional abundances. Let us also assume that the pixel vector at spatial coordinates (i, j) in the original hyperspectral scene is given by $\mathbf{X}^{(\mathrm{O})}(i, j) = [x_1^{(\mathrm{O})}(i, j), x_2^{(\mathrm{O})}(i, j), ..., x_n^{(\mathrm{O})}(i, j)]$, while the corresponding pixel vector at the same spatial coordinates in the reconstructed hyperspectral scene is given by $\mathbf{X}^{(\mathrm{R})}(i, j) = [x_1^{(\mathrm{R})}(i, j), x_2^{(\mathrm{R})}(i, j), ..., x_n^{(\mathrm{R})}(i, j)]$. With the above notation in mind, the RMSE between the original and the reconstructed hyperspectral scenes is calculated as follows:

$$\mathrm{RMSE}(\mathbf{I}^{(\mathrm{O})}, \mathbf{I}^{(\mathrm{R})}) = \frac{1}{s \times l} \sum_{i=1}^{s} \sum_{j=1}^{l} \left(\frac{1}{n} \sum_{k=1}^{n} [x_k^{(\mathrm{O})}(i,j) - x_k^{(\mathrm{R})}(i,j)]^2 \right)^{1/2}. \quad (8)$$

Table 1 tabulates the SAD scores (in degrees) obtained after comparing the USGS library spectra of five highly representative minerals in the Cuprite mining district (*alunite*, *buddingtonite*, *calcite*, *kaolinite* and *muscovite*) with the corresponding endmembers extracted by different algorithms from the AVIRIS Cuprite scene. In all cases, the input parameters of the different endmember extraction methods tested have been carefully optimized so that the best performance for each method is reported. The smaller the SAD values across the five minerals in Table 1, the better the results. It should be noted that Table 1 only displays the smallest SAD scores of all endmembers with respect to each USGS signature for each algorithm. For reference, the mean SAD values across all five USGS

Table 1 SAD-based spectral similarity scores (in degrees) between the USGS mineral spectra and their corresponding endmember pixels produced by several endmember extraction algorithms

Algorithm	Alunite	Buddigntonite	Calcite	Kaolinite	Muscovite	Mean
N-FINDR	9.96°	7.71°	12.08°	13.27°	5.24°	9.65°
OSP	4.81°	4.16°	9.62°	11.14°	5.41°	7.03°
VCA	10.73°	9.04°	6.36°	14.05°	5.41°	9.12°
AMEE	4.81°	4.21°	9.54°	8.74°	4.61°	6.38°
SSEE	4.81°	4.16°	8.48°	11.14°	4.62°	6.64°
SPP+N-FINDR	12.81°	8.33°	9.83°	10.43°	5.28°	9.34°
SPP+OSP	4.95°	4.16°	9.96°	10.90°	4.62°	6.92°
SPP+VCA	12.42°	4.04°	9.37°	7.87°	6.18°	7.98°



N-FINDR(5.90) OSP(9.18) VCA(9.22) AMEE(5.44)

SSEE(5.53) SPP+N-FINDR(5.44) SPP+OSP(4.81) SPP+VCA(6.43)

Fig. 12 RMSE reconstruction errors (in percentage) for various endmember extraction algorithms after reconstructing the AVIRIS Cuprite scene

signatures is also reported. In all cases, the number of endmembers to be extracted was set to $p = 14$ after using the VD concept in [10]. Table 1 reveals that the AMEE provides very good results (all SAD scores below 10°), with the SSEE and the SPP+OSP being the algorithms that can provide comparable—but slightly worst—results. Table 1 also reveals that, in this real example, spatial preprocessing generally improves the signature purity of the endmembers extracted by spectral-based algorithms.

On the other hand, Fig. 12 graphically represents the per-pixel root mean square error (RMSE) obtained after reconstructing the AVIRIS Cuprite scene using $p = 14$ endmembers extracted by different methods. It can be seen that the methods using spatial preprocessing (SPP+OSP, SPP+N-FINDR, SPP+VCA) improve their respective spectral-based versions in terms of the quality of image reconstruction, while both AMEE and SSEE also provide lower reconstruction errors than OSP, N-FINDR and VCA. These results suggest the advantages of incorporating spatial information into the automatic extraction of image endmembers from the viewpoint of obtaining more spatially representative spectral signatures which can be used to describe other mixed signatures in the scene.

4.2 *Second Experiment: DAIS 7915 and ROSIS Hyperspectral Data*

In this section, a set of scenes collected over a so-called *Dehesa* semi-arid ecosystem (formed by *quercus ilex* or cork-oak trees, soil and pasture) is used as a case study to illustrate the applicability of nonlinear neural network-based unmixing to a real problem. In the Iberian Peninsula, Dehesa systems are used for a combination of livestock, forest and agriculture activity [52]. The outputs of these systems include meat, milk, wool, charcoal, cork bark and grain. Around 12–18% of the area is harvested on a yearly basis. The crops are used for animal feed or for cash cropping, depending on the rainfall of the area. Determination of fractional land-cover using remote sensing techniques may allow for a better monitoring of natural resources in Dehesa agro-ecosystems. Our choice of this type of landscape for evaluating spectral unmixing techniques was made on several accounts. The first one is the availability of hyperspectral image data sets with accurate geo-registration for a real Dehesa test site in Caceres, SW Spain, collected simultaneously in July 2001 by two instruments operating at multiple spatial resolutions: DAIS 7915 and ROSIS, operated by the German Aerospace Agency (DLR). A second major reason is the simplicity of the Dehesa landscape, which greatly facilitates the collection of reliable field data for model validation purposes. It is also important to emphasize that the scenes were collected in summertime, so atmospheric interferers were greatly minimized. Before describing our experiments, we first provide a comprehensive description of the data sets used and ground-truth activities in the study area.

4.2.1 Data Description

The data used in this study consisted of two main components: image data and field measurements of land-cover fractions, collected at the time of image data acquisition. The image data is formed by a ROSIS scene collected at high spatial resolution, with 1.2-m pixels, and its corresponding DAIS 7915 scene, collected at low spatial resolution with 6-m pixels. The spectral range from 504 to 864 nm was selected for experiments, not only because it is adequate for analyzing the spectral properties of the landscape under study, but also because this spectral range is well covered by the two considered sensors through narrow spectral bands. Figure 13 shows the full flightline of the ROSIS scene, which comprises a Dehesa area located between the facilities of University of Extremadura in Cáceres (leftmost part of the flightline) and Guadiloba water reservoir at the center of the flightline. Figure 14a shows the Dehesa test site selected for experiments, which corresponds to a highly representative Dehesa area that contains several cork-oak trees (appearing as dark spots) and several pasture (gray) areas on a bare soil (white) background. Several field techniques were applied to obtain reliable estimates of the fractional land cover for each DAIS 7915 pixel in the considered Dehesa test site:

1. First, the ROSIS image was roughly classified into the three land-cover components above using a maximum-likelihood supervised classification approach based on image-derived spectral endmembers, where Fig. 14b shows the three endmembers used for mapping that were derived using the AMEE algorithm. Our assumption was that the pixels in the ROSIS image were sufficiently small to become spectrally simple to analyze.

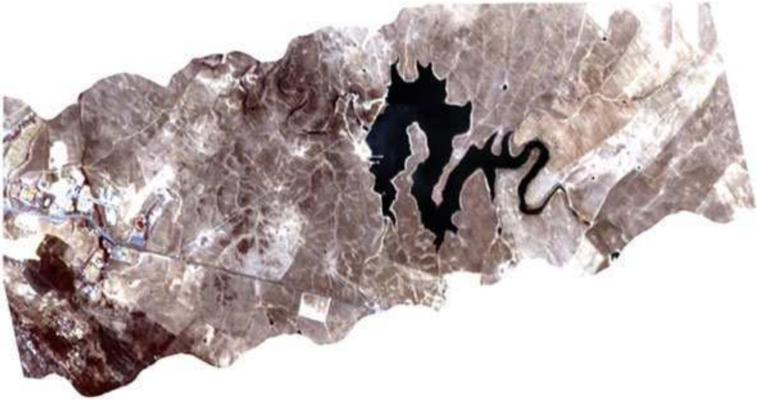


Fig. 13 Flightline of a ROSIS hyperspectral scene collected over a Dehesa area in Cáceres, Spain

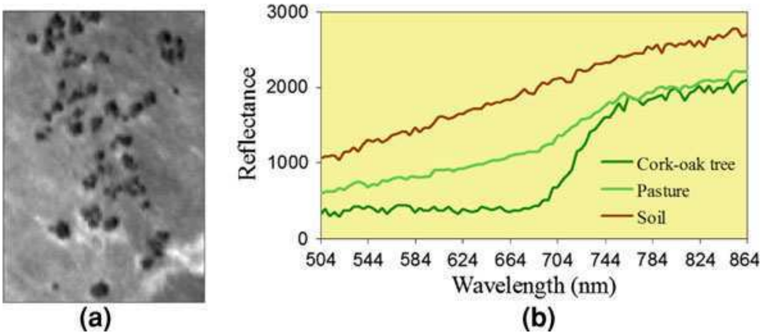


Fig. 14 **a** Spectral band (584 nm) of a ROSIS Dehesa subset selected for experiments. **b** Endmember signatures of soil, pasture and cork-oak tree extracted by the AMEE algorithm, where scaled reflectance values are multiplied by a constant factor

2. Then, the classified ROSIS image was registered with the DAIS 7915 image using a ground control point-based method with sub-pixel accuracy [53].
3. The classification map was then associated with the DAIS 7915 image to provide an initial estimation of land cover classes for each pixel at the DAIS 7915 image scale. For that purpose, a 6 × 6-m grid was overlaid on the 1.2 × 1.2-m classification map derived from the ROSIS scene, where the geographic coordinates of each pixel center point were used to validate the registration with sub-pixel precision.
4. Next, fractional abundances were calculated within each 6 × 6-m grid as the proportion or ROSIS pixels labeled as cork-oak tree, pasture and soil located within that grid, respectively.
5. Most importantly, the abundance maps at the ROSIS level were thoroughly refined using field measurements (see Fig. 15a) before obtaining the final

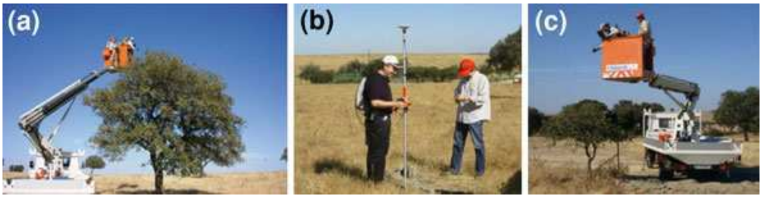


Fig. 15 Ground measurements in the area of study. **a** Spectral sample collection using an ASD FieldSpec Pro spectroradiometer. **b** High-precision GPS geographic delimitation. **c** Field spectral measurements at different altitudes

proportions. Several approaches were developed to refine the initial estimations:

- Fractional land cover data were collected on the ground at more than thirty evenly distributed field sites within the test area. These sites were delineated during the field visit as polygons, using high-precision GPS coordinates (see Fig. 15b).
- Land cover fractions were estimated at each site using a combination of various techniques. For instance, field spectra were collected for several areas using an ASD FieldSpec Pro spectro-radiometer. Of particular interest were field measurements collected on top of tree crowns (Fig. 15c), which allowed us to model different levels of tree crown transparency.
- On the other hand, the early growth stage of pasture during the summer season allowed us to perform ground estimations of pasture abundance in selected sites of known dimensions, using pasture harvest procedures supported by visual inspection and laboratory analyses.

After following the above-mentioned sequence of steps, we obtained a set of approximate fractional abundance labels for each pixel vector in the DAIS 7915 image. Despite our effort to conduct a reliable ground estimation of fractional land-cover in the considered semi-arid environment, absolute accuracy is not claimed. We must emphasize, however, that the combined use of imagery data at different resolutions, sub-pixel ground control-based image registration, and extensive field work including high-precision GPS field work, spectral sample data collection and expert knowledge, represents a novel contribution in the area of spectral mixture analysis validation, in particular, for Dehesa-type ecosystems.

4.2.2 Fractional Abundance Estimation Results

In order to evaluate the accuracy of linear mixture modeling in the considered application, Fig. 16 shows the scatterplots of measured versus FCLSU-estimated fractional abundances for the three considered land-cover materials in the DAIS 7915 (low spatial resolution) image data set, where the diagonal represents perfect match and the two flanking lines represent plus/minus 20% error bound. Here, the

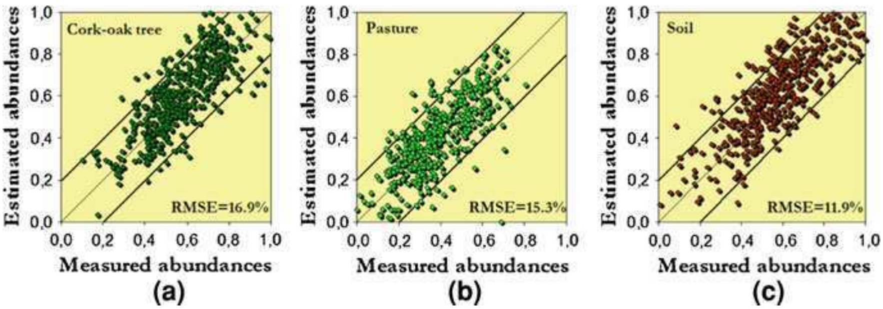


Fig. 16 Abundance estimations of cork-oak tree (**a**), pasture (**b**) and soil (**c**) by the fully constrained linear mixture model from the DAIS 7915 image

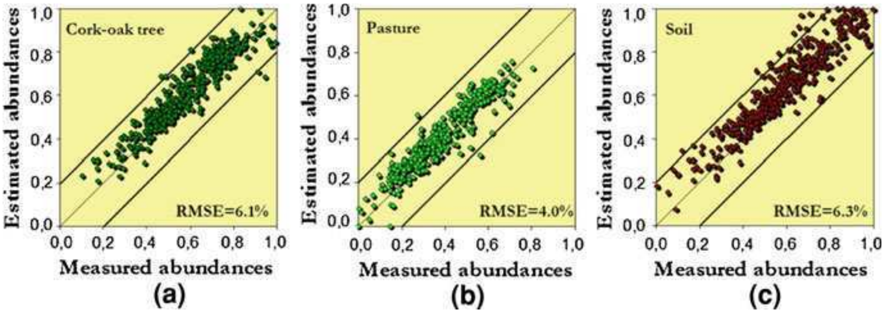


Fig. 17 Abundance estimations of cork-oak tree (**a**), pasture (**b**) and soil (**c**) by the MLP-based mixture model (trained using MTA) from the DAIS 7915 image

three spectral endmembers were derived using the AMEE algorithm, which incorporates spatial information into the endmember extraction process. As expected, the flatness of the test site largely removed topographic influences in the remotely sensed response of soil areas. As a result, most linear predictions for the soil endmember fall within the 20% error bound (see Fig. 16a). On the other hand, the multiple scattering within the pasture and cork-oak tree canopies (and from the underlying surface in the latter case) complicated the spectral mixing in nonlinear fashion, which resulted in a generally higher number of estimations lying outside the error bound, as illustrated in Fig. 16b, c. Also, the RMSE scores in abundance estimation for the soil (11.9%), pasture (15.3%) and cork-oak tree (16.9%) were all above 10% estimation error in percentage, which suggested that linear mixture modeling was not flexible enough to accommodate the full range of spectral variability throughout the landscape.

In order to characterize the Dehesa ecosystem structure better than linear models do, we hypothesized that intelligently selected training data might be required to better characterize nonlinear mixing effects. For this purpose, we applied the MTA algorithm to automatically locate highly descriptive training sites in the DAIS 7915 scene and then used the obtained samples (and the ground-truth information associated to those samples) to train the proposed MLP-based neural network. Figure 17 shows the scatter plots of measured versus predicted fractional abundances for soil, pasture and cork-oak tree by the proposed MLP-based model, trained with the three endmembers derived by AMEE (see Fig. 14b) plus 40 additional training samples selected by MTA, which represent less than 1% of the total number of pixels in the DAIS 7915 scene. These samples were excluded from the testing set made up of all remaining pixels in the scene. From Fig. 17, it is clear that the utilization of intelligently selected training samples resulted in fewer points outside the two 20% difference lines, most notably, for both pasture and cork-oak abundance estimates. The pattern of the scatter plots obtained for the soil predictions (see Fig. 17a) was similar (in particular, when the soil abundance was high). Most importantly, the RMSE scores in abundance estimation were significantly reduced (with regards to the experiment using FCLSU) for the soil (6.1%), pasture (4%) and cork-oak tree (6.3%). These results confirm our intuition that nonlinear effects in Dehesa landscapes mainly result from multiple scattering effects in vegetation canopies.

Before concluding the chapter it is worth noting that, although abundance sum-to-one and abundance non-negativity constraints were not imposed in our proposed MLP-based learning stage, negative and/or unrealistic abundance estimations (which usually indicate a bad fit of the model and reveal inappropriate endmember/training data selection) were very rarely found in our experiments. Summarizing, the experimental validation carried out in this section indicated that the intelligent incorporation of mixed training samples can enable a more accurate representation of nonlinearly mixed signatures. It was apparent from experimental results that the proposed neural network-based model was able to generate abundance estimates that were close to abundance values measured in the field, using only a few intelligently generated training samples. The need for mixed training data does, however, require detailed knowledge on abundance fractions for the considered training sites. In practice, these data are likely to be derived from imagery acquired at a finer spatial resolution than the imagery to be classified, e.g., using data sets acquired by sensors operating simultaneously at multiple spatial resolutions as it is the case of the DAIS 7915 and ROSIS instruments considered in this experiment. Such multi-resolution studies may also incorporate prior knowledge or ancilliary information, which can be used to help target the location of training sites, and to focus training site selection activities on regions likely to contain the most informative training samples.

5 Parallel Implementation Case Study

The endmember extraction and spectral unmixing techniques introduced in previous sections of this chapter introduce new processing challenges, in particular, for very high-dimensional data sets [54]. From a computational perspective, these algorithms can be extremely time consuming when applied to real hyperspectral data sets such as the AVIRIS scene in Fig. 11a, with 137 MB in size, or the ROSIS scene in Fig. 13, with about 1 GB of size for the full flightline. At the same time, these techniques exhibit inherent parallelism at multiple levels [55]: across pixel vectors (coarse grained pixel-level parallelism), across spectral information (fine grained spectral-level parallelism), and even across tasks (task-level parallelism). As a result, they map nicely to massively parallel systems such as clusters of computers or heterogeneous networks of workstations [56]. Unfortunately, these systems are expensive and difficult to adapt to on-board data processing scenarios, in which low-weight and low-power integrated components are mandatory to reduce mission payload [57].

An exciting recent development in the field of commodity computing is the emergence of programmable graphics processing units (GPUs) [58, 59], mainly due to the advent of video-game industry. The speed of graphics hardware doubles approximately every six months, which is much faster than the improving rate of the CPU. The ever-growing computational requirements introduced by hyperspectral imaging applications can benefit from this kind of commodity hardware and take advantage of the compact size and relatively low cost of these units, which make them appealing for on-board data processing at much lower costs than those introduced by other hardware devices such as clusters. In the following, we develop a GPU-based implementation of a spectral unmixing chain made up of spatial–spectral endmember extraction using the AMEE algorithm followed by unconstrained linear spectral unmixing (LSU). The chain was implemented using NVidia$^{\mathrm{TM}}$ CUDA,[3] a collection of extensions to the C programming language and a runtime library. CUDA's functionality primarily allows a developer to write C functions to be executed on the GPU. CUDA also includes memory management and execution configuration, so that a developer can control the number of GPU processors and threads that are to be invoked during a function's execution. GPU-based algorithms developed in CUDA are constructed by chaining so-called *kernels*, which take one or more streams as inputs and produce one or more streams as outputs.

The first issue that needs to be addressed when porting a hyperspectral imaging algorithm to a GPU is how to map a hyperspectral image onto the GPU memory. Since the size of hyperspectral images usually exceeds the capacity of such memory, we split them into multiple spatial-domain partitions [56] made up of entire pixel vectors (see Fig.18), i.e., each spatial-domain partition incorporates all the spectral information on a localized spatial region and is composed of spatially adjacent pixel vectors. Once the hyperspectral image has been allocated onto the GPU memory, a set of kernels are applied to perform the desired operations. In our case, the kernels needed to implement the AMEE algorithm for endmember extraction followed by LSU for linear spectral unmixing can be summarized as follows:

[3] http://www.nvidia.com/object/cuda_home.html

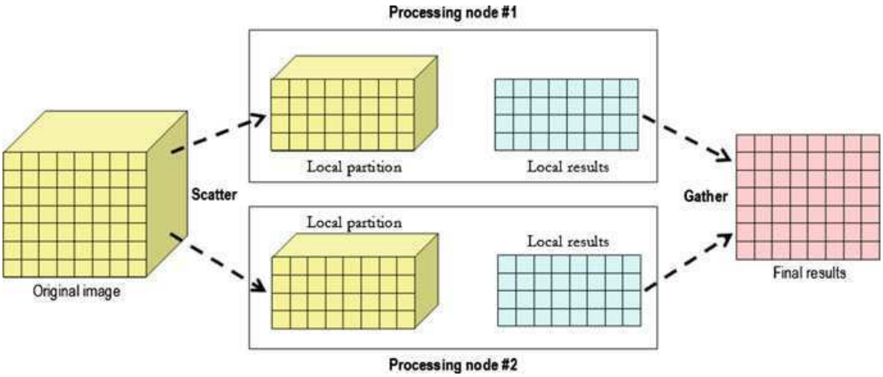


Fig. 18 Spatial-domain decomposition for parallelization of hyperspectral imaging algorithms

- *Cumulative distance*. For each pixel vector, this kernel accumulates the SAD with all the neighboring pixels in order to complete a core operation in the AMEE endmember extraction algorithm. It is based on a single-pass kernel that computes the SAD between two pixel vectors using the inner products and norms produced by the previous kernel. Finally, the kernel calculates, for each pixel vector, the cumulative spectral angle between the pixel and all its neighbors.
- *Max/min finding*. Extended morphological erosion and dilation used by the AMEE algorithm are implemented at this stage through a kernel that applies minimum and maximum reductions. This kernel uses as inputs the cumulative values generated in the previous stage and produces a stream containing (for each pixel) the relative coordinates of the neighboring pixels with maximum and minimum cumulative distance.
- *Eccentricity update*. This kernel updates the morphological eccentricity scores using the maximum/minimum and point-wise distance streams. A complementary kernel applies a threshold to select a set of final AMEE-derived endmembers at the end of the process.
- *Spectral unmixing*. Finally, this kernel uses as inputs the final endmembers selected in the previous stage and produces a set of endmember fractional abundances for each pixel using the unconstrained inversion process in Eq. 2.

The proposed endmember extraction algorithm has been implemented using the Intel C/C++ compiler. The system used in experiments is based on an Intel Core 2 Quad Q6600 CPU running at 2.4 GHz and with 4 GB of RAM. The computer is equipped with an NVidia$^{\text{TM}}$ GeForce 8800 GTX with 16 multiprocessors, each composed of eight SIMD processors operating at 1,350 Mhz. Each multiprocessor has 8,192 registers, a 16 KB parallel data cache of fast shared memory, and access to 768 MB of global memory. The GPU architecture is graphically illustrated in Fig. 19. The hyperspectral data set used in our experiments is the AVIRIS Cuprite scene.

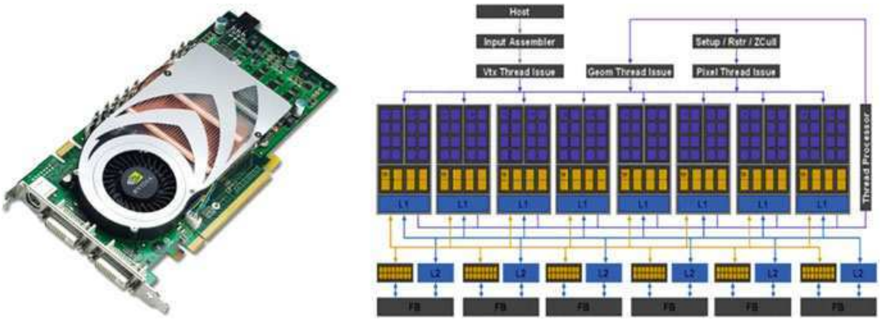


Fig. 19 Architecture of the NVidia$^{\text{TM}}$ GeForce 8800 GTX graphics card used in experiments

Table 2 Processing time (seconds) and speedups for the dual-core CPU and GPU implementations

Algorithm	Processing time (CPU)	Processing time (GPU)	Speedup
AMEE	42.797	1.678	25.504
LSU	4.953	1.297	3.818

Table 2 shows the execution times and speedups measured for the GPU-based implementations of the AMEE and unconstrained LSU algorithms compared to their execution in the quad-core CPU of the system in which the GPU was integrated. The speedup achieved by the GPU implementation of the AMEE algorithm over its respective CPU implementations is close to 25. It should be noted that the speedup achieved for the GPU implementation of AMEE was independent of the structuring element size (the results displayed in Table 2 correspond to a structuring element of 5 × 5 pixels in size which appropriate for endmember extraction from the AVIRIS Cuprite scene, but similar speedups were achieved with other structuring element sizes). On the other hand, Table 2 indicates that the speedup achieved for the parallel implementation of the LSU stage was lower. This is mainly due to the fact that the serial version of LSU is only takes around 5 s to be completed in the quad-core CPU, and it is more difficult to achieve significant speedups in this case since the communication time needed to transfer the data from the CPU to the GPU is more relevant in this case when compared to the total time to finalize the computations in the GPU. As a result, the ratio of computations to communications is smaller for the parallel version of LSU than for the parallel version of AMEE, which has an effect on the achieved speedup. Despite these observations, it can be seen from Table 2 that the considered AVIRIS data cube could be processed in parallel by a full unmixing chain made up of spatial–spectral endmember extraction followed by linear spectral unmixing in just 2.975 s. This response is not strictly in real-time since the cross-track line scan time in AVIRIS, a push-broom instrument, is quite fast (8.3 ms to collect 512 full pixel vectors), which introduces the need to process the considered scene (350 × 350 pixels) in 1.985 s to fully achieve real-time performance. However, we believe that the achieved (near) real-time response time would be relevant in many application domains. Further developments will be pursued in future work in order to approximate real-time performance for on-board data exploitation.

6 Conclusions and Future Research

Endmember extraction is the process of selecting a collection of pure signature spectra of the materials present in a remotely sensed hyperspectral scene. These pure signatures are then used to decompose the scene into abundance fractions by means of a spectral unmixing algorithm. Most techniques available in the endmember extraction literature rely on exploiting the spectral properties of the data alone. As a result, the search for endmembers in a scene is conducted by treating the data as a collection of spectral measurements with no spatial arrangement. In this chapter, we have discussed the role of spatial information in the search for spectral endmembers and further demonstrated via experimental results, using AVIRIS hyperspectral data collected in the framework of a mineral mapping application, that the linear mixture model can benefit from the integration of spatial and spectral information in the task of selecting endmembers. An investigation on the use of the considered spatial–spectral endmember extraction algorithms in conjunction with source separation techniques, such as those described in [60], is a topic deserving future research in this context.

When complex mixtures are present in hyperspectral scenes, nonlinear mixture models may best characterize the resultant mixed spectra for certain endmember distributions. In order to address this issue, we have developed a nonlinear, neural network-based mixture model which is initialized using linear spectral unmixing concepts. The proposed approach is trained with highly representative training sets which can accurately explain the complex nature of the data using only a few training samples. Our study reveals that the most informative training samples for nonlinear mixture characterization are the most highly mixed signatures in the input data set. This observation is in contrast with the overall approach in linear spectral unmixing in which only the purest spectral signatures are used to characterize and decompose spectral mixtures. Critically, if the regions expected to contain the most highly informative training samples for spectral mixture modeling can be identified in advance, then it is possible to direct the training data acquisition procedures to these regions, and thus reduce the number of required training sites without loss of prediction accuracy. This issue is of particular importance in real applications based on the use of airborne/satellite images, in which the acquisition of large training sets is generally very costly in terms of time and finance. To illustrate the concepts above, we have conducted experiments using a set of real hyperspectral images, collected at different altitudes by the DAIS 7915 and ROSIS imaging spectrometers in the framework of an agriculture and farming application in the region of Extremadura, Spain. Although the reported results are promising, it would be also useful to explore in future work the behaviour of spatial–spectral methods in cases where the linear mixture model assumption is no longer valid to describe the mixing systematics of the observed materials, thus conducting a more detailed evaluation of linear versus nonlinear mixture models in different application domains.

Finally, in order to address the extremely high computational requirements introduced by endmember extraction and spectral unmixing applications, this chapter has also presented a parallel implementation case study in which an unmixing chain made up of spatial–spectral endmember extraction followed by unconstrained linear spectral unmixing has been implemented on a specialized graphics processor (GPU). Our experimental results indicate that a low-weight and low-power specialized hardware device such as a GPU has the potential to bridge the gap towards real-time analysis of high dimensional data. This kind of specialized, on-board processing devices are essential to reduce mission payload and obtain analysis results quickly enough for practical use in real applications. Further experimentation will additional hyperspectral scenes will be pursued in future work in order to approximate real-time performance of endmember extraction and spectral unmixing applications for on-board data exploitation.

Acknowledgements This work has been supported by the European Community's Marie Curie Research Training Networks Programme under reference MRTN-CT-2006-035927, Hyperspectral Imaging Network (HYPER-I-NET). This work has also been supported by the Spanish Ministry of Science and Innovation (HYPERCOMP/EODIX project, reference AYA2008-05965-C04-02). Gabriel Martín and Sergio Sánchez are sponsored by research fellowships with references BES-2009-017737 and PTA2009-2611-P, respectively, both associated to the aforementioned project. Funding from Junta de Extremadura (local government) under project PRI09A110 is also gratefully acknowledged. The authors thank Andreas Mueller for his lead of the DLR project that allowed us to obtain the DAIS 7915 and ROSIS hyperspectral datasets over Dehesa areas in Extremadura, Spain, and Robert O. Green at NASA/JPL for making the AVIRIS Cuprite scene available to the scientific community. Last but not least, the authors would like to take this opportunity to gratefully acknowledge the Editors of this volume for their very kind invitation to contribute a chapter on the topic of endmember extraction and spectral unmixing, and for all their support and encouragement during all the stages of the production process for this monograph.

References

1. Goetz, A.F.H., Vane, G., Solomon, J.E., Rock, B.N.: Imaging spectrometry for earth remote sensing. Science **228**, 1147–1153 (1985)
2. Plaza, A., Benediktsson, J.A., Boardman, J., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, J., Marconcini, M., Tilton, J.C., Trianni, G.: Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. **113**, 110–122 (2009)
3. Schaepman, M.E., Ustin, S.L., Plaza, A., Painter, T.H., Verrelst, J., Liang, S.: Earth system science related imaging spectroscopy—an assessment. Remote Sens. Environ. **113**, 123–137 (2009)
4. Adams, J.B., Smith, M.O., Johnson, P.E.: Spectral mixture modeling: a new analysis of rock and soil types at the Viking Lander 1 site. J. Geophys. Res. **91**, 8098–8112 (1986)
5. Keshava, N., Mustard, J.F.: Spectral unmixing. IEEE Signal Process. Mag. **19**(1), 44–57 (2002)

6. Green, R.O., Eastwood, M.L., Sarture, C.M., Chrien, T.G., Aronsson, M., Chippendale, B.J., Faust, J.A., Pavri, B.E., Chovit, C.J., Solis, M., et al.: Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). Remote Sens. Environ. **65**(3), 227–248 (1998)
7. Ball, J.E., Bruce, L.M., Younan, N.: Hyperspectral pixel unmixing via spectral band selection and dc-insensitive singular value decomposition. IEEE Geosci. Remote Sens. Lett. **4**(3), 382–386 (2007)
8. Heinz, D., Chang, C.-I.: Fully constrained least squares linear mixture analysis for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **39**, 529–545 (2001)
9. Plaza, A., Martinez, P., Perez, R., Plaza, J.: A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. IEEE Trans. Geosci. Remote Sens. **42**(3), 650–663 (2004)
10. Chang, C.-I.: Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Kluwer Academic Publishers, New York (2003)
11. Guilfoyle, K.J., Althouse, M.L., Chang, C.-I.: A quantitative and comparative analysis of linear and nonlinear spectral mixture models using radial basis function neural networks. IEEE Trans. Geosci. Remote Sens. **39**, 2314–2318 (2001)
12. Plaza, J., Plaza, A., Perez, R., Martinez, P.: On the use of small training sets for neural network-based characterization of mixed pixels in remotely sensed hyperspectral images. Pattern Recognit. **42**, 3032–3045 (2009)
13. Harsanyi, J.C., Farrand, W., Chang, C.-I.: Detection of subpixel spectral signatures in hyperspectral image sequences. Proc. American Society on Photogrammetry and Remote Sensing Annual Meeting, pp. 236–247 (1994).
14. Boardman, J.W., Kruse, F.A., Green, R.O.: Mapping target signatures via partial unmixing of AVIRIS data. In: Proceedings of JPL Airborne Earth Science Workshop, pp. 23–26 (1995)
15. Winter, M.E.: N-FINDR: an algorithm for fast autonomous spectral endmember determination in hyperspectral data. Proc. SPIE **3753**, 266–277 (1999)
16. Winter, M.E.: A proof of the N-FINDR algorithm for the automated detection of endmembers in a hyperspectral image. Proc. SPIE Algorithms Technol. Multispectral Hyperspectral Ultraspectral Imagery X **5425**, 31–41 (2004)
17. Zortea, M., Plaza, A.: A quantitative and comparative analysis of different implementations of N-FINDR: a fast endmember extraction algorithm. IEEE Geosci. Remote Sens. Lett. **6**, 787–791 (2009)
18. Neville, R.A., Staenz, K., Szeredi, T., Lefebvre, J., Hauff, P.: Automatic endmember extraction from hyperspectral data for mineral exploration. In: Proceedings of 21st Canadian Symposium Remote Sensing, pp. 21–24 (1999)
19. Bowles, J.H., Palmadesso, P.J., Antoniades, J.A., Baumback, M.M., Rickard, L.J.: Use of filter vectors in hyperspectral data analysis. Proc. SPIE Infrared Spaceborne Remote Sens. III **2553**, 148–157 (1995)
20. Ifarraguerri, A., Chang, C.-I.: Multispectral and hyperspectral image analysis with convex cones. IEEE Trans. Geosci. Remote Sens. **37**(2), 756–770 (1999)
21. Nascimento, J.M.P., Bioucas-Dias, J.M.: Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Trans. Geosci. Remote Sens. **43**(4), 898–910 (2005)
22. Harsanyi, J.C., Chang, C.-I.: Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection. IEEE Trans. Geosci. Remote Sens. **32**(4), 779–785 (1994)
23. Berman, M., Kiiveri, H., Lagerstrom, R., Ernst, A., Dunne, R., Huntington, J.F.: ICE: a statistical approach to identifying endmembers in hyperspectral images. IEEE Trans. Geosci. Remote Sens. **42**(10), 2085–2095 (2004)
24. Miao, L., Qi, H.: Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. IEEE Trans. Geosci. Remote Sens. **45**(3), 765–777 (2007)

25. Chang, C.-I., Wu, C.-C., Liu, W., Ouyang, Y.-C.: A new growing method for simplex-based endmember extraction algorithm. IEEE Trans. Geosci. Remote Sens. **44**(10), 2804–2819 (2006)
26. Wang, J., Chang, C.-I.: Applications of independent component analysis in endmember extraction and abundance quantification for hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **44**(9), 2601–2616 (2006)
27. Plaza, A., Chang, C.-I.: Impact of initialization on design of endmember extraction algorithms. IEEE Trans. Geosci. Remote Sens. **44**(11), 3397–3407 (2006)
28. Chang, C.-I., Plaza, A.: A fast iterative algorithm for implementation of pixel purity index. IEEE Geosci. Remote Sens. Lett. **3**(1), 63–67 (2006)
29. Zare, A., Gader, P.: Hyperspectral band selection and endmember detection using sparsity promoting priors. IEEE Geosci. Remote Sens. Lett. **5**(2), 256–260 (2008)
30. Plaza, A., Martinez, P., Plaza, J., Perez, R.: Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. IEEE Trans. Geosci. Remote Sens. **43**(3), 466–479 (2005)
31. Plaza, A., Martinez, P., Perez, R., Plaza, J.: Spatial/spectral endmember extraction by multidimensional morphological operations. IEEE Trans. Geosci. Remote Sens. **40**(9), 2025–2041 (2002)
32. Rogge, D.M., Rivard, B., Zhang, J., Sanchez, A., Harris, J., Feng, J.: Integration of spatial–spectral information for the improved extraction of endmembers. Remote Sens. Environ. **110**(3), 287–303 (2007)
33. Zortea, M., Plaza, A.: Spatial preprocessing for endmember extraction. IEEE Trans. Geosci. Remote Sens. **47**, 2679–2693 (2009)
34. Chang, C.-I.: Hyperspectral Data Exploitation: Theory and Applications. Wiley, New York (2007)
35. Richards, J.A., Jia, X.: Remote Sensing Digital Image Analysis: An Introduction. Springer, Berlin (2006)
36. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley, New York (2003)
37. Green, A.A., Berman, M., Switzer, P., Craig, M.D.: A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Trans. Geosci. Remote Sens. **26**, 65–74 (1988)
38. Ren, H., Chang, C.-I.: Automatic spectral target recognition in hyperspectral imagery. IEEE Trans. Aerosp. Electron. Syst. **39**(4), 1232–1249 (2003)
39. Chang, C.-I., Heinz, D.: Constrained subpixel target detection for remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. **38**, 1144–1159 (2000)
40. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
41. Baraldi, A., Binaghi, E., Blonda, P., Brivio, P.A., Rampini, P.: Comparison of the multilayer perceptron with neuro-fuzzy techniques in the estimation of cover class mixture in remotely sensed data. IEEE Trans. Geosci. Remote Sens. **39**, 994–1005 (2001)
42. Plaza, J., Plaza, A.: Spectral mixture analysis of hyperspectral scenes using intelligently selected training samples. IEEE Geosci. Remote Sens. Lett. **7**, 371–375 (2010)
43. Liu, W., Wu, E.Y.: Comparison of non-linear mixture models. Remote Sens. Environ. **18**, 1976–2003 (2004)
44. Zhuang, X., Engel, B.A., Lozano, D.F., Fernßndez, R.B., Johannsen, C.J.: Optimization of training data required for neuro-classification. Int. J. Remote Sens. **15**, 3271–3277 (1999)
45. Prasad, S., Bruce, L.M.: Overcoming the small sample size problem in hyperspectral classification and detection tasks. Proc. IEEE Int. Geosci. Remote Sens. Symp. **5**, 381–384 (2008)
46. Chi, M., Bruzzone, L.: A semilabeled-sample-driven bagging technique for ill-posed classification problems. IEEE Geosci. Remote Sens. Lett. **2**, 69–73 (2005)
47. Foody, G.M.: The significance of border training patterns in classification by a feedforward neural network using backpropagation learning. Int. J. Remote Sens. **20**, 3549–3562 (1999)

48. Plaza, J., Plaza, A., Martinez, P., Perez, R.: Nonlinear mixture models for analyzing laboratory simulated-forest hyperspectral data. Proc. SPIE **5508**, 660–670 (2003)
49. Borel, C.C., Gerslt, S.A.W.: Nonlinear spectral mixing models for vegetative and soil surfaces. Remote Sens. Environ. **47**, 403–416 (1994)
50. Clark, R.N., Swayze, G.A., Livo, K.E., Kokaly, R.F., Sutley, S.J., Dalton, J.B., McDougal, R.R., Gent, C.A.: Imaging spectroscopy: earth and planetary remote sensing with the USGS tetracorder and expert systems. J. Geophys. Res. **108**, 1–44 (2003)
51. Swayze, G., Clark, R.N., Kruse, F., Sutley, S., Gallagher, A.: Ground-truthing AVIRIS mineral mapping at Cuprite, Nevada. In: Proceedings of JPL Airborne Earth Science Workshop, pp. 47–49 (1992)
52. Pulido, F.J., Diaz, M., Hidalgo, S.J.: Size structure and regeneration of spanish holm oak quercus ilex forests and dehesas: effects of agroforestry use on their long-term sustainability. For. Ecol. Manage. **146**, 1–13 (2001)
53. Plaza, A., Moigne, J.L., Netanyahu, N.S.: Morphological feature extraction for automatic registration of multispectral scenes. Proc. IEEE Int. Geosci. Remote Sens. Symp. **1**, 421–424 (2007)
54. Plaza, A., Chang, C.-I.: High Performance Computing in Remote Sensing. CRC Press, Boca Raton (2007)
55. Plaza, A., Plaza, J., Valencia, D.: Impact of platform heterogeneity on the design of parallel algorithms for morphological processing of high-dimensional image data. J. Supercomput. **40**, 81–107 (2007)
56. Plaza, A., Valencia, D., Plaza, J., Martinez, P.: Commodity cluster-based parallel processing of hyperspectral Imagery. J. Parallel Distrib. Comput. **66**(3), 345–358 (2006)
57. Plaza, A., Chang, C.-I.: Clusters versus FPGA for parallel processing of hyperspectral imagery. Int. J. High Perform. Comput. Appl. **22**(4), 366–385 (2008)
58. Setoain, J., Prieto, M., Tenllado, C., Plaza, A., Tirado, F.: Parallel morphological endmember extraction using commodity graphics hardware. IEEE Geosci. Remote Sens. Lett. **43**, 441–445 (2007)
59. Tarabalka, Y., Haavardsholm, T.V., Kasen, I., Skauli, T.: Real-time anomaly detection in hyperspectral images using multivariate normal mixture models and GPU processing. J. Real-Time Image Process. **4**, 1–14 (2009)
60. Moussaoui, S., Hauksdottir, H., Schmidt, F., Jutten, C., Chanussot, J., Brie, D., Doute, S., Benediktsson, J.: On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation. Neurocomputing **71**, 2194–2208 (2008)

Change Detection in VHR Multispectral Images: Estimation and Reduction of Registration Noise Effects

Lorenzo Bruzzone, Silvia Marchesi and Francesca Bovolo

Abstract In this chapter we address the problem of change detection (CD) in very high geometrical resolution (VHR) optical images by studying the effects of residual misregistration (registration noise) between images acquired over the same geographical area at different times. According to an experimental analysis driven from a theoretical study, we identify the main effects of RN in VHR images and derive some important properties exploiting a polar framework for change vector analysis (CVA). On the basis of the identified properties, we propose: (i) a technique for an adaptive and unsupervised explicit estimation of the RN distribution based on a multiscale analysis of the behavior of spectral change vectors in the polar domain and the Parzen window method; and (ii) an automatic context-sensitive technique robust to registration noise (RN) for CD based on a multiscale analysis in a quantized polar domain. Experimental results obtained on simulated and real VHR multitemporal images confirm the validity of the proposed analysis on RN, the reliability of the derived properties and the effectiveness of the proposed techniques for the estimation of RN distribution and change detection.

Parts of this chapter have been taken from:Bovolo, F., Bruzzone, L., Marchesi, S.: Analysis and adaptive estimation of the registration noise distribution in multitemporal VHR images. IEEE Trans. Geosci. Remote Sens. **47**(9), 3180–3191 (2009). Marchesi, S., Bovolo, F., Bruzzone, L.: A context-sensitive technique robust to registration noise for change detection in VHR images. IEEE Trans. Image Process. **19**(7),1877-1889 (2010).

L. Bruzzone (✉), S. Marchesi and F. Bovolo
Department of Information Engineering and Computer Science, University of Trento,
via Sommarive, 14-38123 Trento, Italy
e-mail: lorenzo.bruzzone@ing.unitn.it

S. Marchesi
e-mail: silvia.marchesi@disi.unitn.it

F. Bovolo
e-mail: francesca.bovolo@disi.unitn.it

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_13,

Keywords Change detection · Very high geometrical resolution images · Registration noise · Change vector analysis · Remote sensing

1 Introduction

Remote sensing images regularly acquired by satellites over the same geographical area, make the analysis of multitemporal data, one of the most interesting research topics for the remote sensing community.

In the last years, the ever increasing availability of multitemporal very high geometrical resolution (VHR) (i.e. 0.6–2.05 m) remote sensing images resulted in new potentially relevant applications related to environmental monitoring and land control and management. Most of these applications are associated with the analysis of dynamic phenomena that occur at different scales and result in changes on the Earth surface. The effects of these phenomena can be detected developing change-detection (CD) techniques capable to automatically identify changes occurred between two VHR images acquired at different times. Several different automatic CD techniques have been proposed in the image processing and remote sensing literature [1–4]. These techniques have been successfully employed in many different application domains, like analysis of growth of urban areas, cadastral map updating, risk analysis, damage assessment, etc. However, most of the available methods are oriented to the analysis of images acquired by medium resolution (MR) sensors and result completely ineffective when dealing with images showing metric resolution (e.g., Ikonos, QuickBird, EROS, SPOT-5, GeoEye-1, World View-2). Therefore it is necessary to develop novel methodologies capable to exploit the properties of VHR images in detecting changes between multitemporal images.

Change-detection techniques generally compare two images acquired at different times by assuming that they are similar to each other except for the presence of changes occurred on the ground. Unfortunately, this assumption is seldom completely satisfied due to differences in atmospheric and sunlight conditions, as well as in the sensor acquisition geometry. In order to satisfy the similarity assumption, pre-processing steps are required, including: image co-registration, radiometric and geometric corrections, and noise reduction. Among the others, co-registration plays a fundamental role as it allows one to obtain a pair of images where corresponding pixels are associated with the same position on the ground. However, in practice, it is not possible to obtain a perfect alignment between images acquired at different times. This may significantly affect the accuracy of the change-detection process. The co-registration procedure becomes more complex and critical (and therefore intrinsically less accurate) when VHR images acquired by the last generation sensors (e.g. Ikonos, QuickBird, EROS, SPOT-5, GeoEye-1, and World View-2) are considered. These images can be acquired with different view angles and often show different geometrical distortions that, even after proper geometric corrections, strongly affect the precision of the registration process, thus resulting in a significant residual registration noise (RN). This noise sharply decreases the accuracy of the change-detection process [5–7].

Another important problem in change detection on VHR images concerns the modeling of the spatial context information of the scene. Most of the classical change-detection techniques generally assume spatial independence among pixels, which is not reasonable in high geometrical resolution data. In order to better exploit the spatial correlation among neighboring pixels and to get accurate and reliable CD maps (both in regions corresponding to border or geometrical details and in homogeneous areas), it is necessary to integrate the spectral information with the spatial one and to model the multiscale properties of the scene. In the literature only few techniques capable to exploit the above-mentioned concepts [8–10] are available.

This chapter aims at analyzing the properties of RN in multitemporal VHR images in order to develop: (i) an adaptive technique for the explicit estimation of the RN distribution; and (ii) an adaptive context-sensitive technique, which: (a) reduces the impact of registration noise in CD on VHR images through a multiscale strategy; (b) considers the spatial dependencies of neighborhood pixels by the definition of multitemporal parcels. The whole analysis is developed in the context of a polar framework for change vector analysis (CVA) recently introduced in the literature for change detection in medium resolution multispectral images [11]. The definition of this framework is based on the analysis of the distribution of spectral change vectors (SCVs) computed according to the CVA technique in the polar domain. The experiments carried out on multitemporal VHR images confirm the validity of the theoretical analysis and the effectiveness of the proposed techniques both in the estimation of the registration noise distribution and in the change-detection approach.

The chapter is organized into seven sections. The next section introduces the notation and background of the polar framework proposed in [11]. Section 3 describes the experimental setup for the study of the properties of RN on simulated multitemporal VHR images and derives the properties of RN. Sections 4 and 5 illustrate an approach to the estimation of the distribution of RN in the polar domain and an adaptive multiscale and context-based technique for CD on VHR images, respectively. Section 6 presents the experimental results obtained on real multitemporal Quickbird images. Finally, Sect. 7 draws the conclusion of this work.

2 Notation and Background

In order to analyze the effects of the RN and to develop a change-detection technique robust to such kind of noise, we take advantage from the theoretical polar framework defined for unsupervised change detection based on CVA proposed in [11]. According to the behavior of SCVs in such polar domain we derive the properties and adaptively estimate the distribution of registration noise. In the following we briefly recall the main concepts of this framework.

Let us consider two VHR multispectral images $\mathbf{X}_1$ and $\mathbf{X}_2$ (e.g. Ikonos, QuickBird, EROS, SPOT-5, GeoEye-1, and World View-2 images) acquired on the same geographical area at different times t_1 and t_2, respectively. Let us assume that these images do not show significant radiometric differences; in particular, let us consider that the spectral channels at the two times have the same mean values (this can be easily obtained with very simple radiometric correction procedures). Let $\Omega = \{\omega_n, \Omega_c\}$ be the set of classes of changed and no-changed pixels to be identified. In greater detail, ω_n represents the class of no-changed pixels, while $\Omega_c = \{\omega_{c_1}, \ldots, \omega_{c_K}\}$ the set of the K possible classes (kinds) of change occurred in the considered area. For simplicity, the polar framework as well as the whole analysis on the registration noise properties is presented considering a two-dimensional feature space (however it can be generalized to the case of more features, see [11] for details). In this manner, it is possible to represent the information in a 2-D domain and to better understand the implications of the analysis. Let $\mathbf{X}_D$ be the multispectral difference image computed according to the CVA technique by subtracting the spectral feature vectors associated with each corresponding spatial position in the two considered images. $\mathbf{X}_D$ is a multidimensional image made up of SCVs defined as:

$$\mathbf{X}_D = \mathbf{X}_2 - \mathbf{X}_1 \tag{1}$$

Under the assumption of 2-D feature vectors, the change information contained in the SCVs can be univocally described by the change vector magnitude ρ and direction ϑ defined as:

$$\vartheta = \tan^{-1}\left(\frac{X_{1,D}}{X_{2,D}}\right) \quad \text{and} \quad \rho = \sqrt{\left(X_{1,D}\right)^2 + \left(X_{2,D}\right)^2} \tag{2}$$

where $X_{b,D}$ is the random variable representing the bth component (spectral channel) of $\mathbf{X}_D$ ($b = \{1, 2\}$). The magnitude-direction domain MD (in which all the SCVs of a given scene are included) can be defined as:

$$MD = \{\rho \in [0, \rho_{\max}] \text{ and } \vartheta \in [0, 2\pi]\} \tag{3}$$

where $\rho_{\max}$ is the highest magnitude of SCVs in the considered images.

According to the previous definitions, the change information for a generic pixel in spatial position (i,j) can be represented in the magnitude-direction domain with a vector z_{ij} having components ρ_{ij} and ϑ_{ij} computed according to (2). From the theoretical analysis reported in [11] and under the above-mentioned assumptions, it is expected that in the polar representation no-changed and changed SCVs result in separate clusters. Unchanged SCVs show a low magnitude and are uniformly distributed with respect to the direction variable. In the polar domain the region associated with them is the *circle of no-changed pixels* C_n, defined as:

$$C_n = \{\rho, \vartheta : 0 < \rho \leq T \text{ and } 0 \leq \vartheta < 2\pi\} \tag{4}$$

This circle is centered at the origin and has a radius equal to the optimal (in the sense of the theoretical Bayesian decision theory) threshold T that separates no-changed from changed pixels. On the opposite, changed SCVs are expected to show a high magnitude. The region associated with them in the polar domain is the *annulus of changed pixels* A_c, which is defined as:

$$A_c = \{\rho, \vartheta : T \leq \rho < \rho_{\max} \text{ and } 0 \leq \vartheta < 2\pi\} \tag{5}$$

This annulus has inner radius T and outer radius given by the maximum among all possible magnitudes for the considered pair of images ($\rho_{\max}$). As no-changed SCVs show preferred directions according to the kind of change occurred on the ground, different kinds of changes can be isolated with a pair of threshold values (ϑ_{k_1} and ϑ_{k_2}) in the direction domain. Each pair of thresholds identifies an *annular sector* S_k *of change* $\omega_k \in \Omega_c$ in the *annulus of changed pixels* A_c defined as:

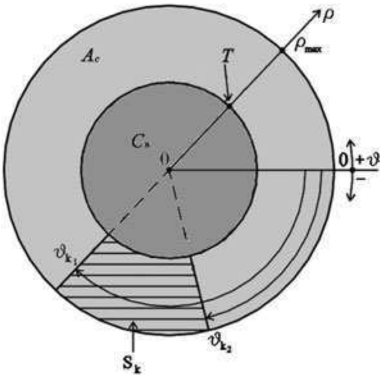
$$S_k = \{\rho, \vartheta : \rho \geq T \text{ and } \vartheta_{k_1} \leq \vartheta \leq \vartheta_{k_2}, 0 \leq \vartheta_{k_1} < \vartheta_{k_2} \leq 2\pi\} \tag{6}$$

All the mentioned regions are depicted in Fig. 1. The reader is referred to [11] for further details on both the polar framework and the general properties of SCVs in this kind of representation.

3 Analysis of Registration Noise Properties

As previously mentioned, residual misregistration affects multitemporal data and represents an important source of noise. In particular, this noise becomes more relevant when dealing with VHR images, as the process of co-registration is more complex and critical. Indeed, images acquired by VHR sensors of the last generation can be acquired with different view angles and often show different geometrical distortions that strongly affect the registration process. Thus, they result in a significant amount of residual registration noise. For this reason, it is very

Fig. 1 Representation of the regions of interest in the CVA polar framework



important to study the properties of RN and to define CD techniques robust to such kind of noise.

The residual registration noise can be modeled as the effect of different types of transformations between the images, such as scale variation, rotation, translation and skew [6]. In this section, for space constraints, only examples modeling the registration noise as a translational effect are reported; however this choice is reasonable as, according to [6], non-translational effects show (from a statistical viewpoint) a behavior similar to that of the translational ones. This behavior is confirmed by experimental results obtained with misregistered data sets generated considering relative rotation and roto-translation, which are not reported here for space constraints.

3.1 Experimental Setup

In order to study the registration noise in the polar CVA domain several data sets have been selected by considering: (i) very high geometrical resolution images acquired by different sensors (i.e., Quickbird, Ikonos, and Pleiades simulator); and (ii) areas with different characteristics, representative of the most frequent land-cover types (i.e., urban, rural, and forestry). Three different experiments have been defined to understand the behavior of RN on unchanged and changed pixels when the misalignment between images increases and the resolution level decreases. To avoid intrinsic differences between images typical of real multitemporal data sets (e.g., atmospheric differences, etc.), in the first phase of the analysis a single-date image has been considered for each data set, while the second acquisition has been simulated.

In the following we describe the experiments considering the analysis conducted on a Quickbird image acquired on the city of Trento (Italy) in July 2006

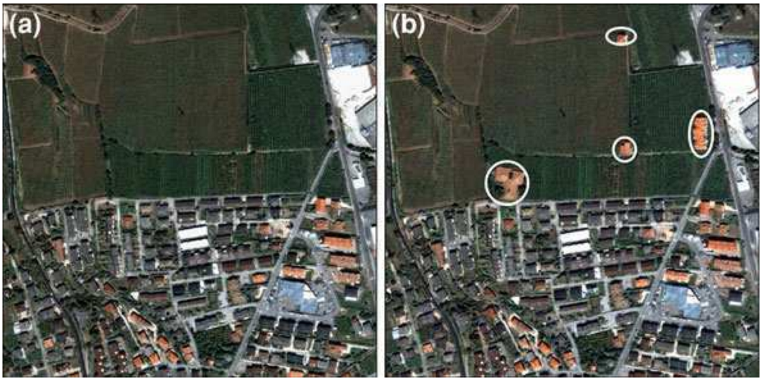


Fig. 2 Channel 4 of pan-sharpened image of the city of Trento (Italy) acquired by the Quickbird VHR multispectral sensor in July 2006 **a** original image without simulated changes, **b** original image with simulated changes (pointed out with *white circles*)

($\mathbf{X}_1$). The selected test site is a section of a full scene including both rural and urban areas (Fig. 2a). Results obtained on other data sets (which contain areas with other characteristics and images acquired by other sensors) are very similar to those reported here, and thus omitted for space constraints. In the following, after an accurate preliminary analysis, among the four available spectral channels, only the red and the near-infrared ones were considered for analyzing the distributions in the polar domain, as they demonstrated to be the most effective in emphasizing the properties of RN (with respect to both changed and unchanged pixels) on the adopted data set. Different choices led to poorer visual representations but to similar conclusions.

3.1.1 Experiment 1: Effects of Increasing Misregistration on Unchanged Pixels

From the considered image $\mathbf{X}_1$ different simulated images $\mathbf{X}_2$ have been generated introducing some pixels of misregistration according to translations in several directions. This resulted in different multitemporal data sets made up of the original image $\mathbf{X}_1$ and of its shifted versions $\mathbf{X}_2$. In particular, we considered misregistration between 1 and 6 pixels, which are possible values when taking into account large VHR images acquired with different view angles and/or in complex areas. After the application of the CVA, the SCV distributions were analyzed in the polar scatterograms in order to derive the properties of RN on unchanged pixels. It is worth noting that the application of the CVA technique to $\mathbf{X}_1$ and a copy of itself when images are perfectly co-registered leads to a multispectral difference image made up of SCVs with all zero components. Thus the representation in polar

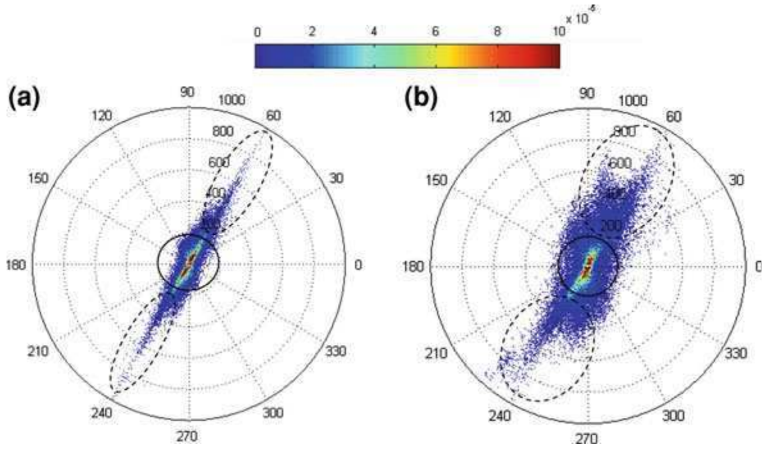


Fig. 3 Scatterograms in the polar coordinate system obtained by applying CVA to the simulated multitemporal data sets (which do not contain any change) that show **a** 2 pixels, and **b** 6 pixels of residual misregistration (Experiment 1)

coordinates of SCVs collapses in a single point at the origin. This is no longer valid if the CVA is applied to misregistered images; in this case the distribution of SCVs in the polar domain corresponds to the distribution of registration noise (as no changes are present in the considered data set). Figure 3 shows an example of the behaviors of scatterograms obtained by applying the CVA technique to $\mathbf{X}_1$ and its 2- and 6-pixels shifted versions, respectively. An analysis of these scatterograms allows us to derive the properties of registration noise when no changes are present between the considered images (see Sect. 3.2).

3.1.2 Experiment 2: Effects of Increasing Misregistration on Changed Pixels

From the considered image $\mathbf{X}_1$ a new image $\mathbf{X}_2$ has been generated by adding simulated changes. These changes have been accurately introduced in order to be as similar as possible to real changes. In particular, some buildings have been added to the scene (see regions marked with white circles in Fig. 2b) taking their geometrical structures and spectral signatures from other real buildings present in the image. All the mentioned buildings have similar spectral signatures and are located on agricultural fields. Therefore the solution to the simulated change-detection problem requires the identification of a single class of changed pixels (ω_{c_1}). As in the first experiment, from the simulated image six new images have been generated introducing some pixels of residual misregistration. This resulted in seven multitemporal data sets made up of the original image ($\mathbf{X}_1$) and one of the simulated images ($\mathbf{X}_2$). In particular, the two images in the first data set are perfectly aligned and differ only for the simulated changes, while the images in the other data sets show also a residual misregistration between 1 and 6 pixels. It is worth noting that when the images are perfectly co-registered the application of the CVA technique to $\mathbf{X}_1$ and to the image obtained introducing simulated changes leads to a multispectral difference image made up of SCVs with non-zero values only for the simulated changes. Other non-zero SCVs (associated with RN) appear if we compute the scatterograms of pair of misregistered images. Figure 4 shows an example of the behaviors of such scatterograms obtained by applying the CVA technique to the image $\mathbf{X}_1$ and: (a) the simulated image perfectly aligned; (b) the simulated image with 2 pixels of residual misregistration; and (c) the simulated image with 6 pixels of residual misregistration. An analysis of these scatterograms (and of the others obtained for different values of misregistration) allowed us to derive the properties of the registration noise on the class of changed pixels (see Sect. 3.2).

3.1.3 Experiment 3: Effects of Misregistration at Different Scales

Further data sets have been generated from the considered image ($\mathbf{X}_1$) and the simulated image including changes with a 4-pixel misregistration ($\mathbf{X}_2$) by applying

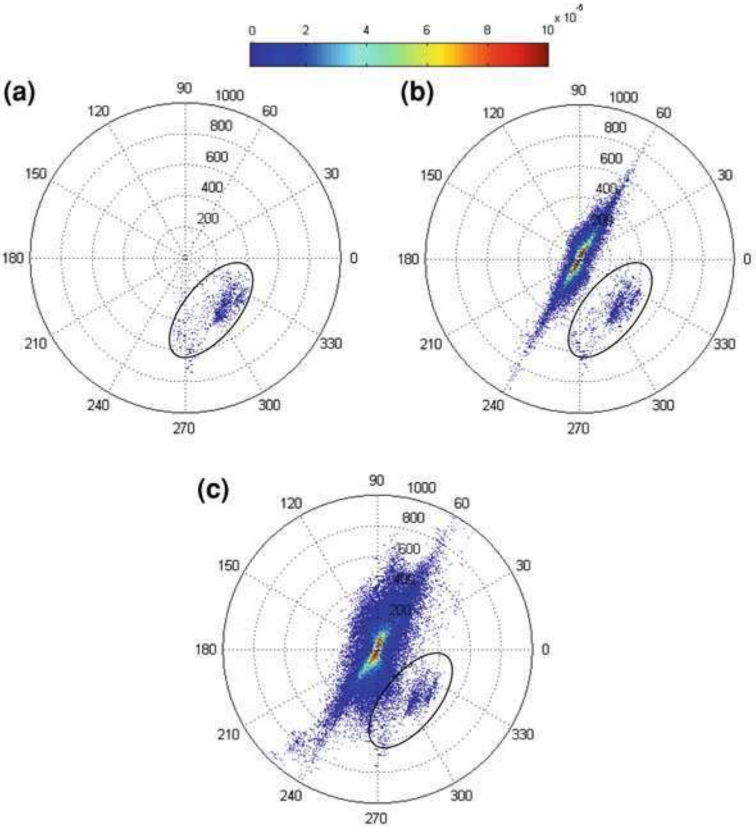


Fig. 4 Scatterograms in the polar coordinate system obtained by applying CVA to the simulated data sets containing changes in the case of **a** perfect alignment between images, **b** 2 pixels of residual misregistration, and **c** 6 pixels of residual misregistration (Experiment 2)

to them a decomposition filter. In this manner two sets of images ($\mathbf{X}_1^n$ and $\mathbf{X}_2^n$, $n = 1, 2, \ldots, N$) have been generated that have lower scale (resolution) than the original ones. These images show a consistent decrease in detail content. In order to obtain the multiscale representation of the images, in the experimental phase different decomposition approaches have been used, as Laplacian/Gaussian pyramid decomposition, iterative sliding window low pass filter, recursively upsampled bicubic filter, wavelet transform. All these approaches provided similar results. For this reason we report only the analysis obtained by applying to $\mathbf{X}_1$ and $\mathbf{X}_2$ the *Daubechies-4* stationary wavelet transform [12, 13]. In the following, as an example, the results achieved considering the pair of images obtained at the third decomposition level ($n = 3$) are reported. It is worth noting that the choice of the level of decomposition is strictly data and application dependent. Figure 5 reports the scatterograms obtained by applying the CVA technique to images $\mathbf{X}_1$ and $\mathbf{X}_2$ (full resolution) and to $\mathbf{X}_1^3$ and $\mathbf{X}_2^3$, respectively. By comparing these scatterograms

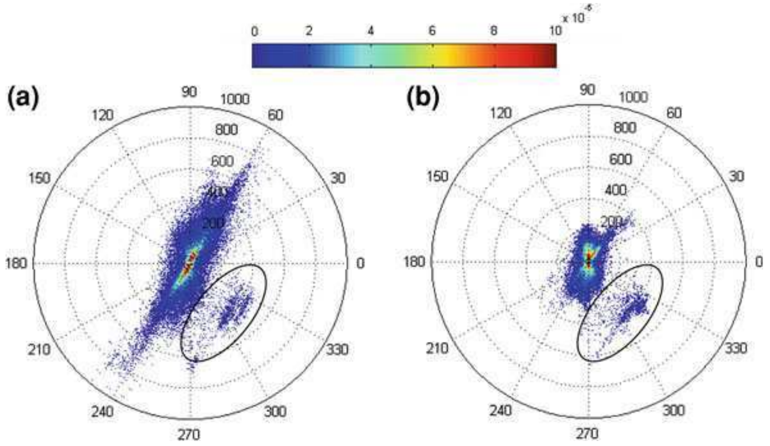


Fig. 5 Scatterograms in the polar coordinate system obtained by applying the CVA technique to the simulated data sets containing changes **a** at full resolution, and **b** at a lower scale (level 3) (Experiment 3)

(and the others obtained for different values of misregistration and at different resolution levels, which are not reported for space constraints) it is possible to study the effects of multiscale decomposition on the distribution of registration noise and of real changes (see Sect. 3.2).

3.2 Properties of RN in VHR Images

An analysis of the scatterograms obtained from the three sets of previously described experiments, and a study on the behavior of SCVs in the polar domain for each investigation setup allowed us to derive some important properties of the registration noise on both unchanged and changed pixels.

Property 1 *RN affects unchanged pixels by: (a) increasing the spread of the cluster in the circle of unchanged pixels C_n with respect to the case of perfectly aligned images; (b) generating clusters of dominant registration noise in the annulus of changes A_c that have properties very similar to those of changed pixels.*

Experiment 1 makes it possible the study of the behavior of the distribution of registration noise (associated with the distribution of SCVs) versus different amounts of misregistration in the polar domain. As the misalignment increases, the number of multitemporal pixels having the same coordinates but that do not correspond to the same position on the ground at the two dates increases. Therefore, the CVA technique performs a comparison between pixels that are not associated to the same area on the ground due to the misalignment. This results in two different contributions to the distribution of RN in the polar domain: (i) the first one is related to the comparison of pixels that belong to the same object in the two images, but that are not associated with the same position on the ground due to misregistration (slightly different spectral signatures due to the heterogeneity of objects in VHR images); (ii) the second one comes from the comparison between pixels that belong to different objects in the two images (pixels associated with details and border regions). These contributions result in: (a) an increase of the standard deviation of the cluster of unchanged pixels when RN increases, and (b) the generation of cluster of unchanged pixels with properties very similar to those of real changes.

Sub-property 1.a *The spread of the cluster in C_n increases by increasing the misalignment.*

Let us consider at first only the effect of the spectral differences between misaligned pixels of the same object. This effect can be observed in the scattergrams of Fig. 3, where some SCVs associated with unchanged pixels that should stay in C_n fall in A_c. Nevertheless, they still show a relative low magnitude and a rather uniform distribution along the direction variable, as it happens for medium resolution images [11] (see regions marked with the continuous line circle in Fig. 3). We can observe that the spread of the cluster of unchanged pixels increases, exhibiting an effect that is sharply amplified with respect to medium resolution images, due to the higher spectral heterogeneity within the objects. It is worth noting that the rather uniform distribution of SCVs along the direction is due to the fact that the structure of objects are usually different for different elements in the scene.

A quantitative analysis carried out on both the magnitude and the direction of SCVs shows that the standard deviation σ_{ω_n} of the class of unchanged patterns ω_n increases in a nonlinear way by increasing the misalignment (see Fig. 6) and, as expected, it tends to saturate when the residual registration noise is over a given threshold.

Statistically, as reported in [11] for the class of unchanged pixels, registration noise generated by the comparison of pixels that belong to the same object can be modeled as a mixture of Gaussian distributions with the same mean values (as the distributions at the two dates are related to the same class) in the Cartesian domain,

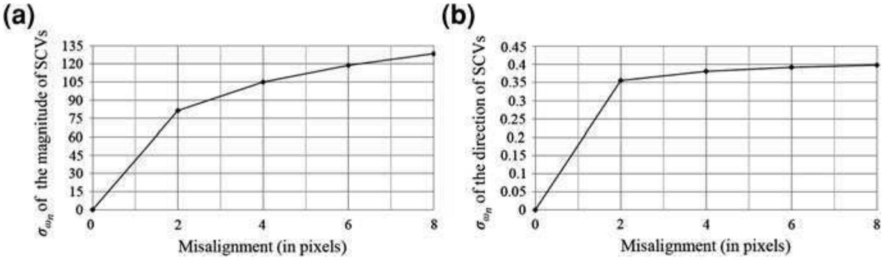


Fig. 6 Behaviors of the standard deviation of **a** the magnitude and **b** the direction of the SCVs in the cluster of unchanged pixels versus the number of pixels of misalignment (Experiment 1)

which corresponds to a Rayleigh distribution along the magnitude variable of the polar domain and to a uniform distribution along the direction variable.

Sub-property 1.b *The clusters of dominant registration noise in A_c have properties very similar to those of real changes and are made up of a number of patterns that increases by increasing the misalignment.*

Let us now consider the effects of pixels that at the two acquisition dates belong to different objects on the ground. In this case significantly different spectral signatures are compared leading to SCVs with large magnitude values. This behavior can be observed in the scatterograms of Fig. 3 where it is possible to note that a large number of unchanged SCVs show a magnitude significantly higher than expected, thus falling in A_c (see regions marked with dashed circles in Fig. 3). In the medium resolution case the distribution of such SCVs is nearly uniform along the direction [11]. On the contrary, when dealing with VHR images, their distribution has preferential directions, resulting in clusters of pixels of registration noise in A_c that exhibit properties very similar to those of changed pixels. Such an effect is mainly due to the comparison of misaligned pixels belonging to different objects with similar structures in different positions of the images. This can be explained, for example, with the regular structure of the urban areas and of the crop rows, as well as with the high frequency content of the VHR images. The number of SCVs composing these clusters increases by increasing the amount of RN. It is worth noting that, on the contrary, when dealing with medium resolution images, the number of misregistered pixels belonging to different objects is small and the effects of registration noise less evident and more uniformly distributed along the direction variable. This is due to both the small amount of geometrical details contained in such images, and the intrinsic effectiveness of classic registration algorithm on medium resolution data. We define the annular sectors in the polar domain associated with these clusters as *sectors of dominant registration noise* $S_{RN_i}^D$:

$$S_{RN_i}^D = \{\rho, \vartheta : \rho \geq T \text{ and } \vartheta_{i_1} \leq \vartheta \leq \vartheta_{i_2}, 0 \leq \vartheta_{i_1} < \vartheta_{i_2} < 2\pi\} \tag{7}$$

Each $S_{RN_i}^D$ can be represented in the polar domain as a sector within A_c bounded from two angular thresholds ϑ_{i_1} and ϑ_{i_2}. This is not surprising as SCVs due to misregistration, exactly as SCVs of true changes, are originated from the comparison of pixels that are associated with different objects on the ground at the two acquisition dates. It follows that sectors of dominant registration noise are very critical because at full resolution they cannot be distinguished from sectors of true changes, resulting in a significant false alarm rate in the change-detection process. Statistically, as reported in [11] for the class of changed pixels, registration noise generated by the comparison of pixels that belong to different classes can be modeled as a mixture of Gaussian distributions with different mean values in the Cartesian domain which corresponds to a Ricean distribution along the magnitude variable of the polar domain and to a non-uniform distribution along the direction variable.

Property 2 *Statistical properties of clusters associated with changed pixels in A_c slowly vary with the amount of misalignment.*

Experiment 2 points out the behaviors of SCVs associated with changed pixels versus the amount of misalignment that affects the considered simulated data sets. Observing Fig. 4 it is possible to note that SCVs associated with the class of changed pixels $\omega_{c_1}$are not significantly affected by an increase of the amount of misregistration between images. Indeed, the cluster of changed pixels can be easily identified in all the three scatterograms and shows quite stable behaviors (see regions marked with circles in Fig. 4). The position of the annular sector S_1 (which identifies pixels belonging to ω_{c_1}) is almost invariant with the misregistration. This behavior allows one to conclude that the registration noise does not affect significantly the properties of the cluster of changed pixels. This is confirmed from a quantitative analysis of the behavior of the mean value μ_{ω_c} and standard deviation σ_{ω_c} of the magnitude of SCVs in the cluster of changed pixels ω_c (for simplicity of notation in the following ω_{c_1} will be indicated as ω_c) versus the amount of misregistration (in pixels). As one can see from Fig. 7, these behaviors do not show significant variations by increasing misregistration.

Nonetheless, the RN indirectly affects the detection of changed pixels (see Property 1) as: (i) the overlap between clusters of changed and unchanged pixels increases when the standard deviation of the patterns in C_n increases; (ii) the presence of sectors of dominant RN in A_c results in false alarms.

Property 3 *Clusters of dominant registration noise in A_c exhibit significant variations of properties versus the scale (resolution) of the images.*

From Experiment 3 we can observe the effects of a multiscale decomposition of the images on pixels associated with both changed and unchanged areas. Let us first consider only unchanged pixels (changed pixels will be discussed in Property 4). As the resolution of the images decreases the presence of small and thin structures diminishes. This results in a reduced impact of registration noise at lower scales (resolutions) as the details and border regions are smoothed out from the low-pass effects associated with scale reduction. Comparing the scatterograms

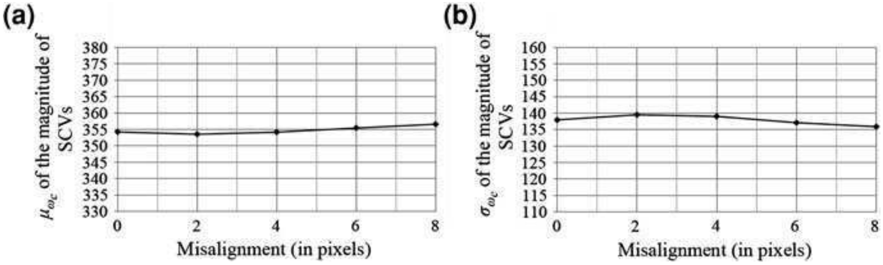


Fig. 7 Behaviors of **a** the mean value μ_{ω_c} and **b** the standard deviation σ_{ω_c} of the magnitude of SCVs in the cluster of changed pixels versus registration noise in the considered images (Experiment 2)

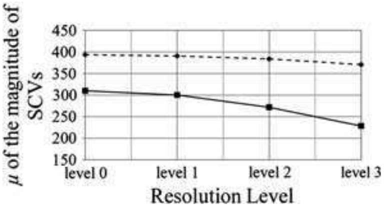


Fig. 8 Behavior of the mean value of the magnitude of SCVs versus the resolution levels (scale) for clusters of change (*dashed line*) and of registration noise (*continuous line*)

of Fig. 5 (derived from Experiment 3) it can be observed that reducing the scale, SCVs associated with registration noise tend to disappear. In other words, decreasing the resolution sectors of dominant registration noise tend to disappear, thus exhibiting a non-stationary behavior with respect to the scale. In particular, such SCVs tend to collapse within C_n. This is confirmed from Fig. 8, which reports the behavior of the mean value of the magnitude of SCVs associated with RN versus the resolution level (scale). As can be seen from the continuous line in the diagram, the mean value of RN clusters rapidly decreases by reducing the resolution.

Property 4 *Clusters associated with changed pixels in A_c exhibit slow varying statistical properties versus the scale (resolution) of the images.*

From Experiment 3 it is also possible to observe the behavior of the cluster of changed pixels when the scale decreases. Observing regions marked with circles in Fig. 5, it is possible to note that the cluster of pixels associated with true changes reduces its spread, but it is not completely smoothed out when the resolution decreases. In other words, it shows a nearby stationary behavior versus the resolution. This is confirmed by an analysis of the behavior of the mean value of the magnitude of SCVs associated with true changes versus the scale. As it can be seen from the dashed line in Fig. 8, the mean value slightly varies with the resolution, but it decreases slower than the one of SCVs associated with registration noise (continuous line in Fig. 8).

From Properties 3 and 4 it follows that the behaviors of changed and unchanged (i.e., the ones due to RN) SCVs that fall in A_c versus the resolution are different: decreasing the resolution, sectors of changes, unlike sectors of dominant registration noise, are preserved. It is worth noting that this property is true under the reasonable and realistic assumption that given the very high geometrical resolution of the sensor, the true significant changes are associated with objects with a non-negligible size. This results in an intrinsic robustness of changes to the scale. On the contrary, misregistration appears in the difference image with linear (or nonlinear) and relatively thin structures having different orientations, that are smoothed out from the scale reduction process. These properties suggest us a multiscale strategy for developing: (i) the adaptive technique for the estimation of registration noise distribution described in the next section; and (ii) the change-detection technique robust to such kind of noise, described in Sect. 5.

4 Proposed Technique for the Adaptive Estimation of the Registration Noise Distribution

As pointed out in the previous section, the properties of RN suggest us to exploit the behaviors of SCVs in the polar domain at different resolution levels (scales) for explicitly estimating the statistical distribution of RN. Properties 3 and 4, in fact, clearly show the usefulness of a multiresolution decomposition in identifying and separating annular sectors of dominant registration noise from annular sectors of real changes. If we reduce the resolution of images, we implicitly decrease the impact of the registration noise with respect to that on the original scene (Property 3), while true changes maintain a good stability (Property 4). In other words, the lower is the geometrical resolution, the lower is the probability of identifying in the polar representation annular sectors of dominant registration noise. This means that at low resolution, in the *annulus of changed pixels* mainly sectors (i.e., clusters) due to the presence of true changes on the ground are detected. Thus, by comparing the clusters present in the polar domain at full resolution and at reduced resolution, it is possible to identify annular sectors dominated from registration noise and separate them from annular sectors of changes. It is worth noting that this is made possible from the thin structures associated with RN that result in strong changes in the corresponding SCV clusters when the low pass effect of the scale reduction is considered.

On the basis of the aforementioned analysis, we propose an adaptive multiscale strategy that exploits the behaviors of SCVs to identify the distribution of the registration noise. The proposed technique compares the distribution of the SCVs at the highest resolution level with the one at a lower level in order to derive the distribution of registration noise at full resolution. In particular, first of all the two multitemporal images are decomposed according to a multiscale transformation (as described in Sect. 3 different algorithms can be used, like stationary wavelet transform, recursively upsampled bicubic filter, etc.). In greater detail we applied the two-dimensional discrete stationary wavelet transform (2D-SWT); this decomposition technique is obtained as an extension of the one-dimensional discrete stationary wavelet transform by applying one-dimensional filters independently along both dimensions of the considered image. In particular, two filters with different impulse responses are considered to built up the SWT filter bank: (i) a low-pass filter with impulse response $l(.)$; and (ii) a high-pass filter with impulse response $h(.)$. A one-step wavelet decomposition applies both filters separately, first along columns and then along rows. The original image $\mathbf{X}_i$ ($i = 1, 2$) is decomposed into a low resolution image (the approximation sub-band X_i^{LL}), containing low spatial frequencies in both the horizontal and the vertical direction, and three detail images X_i^{LH}, X_i^{HL} and X_i^{HH}, which correspond to the horizontal, vertical and diagonal detail sub-bands at resolution level 1, respectively. Note that, superscripts *LL*, *LH*, *HL* and *HH* specify the order on which high- and low-pass filters have been applied to obtain the considered sub-band. The multiscale decomposition is obtained by recursively applying the described procedure to the approximation sub-band obtained at each scale 2^n. Thus the output at a generic resolution level n can be express analytically as follows:

$$\begin{aligned}
\mathrm{X}_i^{LL(n+1)}(i,j) &= \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} l^n[p]l^n[q]\mathrm{X}_i^{LLn}(i+p,j+q)\\
\mathrm{X}_i^{LH(n+1)}(i,j) &= \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} l^n[p]h^n[q]\mathrm{X}_i^{LLn}(i+p,j+q)\\
\mathrm{X}_i^{HL(n+1)}(i,j) &= \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} h^n[p]l^n[q]\mathrm{X}_i^{LLn}(i+p,j+q)\\
\mathrm{X}_i^{HH(n+1)}(i,j) &= \sum_{p=0}^{D^n-1}\sum_{q=0}^{D^n-1} h^n[p]h^n[q]\mathrm{X}_i^{LLn}(i+p,j+q)
\end{aligned} \tag{8}$$

where D^n is the length of the wavelet filters at resolution level n. At each decomposition step, the length of the impulse response of both high- and low-pass filters is upsampled by a factor 2. Thus, filter coefficients for computing sub-bands at resolution level $n + 1$ can be obtained by applying a dilation operation to the filter coefficients used to compute level n. In particular, 2^{n-1} zeros are inserted between the filter coefficients used to compute sub-bands at the lower resolution level. This allows a reduction in the bandwidth of the filters by a factor 2 between subsequent resolution levels. Filter coefficients of the first decomposition step for $n = 0$ depend on the selected wavelet family and on the length of the chosen wavelet filter. To this purpose, we selected the *Daubechies* wavelet family and set the filter length to 8. The finite impulse response of the high-pass filter for the decomposition step is obtained by satisfying the properties of the quadrature mirror filters. This is done by reversing the order of the low-pass decomposition filter coefficient and by changing the sign of the even indexed coefficients [13].

In order to perform the proposed analysis, one must return to the original image domain. This is done by applying only to the approximation sub-bands the two-dimensional inverse discrete stationary wavelet transform (2D-ISWT) at each resolution level independently. In this manner we obtain two sets of images $\mathbf{X}_{MS_i} = \left\{\mathbf{X}_i^0, \ldots, \mathbf{X}_i^n, \ldots, \mathbf{X}_i^{N-1}\right\}$ where the subscript i ($i = 1, 2$) denotes the acquisition date, and the superscript n ($n = 0, 1,\ldots, N - 1$) indicates the resolution level (note that $\mathbf{X}_i^0 \equiv \mathbf{X}_i$). Then the CVA technique is applied to each corresponding pair of images $\left(\mathbf{X}_1^n, \mathbf{X}_2^n\right)$ and the distributions of the direction of SCVs at different resolution levels are analyzed. In particular, the behaviors of SCVs in A_c are studied. To this purpose, we compute the conditional density of the direction of pixels in A_c. In order to estimate this distribution we take advantages from the Parzen windows technique [14–17], which is a basic and effective estimation method for one dimensional problems. According to this technique the density estimation can be computed as:

$$\hat{p}_n(\vartheta|\rho \geq T) = \frac{1}{M_n}\sum_{m=1}^{M_n}\frac{1}{h_n}\gamma\left(\frac{\vartheta - \vartheta_m}{h_n}\right) \tag{9}$$

where T is the threshold value that separates the circle of unchanged pixels from the annulus of changed pixels (it can be retrieved either manually or in an automatic way through one of the algorithms proposed in the literature [18, 19], see Sect. 2), n ($n = 0, 1, \ldots, N - 1$) denotes the resolution level at which the estimation is computed, ϑ_m represents the direction value of the mth SCV in A_c, M_n is the number of SCVs in A_c at scale n, $\gamma(.)$ is the kernel function used in the estimation process and h_n is the width of the kernel window (smoothing parameter) at scale n.

In particular, we used Gaussian kernel, so that the final estimation is given by:

$$\hat{p}_n(\vartheta|\rho \geq T) = \frac{1}{M_n}\sum_{m=1}^{M_n}\frac{1}{h_n\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\vartheta - \vartheta_m}{h_n}\right)^2\right] \tag{10}$$

For what concerns the smoothing parameter, which in our case is represented by the standard deviation of the Gaussian function, we propose to compute it as a function of the number of pixels that fall in A_c. In particular, considering a Gaussian kernel, the width value at scale n can be derived as in [14]:

$$h_n = sig * \left(\frac{4}{3M_n}\right)^{1/5} \tag{11}$$

where:

$$sig = \underset{m=1,\ldots,M_n}{median}\left|\vartheta_m - \underset{m=1,\ldots,M_n}{median}(\vartheta_m)\right|/0.6745 \tag{12}$$

Then we observe the behaviors of $\hat{p}_n(\vartheta|\rho \geq T)$ versus the scale. According to the properties of RN, this density decreases at lower resolutions in the *annular sectors of dominant registration noise* $S^D_{RN_i}$, whereas it remains nearby constant in the *annular sectors of true changes* S_k. On the basis of this analysis, we propose to estimate the conditional density of registration noise in the direction domain $\hat{p}_{RN}(\vartheta|\rho \geq T)$ as:

$$\hat{p}_{RN}(\vartheta|\rho \geq T) = C[P_0(\rho \geq T)\hat{p}_0(\vartheta|\rho \geq T) - P_{N-1}(\rho \geq T)\hat{p}_{N-1}(\vartheta|\rho \geq T)] \tag{13}$$

where $P_n(\rho \geq T)$ is the probability of SCVs to be in A_c at scale n, $\hat{p}_0(\vartheta|\rho \geq T)$ and $\hat{p}_{N-1}(\vartheta|\rho \geq T)$ are the marginal conditional densities of the direction of pixels in A_c at the full resolution and at the lowest considered resolution level ($N - 1$), respectively, and C is a constant defined such that $\int_{-\infty}^{+\infty} \hat{p}_{RN}(\vartheta|\rho \geq T)d\vartheta = 1$.

The term $P_n(\rho \geq T)$ in (13) is necessary in order to obtain a reliable comparison between distributions at different resolution levels.

In this way we obtain an explicit estimation of the distribution of registration noise that is adaptive (in the sense that it intrinsically takes into account the properties of the considered images). It is worth noting that this estimated distribution represents the behavior of RN at full scale (resolution). In the proposed technique the analysis at the lowest resolution is only used for separating the RN contribution from that of true changes (and of other possible sources of noise).

5 Proposed Change-Detection Technique Robust to Registration Noise

As previously pointed out, the multiscale properties of RN not only allow us to define a strategy for the estimation of RN noise, but also are important for the definition of the proposed change-detection technique robust to such kind of noise. Starting from the same assumption (true significant changes are associated with objects with a non-negligible size, while misregistration appears in the multispectral difference image with relatively thin structures having different orientations) and taking advantages from the multiscale technique for the adaptive estimation of RN distribution, we propose a change-detection technique that exploits a multiscale decomposition in order to automatically extract information about registration noise, and generates the final change-detection map working at full resolution. In this way we preserve the high geometrical detail content of VHR images. In addition, in order to exploit the specific properties of VHR images, the proposed technique adaptively models also the spatial context information.

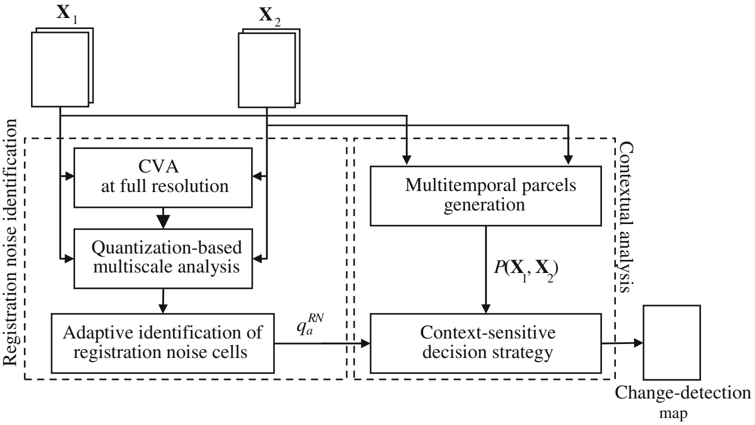

$\mathbf{X}_1$
$\mathbf{X}_2$
Registration noise identification
CVA at full resolution
Quantization-based multiscale analysis
Adaptive identification of registration noise cells
q_a^{RN}
Contextual analysis
Multitemporal parcels generation
$P(\mathbf{X}_1, \mathbf{X}_2)$
Context-sensitive decision strategy
Change-detection map

Fig. 9 General architecture of the proposed multiscale and parcel-based change-detection technique

The proposed method can be divided into two main phases: (i) registration noise identification; and (ii) context-sensitive decision strategy for the generation of the final change-detection map. The main idea of the developed technique is to detect the regions of the polar framework where the registration noise is dominant according to a multiscale strategy, and to consider the spatial-context information through the definition of multitemporal parcels in order to generate the final change-detection map (see Fig. 9). In the following details on the two phases are reported.

5.1 Registration Noise Identification

The first phase of the proposed technique aims at identifying in an automatic way the regions related to registration noise in the polar domain. To this purpose, we take advantages from the technique described in the previous section; in particular, we exploit the multiscale analysis and we add some steps for retrieving in an automatic way a label for each pixel related to its membership to registration noise or not. In order to identify registration noise, we apply an analysis based on the following three steps: (1) CVA at full resolution (identification at full resolution of regions in the polar domain candidate to include registration noise SCVs, i.e. A_c); (2) quantization-based analysis of the SCV distributions at different resolution levels; and (3) adaptive identification of registration noise cells.

In the first step the CVA technique is applied to the original images $\mathbf{X}_1$ and $\mathbf{X}_2$, and the threshold value T that separates the *circle of no-changed pixels* from the *annulus of changed pixels* is estimated. SCVs in C_n are labeled as no-changed SCVs, whereas pixels in A_c should be further analyzed in order to separate SCVs associated with registration noise from pixels of true changes.

To this end, in the second step, A_c is divided into M uniformly distributed quantization cells q_m ($m = 1, \ldots, M$) ($A_c = \{q_1, q_2, \ldots, q_M\}$) of fixed shape and size. Each cell is characterized by its extension $\Delta\rho$ and $\Delta\vartheta$ in the magnitude and in the direction coordinates respectively (see Fig. 10). It is worth noting that the choice of the cell size is an important aspect to consider; however, similar results

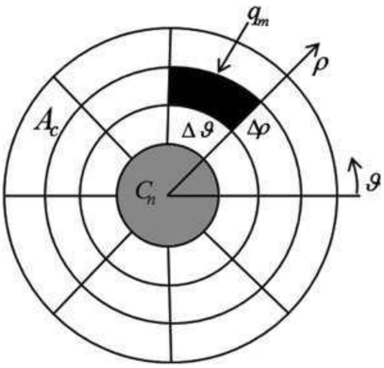


Fig. 10 Quantized magnitude-direction polar domain

can be obtained with different quantization values in consistent ranges of $\Delta\rho$ and $\Delta\vartheta$. Once cells have been defined, the two multitemporal images are decomposed according to a multiscale transformation obtaining two sets of images $\mathbf{X}_{MS_i} = \left\{\mathbf{X}_i^0, \ldots, \mathbf{X}_i^n, \ldots, \mathbf{X}_i^{N-1}\right\}$, as described in the previous section. The CVA technique is applied to each corresponding pair $\left(\mathbf{X}_1^n,\ \mathbf{X}_2^n\right)$, $n = 1, 2, \ldots, N-1$, of low resolution images in $\mathbf{X}_{MS_1}$ and $\mathbf{X}_{MS_2}$. Then the distribution of SCVs within each cell is studied at different scales. In particular, for each set of pixels with SCVs falling in a given cell q_m ($m = 1, 2, \ldots, M$) at full resolution, the behavior of the distribution of the same SCVs at resolution level $N-1$ (i.e., the lowest considered one) is analyzed in order to identify whether the cell is associated with registration noise or not. It is worth noting that the maximum level of decomposition $N-1$ has to be selected according to the size of expected main change structures in the considered images. As for the definition of the registration noise distribution, the main idea of this procedure is to identify cells of registration noise through a comparison between the distribution of the magnitude of SCVs at full resolution and at the lowest considered resolution. In particular, in this procedure the behavior of the mean value of SCVs on the magnitude variable at different resolutions is analyzed, considering the multiscale properties of RN. In the proposed method the mean value μ_{ρ,q_m}^0 of the magnitude ρ of SCVs that fall within a cell q_m at full resolution (level 0) is compared with the mean value μ_{ρ,q_m}^{N-1} that the same SCVs have at resolution level $N-1$.[1] A cell is associated with *RN* or not (*RNfree*) according to the following decision rule:

$$q_m \in \begin{cases} RNfree & \text{if } \left\{\left|\mu_{\rho,q_m}^0 - \mu_{\rho,q_m}^{N-1}\right| < K\right\} \\ RN & \text{if } \left\{\left|\mu_{\rho,q_m}^0 - \mu_{\rho,q_m}^{N-1}\right| \geq K\right\} \end{cases} \tag{14}$$

where K is a threshold value empirically set as equal to the difference between the mean value of all the SCVs falling in A_c at full resolution and the mean value of the corresponding SCVs at the lowest level, i.e.:

$$K = \mu_{\rho,A_c}^0 - \mu_{\rho,A_c}^{N-1} \tag{15}$$

It is worth noting that small variations of the threshold value around the automatic retrieved one do not significantly affect the identification of registration noise clusters. Let q_m^{RN} be a generic cell q_m associated with registration noise according to (14). A generic SCV z_{ij} is associated with registration noise if it falls within a cell q_m^{RN}, i.e.

[1] It is worth noting that in order to identify cells of registration noise we do not analyze the behavior of SCVs that fall within the same cell at different resolution levels, but we consider SCVs that at the highest resolution fall within a cell and the same SCVs at the lowest considered level. This approach allows us to follow the low-pass effect of the decomposition filter, which causes a migration of SCVs toward the origin of the polar domain.

$$z_{ij} \in \begin{cases} RN & \text{if } z_{ij} \in q_m^{RN} \\ RNfree & \text{otherwise} \end{cases} \tag{16}$$

In this way we locate the SCVs affected by registration noise in the polar domain.

5.2 Context-Sensitive Decision Strategy for the Generation of the Final Change-Detection Map

The retrieved information on each adaptive cell is used for properly driving the generation of the final change-detection map according to a context-sensitive parcel-based procedure. Parcels are defined as regions that adaptively characterize the local neighborhood of each pixel in the considered scene and are homogeneous in both temporal images [9, 20]. The adaptive nature of multitemporal parcels allows one to model complex objects in the investigated scene as well as borders of the changed areas and geometrical details. In order to generate multitemporal parcels from the two original images we first compute two segmentation maps $P(\mathbf{X}_1)$ and $P(\mathbf{X}_2)$ applying a segmentation algorithm separately to images $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. In this work a region growing segmentation algorithm was considered, however any different kind of technique can be adopted. Each $P(\mathbf{X}_t)$ represents a partition of image $\mathbf{X}_t$ ($t = 1, 2$) in disjoint regions of spatially contiguous pixels. Each single region in both partitions satisfies a homogeneity measure $H(.)$ that involves spectral and spatial properties [21, 22]. The desired representation of the spatio-temporal context of the considered scene is obtained merging the two segmentations. The final output is a partition $P(\mathbf{X}_1, \mathbf{X}_2)$ shared by both considered images made of N regions p_r ($r = 1, \ldots, R$) called parcels. The defined multitemporal parcels satisfy the following conditions:

$$\begin{aligned} H[\mathbf{X}_1(p_r)] = true \text{ AND } H[\mathbf{X}_2(p_r)] = true \\ H[\mathbf{X}_1(p_r) \cup X_1(p_k)] = false \text{ OR } H[\mathbf{X}_2(p_r) \cup \mathbf{X}_2(p_k)] = false \\ \forall\, r, k = 1, \ldots, R\, and\, r \neq k \end{aligned} \tag{17}$$

where $\mathbf{X}_t(p_r)$ represent the portion of image $\mathbf{X}_t$ ($t = 1, 2$) covered by parcel p_r ($r = 1, \ldots, R$) and p_r and p_k are adjacent.

The spatial-context information associated to each parcel is integrated to the information about presence or absence of registration noise retrieved from the multiscale analysis in the previous phase. Let Z_r be the set of spectral change vectors corresponding to the pixels included in parcel p_r, i.e. $Z_r = \left\{ z_{ij} | z_{ij} \in p_r \right\}$. Each SCV in Z_r can assume one out of three labels. Therefore the SCVs (i.e., the pixels) in a generic parcel p_r can be divided into three subsets: (i) Z_r^{RN}which includes SCVs of registration noise labeled according to (16); (ii) Z_r^{RNfree} which includes SCVs that are not affected by registration noise according to (16); and

(iii) $Z_r^{\omega_n}$ which includes SCVs that fall into C_n. According to this notation, all the SCVs in a generic parcel p_r and thus the parcel itself are classified as changed or no-changed according to the following majority rule:

$$p_r \in \begin{cases} \omega_n & \text{if } \dfrac{\left|Z_r^{RN}\right| + \left|Z_r^{\omega_n}\right|}{\left|Z_r\right|} \geq 0.5 \\ \Omega_c & \text{otherwise} \end{cases} \tag{18}$$

where |.| is the mathematical operator that returns the cardinality of sets. In other words a parcel p_r (and therefore all the pixels in it) is labeled as no-changed if the most of the SCVs belonging to it either have been classified as SCVs affected by registration noise according to (16) or fall into C_n. It is worth noting that the proposed approach allows us to create a relationship between the RN information retrieved in the polar domain (related to spectral change vectors) and the spatial information of the original images (related to pixels and parcels). The final change-detection map is obtained at full resolution, as low resolution components extracted from the multiscale analysis are used only for detecting quantization cells associated with registration noise. Thus the obtained change-detection map adequately models geometrical details present in the analyzed VHR images, reproducing accurately both border and homogeneous changed regions.

6 Experimental Results

In this section the experimental analysis conducted on real data is presented. First of all the data set is described, then the experimental analysis on both the reliability of the derived properties of RN and the effectiveness of the proposed method to estimate the distribution of RN on real multitemporal images is presented. Finally, the proposed multiscale and parcel-based technique is applied to real data.

6.1 Data Set Description

In order to assess the effectiveness of the proposed techniques, a multitemporal data set made of two images acquired on the city of Trento (Italy) by the Quickbird multispectral sensor in October 2005 and July 2006 was considered. In the pre-processing phase the two images were: (i) pan-sharpened; (ii) radiometrically corrected; and (iii) co-registered. In particular, we considered pan-sharpened images as we expect that the pan-sharpening process can improve the results of the change-detection process, as demonstrated in previous work [23]. To this purpose we applied the Gram–Schmidt procedure implemented in the ENVI software package [24] to the panchromatic channel and the four bands of the multispectral images. Concerning radiometric corrections, we simply normalized the images by subtracting from each spectral channel of the two considered images its mean value. The registration process was carried out by using a polynomial function of order 2 according to 14 ground control points (GCPs), and by applying a nearest neighbor interpolation [24]. In our experiments we did not use more advanced registration techniques and procedures for geometric corrections for better assessing the robustness of the proposed method to the residual registration noise. The final data set was made up of two pan-sharpened multitemporal and multispectral images of 984 × 984 pixels (a section of the full scene) with a spatial resolution of 0.7 m on the ground, which have a residual misregistration of about 1 pixel on GCPs. Figure 11a, b shows channel 4 of the pan-sharpened images $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Between the two acquisitions two kinds of changes occurred: (i) simulated changes that consist of new houses introduced on the rural area (white circles in Fig. 11b); and (ii) real changes that consist of some roofs rebuilt in the urban area (black circles in Fig. 11b). It is worth noting that

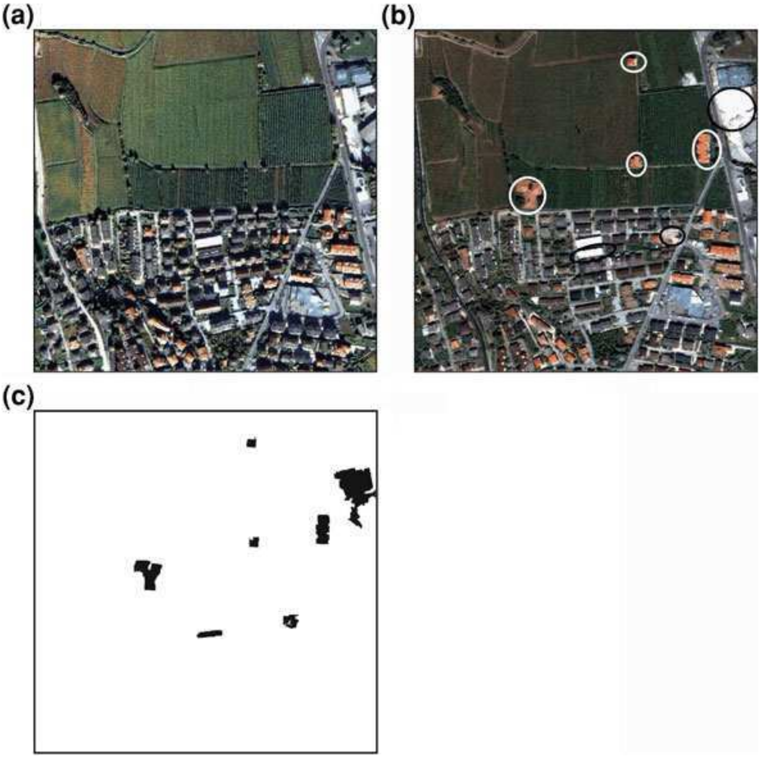


Fig. 11 Channel 4 of pan-sharpened images of the Trento city (Italy) acquired by the Quickbird VHR multispectral sensor in: **a** October 2005; and **b** July 2006 (simulated changes appear in the regions marked with *white circle*, while real changes occurred between the two acquisition dates appear in regions marked with *black circles*). **c** Change-detection reference map

simulated changes have been accurately introduced in order to be as similar as possible to real changes. Simulated buildings have been added to the scene taking their geometrical structures and spectral signatures from other real buildings present in other portions of the available full scene in order to take into account the image dynamic and noise properties. Moreover, between the two dates also significant seasonal differences in the crop rows and in the shape of shadows are present, due to the different acquisition seasons (i.e., summer and autumn) of the considered images. It is worth noting that from the theoretical viewpoint the proposed technique identifies all significant spectral changes occurred between the two images, as no semantic information is exploited for discriminating different kinds of spectral changes.

To perform a quantitative assessment of the effectiveness of the proposed method, a reference map (which includes 20,602 changed pixels and 968,256 no-changed pixels) was defined according to both the available prior knowledge on the considered area and to a visual analysis of images (see Fig. 11c). The reference map only reports changes that are significant with respect to the considered application.

6.2 Estimation Results

For applying the proposed method to the estimation of registration noise, the original images $\mathbf{X}_1$ and $\mathbf{X}_2$ were transformed to lower scales through a four-step stationary wavelet transform [12, 20] using 4th order orthogonal filters of the *Daubechies* family. The maximum level of decomposition was selected according to a tradeoff between the degree of sensitivity desired in the RN estimation and the size of the expected main change structures present in the images. Then the CVA technique was applied to the images at different scales. In order to separate the *circle of unchanged pixels* (C_n) from the *annulus of changed pixels* (A_c), for each data set a proper threshold value T on the magnitude variable was retrieved according to a trial-and-error procedure (we did not use an automatic technique for avoiding biases introduced from the threshold selection method in the evaluation of the effectiveness of the proposed method). However, at an operational level, one of the thresholding algorithms proposed in the literature can be used [18, 19]. In greater detail, in order to find the optimal threshold for our purposes, the whole analysis for the estimation of the RN distribution has been conducted for different values of the thresholds T in a consistent range of the magnitude values. All of them provided similar results in the estimation of registration noise. For space constraints, in the following only the results obtained with a single threshold value for each data set are reported. The marginal conditional densities of the directions of pixels in A_c at the highest resolution and at a lower resolution levels (see Fig. 12) were computed according to (10), and finally the conditional density of registration noise was estimated according to (13) (see Fig. 13). From an analysis

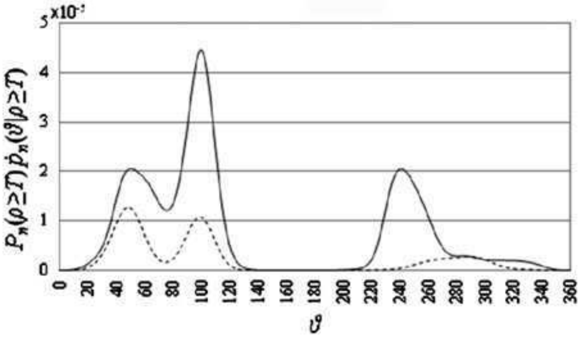


Fig. 12 Marginal weighted conditional densities $P_n(\rho \geq T)\hat{p}_n(\vartheta|\rho \geq T)$ of the direction in A_c at full resolution (*continuous line*) and at level 4 of the *Daubechies* stationary wavelet transform (*dashed line*)

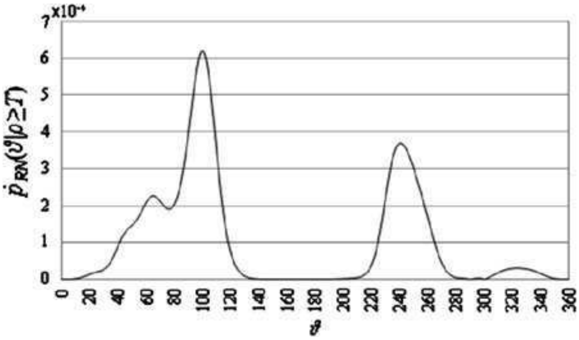


Fig. 13 Estimated conditional density $\hat{p}_{RN}(\vartheta|\rho \geq T)$of registration noise obtained with the proposed technique

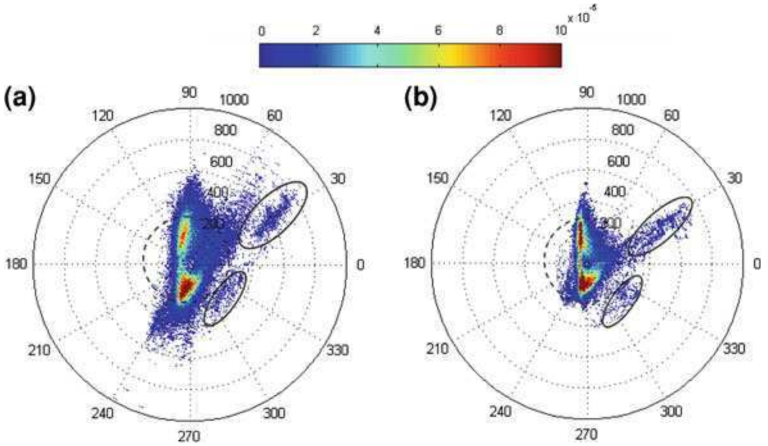


Fig. 14 Scatterograms in the polar coordinate system of **a** the full resolution original difference image$\mathbf{X}_D^0$, and **b** the low resolution image $\mathbf{X}_D^4$ obtained at level 4 of the wavelet decomposition. *Dashed circles* separate C_n from A_c, while *continuous circles* indicate sectors of true changes

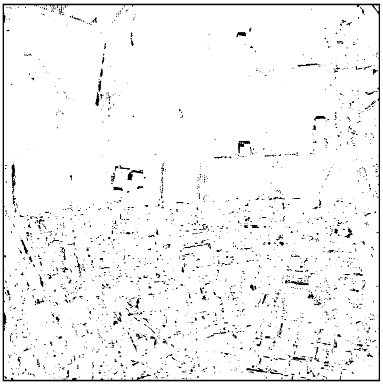


Fig. 15 Registration-noise map obtained by thresholding the $\hat{p}_{RN}(\vartheta|\rho \geq T)$ obtained with the proposed technique

of the behavior of $\hat{p}_{RN}(\vartheta|\rho \geq T)$ it is possible to identify three main modes, which potentially define sectors where the registration noise is dominant. A comparison between the scatterograms at full and at low resolution (see Fig. 14) points out that in the sectors corresponding to the three modes of $\hat{p}_{RN}(\vartheta|\rho \geq T)$ the density of the magnitude of SCVs in the annulus of changed pixel reduces significantly when the resolution decreases, whereas in the others it is nearly constant. In particular, it is possible to verify that the sectors in which the behavior of SCVs is quite stable correspond to sectors of true changes (continuous circles in Fig. 14). This behavior also confirms the properties derived from the simulated data sets.

To further understand the effectiveness of the proposed estimation technique, we identified sectors of dominant registration noise by thresholding the conditional density of registration noise as defined in (13). In other words pixels with a high probability of being of registration noise were isolated. Fig. 15 shows the pixels with a probability of being of registration noise higher than 1×10^{-4}. This threshold was set empirically and led to the definition in the annulus of changed pixels (i.e., Ac=$\{\rho,\vartheta$: $\rho > 310\}$) of two sectors of dominant registration noise. The first sector has direction values between 35° and 115° whereas the second one has direction values between 225° and 265°.

A visual analysis of (Fig. 15) confirms that the regions identified as registration noise by the proposed technique are associated with areas that show the effects of misregistration between the multitemporal images, as they mainly refer to border regions of buildings located in the urban area, to roads and to crop rows. In addition, it is possible to note that the regions identified in the registration-noise map do not belong to areas of changes. This behavior confirms the effectiveness of the proposed technique that properly distinguishes between registration noise and true changes contributions in the estimation of $\hat{p}_{RN}(\vartheta|\rho \geq T)$ (Fig. 15).

6.3 Change-Detection Results

The effectiveness of the proposed change-detection technique was tested on the real data set described in Sect. 6.1. The CVA technique was applied to the red and near-infrared spectral channels of the original images. According to the proposed technique the decision threshold that separates A_c from C_n was computed in an automatic way (T was set equal to 220). SCVs in C_n were labeled as no-changed SCVs, whereas the *annulus of changed pixels* was divided into quantization cells of size $\Delta\rho \times \Delta\vartheta$ in order to further distinguish between registration noise pixels and changed pixels. In the following, for simplicity, the results obtained with a quantization equal to 300 $\times$ 10 are reported. We refer the reader to [25] for a more detailed analysis of the effects of the quantization values on the CD results. The CVA technique was also applied to the low resolution images obtained through the decomposition procedure described in Sect. 4 and the adaptive analysis of the SCVs distribution was performed. For each cell the difference in the mean value of the magnitude of SCVs between the resolution level 0 and 4 was computed and compared with the threshold K derived according to (15) (for T equal to 220 the value of K resulted equal to 190). SCVs falling into cells in which the difference was greater than K were classified as belonging to registration noise according to [14]. At this stage, for comparison purposes, a change-detection map was computed by assigning SCVs in C_n and SCVs of registration noise to the class of no-changed pixels and all the others to the class of changed pixels (see Results for the pixel-based proposed technique in Table 1). Finally, the information about adaptive cells of registration noise was used within the parcel-based decision strategy for computing the final change-detection map according to the proposed strategy. To this end, multitemporal parcels were generated as described in Sect. 5.2 and SCVs in each parcel were labeled according to (18). As one can see from Table 1, the use of the spatial-context information significantly reduces both false and missed alarms. It is worth noting that the use of spatial-context information retrieved according the parcel-based strategy allows one to obtain a regularized change-detection map without affecting the geometrical details content of the map itself. For a further assessment of the effectiveness of the proposed technique, change detection was performed according to the standard pixel-based [18] and parcel-based [20] change vector analysis ignoring the information about registration noise. In both cases (see Table 1) it is clear that standard methods are sharply affected by the presence of registration noise, which involves a high

Table 1 Change-detection results obtained both at a pixel and at a parcel level by the proposed adaptive and multiscale technique, the standard CVA technique and the manual approach

Technique	False alarms	Missed alarms	Overall error	Overall accuracy (%)
Pixel based				
Proposed	62,867	4,728	67,595	93.02
Standard CVA	173,676	1,470	175,146	81.91
Manual	55,984	5,768	61,752	93.62
Parcel based				
Proposed	29,616	3,382	32,998	96.59
Standard CVA	106,580	734	107,314	88.92
Manual	23,160	4,192	27,352	97.18

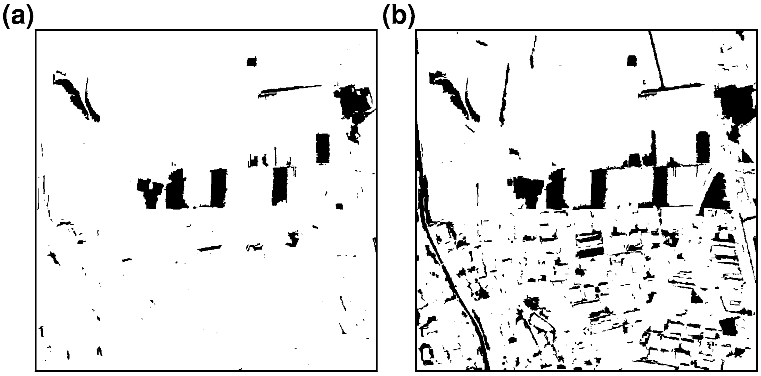


Fig. 16 Change-detection maps obtained with: **a** proposed multiscale approach with the adaptive estimation of the cell dimension at a parcel level; and **b** the standard parcel-based CVA

number of false alarms mainly located in the high frequency regions of the images. On the contrary, the proposed method significantly reduces false alarms both at pixel (from 173,676 to 62,867) and at parcel level (from 106,580 to 29,616), and generates change-detection maps characterized by high accuracy both in homogeneous and border areas. Figure 16 allows one a visual comparison between the change-detection map obtained at parcel level with the proposed technique (Fig. 16a) and the standard CVA (Fig. 16b). A final comparison is made with the results achieved according to a manual trial-and-error approach. In this case the final change-detection map is computed assigning SCVs that fall into C_n to ω_n, and applying manual thresholds for isolating within A_c SCVs associated with changed pixels from those associated with registration noise on the basis of some prior information. Two maps were generated. The first considers the spatial-context information arising from multitemporal parcels while the second does not. Results yielded with this procedure can be considered as an upper bound for the proposed technique. Observing Table 1, one can conclude that the proposed method performs effectively both at a pixel and at a parcel level, as it exhibits overall accuracies that are close to those obtained by the manual (optimal) approach (i.e., 93.02% vs. 93.62% for the pixel-based case and 96.59% vs. 97.18% for the parcel-based one).

As final remark it is important to notice that the change-detection map derived by the proposed approach presents residual false alarms mainly due to the different acquisition seasons of the considered images (i.e., summer and autumn). This characteristic resulted in significant radiometric differences related to seasonal variations in the crop rows and in the shape of shadows. The false alarms due to such acquisition conditions can be reduced only considering additional semantic information associated with changes. However, the overall accuracy achieved by the proposed context-sensitive technique robust to registration noise (i.e., 96.59%) due to sharp reduction of false alarms and the high fidelity in the reproduction of changed objects (both in uniform and contour regions) confirms its validity.

7 Discussion and Conclusion

In this chapter we have analyzed the properties of registration noise on VHR multitemporal remote sensing images. This analysis was carried out in the context of a polar framework for change vector analysis (CVA), where both the magnitude and the direction information of SCVs are represented. On the basis of the derived properties, at first a novel method for an adaptive estimation of the statistical distribution of RN in multitemporal VHR images has been proposed and then a context-sensitive multiscale technique robust to such kind of noise for change detection on VHR multispectral images has been derived.

When dealing with change detection in multitemporal VHR images one of the most significant sources of errors is registration noise. Such kind of noise is due to the impossibility to perfectly align multitemporal images even if accurate co-registration techniques are applied to the data. In order to understand how to reduce the impact of residual misregistration on the change-detection process, in this work we carried out an analysis of the behaviors of registration noise that affect multitemporal VHR data sets. Images acquired by several sensors and with different land-cover types were considered in the analysis. From them, some simulated data sets have been generated in order to study the effects of RN when: (i) the misregistration between the two considered images increases; and (ii) the resolution of the original images decreases. From this analysis four different properties of the RN in VHR images have been derived, associated with both unchanged and changed pixels. These properties point out that misregistration may significantly affect the accuracy of change detection and show some important effects due to this specific kind of noise on VHR images. In particular, it was observed that SCVs that fall into the annulus of changed pixels but are associated with registration noise (and therefore are a possible source of false alarms) exhibit significant variations of statistical properties as the scale is reduced. According to this observation, we defined a novel technique for the estimation of RN. This technique derives the conditional density of RN with respect to the direction variable in the annulus of changed pixels on the basis of a multiscale analysis of the distributions of SCVs. Even if the proposed technique exploits a multiscale decomposition for identifying RN and modeling its conditional distribution, the resulting estimate represents the behavior of the RN at full resolution. Thus the estimated distribution can be used for analyzing the images at full scale, as the low-pass component used in the proposed strategy does not affect the scale of the estimation. This estimation provides us valuable information for the design of a change-detection procedure. The proposed change-detection approach, in fact, at first performs a quantization-based multiscale analysis of SCVs in the magnitude-direction domain in order to identify SCVs associated with registration noise. The retrieved information on registration noise is then exploited in the framework of a parcel-based decision strategy that takes advantage of spatial-context information in defining the final change-detection map. This step is performed at full resolution in order to preserve all the high geometrical detail information characteristic of VHR images.

The effectiveness of both the proposed techniques has been tested on a data set made up of a pair of QuickBird images. Results obtained confirm: (i) the capabilities of the estimation technique in identifying and modeling RN also in presence of real multitemporal noisy images acquired under different conditions; and (ii) the accuracy of the proposed CD technique, which involves a low amount of false alarms in change-detection maps and a high accuracy in modeling both geometrical details and homogeneous areas. The achieved results are significantly better than the ones yielded by standard change-detection techniques. The effectiveness of the proposed techniques was also tested on other data sets acquired by different remote sensing sensors, which confirmed the conclusion drawn for the presented QuickBird data. It is worth noting that despite the proposed analysis was developed for VHR remote sensing images (as the impact of misregistration on this kind of data is more relevant), it can be suitable also for the analysis of optical data at lower resolution.

As future developments of this work we plan to fully exploit both the derived properties and the technique for the estimation of the registration noise distribution to develop adaptive co-registration strategies based on the estimated local behavior of the registration noise. With regard to the change-detection strategy we plan to extensively test the proposed method on other multitemporal images acquired by different sensors representing different change-detection problems.

References

1. Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: a systematic survey. IEEE Trans. Image Proc. **14**(3), 294–307 (2005)
2. Lu, D., Mausel, P., Brondizio, E., Moran, E.: Change detection techniques. Int. J. Remote Sens. **25**(12), 2365–2407 (2004)
3. Singh, A.: Digital change detection technique using remotely senses data. Int. J. Remote Sens. **10**(6), 989–1003 (1989)
4. Coppin, P.R., Jonckheere, I., Nachaerts, K.: Digital change detection in ecosystem monitoring: a review. Int. J. Remote Sens. **25**(9), 1565–1596 (2004)
5. Townshend, J.R.G., Justice, C.O., Gurney, C.: The impact of misregistration on change detection. IEEE Trans. Geosci. Remote Sens. **30**, 1054–1060 (1992)
6. Dai, X., Khorram, S.: The effects of image misregistration on the accuracy of remotely sensed change detection. IEEE Trans. Geosci. Remote Sens. **36**, 1566–1577 (1998)
7. Bruzzone, L., Cossu, R.: An adaptive approach for reducing registration noise effects in unsupervised change detection. IEEE Trans. Geosci. Remote Sens. **41**(11), 2455–2465 (2003)
8. Li, J., Qian, S., Chen, X.: Object-oriented method of land cover change detection approach using high spatial resolution remote sensing data. IEEE Trans. Geosci. Remote Sens. **5**, 3005–3007 (2003)
9. Bovolo, F.: A multilevel parcel-based approach to change detection in very high resolution multitemporal images. IEEE Geosci. Remote Sens. Lett. **6**(1), 33–37 (2009)
10. Niemeyer, I., Marpu, P.R., Nussbaum, S.: Change detection using the object features. In: IEEE International Geoscience & Remote Sensing Symposium, pp. 2374–2377, 2007
11. Bovolo, F., Bruzzone, L.: A theoretical framework for unsupervised change detection based on change vector analysis in polar domain. IEEE Trans. Geosci. Remote Sens. **45**(1), 218–236 (2007)

12. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Geosci. Remote Sens. **PAMI-11**(7), 674–693 (1989)
13. Bovolo, F., Bruzzone, L.: A detail preserving scale-driven approach to change detection in multitemporal SAR images. IEEE Trans. Geosci. Remote Sens. **43**(12), 2963–2972 (2005)
14. Bowman, A.W., Azzalini, A.: Applied Smoothing Techniques for Data Analysis: Kernel Approach with S-plus Illustrations. Clarendon Press, Oxford (1997)
15. Parzen, E.: On estimation of a probability density function and mode. Ann. Math. Stat. **33**, 1065–1077 (1962)
16. Patrick, E.A., Fischer, F.P.: III. A generalized k-nearest neighbor rule. Inf. Control **16**, 128–152 (1970)
17. Reilly, D.L., Cooper, L.N., Elbaun, C.: A neural model for category learning. Biol. Cybern. **45**, 35–41 (1982)
18. Bruzzone, L., Prieto, D.F.: Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Remote Sens. **38**, 1171–1182 (2000)
19. Bruzzone, L., Prieto, D.F.: An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images. IEEE Trans. Image Process. **11**(4), 452–466 (2002)
20. Bruzzone, L., Prieto, D.F.: An adaptive parcel-based technique for unsupervised change detection. Int. J. Remote Sens. **21**(4), 817–822 (2000)
21. Bruzzone, L., Carlin, L.: A multilevel context-based system for classification of very high spatial resolution images. IEEE Trans. Geosci. Remote Sens. **44**(9), 2587–2600 (2006)
22. Baatz, M., Benz, U., Dehghani, S., Heynen, M., Höltje, A., Hofmann, P., Lingenfelder, I., Mimler, M., Sohlbach, M., Weber, M., Willhauck, G.: eCognition user guide 4. Definiens Imaging (2004)
23. Bovolo, F., Bruzzone, L., Capobianco, L., Garzelli, A., Marchesi, S., Nencini, F.: Analysis of the effects of pansharpening in change detection on VHR images. IEEE Geosci. Remote Sens. Lett. **7**(1), 53-57 (2010)
24. ENVI User Manual. RSI, Boulder, CO (2003) http://www.RSInc.com/envi
25. Marchesi, S., Bovolo, F., Bruzzone, L.: A context-sensitive technique robust to registration noise for change detection in VHR images. IEEE Trans. Image Process. **19**(7), 1877-1889 (2010)

Effects of the Spatial Enhancement of Hyperspectral Images on the Distribution of Spectral Classes

Andrea Garzelli and Luca Capobianco

Abstract In this chapter, we present a study on the effects of the spatial enhancement of hyperspectral (HS) images on the distribution of spectral classes. The analysis is based on the concept of dimensionality reduction, the transformation of high-dimensional data into a meaningful representation of reduced dimensionality which may favor visualization and understanding of high-dimensional data. Nonlinear techniques of dimensionality reduction are applied to original Hyperion HS data (30 m) and to fusion products with the panchromatic channel of ALI (10 m) obtained from different sharpening methods, in order to evaluate possible advantages or critical situations deriving from multi-sensor, multi-resolution data fusion.

Keywords Image fusion · Hyperspectral · Spectral signatures

1 Introduction

Hyperspectral (HS) spatial enhancement refers to the joint processing of HS imagery along with panchromatic (Pan) or multispectral (MS) imagery of higher spatial resolution in order to obtain a hyperspectral image product that exhibits, ideally, the spectral characteristics of the observed hyperspectral image at the spatial resolution and sampling of the higher spatial resolution image [1]. A sensing platform that has the capability to concurrently capture HS and MS+Pan

A. Garzelli (✉) and L. Capobianco
Department of Information Engineering, University of Siena, Via Roma, 56, 53100
Siena, Italy
e-mail: garzelli@dii.unisi.it

L. Capobianco
e-mail: capobianco@dii.unisi.it

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_14,

data is the NASA Earth Observing 1 (EO-1) satellite through its optical sensors Hyperion (220 bands at 30 m resolution) and Advanced Land Imager (ALI) (6 bands at 30 m and a 10 m Pan band). The difference in spatial resolution is generally a result of the fundamental tradeoff between spatial resolution, spectral resolution, and radiometric sensitivity in the design of electro-optical sensor systems.

Most of the approaches that can be used for hyperspectral resolution enhancement have heritage in the sharpening of multispectral imagery based on higher resolution panchromatic imagery (pan-sharpening). Spatial enhancement of hyperspectral (HS) imagery, however, is more complex than pan-sharpening of multispectral (MS) data, for three main reasons:

1. the huge number of HS bands normally does not allow to apply fusion methods based on local context, either in the original spatial domain or in a transformed (multiresolution) domain;
2. the spectral coverage of the panchromatic (Pan) image does not match the wavelength acquisition range of the HS bands;
3. the spatial scale ratio (SR) between HS and Pan may not be a power of two, e.g., SR $= 3$ in the case of Hyperion data (30 m) and the panchromatic channel of ALI (10 m).

A specific spectral fusion model is required to preserve the spectral information of the data with lower spatial resolution, or even to enhance it through the unmixing of the coarse-resolution HS pixels, based on information extracted from the high-resolution Pan data. Spatial details that are not available for HS bands have to be inferred through the model, starting from the high spatial frequency components of Pan. The fusion model should be as simple as possible, in order to limit the computational complexity, and the model parameters should be spatially invariant, band dependent, and should be easily, yet accurately, estimated from the available dataset.

The Chapter presents a study on the effects of the spatial enhancement of HS images on the distribution of spectral classes. The analysis is based on the concept of dimensionality reduction, the transformation of high-dimensional data into a meaningful representation of reduced dimensionality which may favor visualization and understanding of high-dimensional data. Non-linear techniques of dimensionality reduction applied to original Hyperion HS data (30 m) and to fusion products with the panchromatic channel of ALI (10 m) obtained from different sharpening methods are investigated. The goal is to evaluate possible advantages (unmixing capabilities) or critical situations (reduced class separability) deriving from multi-sensor, multi-resolution data fusion.

Different spatial-resolution enhancement algorithms are tested on Hyperion HS and ALI Pan data in order to compare their performances and investigate on potential drawbacks of pan-sharpening of HS images. To this aim, the paper focuses on the objective analysis of the intrinsic spectral information of the fusion products.

Section 2 introduces the problem of spatial enhancement of HS images and indicates different approaches and practical solutions. Section 3 describes the dimensionality reduction methods applied to assess the performances of different pan-sharpening algorithms. The experimental results obtained on true HS/Pan data are reported in Sect. 4 and conclusions are drawn in Sect. 5

2 Spatial Enhancement of Hyperspectral Images

Few algorithms, mainly derived from methods for pan-sharpening of MS images, may be successfully applied to HS spatial enhancement [1, 2].

An extensive number of pan-sharpening methods for MS data have been proposed in the literature, starting from the second half of the 1980s. Most of them are based on a general protocol in which high-frequency spatial information is extracted from the Pan image and injected into the resampled MS bands by exploiting different models. In general, the image fusion methods described by this protocol can be divided into two main families: the techniques based on a linear spectral transformation followed by the substitution of a component in the transformed domain (component substitution, CS, methods), and the algorithms that perform spatial injection after applying a spatial frequency decomposition usually performed by means of multiresolution analysis (MRA).

2.1 Component Substitution Methods

Basically, CS techniques linearly transform the MS data set into a more uncorrelated vector space. Then, one of the transformed bands, usually the low-resolution intensity I, is replaced by the sharp panchromatic image P, histogram-matched to the I component itself, before the inverse transformation (Intensity–Hue–Saturation, IHS) is applied. This procedure is equivalent to inject, i.e., add, the difference between P and I into the resampled MS data set [3, 4]. The intensity image I can be obtained by weighting the MS bands with a set of coefficients whose choice is related to the spectral responses of Pan and MS bands [5, 6].

Principal component analysis (PCA) is an alternative to the IHS techniques. It is analogous to the IHS scheme since the Pan image is substituted by the first principal component (PC1). Histogram matching of Pan to PC1 is mandatory before substitution because the mean and variance of PC1 are generally far greater than those of Pan. It is well established that PCA performances are better than those of IHS [7] and, in particular, that the spectral distortion in the fused bands is usually less noticeable, even if it cannot completely be avoided. Generally speaking, if the spectral responses of the MS bands are not perfectly overlapped with the bandwidth of Pan, as it happens with the most advanced very high resolution imaging sensors, IHS- and PCA-based methods may yield poor results in terms of spectral fidelity [8]. Another CS technique reported in the literature is Gram-Schmidt (GS) spectral sharpening, which was invented by Laben and

Brower in 1998 and patented by Eastman Kodak [9]. The GS method is widely used since it has been implemented in the Environment for Visualizing Images (ENVI) software. The GS method is efficient since it benefits from a detail injection rule by which, for each MS band, the injection gain is proportional to the covariance value between the synthesized intensity and the expanded MS band as reported in [10]. As a matter of fact, since the sharp P and the smooth I have generally a different local radiometry, spectral distortions can arise in the fusion results. A mitigation of the consequent color changes can be obtained if I matches as much as possible the spectral response of Pan. This result is achieved by designing I as a linear combination of the MS bands which is based on the spectral responses of the MS and Pan image sensors [5]. Such coefficients can be further optimized by minimizing the distance between P and I, for example in the minimum mean square error sense [10], applying a genetic algorithm [11] that optimizes the Q4 score parameter defined in [12] or imposing MMSE constraints on the multispectral images [13]. The last method may be considered as a hybrid CS-MRA pan-sharpening technique.

2.2 Multiresolution Methods

The spectral quality of CS fusion results may be sufficient for most applications and users. Generally, lower spectral distortion may be obtained by injecting zero-mean high-pass spatial details, taken from the Pan image without resorting to any transformation. In fact, since the pioneering high-pass filtering (HPF) technique [7], fusion methods based on injecting high-frequency components into resampled versions of the MS data have demonstrated a superior spectral fidelity [14–16]. HPF basically consists of an addition of spatial details, taken from a high-resolution Pan observation, into a bicubically resampled version of the low resolution MS image. Such details are obtained by taking the difference between the Pan image and its low-pass version achieved through a simple local pixel averaging, i.e., a box filtering. Later improvements have been obtained with the introduction of multiresolution analysis (MRA), by employing several decomposition schemes, specially based on the discrete wavelet transform (DWT) [17, 18], uniform rational filter banks (borrowed from audio coding) [19], and Laplacian pyramids (LP) [20, 21]. The DWT has been extensively employed for remote sensing data fusion [22–24]. According to the basic DWT fusion scheme, couples of subbands of corresponding frequency content are merged together. Afterwards the fused image is synthesized by taking the inverse transform. Fusion schemes based on the “à trous” wavelet algorithm and Laplacian pyramids were successively proposed [25, 26]. Actually, unlike the DWT which is critically subsampled, the “à trous” wavelet and the LP are overcomplete representations. The missing of the decimation step allows an image to be decomposed into nearly disjointed bandpass channels in the spatial frequency domain, without losing the spatial connectivity (translation invariance property) of its high-pass details, e.g., edges and textures. This property is fundamental because, for critically sub-sampled schemes, spatial distortions, typically ringing or aliasing effects may be present in the fused products and originate shifts or blur of contours and textures.

Data-fusion methods require the definition of a model that establishes how the missing high-pass information is injected into the resampled MS bands [23]. In other words, the model, referred to as interband structure model (IBSM), deals with the radiometric transformation (gain and offset) of spatial structures (edges and textures) when passing from the Pan to MS images. The model is generally inferred at the coarser resolution and extrapolated to the finest resolution. This condition has been proven to be satisfactory for the MS and Pan data whose scale ratio is equal to four by investigating a Kalman-based fusion method which performs a prediction of fusion parameters across scales [27]. It should also be advisable to compute a high-resolution IBSM (HRIBSM) by considering additional information on the MS-imaging system. Notable examples of injection models are additive combination of 'à-trous' wavelet frames, as in the additive wavelet to the luminance component (AWL) technique [25], the injection of wavelet details after applying intensity–hue–saturation (IHS) transformation or principal component analysis [28], the spectral distortion minimization (SDM) with respect to the resampled MS data [15], or the spatially adaptive injection, as in the context-based-decision (CBD) algorithm [29] and in the RWM method [30]. More efficient schemes can be obtained by incorporating the Modulation Transfer Functions (MTFs) of the MS scanner and of the Pan sensor in order to design the MRA reduction filters or the decimation filters generating the MS and Pan data at degraded scales. In this way, it is possible to avoid a poor enhancement that sometimes occurs when MTFs are assumed to be ideal filters [31]. Theoretical considerations on injection models and experimental comparisons among MRA-based pan-sharpening methods can be found in [32].

A further issue concerns the adoption of global or local injection models. Computational cost is lower for global models but results are superior in general for local ones even if some caution should be adopted for local models since due their nature they are responsible for local improvements but also for possible local distortions or impairments.

2.3 Selected Methods for Testing on HS+Pan Images

Among the methods introduced in Sect. 2 , we have focused our attention on few sharpening algorithms that can be easily applied to HS images. In order to give evidence to different effects of spatial enhancement on the distribution of spectral classes, we are interested on

- efficient fixed-scheme injection methods, since the computational complexity of the injection strategies based on local statistics is unacceptable for HS image enhancement;
- classical, well-established methods, even if their performances are not excellent.

The following fusion methods are selected and tested on the considered data set:

1. the Global MMSE fusion method with band-dependent generalized intensity (GMMSE) [13];
2. the Generalized Intensity–Hue–Saturation method (GIHS) [3];
3. the High-pass filtering (HPF) method [7];
4. a modified version of the HPF method, referred as HPF-P with a filter characterized by 1/3 cutoff frequency and not introducing any spectral distortion with respect to the original HS image data.

The GMMSE pan-sharpening method [13] is optimal in the minimum mean squared error sense and it is characterized by low computational complexity. This solution adopts a linear injection model in which an optimal detail image extracted from the panchromatic band is calculated for each MS/HS band $(B_k, k = 1, \ldots, N)$ by evaluating a band-dependent generalized intensity from the N original bands. The fusion equations are

$$B_k^F = B_k^{\uparrow} + g_k \left(P - \sum_{j=1}^{N} w_{k,j} B_j^{\uparrow} \right), \quad k = 1, \ldots, N, \tag{1}$$

where B_k^F is the kth pan-sharpened band and $B_k^{\uparrow}$ indicates the kth original band upsampled to the Pan resolution. The N^2 weights $w_{k,j}$, of the linear combination equations which provide the generalized intensity images (one for each band to be fused), and the N gains g_k that regulate the spatial detail injection are *jointly* calculated at degraded resolution according to a minimum mean squared error criterion [13]. This procedure has been demonstrated to be fast and reliable thanks to the computation of N different intensity images.

The Generalized Intensity Hue Saturation (GIHS) fusion method [3] computes a generalized intensity image I by a weighted linear combination of the original MS/HS bands (the generalized Intensity–Hue–Saturation transform), and subtracts it from the Pan image histogram-matched to I, denoted as P_{hm}. Such difference image is added to each MS/HS band:

$$B_k^F = B_k^{\uparrow} + (P_{hm} - I), \quad k = 1, \ldots, N, \tag{2}$$

Its main critical point, due to the generalized intensity generation, is that the fusion products may exhibit important spectral distortions.

The HPF method [7] consists of an addition of spatial details, taken by local pixel averaging (box filtering) from a high-resolution Pan observation, into a bicubically resampled version of the low-resolution MS image. Its modified version, referred as HPF-P, substitutes the box filter with a quasi-ideal zero-phase lowpass filter with 1/3 cutoff frequency: it is therefore more suitable for merging 30 m Hyperion with 10 m ALI image data. In addition, it injects a spatial detail image which is pixel-wise projected along the direction of the original HS pixel in the N-dimensional hyperspectral space.

2.4 Evaluation of Spatial Enhancement Methods

There are several ways for evaluating an algorithm for spatial enhancement of HS images, which may also depend on the particular use of the enhanced data products.

1. Objective quality assessment by score indexes.
   a. The index can be applied to original HS images as reference data for pan-sharpened images obtained from spatially degraded HS and Pan images. The degradation factor equals the spatial resolution ratio between original HS and Pan data. This evaluation protocol is used, among others, by ERGAS [23] and the recent $Q2^n$ index [33].
   b. Quality may be assessed without a reference image, i.e., directly at the spatial resolution of Pan, by evaluating the QNR index [34].
2. Effects on classification. The objective evaluation of a classification phase applied to both original (HS+Pan), and pan-sharpened HS data may provide useful indications on which spatial enhanced method is more suitable for classification. Interesting results have been presented in [35] for pan-sharpened very high resolution (VHR) MS images.
3. Effects on change detection. The impact of pan-sharpening on the accuracy of change detection can be evaluated to investigate whether the improvement in geometrical resolution of change detection maps given by pan-sharpening is affected or not by possible artifacts introduced by the pan-sharpening process. A theoretical and experimental study on VHR MS data has been presented in [36].
4. Effects on target detection. The effects of spatial enhancement through a panchromatic band on target detection applications has been investigated in [37].
5. Effects on the distribution of spectral classes. This approach is adopted in the present Chapter: it allows to verify to which extent the process of spatial enhancement affects the spectral information of the original HS image.

3 Dimensionality Reduction for the Assessment of Pan-Sharpening Algorithms

This section introduces a data fusion case study, where a real hyperspectral image is employed to assess the performance of the pan-sharpening methods indicated in Sect. 2.3.

The analysis is concerned with the effects that the fusion algorithms have in the spectral subspace containing the data when spatial information is injected: in other words, we are interested in the description of the pixel (or) *sample* distribution of the original data and the re-distribution after the fusion process. A complete description of the subspace spanned by data samples or visualization of full data would be prohibitive for two main reasons: on one hand, the processing of the whole hyperspectral data would be computationally very demanding and, on the other hand, the information mining of huge data with no prior information is not feasible. Hence, the reasonable approach used in the chapter involves a 'subspace spectral sampling' by means of the selection of many pixel-based *regions of interest* (ROI's), shown in Fig. 1. The ground truth composed by the ROI's is used to build different datasets to carry out different analysis. In this way, it is possible to understand if and how similar pixels (such as those belonging to the same class) are displaced in the feature space (assessment of local effects) and how spectral information and related properties are modified (assessment of global effects).

The used data set has been acquired over the region of Palo Alto (USA) on June 23, 2002 by the HYPERION and ALI sensors mounted on the EO-1 platform. The 220 HS bands of Hyperspectral sensor span from 0.4 to 2.5 μm with a spatial resolution of 30 m. The PAN band acquired by the ALI scanner approximately covers a short interval of HS (from 0.45 to 0.69 μm) with a spatial resolution of 10 m. After removing low SNR and water-absorption bands, we considered a total of 85 bands for data analysis. The data set has been radiometrically calibrated from digital counts, geocoded, i.e., resampled to uniform ground resolutions of 30 m (HS) and 10 m (PAN) ground spatial distance, and packed in 16-bit words.

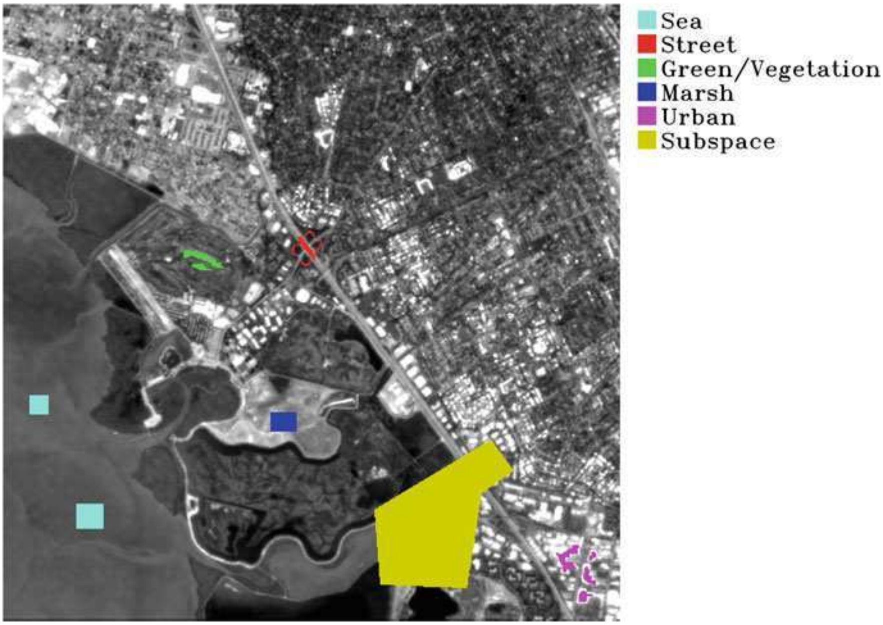


Fig. 1 ALI Pan image. Five near-homogeneous classes are labeled and highlighted in the image. The region "Subspace" represents a near-linear mixture of the five selected classes

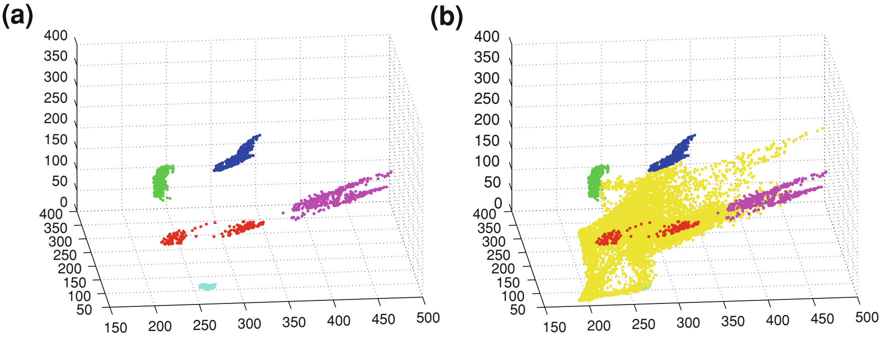


Fig. 2 3D ScatterPlot of radiance values in bands 10,30: **a** five classes; **b** five classes and image subspace

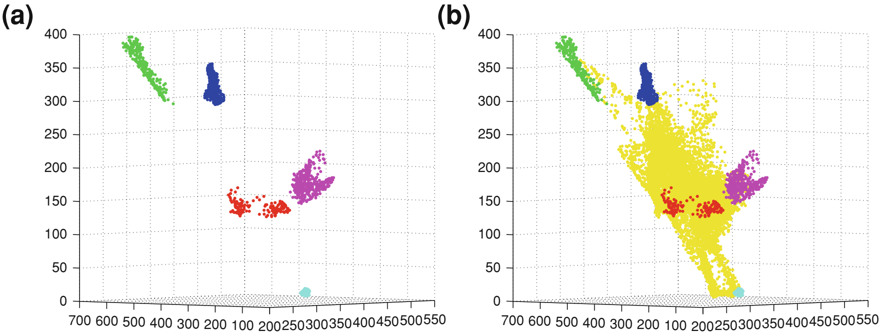


Fig. 3 3D ScatterPlot of radiance values in bands 5, 50, 70: **a** five classes; **b** five classes and image subspace

Hence, in the following, band values are to be intended as being the radiance values, with SI Unit (Watt sr^{-1} m^{-2} nm^{-1}).

Visual analysis of these data reveals the presence of sea, marshes, urban areas, vegetated areas, roofs, streets, and shadows. (more information and images are available at http://eo1.usgs.gov/).

The highlighted classes shown in Fig. 1 were selected for their near-homogeneity properties, through spectral analysis and color photointerpretation of many false RGB compositions, with the help of 3D spectral projections as shown in Figs. 2 and 3.

The class marked in yellow in Fig. 1 and named ‘Subspace’ describes in many three-dimensional projections a geometric shape whose vertices are covered by the other classes: the ‘Subspace’ class plays an important role in the subsequent analysis, since it appears to be a quasi-linear mixture of the five homogeneous classes.

In the following, N_s stands for the number of spectral samples $\mathbf{x}_i \in \mathbf{X} \in \mathbf{R}^N$, $i = 1, \ldots, N_s$, each one with dimension $N = 85$.

By using the above mentioned selection of classes, we are confident that the spectral samples contain different ‘degrees of homogeneity’, hence we will be able to assess the effects of pan-sharpening on ‘pure’ pixels and ‘mixed’ pixels. For example, the 3D scatter plots in Figs. 2 and 3 show that the class ‘Sea/Water’ (Cyan) is very clustered, even if it is composed by disconnected areas of the image, while the class ‘Street’ (Red) is less clustered and contains more spectral variability or, roughly speaking, each class pixel contains many materials at the given sensor resolution.

The results reported in Sect. 4 have general validity: the objective of the analysis is to provide results not concerned with a particular remote sensing application, such as classification or target, anomaly and change detection. Analyzing data with algorithms using a large number of parameters is avoided for the same reason, and also for the complexity in managing high dimensional data.

4 Experimental Results

This section is concerned with the analysis and comparison results of the fusion methods described in Sect. 2.3, obtained by exploiting different methodologies, from the simplest ‘Visual Approach’ to the more sophisticated algorithms, such Kernel PCA, up to the Linear Preserving Projection algorithm, developed for maintaining local and global structure in dimensionality reduction. All the experiments are carried out at the same time on two different kinds of datasets: the former is built by considering only five classes of the user’s ground truth (Dataset #1), while the latter includes also the ‘Subspace’ class, i.e. the whole ground truth, (Dataset #2).

4.1 ‘Visual’ Approach

The analysis begins with the simplest possible approach: original and pan-sharpened data are visualized in three-dimensional scatterplots, as already explained. The high correlation between spectral bands in hyperspectral data allows the usage of this methodology since, even if different triplets of band are chosen for visualization (see as instance Figs. 2 and 3), the results are not surprisingly different. Figure 4 shows the same projection of Fig. 2, but rescaled so that all the results are scattered in a subspace at the same scale.

4.1.1 GIHS

The effects of the GIHS fusion algorithm may be assessed by looking at the shape of the pixel distribution in Fig. 5a. Spectral bands appear to be more correlated due

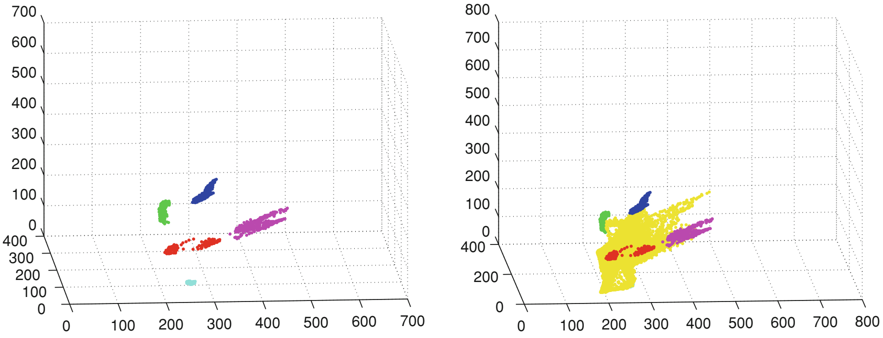


Fig. 4 Data visualization with 3D ScatterPlot radiance values in bands 10, 30. Scattering of original values as contained in the data, Dataset #1 (*left*) and Dataset #2 (*right*)

to the spectral distortion caused by the spatial injection from Pan: note that the pixels belonging to the selected classes appear more linearly scattered, which means that local spatial information prevails upon spectral information or, in other words, that GIHS privileges local spatial information with respect to the spectral one: this appears clear in the cluster formed by pixels belonging to Sea/Water (cyan) since, after the fusion, it splits into two smaller classes that, actually, are the two disconnected regions marked in Fig. 1. However, it is worth noting that the pixels in the classes are still on the edges of the subspace spanned by the pixels in class ‘Subspace’, even if the linearity hypothesis in this case is weaker than in the original case.

4.1.2 HPF

Same considerations can be done for the HPF fusion algorithm, whose results are depicted in Fig. 5b; in this case the unbalanced usage of spatial and spectral information is greater than for the GIHS case, since the phenomena of pixel diffusion in the direction of linear correlation is definitely more relevant. Moreover, the radiance band values of many pixels are set to zero and a relevant amount of information is lost during the fusion process. Besides, the distribution of the classes is completely changed and the linearity hypothesis in this case does not hold at all.

4.1.3 HPF-P

In this case, the pixel distribution follows a ‘cone’ behavior, for the definition itself of the algorithm. In fact, since the fusion is based on projection measures, the evaluation can be considered in some sense ‘angle-based’: pixels with high spectral angles are projected in the new distribution close to zero. The results are

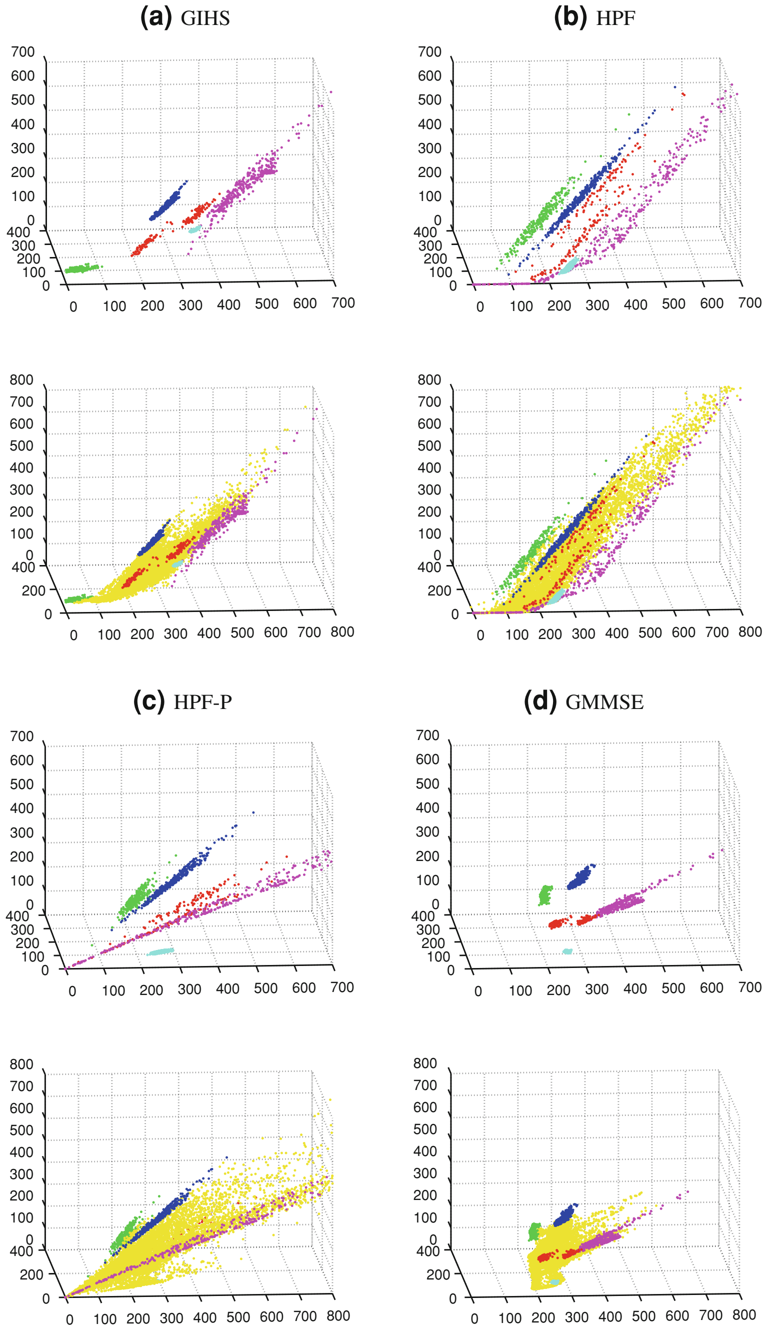


Fig. 5 Data visualization with 3D ScatterPlot radiance values in bands 10,30. Scattering of data values after fusion, Dataset #1 (*upper part*) and Dataset #2 (*lower part*)

shown in Fig. 5c, where each class appears scattered along a cone whose width is given by the projection of the pixel with the maximum angle.

4.1.4 GMMSE

Finally, Fig. 5d shows results for the MMSE fusion process: distribution of classes are almost unchanged and the mixture model is unchanged. Even with a low spectral distortion, the method can be considered to have a good balance between spectral preservation and spatial injection.

It is worth noting that the partial analysis employed in this section is very useful to understand how the algorithms works and to assess how spatial and spectral information are balanced by each algorithm: however, it is useful to remark that the analysis is concerned on the global variations of the spectral distribution: the 'visual' approach used in this section might give an idea about the 'local' effects of pan-sharpening on the hyperspectral data, but it does not allow for a global comprehension of the spectral re-distribution. In the following, different techniques are used to globally capture the subspace transformations.

4.2 Linear and Non-Linear Sample Similarity Measures

Given N_s spectral samples, each one with dimension N, some measures of spectral *'similarity'* between samples are given by kernel matrices K of dimension $N_s \times N_s$, where a value in (i, j), k_{ij} stands for a similarity measure between sample x_i and sample x_j. In the following, the basic principles of kernel theory are recalled to introduce the experimental analysis. For a more complete review of kernel methods for remote sensing, the reader may refer to Chap. 10.

The entries in the kernel matrix are defined by

$$k_{ij} = \kappa(x_i, x_j) \tag{3}$$

where κ is a kernel function. Subsequently, the kernel matrix K is centered using the following modification of the entries

$$k_{ij} = k_{ij} - \frac{1}{N_s}\sum_{l} k_{il} - \frac{1}{N_s}\sum_{l} k_{jl} + \frac{1}{{N_s}^2}\sum_{lm} k_{lm}. \tag{4}$$

The centering operation corresponds—in the original domain—to subtracting the mean of the features. It makes sure that the features in the high-dimensional space defined by the kernel function are zero-mean.

It is possible to define the measure of similarity k_{ij} in Eq. 3 in different ways: if a linear function (or 'linear kernel') is used, i.e.

$$k_{ij} = x_i^T x_j \tag{5}$$

the kernel function is actually the correlation between the two samples. However, many other function can be chosen to evaluate the similarity. Valid kernel function largely used are

$$k_{ij} = e^{(-\|x_i - x_j\|^2/(2\sigma^2))} \text{ Gaussian Kernel} \tag{6}$$

$$k_{ij} = ((x_i^T x_j) + 1)^d \text{ Polynomial Kernel.} \tag{7}$$

Figure 6 reports some measures of correlation between the pixels belonging to the selected classes, except those for the ‘Subspace’, and the linear and Gaussian kernel have been used. The matrix containing the samples is organized with contiguous pixels of each class, so that the ideal matrix K would be composed by high correlation squares on the diagonal and zero outside, and the size of each

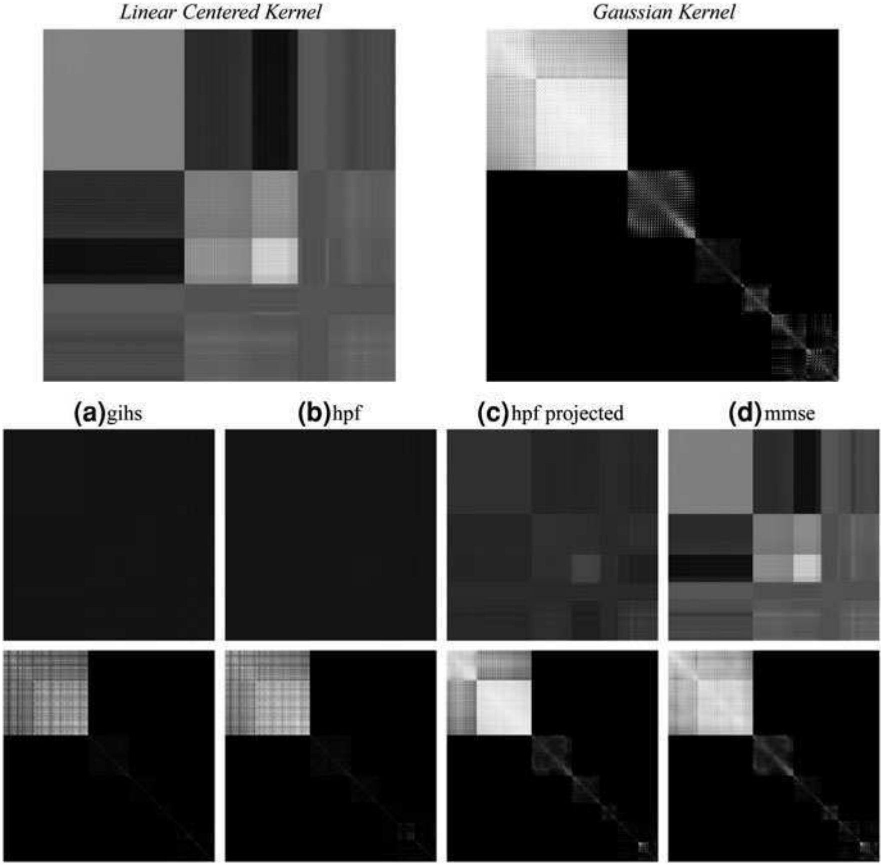


Fig. 6 Kernel matrices on the original data (*first line*) and kernel matrices on the fused data (linear and Gaussian, *second* and *third row*, respectively): samples are sorted so that the ideal matrix K would be composed by high correlation squares on the diagonal and zero outside

square equals the number of the pixels in the corresponding class. In the evaluation of the Gaussian kernel matrices, the same value for the $\sigma = 10^2$ has been used and no centering operation has been performed: this value is justified by the fact that in the experiments only the order of magnitude is set, and with $\sigma = 10$ or $\sigma = 10^3$ results show no variability. The linear kernel matrices instead have been centered with the formula in Eq. 4 to balance the output of the linear function: for this reason, the third class, marked by the third square on the diagonal, appears more clustered (or correlated) than the others. The results related to linear correlation and reported in Fig. 6 confirm and extend the properties of the data described in the previous section: pixels in GIHS, HPF and HPF-P have a linear correlation much lower than those in the original data and in the GMMSE fused dataset, even if results for HPF-P are better than those for GIHS and HPF.

However, if the similarity is measured by means of a Gaussian kernel, pixels still appear to be correlated (even if in a non linear fashion): results show that in the subspace of the fused images the pixels can still be considered as clustered.

To validate this hypothesis, data correlation is analyzed in the following by means of linear and non linear methods.

4.3 PCA

Principal Components Analysis (PCA) [38] is a linear technique for data analysis, and it can also be used for dimensionality reduction: it embeds a linear transformation from the subspace containing the original data into a linear subspace of lower dimensionality. Although there exist various linear techniques, PCA, known also as Karhunen–Loève transform, is by far the most popular (unsupervised) linear technique.

PCA constructs a low-dimensional representation of the data that describes as much of the variance in the data as possible. The new representation is achieved by means of a linear transformation from the original space to a transformed space, having a sorted vector basis of reduced dimensionality so that the amount of variance in the data is shifted and misplaced in descending order.

In mathematical terms, PCA attempts to find a linear mapping $\mathbf{M}$ that maximizes the cost function trace $\mathbf{M}^T \operatorname{cov}(\mathbf{X})\mathbf{M}$ subject to $M^T M = I$, where $\operatorname{cov}(\mathbf{X}) = 1/N_s \sum_i^{N_s} x_i x_i^T$ is the estimate of the covariance matrix of the data $\mathbf{X}$ and I is the identity matrix of size $N_s \times N_s$. It can be shown that this linear mapping is formed by the n principal eigenvectors (i.e., principal components) of the sample covariance matrix of the zero-mean data. Hence, PCA solves the eigenproblem

$$\operatorname{cov}(\mathbf{X})\mathbf{M} = \lambda \mathbf{M}. \tag{8}$$

The eigenproblem is solved for the n principal eigenvalues λ. The low-dimensional data representations $\mathbf{y}_i$ of the datapoints $\mathbf{x}_i$ are computed by mapping them onto the

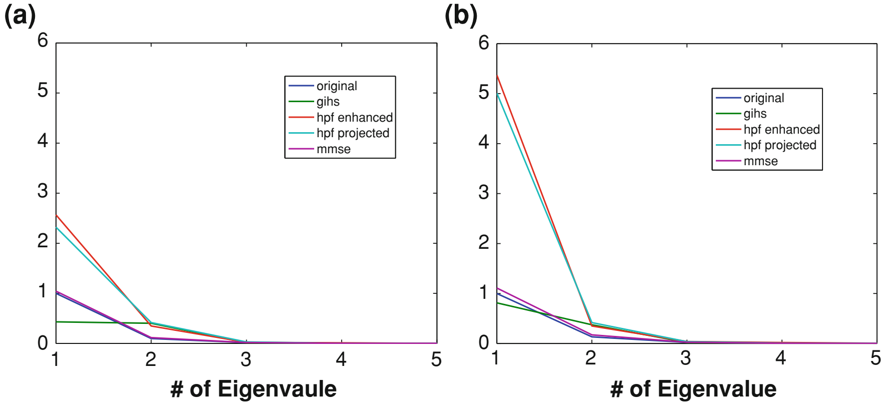


Fig. 7 PCA Eigenvalues, evaluated in the case of only five classes (**a**), and the complete data set (**b**)

linear basis **M**, i.e., **Y** = **XM**. For a deeper and complete insight to PCA transformation refer also to Chap. 10.

Figure 7 shows the scores of the first 5 eigenvalues of the PCA decomposition, related to the principal components: values have been normalized with respect to the first eigenvalue of the decomposition of the original data. In Sect. 4.1 the high correlation due to a strong injection of spatial information in the HPF and HPF-P methods has already been outlined: moreover, the PCA quantitative analysis confirm that in the case of the data sharpened by means of these algorithms, there is a direction (i.e. the first principal component) with a strong correlation (Fig. 8). In fact, the scatter of data reported in Fig. 9b, c reveals by far the strong correlation among the pixels that can not be described by a single direction. This explains why also the second eigenvalue is definitely stronger for these datasets than that for the case of original data and the case of GMMSE, that are mutually very similar (see also Fig. 9d). Quite different is the case of data generated by the

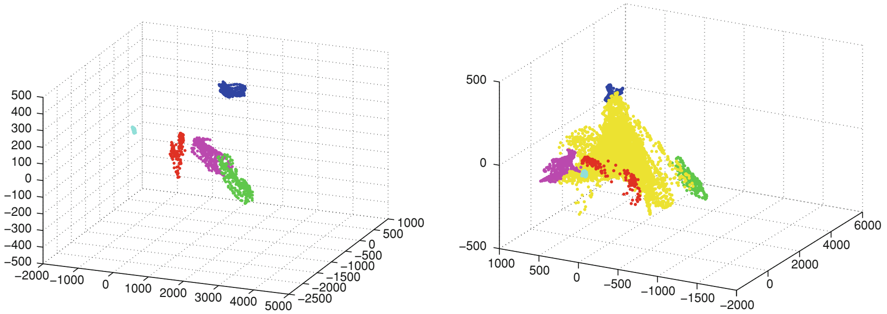
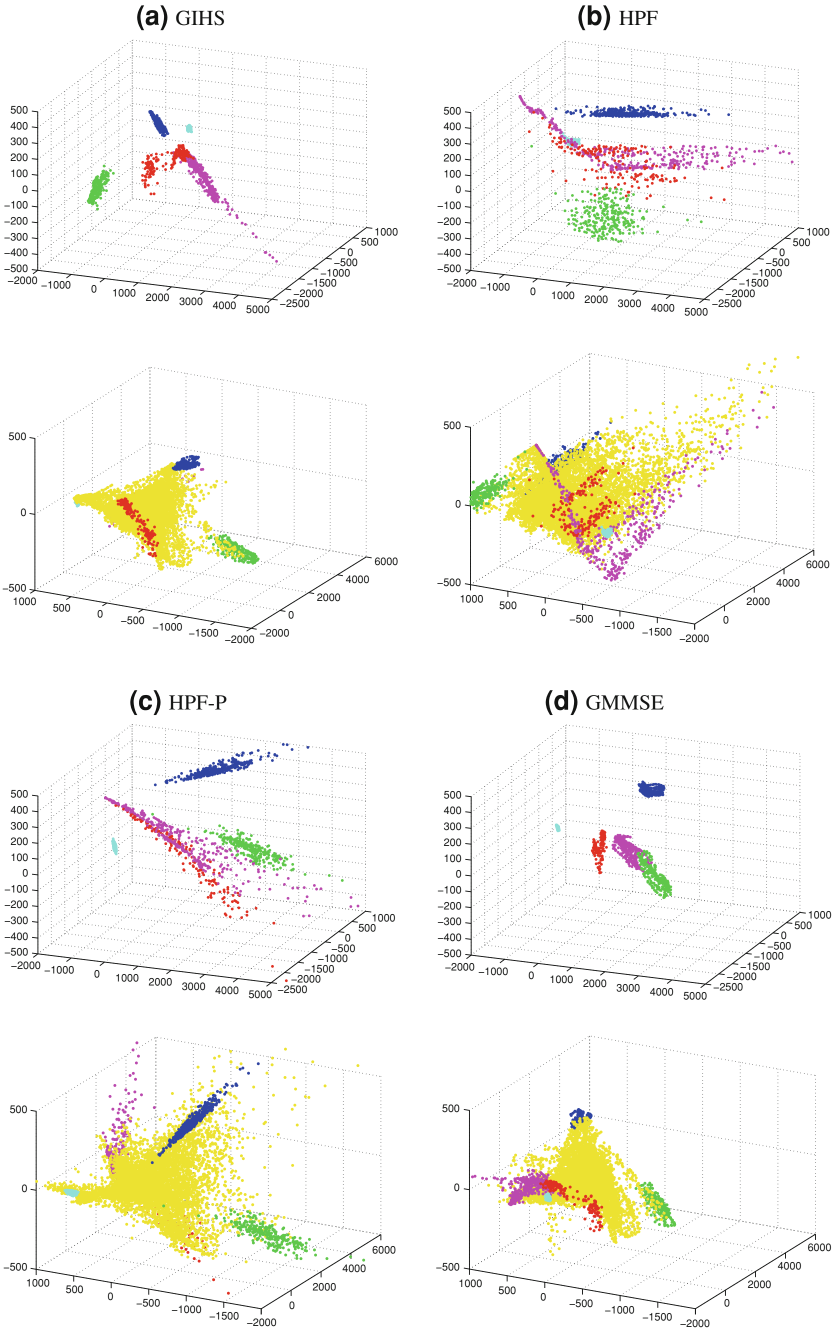


Fig. 8 PCA transformation: 3D ScatterPlot of the first principal components. Scattering of principal components output, Dataset #1 (*left*) and Dataset #2 (*right*)


(a) GIHS

(b) HPF

(c) HPF-P

(d) GMMSE

Fig. 9 PCA transformation: 3D ScatterPlot of the first principal components. Scattering of principal components output, Dataset #1 (*upper part*) and Dataset #2 (*lower part*)

GIHS algorithm: the lower value of the first eigenvalue in the PCA decomposition gives the idea that the data contains less correlation and, hence, more information. Nevertheless the high score of the second eigenvalue almost equals the first one and those of the HPF and HPF-P, while the values of a generic PCA analysis are expected to decrease. However, the analysis in Sect. 4.1 and the matrices in Fig. 6 reveal that the correlation introduced may be supposed to be not linear, thus it can not be caught by a linear method such as PCA. Hence in the following, the effects of the pan-sharpening algorithms are assessed with the help of non-linear techniques.

In recent years in fact, in contrast to traditional linear techniques such as PCA described and applied in this section, nonlinear techniques for dimensionality reduction are often used for the identification of correlation in marginal distribution of high dimensional and complex data.

4.4 Kernel PCA

Kernel PCA (KPCA, see also Chap. 10) is the reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function [39]. Lately, the reformulation of linear techniques based on the dot product using the 'kernel trick', has led to the proposal of many techniques such as kernel ridge regression and Support Vector Machines (SVM) in many research fields and in hyperspectral data analysis as well [40]. Similarly to PCA, Kernel PCA finds the principal components of the distribution, by computing the eigenvector decomposition of the kernel matrix, rather than that of the covariance matrix. In other words, the reformulation of traditional PCA in kernel space is obtained by means of the kernel estimates, that are similar to the inner product of the datapoints in the high-dimensional space that is constructed using the kernel function. The application of PCA in kernel space provides Kernel PCA the property of constructing nonlinear mappings.

The algorithm can be summarized in the following steps [38]: first, kernel PCA computes the kernel matrix K of the datapoints x_i. Subsequently, the principal d eigenvectors v_i of the centered kernel matrix are computed. It can be shown that the eigenvectors of the covariance matrix α_i (in the high-dimensional space constructed by κ) are scaled versions of the eigenvectors of the kernel matrix v_i

$$\alpha_i = \frac{1}{\sqrt{\lambda_i}} v_i. \tag{9}$$

In order to obtain the low-dimensional data representation, the data is projected onto the eigenvectors of the covariance matrix. The result of the projection (i.e., the low-dimensional data representation Y) is given by

$$Y = \left\{ \sum_j \alpha_1 \kappa(x_j, x), \sum_j \alpha_2 \kappa(x_j, x), \ldots, \sum_j \alpha_1 \kappa(x_j, x) \right\} \tag{10}$$

where κ is the kernel function that was also used in the computation of the kernel matrix. Since Kernel PCA is a kernel-based method, the mapping performed by Kernel PCA highly relies on the choice of the kernel function κ.

Figures 10 and 11 reports the results for Kernel PCA: Gaussian kernel has been used, with $\sigma = 10^4$, and transformation has been performed with and without the samples in class ‘Subspace’, which have been subsampled for computational issues. The value selected for σ is significant only for its order of magnitude, similarly to that used in Sect. 4.2. We avoided to apply a tuning operation on each dataset to make possible a straightforward comparison of different data (before and after fusion, for datasets #1 and #2). The three-dimensional projection of samples before pan-sharpening show that the non-linear transformation describes efficiently the information contained in the data and, at the same time, the mixture model effectively holds and class samples are very clustered. Furthermore, column (a) relative to GIHS method reveals that by using the kernel version of PCA, the principal components are much more uncorrelated than the linear case and, in addition, the classes are much more clustered and separated, as in Fig. 11d that shows the results for GMMSE. In both cases, it lightly holds the hypothesis for the class in yellow to be a mixture of the remaining classes, since pixels belonging to the ground truth are scattered on the vertices of the subspace.

Quite different results are provided by the HPF and HPF-P methods: if the KPCA is fed without the samples belonging to the subspace data, the high correlation is poorly exploited, while results obtained in the case of the complete data set appear more consistent. The lines of sight used for visualization of Fig. 11c, d—to directly compare results with other methods—are definitely not clear: same images, but with different angles of sight are reported in Fig. 12. These images, compared to those in Fig. 9 show that the two HPF algorithms introduce a severe spectral distortion. In fact, although KPCA does summarize the non linear-correlation in the HPF and HPF-P data, similarly to PCA, it is not sufficient to describe efficiently the subspace, since the classes do not lie on the vertices but on the edges of the distribution. Eventually, the analysis reveals that in this case, what really matters is the distortion of spectral information.

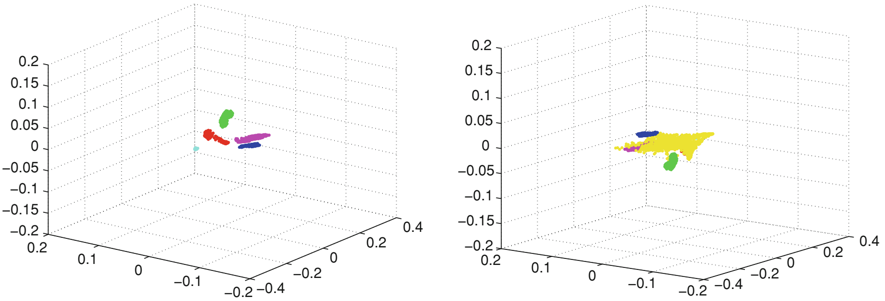


Fig. 10 Kernel PCA transformation: 3D ScatterPlot of the first principal components. Scattering of kernel principal components output, Dataset #1 (*left*) and Dataset #2 (*right*)

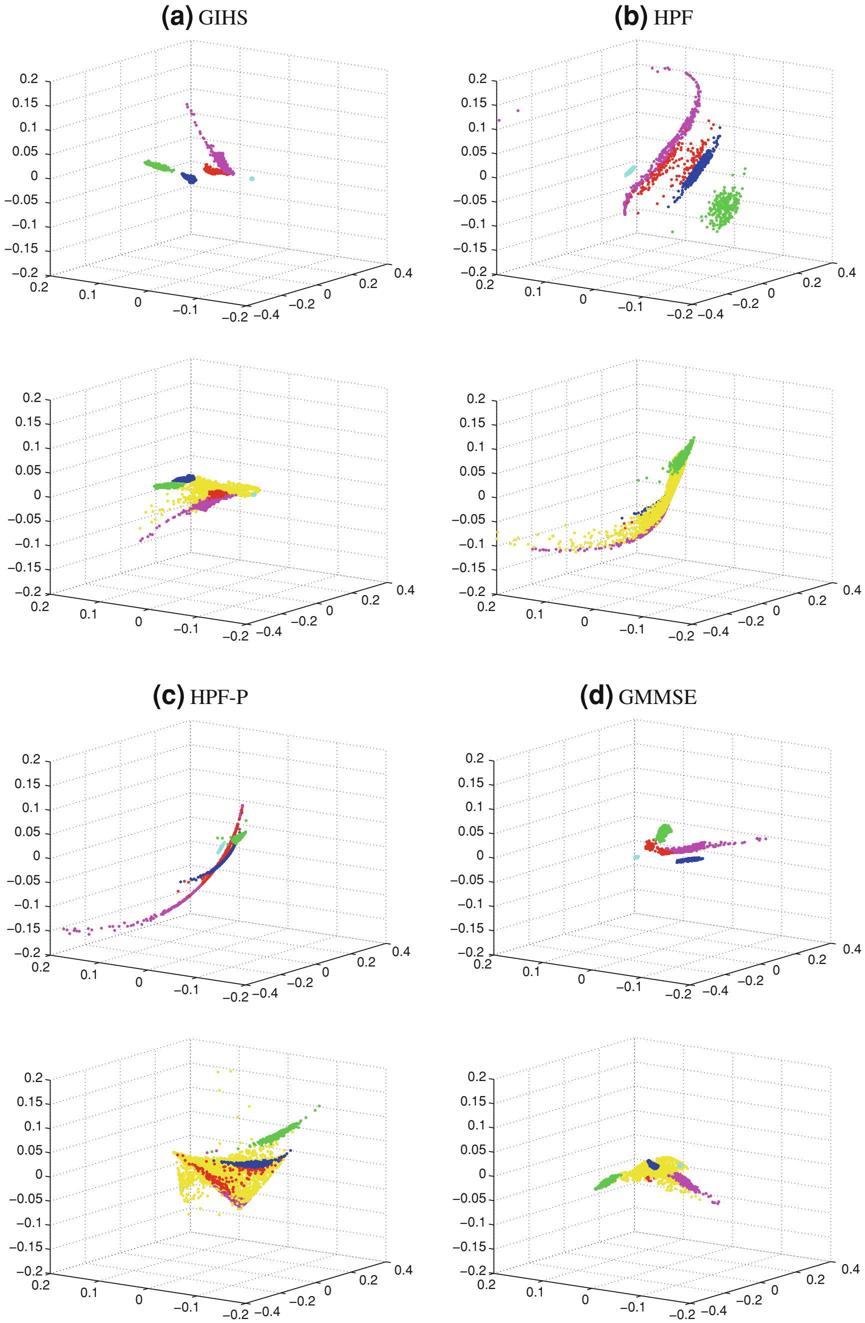


Fig. 11 Kernel PCA transformation: 3D ScatterPlot of the first principal components. Scattering of kernel principal components output, Dataset #1 (*upper part*) and Dataset #2 (*lower part*)

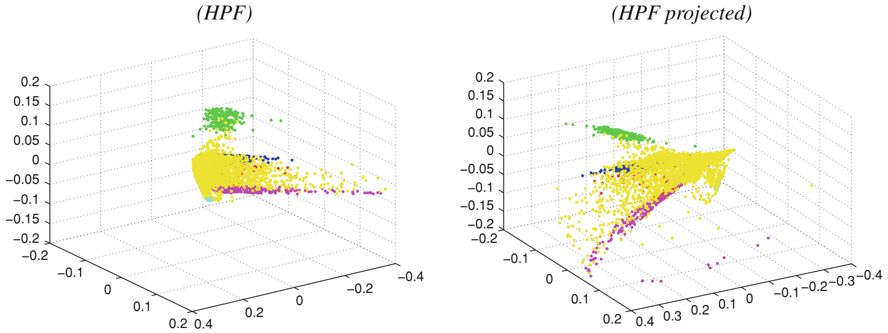


Fig. 12 Kernel PCA transformation: supplementary 3D scatter of the first kernel principal components

4.5 Linearity Preserving Projection (LPP)

Linearity Preserving Projection (LPP) is a technique that aims to combine the benefits of linear techniques and local nonlinear techniques for dimensionality reduction. The task is accomplished through the minimization of a cost function defined so that local properties of the data distribution are preserved based on the pairwise distances between near neighbors. Briefly, LPP computes a low-dimensional representation of the data in which the distances between a sample and its k_{NN} nearest neighbors are minimized by finding a linear mapping that minimizes the cost function defined as [38, 41]:

$$\phi(Y) = \sum_{ij} (y_i - y_j) w_{ij} \tag{11}$$

where y_i represents the samples in the low-dimensional representations and w_{ij} are the weights of the edges. In the cost function, large weights w_{ij} correspond to small distances between the samples x_i and x_j. Hence, the difference between their low-dimensional representations y_i and y_j highly contributes to the cost function. As a consequence, nearby points in the high-dimensional space are closer in the low-dimensional representation.

In detail, similar to many other methods for dimensionality reduction such as Laplacian Eigenmaps, LPP starts with the construction of a nearest neighbor graph $G(V, E)$, defined subsequently, in which every sample x_i is connected to its k_{NN} nearest neighbors x_{ij}. For all points in graph $G(V, E)$ that are connected by an edge, the weight of the edge in the graph is computed using the Gaussian kernel function, as defined in (3), leading to a matrix W with entries

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}. \tag{12}$$

The minimization is achieved by rewriting the problem in terms of graph Laplacian L, introduced in the following through the definition of many mathematical issues.

First, define the graph $G(V, E)$ with a set of n nodes, V, connected by a set of edges, E. The edge connecting nodes i and j has an associated weight, W_{ij} [42]. In this framework, the nodes are the samples, and the edges represent the similarity among samples in the data. A proper definition of the graph is the key to accurately introduce data structure in the machine.

Two mathematical tools have to be introduced to understand how matrix L is constructed [43, 44]:

- D is the *degree* matrix of size $n \times n$. Basically, D is a diagonal matrix $D = [d_1, \ldots, d_n]$ containing the number of connections to a node (degree);
- A is the adjacency matrix of size $n \times n$, where the nondiagonal entry is the number of connection from node i to node j, and the diagonal entry is either twice the number of loops at vertex i or just the number of loops. In our case, it is a matrix containing only (0, 1).

Finally, the Laplacian matrix L is defined as $L = D - W$, where W is obtained from A, the adjacency matrix, by assigning weights to each connection. Also, a normalized version of L can be obtained as

$$L_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } d_j \neq 0 \\ -\frac{1}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise.} \end{cases}$$

where subscripts i and j stand for the row and column indexes as well as the edges as defined before.

Once found the matrix L, LPP solves the generalized eigenproblem

$$(X - \bar{X})^T L (X - \bar{X}) \upsilon = \lambda (X - \bar{X})^T D (X - \bar{X}) \upsilon. \tag{13}$$

It can be shown that the eigenvectors v_i corresponding to the d smallest nonzero eigenvalues form the columns of the linear mapping T that minimizes the cost function in Eq. 11. The low-dimensional data representation Y is thus given by $Y = (X - \bar{X})T$.

Figure 13 shows the scatter plots of samples from data before and after transformation with a graph built by setting $\sigma = 1$ in the Gaussian kernel, while for the construction of the graph there is no need to set the number of nearest neighbors since the matrix D is filled straightforwardly with the distances. The relative low value of σ is justified by a simple consideration: the higher is the value introduced, the more the algorithm would inject increasing information about the structure of the distribution through the graph Laplacian. However, the analysis has revealed that the usage of spatial information does not change coherently the marginal distribution of the classes, or in other words, it produces

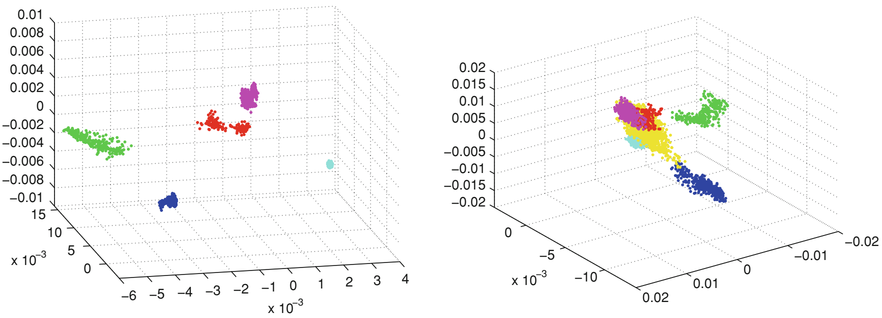


Fig. 13 Kernel LPP transformation: 3D ScatterPlot of the results. Scattering of the three components of the output, Dataset #1 (*left*) and Dataset #2 (*right*)

a diffusion effect whose entity depends on the local spatial distribution and pureness of pixels. Hence, the analysis has to be concerned on the spreading of the relative distances among nearby samples, and not on the global structure of classes distributions.

Regarding the results on the original data it is worth noting that the LPP algorithm produces excellent results since, even if not fed with supervised information, classes appear definitely clustered and separated. Hence, the task is to understand by means of the LPP algorithm, how local distances are maintained.

The upper parts in Fig. 14 report the results of LPP carried out without considering the class ‘Subspace’ that, in some sense, represents the connection between all the other classes. In the case of GIHS and GMMSE both algorithms performs very well and with similar results (except the fact that the features are flipped); moreover, classes appear more clustered than in the original case (see Class ‘Vegetation’, in green). Similar considerations can be done for the GIHS and GMMSE in the case of full data LPP processing.

As well as the case of the KPCA method, HPF and HPF-P behave differently. In particular, when the LPP is fed without the ‘Subspace’ Class, lack or weakness of the graph connections produce poor results; the spatial injection introduces a spectral distortion with a spreading effect on the samples, that can be captured by using higher values of σ in the Gaussian kernel.

When the samples belonging to the ‘Subspace’ class contribute to the input of the LPP algorithm (see lower part in fig. 14), a connecting effect is introduced in the samples graph, hence the results appear more consistent: given the same values of σ, the introduction of a greater number of samples fills the empty space between them, so the weights contributing to the graph creation are higher and consequently the algorithm can estimate better the main directions of the data scatter. The analysis reveals that the distances between samples are higher in the case of HPF and its HPF-P version.

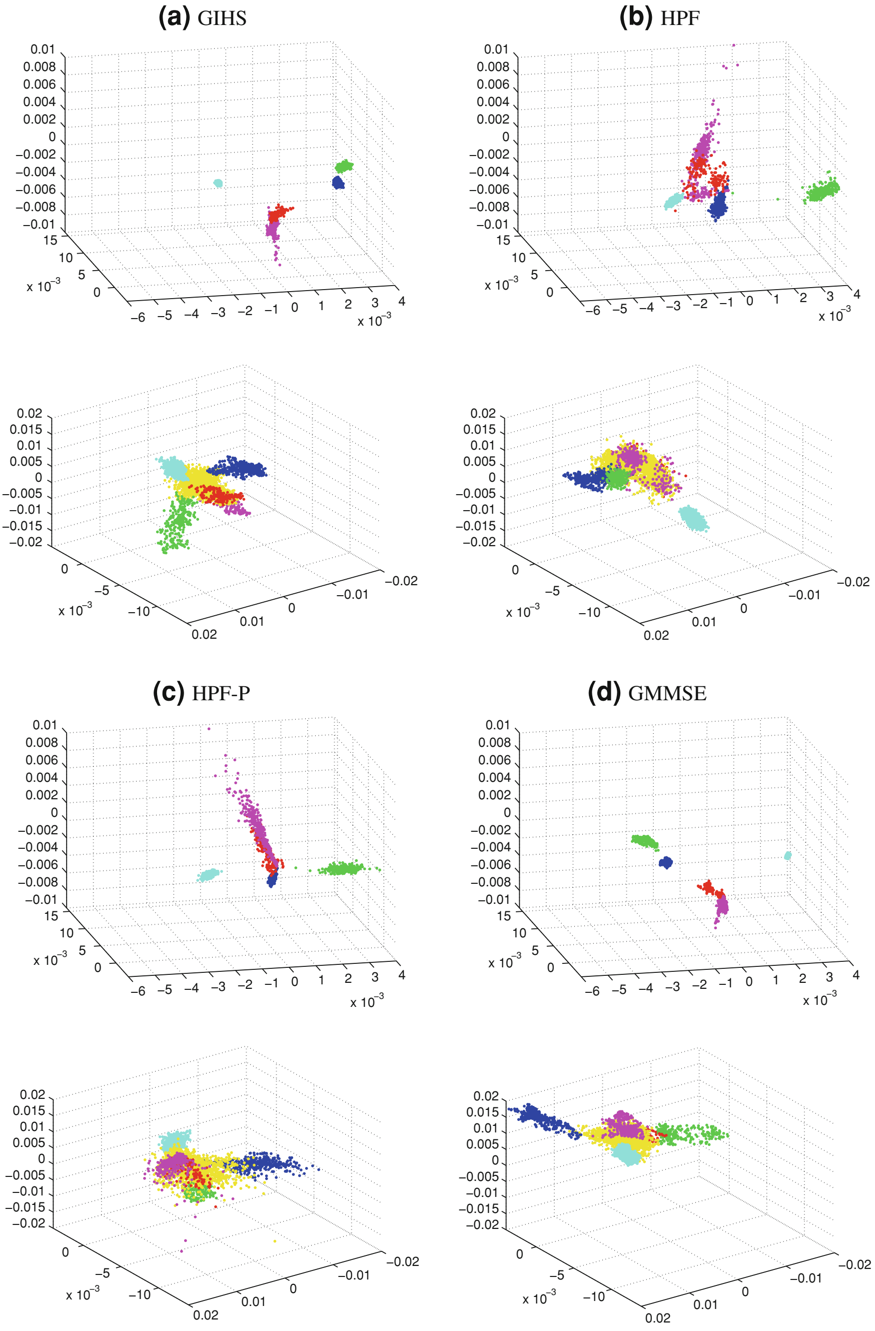


Fig. 14 Kernel LPP transformation: 3D ScatterPlot of the results. Scattering of the three components of the output, Dataset #1 (*upper part*) and Dataset #2 (*lower part*)

5 Conclusions

We have presented an experimental study on the effects of spatial enhancement of HS images on the distribution of spectral classes. The comparative analysis has been performed on original and pan-sharpened HS images by means of both linear and non-linear dimensionality reduction methods. The methodology for data analysis has proven to be consistent with traditional quality assessment of pan-sharpened data, and to be useful for understanding whether a particular fusion algorithm may improve or not application-specific HS processing. The results show that, as expected, the GIHS fusion algorithm privileges local spatial information with respect to the spectral one, since it introduces spectral distortion. The HPF and, to a lesser extent, the HPF-P methods suffer from weak preservation of the spectral correlation of the original data. Among the four pan-sharpening algorithms considered, GMMSE guarantees the best preservation of spectral distribution, almost unchanged class mixture model, good preservation of data correlation in the spectral dimension, and capability of maintaining or even improving clustering of data in the hyperspectral space.

References

1. Eismann, M., Hardie, R.: Hyperspectral resolution enhancement using high-resolution multispectral imagery with arbitrary response functions. IEEE Trans. Geosci. Remote Sens. **43**(3), 455–465 (2005)
2. Capobianco, L., Garzelli, A., Nencini, F., Alparone, L., Baronti, S.: Spatial enhancement of Hyperion hyperspectral data through ALI panchromatic image. In Proceedings of IEEE International Geoscience and Remote Sensing Symposium, IGARSS'07, pp. 5158–5161 (2007)
3. Tu, T., Su, S., Shyu, H., Huang, P.: A new look at IHS-like image fusion methods. Inf. Fusion **2**(3), 177–186 (2001)
4. Carper, W., Lillesand, T., Kiefer, R.: The use of intensity–hue–saturation transformations for merging spot panchromatic and multispectral image data. Photogramm. Eng. Remote Sens. **56**(4), 459–467 (1990)
5. Tu, T., Huang, P., Hung, C., Chang, C.: A fast intensity–hue–saturation fusion technique with spectral adjustment for IKONOS imagery. IEEE Geosci. Remote Sens. Lett. **1**(4), 309–312 (2004)
6. González-Audicana, M., Otazu, X., Fors, O., Álvarez-Mozos, J.: A low computational-cost method to fuse IKONOS images using the spectral response function of its sensors. IEEE Trans. Geosci. Remote Sens. **44**(6), 1683–1691 (2006)
7. Chavez, P.S., Sides, S.C., Anderson, J.A.: Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic. Photogramm. Eng. Remote Sens. **57**(3), 295–303 (1991)
8. Zhang, Y.: Understanding image fusion. Photogramm. Eng. Remote Sens. **70**(6), 653–760 (2004)
9. Laben, C.A., Brower, B.V.: Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. Eastman Kodak Company, Tech. Rep. US Patent #6, 011, 875 (2000)

10. Aiazzi, B., Baronti, S., Selva, M.: Improving component substitution pansharpening through multivariate regression of MS+Pan data. IEEE Trans. Geosci. Remote Sens. **45**(10), 3230–3239 (2007)
11. Garzelli, A., Nencini, F.: Fusion of panchromatic and multispectral images by genetic algorithms. In Proceedings IEEE International Conference on Geoscience and Remote Sensing Symposium, IGARSS'06, pp. 3810–3813 (2006)
12. Alparone, L., Baronti, S., Garzelli, A., Nencini, F.: A global quality measurement of pan-sharpened multispectral imagery. IEEE Geosci. Remote Sens. Lett. **1**(4), 313–317 (2004)
13. Garzelli, A., Nencini, F., Capobianco, L.: Optimal MMSE pan sharpening of very high resolution multispectral images. IEEE Trans. Geosci. Remote Sens. **46**(1), 228–236 (2008)
14. Wald, L., Ranchin, T., Mangolini, M.: Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. Photogramm. Eng. Remote Sens. **63**(6), 691–699 (1997)
15. Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A.: Sharpening of very high resolution images with spectral distortion minimization. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, IGARSS '03, vol. 1, pp. 458–460 (2003)
16. Alparone, L., Wald, L., Chanussot, J., Thomas, C., Gamba, P., Bruce, L.: Comparison of pansharpening algorithms: outcome of the 2006 GRS-S data fusion contest. IEEE Trans. Geosci. Remote Sens. **45**(10), 3012–3021 (2007)
17. Garguet-Duport, B., Girel, J., Chassery, J., Patou, G.: The use of multiresolution analysis and wavelets transform for merging SPOT panchromatic and multispectral image data. Photogramm. Eng. Remote Sens. **62**(9), 1057–1066 (1996)
18. Yocky, D.A.: Multiresolution wavelet decomposition image merger of Landsat Thematic Mapper and SPOT panchromatic data. Photogramm. Eng. Remote Sens. **62**(9), 1067–1074 (1996)
19. Aiazzi, B., Alparone, L., Argenti, F., Baronti, S., Pippi, I.: Multisensor image fusion by frequency spectrum substitution: subband and multirate approaches for a 3:5 scale ratio case. In Proceedings of IEEE International Geoscience and Remote Sensing Symposium, IGARSS'00, vol. 6, pp. 2629–2631 (2000)
20. Wilson, T., Rogers, S., Kabrisky, M.: Perceptual-based image fusion for hyperspectral data. IEEE Trans. Geosci. Remote Sens. **35**(4), 1007–1017 (1997)
21. Alparone, L., Cappellini, V., Mortelli, L., Aiazzi, B., Baronti, S., Carlà, R.: A pyramid-based approach to multisensor image data fusion with preservation of spectral signatures. In: Future Trends in Remote Sensing, Rotterdam, The Netherlands, Balkema (1998)
22. Zhou, J., Civco, D., Silander J., et al.: A wavelet transform method to merge Landsat TM and SPOT panchromatic data. Int. J. Remote Sens. **19**(4), 743–758 (1998)
23. Ranchin, T., Wald, L.: Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation. Photogramm. Eng. Remote Sens. **66**(1), 49–61 (2000)
24. Scheunders, P., De Backer, S.: Fusion and merging of multispectral images with use of multiscale fundamental forms. J. Opt. Soc. Am. A **18**(10), 2468–2477 (2001)
25. Núñez, J., Otazu, X., Fors, O., Prades, A., Palà, V., Arbiol, R.: Multiresolution-based image fusion with additive wavelet decomposition. IEEE Trans. Geosci. Remote Sens. **37**(3), 1204–1211 (1999)
26. Garzelli, A., Benelli, G., Barni, M., Magini, C.: Improving wavelet-based merging of panchromatic and multispectral images by contextual information. Proc. SPIE **4170**, 82 (2003)
27. Garzelli, A., Nencini, F.: Panchromatic sharpening of remote sensing images using a multiscale Kalman filter. Pattern Recognit. **40**(12), 3568–3577 (2007)
28. González-Audicana, M., Saleta, J.L., Catalán, R.G., García, R.: Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. IEEE Trans. Geosci. Remote Sens. **42**(6), 1291–1299 (2004)
29. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A.: Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. IEEE Trans. Geosci. Remote Sens. **40**(10), 2300–2312 (2002)

30. Ranchin, T., Aiazzi, B., Alparone, L., Baronti, S., Wald, L.: Image fusion—the ARSIS concept and some successful implementation schemes. ISPRS J. Photogramm. Remote Sens. **58**, pp. 4–18 (2003)
31. Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Selva, M.: MTF-tailored multiscale fusion of high-resolution MS and pan imagery. Photogramm. Eng. Remote Sens. **72**(5), 591–596 (2006)
32. Garzelli, A., Nencini, F.: Interband structure modeling for pan-sharpening of very high-resolution multispectral images. Information Fusion.**6**(3), 213–224 (2005)
33. Garzelli, A., Nencini, F.: Hypercomplex quality assessment of multi/hyper-spectral images. IEEE Geosci. Remote Sens. Lett. **6**, 662–665 (2009)
34. Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., Selva, M.: Multispectral and panchromatic data fusion assessment without reference. Photogramm. Eng. Remote Sens. **74**, 193–200 (2008)
35. Bruzzone, L., Carlin, L., Alparone, L., Baronti, S., Garzelli, A., Nencini, F.: Can multiresolution fusion techniques improve classification accuracy? In: Bruzzone, L. (ed.) Image and Signal Processing for Remote Sensing XII, p. 636509. SPIE, Bellingham (2006)
36. Bovolo, F., Bruzzone, L., Capobianco, L., Marchesi, S., Nencini, F., Garzelli, A.: Analysis of the effects of pansharpening in change detection on VHR images. IEEE Geosci. Remote Sens. Lett. **6**(1), 53–57 (2010),
37. Garzelli, A., Capobianco, L., Nencini, F.: On the effects of pan-sharpening to target detection. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, IGARSS'09, 136–139 (2009)
38. van der Maaten, L.J.P., Postma, E.O., van den Herik, H.J.: “Dimensionality reduction: a comparative review,” 2007. Tilburg University Technical Report TICC-TR 2009-005, 2009
39. Scholkopf, B., Mika, S., Burges, C., Knirsch, P., Muller, K., Ratsch, G., Smola, A.: Input space versus feature space in kernel-based methods. IEEE Trans. Neural Netw. **10**(5), 1000–1017 (1999)
40. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge (2004)
41. He, X., Niyogi, P.: Locality preserving projections. Adv. Neural Inf. Process. Syst. **16**, 37 (2004)
42. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning, 1st edn. MIT Press, Cambridge (2006)
43. Chung, F.: Spectral Graph Theory, 1st edn. ser. CBMS Regional Conference Series in Mathematics, no. 92. American Mathematical Society, Providence (1997)
44. Camps-Valls, G., Bandos, T., Zhou, D.: Semi-supervised graph-based hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **45**(10), 2044–3054 (2007)

Fusion of Optical and SAR Data for Seismic Vulnerability Mapping of Buildings

Diego Polli and Fabio Dell'Acqua

Abstract Seismic risk depends not only on seismic hazard, but also on the vulnerability of exposed elements since it is important in providing the necessary information to policy and decision-makers in order to prevent and mitigate the loss in lives and property. Currently, the estimation of seismic vulnerability of buildings relies on accurate, complex models to be fed with large amounts of in situ data. A limited geographical scope is a natural consequence of such approach, while extensive assessment would be desirable when risk scenarios are concerned. Remote sensing might be fruitfully exploited in this case, if not for a gap between information required by current, accurate, data-hungry vulnerability models and information derivable from remotely sensed data. In this context, naturally the greatest amount of information should be collected, and data fusion is more a necessity than an option. Fusion between optical and radar data allows covering the widest range of information pieces; in this chapter we will describe how such information may be extracted and how it can be profitably fed to simplified seismic vulnerability models to assign a seismic vulnerability class to each building. Some examples of real cases will also be presented with a special focus on the test site of Messina, Italy, a notorious seismic-prone area, where an intensive campaign of data collection is in progress within our research group.

Keywords Data fusion · Very high resolution radar · Building mapping · Seismic vulnerability

D. Polli
Remote Sensing Group, Department of Electronics, University of Pavia, Pavia, Italy

F. Dell'Acqua (✉)
Telecommunications and Remote Sensing Section, the European Centre for Training and Research on Earthquake Engineering (EUCENTRE), Pavia, Italy
e-mail: fabio.dellacqua@eucentre.it

S. Prasad et al. (eds.), *Optical Remote Sensing*,
Augmented Vision and Reality, 3, DOI: 10.1007/978-3-642-14212-3_15,

1 Introduction

Optical remote sensing has a long story [1] of success in wide-scale classification of land cover as well as in retrieving features and characteristics of selected items such as vegetation [2] and water [3]. Yet some pieces of information or conditions of operation are definitely out of the reach for very-short-wavelength remote sensing, such as directly detecting conductive or moving objects, or operating in poor weather conditions.

Apart from these extreme cases, it is a well known fact that optical and radar remote sensing can complement each other very well and provide, when exploited together, more information than the sheer sum of single contributions. In general, it is correct to assume that improvements in terms of classification accuracy, rejection rate, and interpretation robustness can only be achieved at the expense of additional independent data delivered by sensors. Data fusion is a concept that formalizes the combination of these measurements. In this chapter a review will be provided on the fusion of optical and radar data, with a specific attention to fusion between very-high-resolution data from the two realms.

2 Fusion of Optical and Radar Data

Data fusion [4] gathers together a large number of methods and mathematical tools, ranging from spectral analysis to plausibility theory. Fusion is not specific to a theme or an application; tools used in a data fusion process for a given application may instead be tailored to the case at hand.

Despite the fact the fusion of optical and radar data is potentially very advantageous, the difficulty inherent in combining so largely different types of data prevented it from becoming commonplace. Optical and radar data may not be both available with the given characteristics at the target site; or they may be available, but with such a long time span between them that some relevant information may become uncorrelated. Even when suitable data has been retrieved from both sources, the image pair needs to be accurately co-registered, which is not a painless procedure. Traditional, correlation-based methods [5], which used to work for optical-to-optical image registration, are not applicable when optical-to-radar image registration is concerned.

Correlation-based methods indeed assume similar types of sensors and tend to fail for registration of optical and radar images, because those two images have no radiometric correlation at all, due to the extremely different wavelengths.

Other approaches were then developed which do not assume radiometric correlation: matching connected-groups of pixels (blobs) in the two images [6, 7]; chain-codes description of contours [8]; application of active contour models [9–11].

Even these methods will often fail to accurately register optical and radar images for at least two reasons. The first is—again—that the two images have different radiometric characteristics, and in many cases the contrast between objects can be even reversed. Correlations between such dissimilar images will rarely yield a peak, even when correlation is computed locally. The second important reason for failure lies in speckle noise, which introduces strong distortions in the apparent shape of the areas found in the radar image with respect to the optical one, and this may be sufficient to prevent matching of the corresponding areas. More recently [12], edge-based methods have been proposed which get around the problem of radiometric correlation and are capable of providing good geometric agreement between the registered images, at least at the resolutions typical for the elder generation of Earth Observation (EO) satellites (i.e., on the order of 10 m).

Nowadays, however, we are witnessing a turnover from the old generation of 10-m radar satellites (ERS, ASAR, JERS) to the new, meter-resolution synthetic aperture radar, with the launch of satellites like COSMO/SkyMed [13], TerraSAR-X [14], and RADARSAT-2 [15]. The new generation of very high resolution (VHR) radar satellites brings finest achievable radar resolution once more very close to that of optical satellites, thus making the scales of the two types of data comparable again. It is on urban areas that the finest spatial resolution of such data is best appreciated, given the extreme spatial variability of the urban environment. At these resolutions, details of the buildings can be seen on both types of data, and their fusion can theoretically achieve the best results.

Sportouche et al. [16] have presented a method for building information extraction with the purpose of a 3D reconstruction exploiting data fusion. They use the optical image to obtain the footprint of the building; later they validate building detection and extract height information exploiting SAR data. Other methods employ high-resolution In-SAR data and optical imagery to extract facilities such as buildings or bridges [17, 18]. In case of a seismic event, damage mapping can be very useful and data fusion is still a very powerful tool for this purpose [19, 20]. It is possible to exploit both the high repetition observation rate available with the new generation of SAR systems and the fine level of detail available even in a single multiband optical image with the aim of change detection [21]. Again for the same purpose, i.e. change detection, images are used, acquired at different times during the process of construction of a city or reconstruction of an urban area stricken by a natural disaster [22].

Although still limited in its extent by the relative novelty of the data, fusion between very high-resolution optical and radar data clearly represent a very fertile terrain for the construction of powerful tools for information extraction through remote sensing. Even more so for urban areas, whose inherent complexity makes the fine spatial discrimination granted by these data highly desirable for enabling large-scale exploitation of the wealth of information contained in the acquired data.

In the next section, we will illustrate some of the issues raised by the optical and radar data fusion at very high resolution by analyzing a concrete example.

3 Data Fusion for Vulnerability Assessment

In order to illustrate the usefulness of data fusion, we focus our attention on a particular application (seismic vulnerability assessment), which is particularly interesting as it is relatively new.

3.1 The Aim

Seismic risk depends on both seismic hazard (i.e. how likely an earthquake of given intensity is to occur) and vulnerability of exposed elements (i.e. how likely is a building to suffer damage of a given extent as a consequence of a seismic input of a given intensity), although it is more commonly thought of in terms of hazard alone.

The contribution of remote sensing to seismic hazard computation is generally indirect, as it consists of collecting clues on, e.g. seismic faults location and patterns, to be used as input to probabilistic models, which in turn provide an estimate of the earthquake probabilities. The information fed in through remote sensing is often replaceable with input from other sources, like, e.g. global fault models.

On the other side, the contribution of remote sensing to seismic vulnerability estimation can be substantial. At a very different scale with respect to the factors connected with seismic hazard, vulnerability assessment can help mapping seismic risk at a deep detail level. Thus, a capability to map vulnerability on a wide geographical scope can be very beneficial to improve disaster preparedness on the one side, and to make early-stage damage estimation more precise and reliable by incorporating vulnerability models into damage estimation algorithms.

As already mentioned, the seismic vulnerability of a structure can be defined as its susceptibility to be damaged from ground shaking of a given intensity, usually described in terms of probability of damage and discrete levels of damage, respectively. Evaluating the vulnerability of existing building stock is certainly pivotal in this framework and indeed it has a long history of method proposed along the years [23], based either on empirical, analytical or even hybrid approaches. In general the various methods proposed need a considerable amount of information to be collected; for example, when the response of a single building is considered, existing approaches essentially require several studies on the structure as an accurate examination of the possible local mechanisms of damage and collapse, the selection of a probable non linear response mechanism, and so on. This may represent a severe limitation on the geographic scope of the vulnerability estimation procedure, either because historical data are unavailable at the desired precision or format, or because the in situ collection of data is too expensive and time-consuming to make it practical to collect the required information. Though, it may become feasible once suitable methods become available and trading precision for geographical scope is a viable option. Recently, new algorithms have been developed for vulnerability assessment, which require fewer data, normally available from census on the building stock, e.g. year of construction, number of storeys, materials, etc. One of such methods, termed Simplified Pushover-Based Earthquake Loss Assessment (SP-BELA) [24] can provide a sensible output for comparison purposes even with a very limited set of inputs. These include the footprint of the building and the number of storeys—the latter parameter being more important than the total height of the structure. Remote sensing techniques, which by definition can operate on far larger scales than in situ data collection, are in a position to complete the framework [25]. The 3D shape of the building is a most relevant input item. In literature it is possible to find lots of building height extraction methods, both for optical and SAR imagery. Existing methodologies are either based on shadow analysis or on interferometric data [26, 27]. However, the calculation of the interferogram fails if all of the roof backscattering is sensed before the double bounce area and therefore superimposes with the ground scattering in the layover region, which is usually the case for high buildings. In order to tackle the problem of signal mixture from different altitudes methods founded on interferometric or polarimetric data or stereoscopic SAR are proposed [28, 29]. Recently, methods based on multi-aspect data where the same area is measured from different flight paths, were proposed [30]. Generally speaking, as testified by the amount of relevant literature, the problem of extracting a building 3D shape is quite a complex one. For our purposes, however, such problem can be split into two sub-problems, namely footprint extraction and determination of the number of storey. This latter problem is quite a new one in the remote sensing research scenario, and a simpler one with respect to traditional building height extraction. Our final intent is a wide range scanning of the urban environment, using optical data to extract footprints of buildings and, due to its side-looking nature, using SAR data to extract the number of storey. These pieces of information will then represent the basic input to the vulnerability model.

3.2 Remote Sensing as a Tool

It is thus clear that a combination of optical and radar data, both at a very high resolution, can satisfy the information needs related to wide-scale vulnerability assessment in urban areas.

High-resolution (HR) optical data seems to be a good means to determine items such as shape and size, footprint of the building, relative location and orientation of neighbouring buildings. The main issue with HR optical data is related to its cost, currently around 20 € per square kilometre for archive data, rising up to 40–50 € per square kilometre if multi-vantage point acquisition is involved, useful for, e.g. cross-checking the height of the building with the value determined from shadow length or from estimation of the number of floors.

High-resolution SAR data, as already mentioned, is starting to become more widely available thanks to the launch and activation of a new generation of satellites with ground resolution around 1 m. Such systems have started producing radar images of the Earth surface at an unprecedented spatial resolution, at least for spaceborne systems. This is opening up new possibilities, as these systems combine the all weather, night and day operation typical of radar systems with a fine geometrical resolution, which allows sensing details of the scene previously concealed. Such ability allows for example an accurate updating of the disaster-prone areas, because mapping of the significant elements can be performed as soon as the acquired data becomes available, through a mapping process. This is connected with vulnerability, in the sense of affording an updated scenario of possible life-lines, escape routes and population distribution. It is however difficult to estimate the cost of using such images as most of the data distribution is still made for scientific purposes only, at subsidized prices.

In order to better illustrate the issues involved in seismic vulnerability determination from combined optical and radar satellite data, we will focus on a specific test site, i.e. the city of Messina, Italy. This city is famous to the earthquake scientist community because of the disastrous 1908 event, which triggered also a tsunami resulting in its almost complete destruction. Several studies are underway on this test site and the 2008 Applied Geophysics Conference took place in Messina to celebrate 100 years of progress in disaster mitigation and management. The vulnerability of Messina building stock was analysed through a statistical approach where the assessment unit was the census tract.

Extraction of building footprint, as well as extraction of the number of storeys, is performed relying extensively on a linear feature extractor termed W-Filter which is part of a feature extraction software named BREC [31]. The footprint of the building (Fig. 2) was extracted by applying the linear feature extractor to an optical, very-high-resolution image. This latter consisted of the panchromatic band of a Quickbird image, purchased for this specific purpose, whose features are reported in Table 1. A quick look of the image is visible in Fig. 1.

A procedure has been set up, capable of connecting the extracted linear segments into a "reasonable" footprint for the considered building. This procedure allows to outline the building footprint shape and size and to determine its across and along size, two most important parameters for vulnerability assessment.

The following step is the SAR image analysis: as we can see (Fig. 3a) radar images feature quite apparent rows of scatterers, probably originated by the corner structures constituted by the protruding balconies, in addition to the corner reflector structure at the pavement/façade meeting point. If we assume the footprint of the building is available, so is also the dominant direction of the façade in

Table 1 Information on the quickbird image

Sensor vehicle	Acquisition date	Total off nadir angle	Area max off nadir angle	Area max sun elevation	Total cloud cover pct.	Area cloud cover pct.	Imaging bands
QB2	28/07/2006	18.63°	17.96°	67.01°	4%	5%	Pan + MS1

Fig. 1 Preview of the purchased image ©DigitalGlobe

the image. Directional filtering enables turning such rows of scatterers into a more homogeneous, linear bright area, which can be easily detected by the linear feature extractor, as seen in Fig. 3. Quite apparent here are the three parallel lines which mark the associated three storeys. Counting the longest parallel lines extracted from the image results in determining the number of storeys in the building. The overall information flow is shown in Fig. 4.

The experiments have shown that, unfortunately, although apparently the information on the number of floors can be extracted from visual interpretation, the procedure set up seems to be somehow too simplistic and sometimes it fails to deliver the correct number of floors (Fig. 5).

The main problem seems to be in the directional filtering, failing to sufficiently highlight the edges between reflector rows for the extractor to work correctly.

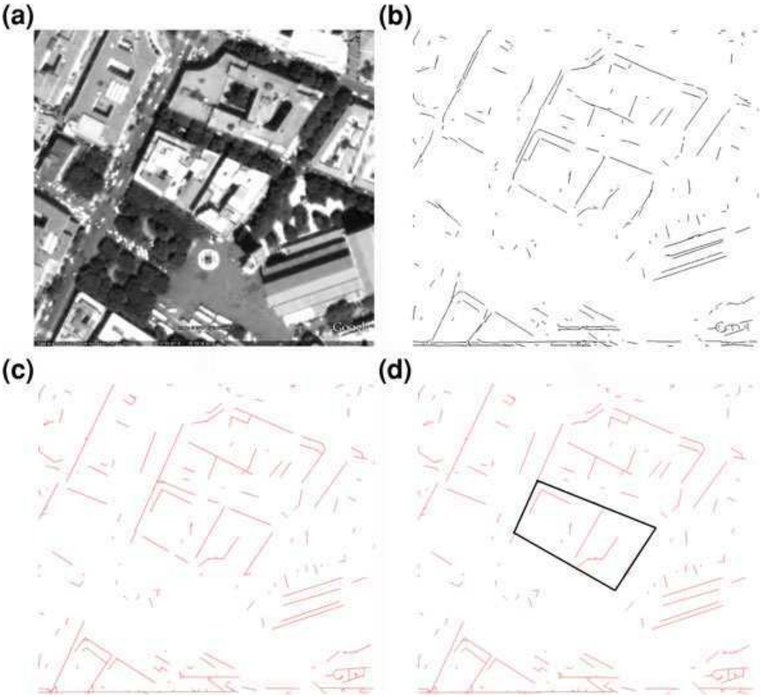


Fig. 2 Steps in generation of building footprint estimate: **a** the original grayscale image, **b** preliminary feature extraction, **c** feature merging, **d** footprint hypothesis

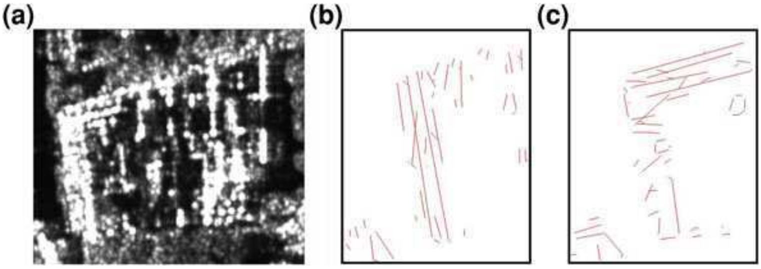


Fig. 3 **a** SAR image of the selected building, **b** segments extracted from north-west façade, **c** segments extracted from north facade

This issue was addressed by introducing two important novelties:

- Use of hard decision (strong scatterer/no strong scatterer) on each pixel.
- SAR + SAR + optical fusion instead of SAR + optical alone.

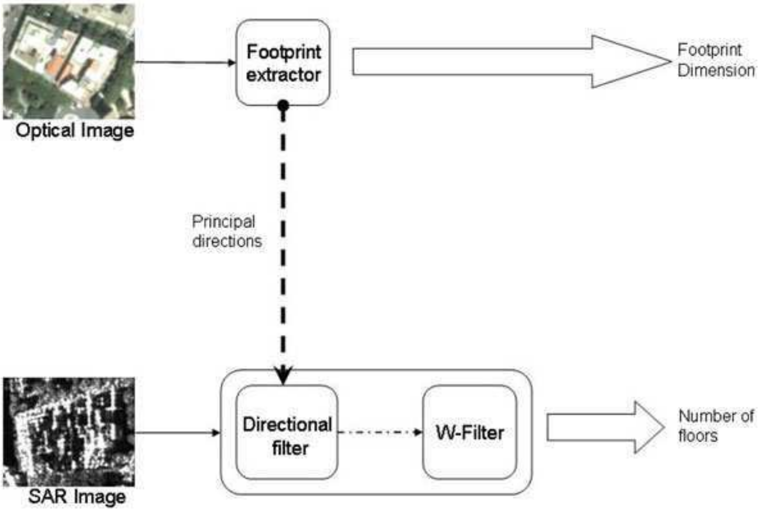


Fig. 4 Flow-chart of the applied method

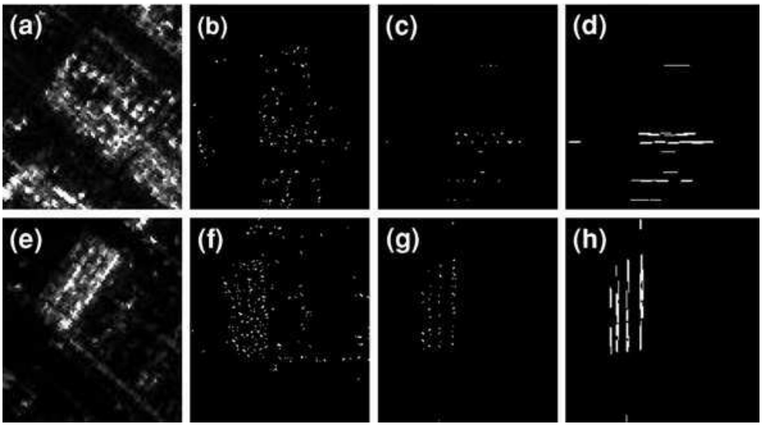


Fig. 5 Steps in number-of-storeys-extraction (a) and (e): original images, (b-c), (f-g): after rotation to align reflector lines with principle direction and filtering; (d) and (h): examples of reflector row extractions.

The first modification was introduced to account for the insufficient contrast created by the directional filter. Instead of attempting to make the impulse response of the filter sharper and sharper, a strategy that has proved to be basically ineffective, a binary logic was introduced. A preliminary step is introduced, in which pixels contained in the image are tested for being local maxima. If so, they are marked with a "1" on a resulting mask image, "0" otherwise. Strong scatterers, despite their spatially spread response probably due to the SAR distributed

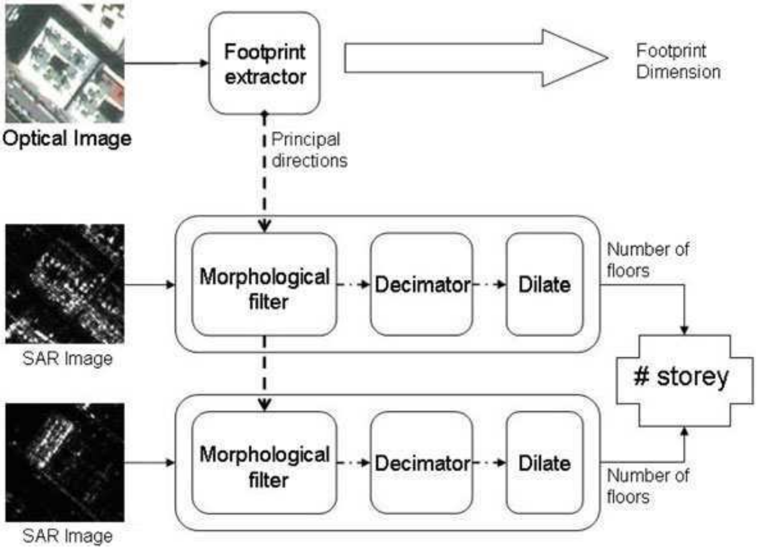


Fig. 6 Flow-chart of the second method

impulse response, are turned into single 1's in the mask image. The mask image is then rotated by the orientation angle retrieved from the optical image.

At this stage, a morphological dilation is performed using a constituting element whose shape is that of a row of pixels—equivalent to extending the "1" marked area along rows, given the rotation of the mask image. This results in merging together the scatterers constituting a row marking a floor boundary. A final stage consists of counting the number of 0–1–0 transitions along each column, as this is expected to be connected to the number of floors. Isolated transitions are not taken into account as they may be connected with speckle spikes.

The second modification was introduced to make the overall procedure more robust. On the test site, as already mentioned, more than one SAR image was available from different vantage points. Thus, a second image of the same building from a more favourable azimuth to see a different façade of the building was considered, and underwent the same procedure.

Figure 6 shows a flowchart representing this second method used to assess then number of floors of a given building:

3.3 Decision-Level Fusion

A final fusion step between the estimates of the floor number is then performed as visible in Fig. 7. The number of floors results from majority voting between the numbers of transitions extracted from the mask image along its columns,

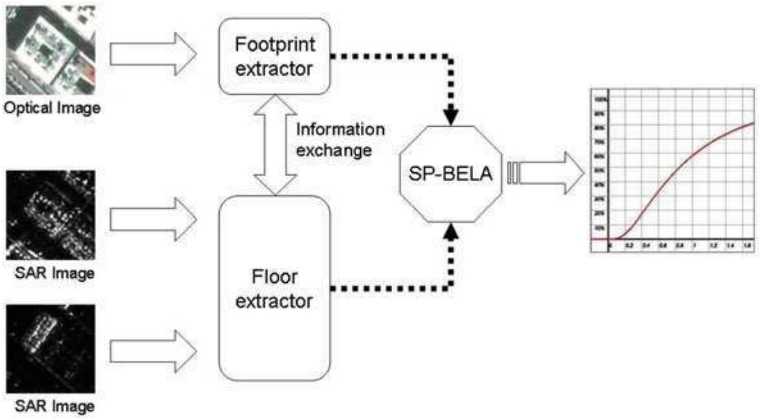


Fig. 7 Flow-chart of the final data fusion

according to the criteria discussed in the former subchapter. The experiments report a large number of errors on single columns yet with a large majority of correct counts.

It can be argued that the method developed is very case-specific as on the particular site of Messina several airborne radar images were available along different flight lines and thus with different azimuth view angles. This naturally makes exploration of scatterers from different sides of the buildings easier. This situation can however be effectively simulated through the use of spaceborne radar images acquired on ascending and descending orbits on the same site. If left- and right-looking capabilities are also available, the total number of images available at different azimuth vantage points rises to 4, which is probably sufficient for many sites.

This method seems to have marked a step forward in reliability of the floor number estimation.

4 Conclusions

In this chapter, the topic of optical and radar data fusion at a very high resolution has been discussed. Fusion of HR SAR and HR optical data has been shown to be useful to make each type of data fill in the other's gaps. Just to mention a few basic examples, severe geometric distortions in radar data may be inverted where near-nadir HR optical data are available, faithfully reproducing the shape of the objects. On the other hand, height information may be more easily extracted from radar shadows than from nadir HR optical data.

In order to discuss more specifically the issues related to optical and radar data fusion, a particular application, i.e. seismic vulnerability assessment, has been addressed. It has been shown in a practical case how the optical and radar image can complement the information one may extract from the two types of data, together providing a fairly complete set of features of an observed building.

Still, the usefulness of VHR optical + radar data fusion is still somehow hindered by the complex behavior of responses from objects observed at those finest resolutions. The literature on this sort of data fusion is still somehow scarce, although it is expected to increase considerably in the coming years thanks to the ever increasing availability of this type of data.

Acknowledgments The authors wish to acknowledge the support of the Italian Civil Protection Department ("Programma Quadro" 2009–2011 funding of the European Centre for Training and Research in Earthquake Engineering, EUCENTRE, Pavia) and the European Commission (funding of project SAFER, 2009). They also wish to thank the colleagues at the Seismic Risk Section of EUCENTRE, particularly Helen Crowley and Barbara Borzi for their help with the SP-BELA model.

References

1. Thrower, N.J.W.: Land use in the Southwestern United States from Gemini and Apollo imagery (map suppl. no. 12). Ann. Assoc. Am. Geogr. **60**(1), 208–209 (1970)
2. Myneni, R.B., Pinty, B., Maggion, D.S. Kimes, S., Iaquinta, J. Privettet, J.L., Gobron, N., Verstraetett, M., Williams, D.L.: Optical remote sensing of vegetation: modeling, caveats, and algorithms. Remote Sens. Environ. **51**, 169–188 (1995)
3. Smith, R.C., Baker, K.S.: The bio-optical state of ocean waters and remote sensing. Limnol. Oceanogr. **23**(2), 247–259 (1978)
4. Wald, L.: A conceptual approach to the fusion of earth observation data. Surv. Geophys. **21**, 177–186 (2000)
5. Fonseca, L.M.G., Manjunath, B.S.: Registration techniques for multisensor remotely sensed imagery. Photogr Eng Remote Sens **62**, 1049–1056 (1996)
6. Ali, M.A., Clausi, D.A.: Automatic registration of SAR and visible band remote sensing images. In: Proceedings of the Geoscience and Remote Sensing Symposium IGARSS '02, IEEE International, pp. 1331–1333 (2002)
7. Dare, P., Dowman, I.: A new approach to automatic feature based registration of SAR and SPOT images. Int. Arch. Photogr. Remote Sens. **XXXIII**, 125–130 (2000)
8. Dai, X., Khorram, S.: A feature-based image registration algorithm using improved chain-code representation combined with invariant moments. IEEE Trans. Geosci. Remote Sens. **37**(5), 2351–2362 (1999)
9. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comp. Vis. **1**(4), 321–331 (1987)
10. Li, H., Manjunath, B.S., Mitra, S.K.: A contour-based approach to multisensor image registration. IEEE Trans. Image Process. **4**(3), 320–334 (1995)
11. Maitre, H., Wu, Y.: A dynamic programming algorithm for elastic registration of distorted pictures based on autoregressive models. IEEE Trans. Acoust. Speech Signal Process **37**, 288–297 (1989)
12. Hong, T.D., Schowengerdt, R.A.: A robust technique for precise registration of radar and optical satellite images. Photogr. Eng. Remote Sens. **71**(5), 585–593 (2005)
13. Impagnatiello, F., Bertoni, R., Caltagirone F.: The SkyMed/COSMOsystem: SAR payload characteristics. In: Proceedings of IGARSS'98, vol. 2, pp. 689–691, 6–10 July 1998, Seattle (WA) (1998)

14. Roth, A.: TerraSAR-X: a new perspective for scientific use of high resolution spaceborne SAR data. In: Proceedings of 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, pp. 4–7, 22–23 May 2003, Berlin, Germany (2003)
15. Morena, L.C., James, K.V., Beck, J.: An introduction to the RADARSAT-2 mission. Can. J. Remote Sens. **30**(3), 221–234 (2004). ISSN 1712-7971
16. Sportouche, H., Tupin, F., Denise, L.: Building extraction and 3D reconstruction in urban areas from high-resolution optical and SAR imagery. Urban Remote Sensing Event, 2009 Joint, 20–22 May, pp. 1–11 (2004)
17. Wegner, J.D., Soergel, U., Thiele, A.: Building extraction in urban scenes from high-resolution InSAR data and optical imagery. Urban Remote Sensing Event, 2009 Joint, 20–22 May, pp. 1–6 (2009)
18. Soergel, U., Thiele, A., Gross, H., Thoennessen, U.: Extraction of bridge features from high-resolution InSAR data and optical images. Urban Remote Sensing Joint Event 11–13 April 2007 pp. 1–6 (2007)
19. Stramondo, S., Bignami, C., Pierdicca, N., Chini, M.: SAR and optical remote sensing for urban damage detection and mapping: case studies. Urban Remote Sensing Joint Event, 11–13 April 2007, pp. 1–6 (2007)
20. Chini, M., Pierdicca, N., Emery, W.J.: Exploiting SAR and VHR optical images to quantify damage caused by the 2003 Bam Earthquake. Geosci. Remote Sens. IEEE Trans. **47**(1), Part 1, 45–152 (2009)
21. Orsomando, F., Lombardo, P., Zavagli, M., Costantini, M.: SAR and optical data fusion for change detection. Urban Remote Sensing Joint Event 11–13 pp. 1–9 (2007)
22. Zhang, J., Wang, X., Chen, T., Zhang, Y.: Change detection for the urban area based on multiple sensor information fusion. Geoscience and Remote Sensing Symposium, 2005. IGARSS ‘05. Proceedings 2005 IEEE International, vol. 1, 25–29, p 4 , July 2005
23. Calvi, G.M., Pinho, R., Bommer, J.J., Restrepo-Vélez, L.F., Crowley, H.: Development of seismic vulnerability assessment methodologies over the past 30 years. ISET J. Earthq. Technol Paper No. 472 **43**(3), 75–104 (2006)
24. Borzi, B., Crowley, H., Pinho, R.: Simplified pushover-based earthquake loss assessment (SP-BELA) method for masonry buildings. Int. J. Archit. Heritage **2**(4), 353–376 (2008)
25. Polli, D., Dell’Acqua, F., Gamba, P., Lisini, G.: Remote sensing as a tool for vulnerability assessment. In: Proceedings of the 6th International Workshop on Remote Sensing for Disaster Management Applications, Pavia, Italy, 11–12 September 2008
26. Hill, R., Moate, C., Blacknell, D.: Estimating building dimensions from synthetic aperture radar image sequences. IET Radar Sonar Navig. **2**(3), 189–199 (2008)
27. Bennett, A.J., Blacknell, D.: Infrastructure analysis from high resolution sar and insar imagery. In: 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas. Berlin, Germany (2003)
28. Cellier, F., Colin, E.: Building height estimation using fine analysis of altimetric mixtures in layover areas on polarimentric interferometric x-band sar images. In: International Geoscience and Remote Sensing Symposium (IGARSS). Denver, CO, USA (2006)
29. Simonetto, E., Oriot, H., Garello, R.: Rectangular building extraction from stereoscopic airborne radar images. IEEE Trans. Geosci. Remote Sens. **43**(10), 2386–2395 (2005)
30. Xu, F., Jin, Y.Q.: Automatic reconstruction of building objects from multiaspect meter-resolution sar images. IEEE Trans. Geosci. Remote Sens. **45**(7), 2336–2353 (2007)
31. Gamba, P., Dell’Acqua, F., Lisini, G.: BREC: the Built-up area RECognition tool. In: Proceedings of the 2009 Joint Urban Remote Sensing Event (JURSE 2009)