

UNIVERSITY OF GENOA



XXX CYCLE PH.D PROGRAM IN
SCIENCE AND TECHNOLOGIES FOR ELECTRONIC
AND TELECOMMUNICATION ENGINEERING

March 2018

**Novel pattern recognition methods
for classification and detection in
remote sensing and power generation
applications**

Author

Andrea De Giorgi

Tutor

Dr. Gabriele Moser

Supervisor

Prof. Andrea Trucco

Ph.D program chair: Prof. Mario Marchese

Contents

Introduction	1
1 Pattern recognition principles	8
1.1 Supervised classification	10
1.1.1 Key ideas and terminology	10
1.1.2 Support vector machines	13
1.2 Unsupervised classification	16
1.3 Contextual and non-contextual classification	17
1.3.1 Key ideas and terminology	17
1.3.2 Markov random fields	19
2 Spatial-contextual kernel-based classification of remote sensing images	22
2.1 Introduction	22
2.2 Markovian support vector classification	25
2.2.1 Proposed formulation of the energy function	25
2.2.2 The proposed classification method	29
2.2.3 Parameter estimation and energy minimization	30
2.2.4 Experimental validation	32
2.3 Region-based kernel for optical data	42
2.3.1 Methodology	42
2.3.2 Experimental validation	45
2.4 Conclusion	52
3 Multiresolution classification of panchromatic and multispectral images	61

Contents

3.1	Introduction	61
3.2	Methodology	65
3.2.1	Multiresolution model and assumptions	65
3.2.2	Overview of the proposed method	67
3.2.3	PDF parameter estimation	69
3.2.4	MAP estimation of the virtual feature vectors	72
3.2.5	MRF parameter estimation and energy minimization	74
3.3	Experimental validation	75
3.3.1	Data sets and experimental setup	75
3.3.2	Classification results and experimental comparisons with the semi-simulated data set	79
3.3.3	Classification results and experimental comparisons with the real data sets	84
3.3.4	Dependence on training set size	91
3.4	Conclusion	96
4	Supervised classification of thermal infrared hyperspectral images	101
4.1	Introduction	101
4.2	Methodology	103
4.2.1	Sequential forward band extraction	103
4.2.2	Contextual and non-contextual classifiers	104
4.3	Experimental validation	105
4.4	Conclusion	111
5	Parameter optimization for Markov random field models for remote sensing image classification	113
5.1	Introduction	113
5.2	Methodology	115
5.3	Experimental validation	118
5.4	Conclusion	122
6	Fault diagnosis for power generation plants based on solid oxide fuel cell	124
6.1	Introduction	124

Contents

6.2	Fault detection and isolation	127
6.2.1	Model-based approach	127
6.2.2	Data-driven approach	128
6.2.3	Hybrid approach	129
6.2.4	Plant models simulating healthy and faulty conditions	129
6.3	Measurements and faults in solid oxide fuel cell plant	132
6.3.1	Plant structure and operating conditions	132
6.3.2	Monitored variables and measurement difficulties	133
6.3.3	Fault classes	134
6.3.4	Quantitative plant model	135
6.4	Support vector for fault diagnosis	137
6.4.1	Proposed hybrid approach, parameter optimization and feature selection	137
6.4.2	Experimental validation	140
6.5	Random forests for fault diagnosis	147
6.5.1	Proposed fault diagnosis approach	147
6.5.2	Experimental validation	149
6.6	Conclusion	157
	Conclusion	160
	Bibliography	164

List of Figures

1.1	Pattern recognition process	9
1.2	Hyperplane in bi-dimensional space: non-linear separability (left), linear separability (right).	12
1.3	Examples of neighboring systems	18
1.4	First and second order neighborhood systems (grey) of a site (black)	20
2.1	AVIRIS Indian Pines hyperspectral data. (a) Three band false color composite, (b) Reference data and (c) Color code.	33
2.2	Houston - From top to bottom: A color composite representation of the hyperspectral data using bands 70, 50, and 20, as R, G, and B, respectively; Training samples; Test samples; and legend of different classes.	36
2.3	ROSIS-03 Pavia University hyperspectral data. (a) Three band false color composite, (b) Reference data and (c) Color code.	37
2.4	Classification maps for Indian Pines: (a) SVM, (b) MSVC with graph cuts.	40
2.5	Classification maps for Pavia University: (a) SVM, (b) MSVC with LBP.	41
2.6	Classification maps for Houston:(a) SVM, (b) MSVC with TRW-S.	42

List of Figures

2.7	Port-au-Prince: (a) RGB composition of the optical channels; classification maps generated by the application of proposed method to the multisensor optical/SAR (b) and by same approach to only optical image (c) and only SAR image (d). Legend: water (blue), buildings with red roofs (red), urban and port areas (cyan), vegetation (green), bare soil (yellow).	43
2.8	First synthetic dataset. RGB false-color composition (a), two-dimensional histogram (b), and classification map generated by the proposed method (c)	46
2.9	Second synthetic dataset. RGB false-color composition (a), two-dimensional histogram (b), and classification map generated by the proposed method (c)	46
2.10	Third synthetic dataset. RGB false-color composition (a), two-dimensional histogram (b), and classification map generated by the proposed method (c)	47
2.11	Alessandria data set: RGB true-color composition of the IKONOS channels (left) and classification map (right).	48
2.12	Classification maps obtained, for the Itaipu data set, by SVM classifiers using the region-based kernel (a), a composite kernel combining the region-based and the pixelwise RBF kernels (b), and the purely pixelwise RBF kernel (c). Color legend: see Table 2.9.	51
2.13	Classification maps obtained, for the Itaipu data set, by SVM classifiers using the region-based kernel (a), a composite kernel combining the region-based and the pixelwise RBF kernels (b), and the purely pixelwise RBF kernel (c). Color legend: see Table 2.9.	52
3.1	Flowchart of the proposed method.	70
3.2	An example classification map on a PAN lattice and multispectral observations on an MS lattice in the MAP estimation of the virtual features. In this example, $\rho = 2$. The colors in the classification map denote different class labels.	73

3.3	“Itaipu”: (a) RGB false-color composition of the IKONOS channels; (b) training map; (c) test map; and classification maps generated by the proposed method (d) and by the technique in [135] (e) (both applied with $\rho = 2$) as well as by a single-resolution Markovian classifier (f). The white box in (a) highlights the detail zoomed in Figure 3.4.	80
3.4	“Itaipu”: detail of the RGB false-color composition of the IKONOS channels (a) and of the classification maps generated by the proposed method with $\rho = 2$ (b) and $\rho = 4$ (c), by the technique in [135] with $\rho = 2$ (d), and by a single-resolution Markovian classifier (e).	81
3.5	“Alessandria”: (a) RGB false-color composition of the MS channels, (b) PAN channel, (c) training map, (d) test map, classification maps generated by the proposed method (e), by the technique in [135] (f), by a single-resolution classification of the GS pansharpening result (g), by a single-resolution classification of the BDSF pansharpening result (h), and by a single-resolution classification of the GSA pansharpening result (i).	86
3.6	“Pavia”: (a) RGB false-color composition of the MS channels, (b) PAN channel, (c) training map, (d) test map, and classification maps generated by the proposed method (e), by the technique in [135] (f), and by a single-resolution classification of the pansharpening result (g).	88
3.7	“Alessandria”: detail of the RGB false-color composition of the MS channels (a), of the PAN channel (b), and of the classification maps generated by the proposed method (c) and by the technique in [135] (d). The color legend is the same as in Fig 3.5.	92
3.8	The behavior of the overall (OA) and average (AA) accuracies of the proposed method and the method described in [135], as functions of the number of training samples for (a) the “Pavia” data set; (b) the “Alessandria” data set; and (c) the “Itaipu” data set with $\rho = 2$. The same legend is used for all the plots.	94

List of Figures

4.1	False color composite of bands no. 1, 40, and 75 of the LWIR hyperspectral image (a), training map drawn mostly on the basis of visible data (b), classification map obtained by the MRF-based Gaussian classifier (c), non-contextual Gaussian Bayesian classifier (d), contextual SVM classifier (e), and non-contextual SVM classifier (f). See Tables 4.2 and 4.3 for the color legend.	107
4.2	Behavior of the OA of the non-contextual Gaussian Bayesian and of the MRF-based Gaussian classifiers as a function of the number of s-bands extracted by SFBE (4 classes).	108
4.3	Feature reduction effects on the MRF-based Gaussian classifier (c), non-contextual Gaussian Bayesian classifier (d), contextual SVM classifier (e), and non-contextual SVM classifier (f).	108
5.1	“Feltwell”: color composition of the image channels (a) and classification maps generated by the proposed method (b) and by the technique in [121] (c).	122
5.2	“Alessandria”: RGB false-color composition of the Landsat-5 TM channels (a) and classification maps generated by the proposed method (b) and by the technique in [121] (c).	123
6.1	Schematic of the classical model-based FDI strategy for FC plants based on parity equations with output errors [69], residual binarization, and FSM.	128
6.2	Schematic of the data-driven FDI strategy applied to FC plants.	129
6.3	Schematic of the hybrid FDI strategy for FC plants, where residuals generated by the parity equations are processed by an adequately trained statistical classifier.	130
6.4	Replacement of the real FC plant and related sensors (see Figures 1 to 3) by a plant model, which simulates healthy and faulty conditions and whose output variables are subject to random errors.	130
6.5	Schematic of the SOFC plant.	132

6.6	Probabilities of detection and false alarm of a classifier using the five easy-to-measure residuals and trained with a mixture of error magnitudes: constant-voltage case (diamonds and dashed line) and constant-current case (squares and dotted line). The probability pair is estimated for the 2%, 4%, and 6% error magnitudes.	146
6.7	Binary vectors, from $[0\ 0\ 0\ 0\ 0]$ to $[1\ 1\ 1\ 1\ 1]$, produced by the model-based FDI system (see Figure 6.1) for the SOFC plant operating under healthy and faulty conditions. For a given fault, each asterisk indicates the binary vector produced by a given combination between the operating condition and fault size. The constant-voltage control for the SOFC plant is considered, the maximum percentage error is 4%, and the relative threshold Ω is 4.5%	152

List of Tables

2.1	Indian Pines: Number of Training and Test Samples.	34
2.2	Houston: Number of Training and Test Samples.	35
2.3	Pavia University: Number of Training and Test Samples. . . .	38
2.4	Indian Pines - Classification accuracies [%] obtained by the MSVC approach, by a non-contextual SVM, and by an MRF-based classifier whose unary term is derived through the method in [162]. Graph cuts (GC), TRW-S, and LBP are used to minimize the energies of both Markovian approaches. .	55
2.5	Pavia University - Classification accuracies [%] obtained by the MSVC approach, by a non-contextual SVM, and by an MRF-based classifier whose unary term is derived through the method in [162]. Graph cuts, TRW-S, and LBP are used to minimize the energies of both Markovian approaches.	56
2.6	Houston - Classification accuracies [%] obtained by the MSVC approach, by a non-contextual SVM, and by an MRF-based classifier whose unary term is derived through the method in [162]. Graph cuts, TRW-S, and LBP are used to minimize the energies of both Markovian approaches.	57
2.7	Port-au-Prince - Classification accuracies [%] on the test set obtained by MSVC-GC method applied to optical/SAR, only optical and only SAR image.	58
2.8	Alessandria data set: confusion matrix and classification accuracies of an SVM classifier based on the proposed region-based kernel applied using input multiscale segmentation.	58

List of Tables

2.9	Confusion matrix obtained on the test set of the Itaipu data set by SVM classifiers using the region-based kernel.	59
2.10	Confusion matrix obtained on the test set of the Itaipu data set by a composite kernel combining the region-based and the pixelwise RBF kernels.	59
2.11	Confusion matrix obtained on the test set of the Itaipu data set by the purely pixelwise RBF kernel.	60
3.1	“Itaipu”: numbers of training and test samples on the PAN lattice; classification accuracies on the test set and computation times of the proposed method, of the technique in [135] and of a single-resolution graph-cut-based Markovian classifier.	83
3.2	The results of McNemar’s significance test comparing the classification results in Tables 3.1, 3.3, and 3.4. “SR” denotes the single-resolution classification of the “Itaipu” data set. “PS + SR” denotes the pansharpening and single-resolution classification of the “Alessandria” and “Pavia” data sets. For each data set, the (i, j) entry of the corresponding table denotes the value of McNemar’s statistic Z_{ij} for method i and method j	85
3.3	“Alessandria”: number of training and test samples on the PAN lattice; classification accuracies on the test set and computation times of the proposed method, of the previous technique in [135], and of the application of a single-resolution graph cut classifier after pansharpening.	87
3.4	“Pavia”: number of training samples on the PAN lattice; classification accuracies on the test set and computation times of the proposed method, of the previous technique in [135], and of the application of a single-resolution graph cut classifier after pansharpening.	91

List of Tables

3.5	Behaviors of the OA, AA, and κ of the proposed method and of the technique in [135] after subsampling, for each data set, the size of the set of training samples of each class down to approximately the size of the smallest class. Means and standard deviations over 10 runs are shown.	95
4.1	Confusion matrix of the MRF-based Gaussian classifier on the test set, in the application to 7 classes.	110
4.2	Confusion matrix of the MRF-based Gaussian classifier on the test set, in the application to 4 classes.	110
4.3	Classification accuracies of the considered classifiers on the test set.	111
5.1	Classification accuracy on the test sets (OA = overall accuracy; AA = average accuracy), execution time, and number of violated inequalities (NVI) obtained by applying the proposed method and the previous HK-MSE method in [121]	120
5.2	Classification accuracies and comparison of the execution times of the proposed method and the Goldfarb-Idnani approach (OA = overall accuracy; AA = average accuracy	121
6.1	Monitored variables used for the FDI in the considered SOFC plant.	134
6.2	Performance of the optimal classifier for a given control strategy and error magnitude.	143
6.3	Feature selection results and performance of the classifier trained with a mixture of error magnitude.	145
6.4	FSM of the model-based FDI system for the SOFC plant operating in constant-voltage conditions. The binary digits from R_1 to R_5 represent the residuals of the variables listed in Table 6.1 after the threshold operation.	153

6.5	OA and PA obtained by the hybrid FDI system when the five monitored variables of Table 6.1 are used, when the maximum temperature gradient (MTG) inside the SOFC stack is added, and when the cathodic activation losses (CAL) inside the SOFC stack are additionally introduced.	156
-----	---	-----

Introduction

The human mind continuously implements complex mechanisms to carry out everyday task. Recognizing the face of a schoolmate, identifying a friend's voice in a crowded place, discerning between rotten food and fresh, are all examples of activities that are apparently very simple and do not require considerable efforts to a human. While we are driving a car, our brain processes a large amount of information and data of different nature coming from different sensors such as ears or eyes. These data are used in conjunction with a priori knowledge on the highway code and on mechanical principles in order to decide the correct moment to change gear, to respect the requirement to give way to other cars at a roundabout, or even to choose a proper space to park (possibly avoiding fines). In general terms, these are decision-making processes that include mechanisms such as recognition, interpretation and classification which are often difficult to distinguish and separate. The pattern recognition and signal processing disciplines jointly provide tools for analyzing, modeling and understanding these processes in order to make them implementable by computing machines in different application domains. Quite often the actions mentioned above are performed unconsciously, but we can realize the complexity of each process when the goal is to obtain the same behavior replicated by computers. The autonomous guidance systems, for instance, represents a topical example. Joint venture of leader companies in the pattern recognition-artificial intelligence field and automotive industries have invested huge resources in this area but significant progress have been achieved only recently. Indeed, many of the simple actions we perform come naturally to all of us, but they can be considered as the result of experience and partly even of the evolution of the human being. The "learning" task is the main issue in all these activities, and from

a methodological view point, it represents the key aspect in the development of a pattern recognition approach. This thesis is framed within the pattern recognition area and focuses on the development and the experimental validation of novel methodologies to computationally address the analysis of image and measurement data. In particular, among the numerous applications of pattern recognition, this thesis deals with the development of methods for the two following themes:

- Remote sensing image classification, with a special focus on the challenging case of very high spatial resolution (VHR) images;
- Fault detection and isolation (FDI) for power generation plants based on solid oxide fuel cells (SOFCs), i.e. an electric generation technology characterized by remarkable energy conversion efficiency and environmental compatibility but also by critical sensitivity to faults and damages.

From a methodological perspective, although these two applications are substantially unrelated, the proposed techniques are remarkably homogeneous as they are all rooted in the common methodological areas of stochastic processes, probabilistic modeling, kernel machines, and Markov models. From an application-oriented viewpoint, both addressed themes are highly topical, owing both to the increasing prominence of VHR imagery from satellite, aircraft, and drones, and to the relevance of fuel cell technologies in the sustainable energy field.

In particular, in the first application domain, the current availability of large volumes of data acquired by multiple spaceborne mission for Earth observation (EO), along with the technological progress in the image acquisition systems, claim for the development of effective approaches for the classification of such images. The identification and discrimination of pixel categories (or possibly objects) in the considered scene make the information content of the images clearer and more usable by the users. Moreover, especially with last-generation VHR images, the correlation between neighboring pixels and the information related to spatial-geometrical structures become of paramount importance in order to develop effective approaches. In such a context, most of the methodologies proposed in this manuscript are rooted

in the theory of probabilistic graphical modeling through Markov random fields (MRFs), which have proven to be effective tools able to incorporate contextual and multisource information in the classification process.

The second application field addressed in this thesis concerns the development of FDI strategies for SOFC-based power generation plants. A Fuel cell (FC) is an electrochemical reactor that converts the chemical energy into electricity through a reaction with oxygen. Specifically, SOFCs are currently considered the preferred option for distributed power generation. Although this type of FCs exhibit high energy conversion efficiency and environmental compatibilities, plants based on stacks of SOFCs still suffer from low reliability and limited lifetime. Thus, the development of specific diagnosis strategy is of very high importance for their commercial diffusion. These monitoring problems are formalized in this manuscript within a pattern recognition perspective by exploiting statistical classifiers in conjunction with a quantitative physicochemical model of the fuel cells. In this general framework, the next paragraphs provide a brief summary of the key ideas of the proposed methods along with the description of the content of each chapter of the thesis.

Chapter 1 The first chapter presents an introduction to the pattern recognition field with particular attention to supervised classification methodologies and a special focus on Markov models and support vector kernel-based approaches. Most of the novel methods developed in the thesis take advantage from the theories of these powerful approaches in the application to two aforementioned applicative fields. Given the key ideas and terminology recalled in Chapter 1, starting from Chapter 2, one or more novel approaches are presented in each chapter, each preceded by a specific introduction that provides an overview on the related context.

Chapter 2 In this chapter the problem of spatial-contextual image classification is addressed in general terms and with a special focus on the topical case of VHR remote sensing images. Two novel classification frameworks based on the rigorous integration of kernel machines, MRFs, last-generation energy minimization methods, and region-based analysis are proposed. This integration is methodologically explained by the aim to combine the robust-

ness to dimensionality issues and the non-parametric fusion capability of kernel machines together with the effectiveness of Markov models and of region-based approaches in characterizing the spatial-contextual information associated with an image. The first approach rigorously integrates MRF and support vector machines (SVM) within a unique kernel-based framework. The related energy minimization task is tackled through graph cut algorithms and belief propagation-type methods. The second approach proposed here defines a novel kernel function whose entries are two segments rather than two pixelwise samples as the inner product between probability density functions of two suitable Gaussian processes in the L^2 Hilbert space. This function is an admissible Mercer kernel by construction, and can be incorporated into an arbitrary kernel machine, including of course SVM classifiers. The experimental validation of these methods is carried out with different remote sensing data, including but not restricted to aerial hyperspectral and satellite VHR imagery.

Resulting publications:

- A. De Giorgi, G. Moser, S.B. Serpico, Contextual Remote-Sensing Image Classification through Support Vector Machines, Markov Random Fields and Graph Cuts, *Proc. of IEEE International Geoscience and Remote Sensing Symposium 2014 / 35th Canadian Symposium on Remote Sensing*, Qubec, Canada, July 2014, pp. 3722-3725.
- P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarabalka, G. Moser, A. De Giorgi, S. B. Serpico, J. Benediktsson, Frontiers in spectral-spatial classification of hyperspectral images, submitted to *IEEE Geoscience and Remote Sensing Magazine*.

Chapter 3 The problem of the supervised classification of multiresolution images, such as those composed of a higher resolution panchromatic channel and of several coarser resolution multispectral channels, is addressed in this chapter. Similar to the previous chapter, the concern is again on the analysis of VHR imagery. Differently from the previous chapter, here, the joint analysis data with multiple spatial resolutions is addressed. A novel contextual method based on MRFs, Gaussian processes, and graph cuts is proposed.

The method generates a contextual classification map at the highest spatial resolution available in the input data set. The estimation of the parameters of the method is performed by extending recently proposed techniques based on the expectation-maximization (EM) and Ho-Kashyap algorithms. The convergence properties of EM for this specific parametric model are also analytically investigated. Experiments on semi-simulated and real data involving both IKONOS and Landsat-7 ETM+ images are presented.

Resulting publications:

- G. Moser, A. De Giorgi, S.B. Serpico, Multiresolution supervised classification of panchromatic and multispectral images by Markov random fields and graph cuts, *IEEE Transactions on Geoscience and Remote Sensing*, vol 54, no. 9, pp 5054-5070, 2016.

Chapter 4 This chapter focuses again of supervised image classification, but unlike the two previous chapters, here, the concern is on the role of a new and topical type of remote sensing data. The chapter presents a study on hyperspectral images collected in the thermal infrared (long-wave infrared, LWIR) range, a kind of data which is attracting increasing attention because of their appealing properties of insensitivity to Sun illumination, reduced sensitivity to smoke and mist as compared to data from visible sensors, and complementarity with respect to imagery collected in the visible and near infrared (VNIR) ranges. Here, the aim is to experimentally investigate the potential of supervised classification approaches that has been found effective for VNIR hyperspectral data for land cover mapping from VHR aerial thermal hyperspectral images. The considered methods include both non-contextual and spatial-contextual classifiers, and encompass methodological approaches based on Bayesian decision theory, MRFs, multiscale region-based analysis, and Bayesian feature reduction. Experiments were conducted with a challenging data set associated with a complex urban and vegetated scene.

Resulting publications:

- F. Barisone, D. Solarna, A. De Giorgi, G. Moser, S.B. Serpico, Supervised classification of thermal infrared hyperspectral images through

bayesian, markovian, and region-based approaches, Proc. of *IEEE International Geoscience and Remote Sensing Symposium 2016*, Beijing, China, July 2016, pp. 937-940.

Chapter 5 MRF models of various kinds are involved in the techniques developed and validated within all previous chapters. Indeed, MRF models for classification usually include that generally impact on classification accuracy. Therefore, their automatic optimization is a relevant problem in itself, and is still largely an open issue especially in the supervised case. For this purpose, an automatic parameter optimization method for MRF models is presented in this chapter. The technique combines a least mean square error approach to estimation and the sequential minimal optimization (SMO) quadratic programming algorithm. The method is tested with five data sets, which differ in type of sensor, spectral and spatial resolution, number of classes, and number of channels and is compared with two previous techniques.

Resulting publications:

- A. De Giorgi, G. Moser, S.B. Serpico, Parameter optimization for Markov random field models for remote sensing image classification through sequential minimal optimization, Proc. of *IEEE International Geoscience and Remote Sensing Symposium 2015*, Milan, Italy, July 2015, pp. 2346-2349.

Chapter 6 In this chapter FDI methodologies are developed with the aim to mitigate the reliability issues that characterize power generation plants based on SOFC. A quantitative model of the SOFC-based plant, which is able to simulate regular and faulty conditions, is used to predict the values of several physicochemical variables that characterize the behavior of the non-faulty system under different operating conditions. Specifically, two different approaches are proposed within a hybrid FDI strategy realized by combining a model-based scheme with a statistical classifier. The first technique is developed to jointly perform feature and model selection in the context of SVM classification applied as a diagnosis strategy. In the second proposed technique a hybrid approach based on the random forests (RFs) classification method is introduced. RF is adopted here because it represents an effective

approach to supervised classification. Although it is able to obtain classification accuracies comparable to those of other non-parametric techniques, it offers some advantages that are particularly relevant in FDI systems for FC-based plants. The performance comparison with pure data-driven and model-based approaches is also presented.

Resulting publications:

- L. Pellaco, P. Costamagna, A. De Giorgi, A. Greco, L. Magistri, G. Moser, A. Trucco, Fault diagnosis in fuel cell systems using quantitative models and support vector machines, *Electronics Letters*, vol. 50, no. 11 pp. 824-826, 2014.
- G. Moser, P. Costamagna, A. De Giorgi, A. Greco, L. Magistri, L. Pellaco, A. Trucco, Joint Feature and Model Selection for SVM Fault Diagnosis in Solid Oxide Fuel Cell Systems,” *Mathematical Problems in Engineering*, vol. 2015, 2015.
- P. Costamagna, A. De Giorgi, L. Magistri, G. Moser, L. Pellaco, A. Trucco, A classification approach for model-based fault diagnosis in power generation systems based on solid oxide fuel cells, *IEEE Transactions on Energy Conversion*, vol.31, no. 2, pp 676-687, 2015.
- G. Moser, P. Costamagna, A. De Giorgi, L. Pellaco, A. Trucco, S. B. Serpico, Kernel-based learning for fault detection and identification in fuel cell systems, in *Handbook of Pattern Recognition and Computer Vision*, 5th Edition, editor: C.-H- Chen, World Scientific Publishing, January 2016.
- P. Costamagna, A. De Giorgi, A. Gotelli, L. Magistri, G. Moser, E. Sciacaluga, A. Trucco, Fault diagnosis strategies for SOFC-based power generation plants, *Sensors*, vol 16, no. 8, 2016.

Conclusion The conclusions on the individual proposed methods, already drawn in each chapter separately, are summarized here along with comments on future possible extensions.

Chapter 1

Pattern recognition principles

Pattern recognition and machine learning are the disciplines aimed at the development of computational methodologies for the classification, recognition, and interpretation of data [45]. This concept has its roots in the study of the learning process of human beings and their capability to distinguish among objects and take decisions on the basis of previous observations and experiences [31]. The aim of this discipline is indeed to formalize how to perform an automatic or semi-automatic learning procedure through the use of suitable algorithms. Both phrases pattern recognition and machine learning correspond to this goal and discipline, although they historically stem from different methodological communities, i.e., signal processing and computer science, respectively. A pattern is a description of an object in term of attributes and features representative of the object itself. A recognition task consists of bringing an object back to a known model, namely a membership class where elements share similarities and common attributes. This classification problem is among the most typical ones in the pattern recognition field.

As illustrated in Figure 1.1, the pattern recognition process can be summarized through the following stages:

- Data acquisition, i.e., the measurement of variables which characterize a physical phenomenon, which are obtained by different types of sensors;
- Pre-processing, i.e., a step that includes operations of filtering, normal-

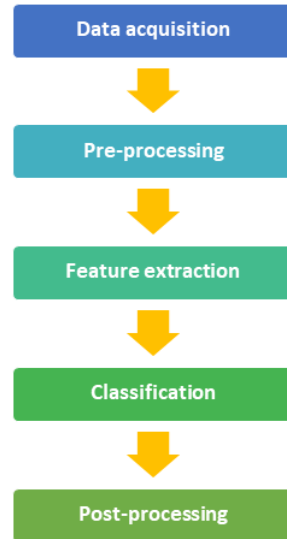


Figure 1.1: Pattern recognition process

ization, transformation, and selection, aimed to ease the classification process;

- Feature extraction, aimed to obtain a proper representation of the dataset and reduce the data computational complexity and the impact of possible curse-of-dimensionality issues. Each sample of the dataset is associated with a feature vector that is representative of the original data. The features should represent an accurate characterization of a sample belonging to a specific category and maximize the difference between distinct classes;
- Classification, which is the identification of the membership class of each sample of the dataset on the basis of learning models and features. The main goal of this task is to give an interpretation to raw data in order to extract relevant informations for decision making in specific applications;
- Post-processing, aimed at performance evaluation in terms of accuracy

1.1. Supervised classification

and error rate of the obtained results. A feedback can be used to modify the system modules in order to improve performances.

Each pattern is represented as a set of values collected in a d dimensional *vector of features* \mathbf{x} where the elements x_1, x_2, \dots, x_d identify the coordinates of a point in the feature space. Samples belonging to the same membership class usually show similar values. The number of features should be chosen in order to avoid the *curse of dimensionality*, more precisely represented by the *Hughes phenomenon* in a classification process [65]: the linear increase of the feature space dimension requires a corresponding (and possibly quadratic or exponential) growth of the amount of data to be used for training.

Classification methods can be divided in two main categories according to the learning approach: supervised and unsupervised [45]. Supervised classification needs prior information about the classes and assign each data to a suitable membership class by processing the associated feature vectors and a set of labeled samples. Unsupervised classification groups data into natural classes (*cluster*) on the basis of similarities of their feature vectors. In this thesis, the focus is on supervised approaches.

1.1 Supervised classification

1.1.1 Key ideas and terminology

Supervised classification represents a process of (semi)automatic learning that makes use of prior knowledge by the analysis about data features and class labels. The learning phase produces a function that is used to compare unknown data and assign them a suitable class label.

Given,

- a column vector of d features $\mathbf{x} = [x_1, x_2, \dots, x_d]^t$;
- a set of information classes $\Omega = \{\omega_1, \omega_2, \dots, \omega_M\}$

the supervised classification process assigns a class label \hat{y} , belonging to Ω to each sample \mathbf{x} . Each class is associated with a region in the feature space

1.1. Supervised classification

which represents the d -dimensional subset corresponding to the values that \mathbf{x} can assume.

In the literature, many algorithms have been proposed in order to address this problem. All of them have in common the following stages:

1. *Training phase:*

- establish the set of information classes able to describe the composition of a dataset;
- choose a training set, i.e. a set of the most representative vectors of each class on the basis of the knowledge of the considered phenomenon. This set should be statistically complete, thus it should be as representative as possible of all the classes in order to obtain a reliable estimation of the classifier parameters;
- optimize the classifier parameters according to the training set. The output of this step is the trained classifier.

2. *Test phase:*

- choose a test set, i.e. a set of labeled samples disjoint from the training set and to be used to verify the accuracy of the results;
- evaluate the performance of the classifier using the test set, before the application to unknown target data.

Each sample of a dataset can be represented in a d -dimensional feature space in which a subregion assigned to each class ω_i can be identified. This subregion is called decision region of ω_i ($i = 1, 2, \dots, M$). The classification process leads to obtain a set of discriminant functions f_{ij} in the feature space between pairs of classes ω_i and ω_j . The discriminant functions are used to label an unknown sample \mathbf{x}^* according to the following rule, with two classes and more than two classes respectively:

$$\begin{cases} f_{12}(\mathbf{x}^*) > 0 \rightarrow \mathbf{x}^* \in \omega_1 \\ f_{12}(\mathbf{x}^*) < 0 \rightarrow \mathbf{x}^* \in \omega_2 \end{cases} \quad (1.1)$$

$$f_{ij}(\mathbf{x}^*) > 0 \quad \forall i \neq j \rightarrow \mathbf{x}^* \in \omega_i \quad (1.2)$$

1.1. Supervised classification

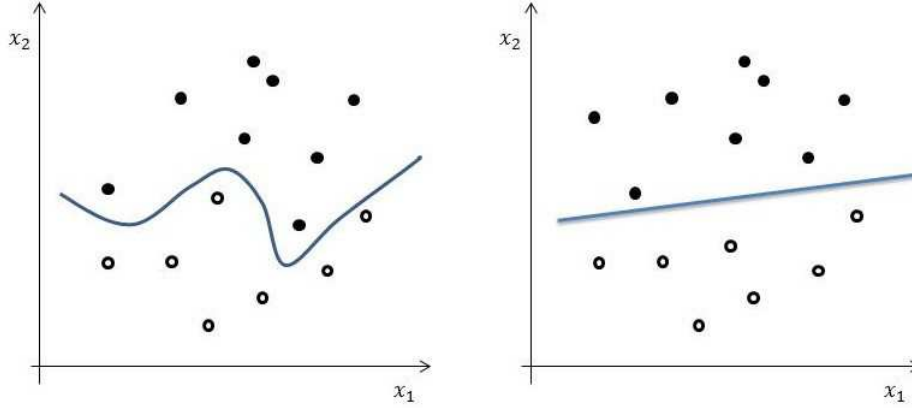


Figure 1.2: Hyperplane in bi-dimensional space: non-linear separability (left), linear separability (right).

In the multiclass case the problem is decomposed into a set of simpler binary decision problems by identifying a discrimination function for each couple of classes and by composing the results (e.g. voting process). The hypersurfaces that separate the decision regions of different classes are named decision boundaries.

Given a training set, a class is defined *separable* if there exist decision boundaries (hypersurfaces) such that all and only the samples belonging to that class are located on the same side of those boundaries. When two classes are linearly separable, this hypersurface can be chosen as a hyperplane (Figure 1.2). Except when there are two training samples that belong to two distinct classes, and yet, they coincide, a suitably complex hypersurface can always be found to separate classes. However the problem of this search could not have a straightforward solution and the possibly excessive complexity of the resulting hypersurface might affect the classifier performances.

Supervised classification algorithms can be parametric or non-parametric [45]. The difference between the two approaches is that the parametric one makes assumptions about the distribution of the feature vectors associated with the classes and estimates the related parameters from the training set. A rather frequent assumption is the normal distribution of each information class. However, in many cases this assumption is not respected and re-

1.1. Supervised classification

sults can be limited. Thus, because of the difficulty of identifying a suitable model that fits data, it is often preferable to use non-parametric methods. Non-parametric algorithms do not assume any predefined parametric model for the class statistics. In particular, Bayesian non-parametric approaches estimate the distribution of the classes through the training set directly. They have remarkable flexibility, however the complexity of the distribution function estimation is sometimes even more than that of the classification problem itself. Non-Bayesian non-parametric approaches are designed in order to exploit the information associated with the training samples to derive decision boundaries and rules directly, without passing through any model for the class statistics. A basic example of the latter category is represented by Nearest Neighbor methods, in which an unknown sample is assigned to the most common class among the nearest samples in the feature space. Other important examples include approaches to decision boundary construction that are based on the idea that the most informative samples are located near the border between two classes, e.g. *support vector machines* [36]. This last approach will be of great importance in several chapters of this thesis and its key ideas will be reviewed in the next section. Here, we only recall that other major non-parametric approaches stem from the theories of neural networks, ensemble learning, fuzzy logic and set theory.

1.1.2 Support vector machines

Support vector machines (SVM) represent a family of automatic learning methods developed since the 60's by the research group led by Vladimir Vapnik at the AT&T Bell in New Jersey [152]. The SVMs are framed within the theory of statistical learning and consist of a supervised learning approach for the resolution of classification problems. They have been also studied for regression [36]. The success obtained by the SVMs for several decades is mainly due to their generalization properties with data not belonging to the training set and to their robustness to the Hughes phenomenon.

In general terms, the SVM approach consists in the identification of a hyperplane through the resolution of a quadratic programming problem and generates a solution based on the samples located at the interface between

1.1. Supervised classification

two classes (*support vectors*). Therefore, the learning phase identifies a subset of the training set whose samples that are the most informative from the viewpoint of separating the classes.

Initially, the SVM were defined for the resolution of problems with linearly separable classes. They were subsequently extended to non-separable and non-linear cases, making it possible to apply them in much more general conditions of real interest. The extension of the SVMs to non-linearly separable classes is made possible thanks to the definition of functions, known as *kernel functions*, which are able to implicitly transform the feature space into a higher dimensional space in which the classes become linearly separable. Indeed, with reference to the Cover theorem [35], a non-linear transformation of data into a higher dimensional space increases the chances to obtain a linear separation between two classes.

Hence, the key ideas of the SVM approach consist in:

- mapping input data in a higher dimensional space \mathcal{H} through a suitable kernel function;
- identifying a hypersurface able to separate samples belonging to non-linearly separable classes with the maximum margin.

Let us first consider first the binary classification case and a training set of ℓ samples \mathbf{x}_i ($i = 1, 2, \dots, \ell$) composed of d features with the associated class label y_i that takes on the value $+1$ or -1 depending on the membership of either class. An SVM classifier assigns a suitable class label to an unknown sample $\mathbf{x}^* \in \mathbb{R}^d$ as $\text{sgn}f(\mathbf{x}^*)$, where the discriminant function $f(\cdot)$ is the following kernel expansion:

$$f(\mathbf{x}^*) = \sum_{i=1}^{\ell} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}^*) + b \quad (1.3)$$

and where $K(\cdot, \cdot)$ is a kernel function and α_i are coefficients to compute. A function K of two vectors is a kernel if it is equivalent to the evaluation of an inner-product in some non-linearly transformed space. Let us define $\Phi(\cdot)$ as a function mapping the feature space (or a compact subset of the feature space) to a vector space \mathcal{H} endowed with an inner product. Thus the

1.1. Supervised classification

kernel function can be expressed as $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$. Note that, \mathcal{H} is never computationally involved because all calculations only use the kernel K .

The coefficients α_i ($i = 1, 2, \dots, \ell$) are determined by solving the following quadratic programming (QP) problem [152]:

$$\begin{cases} \min_{\boldsymbol{\alpha}} \left(\frac{1}{2} \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha} \right) \\ \mathbf{y}^T \boldsymbol{\alpha} = 0 \\ 0 \leq \alpha_i \leq C \\ i = 1, 2, \dots, \ell \end{cases} \quad (1.4)$$

the bias b is derived as a by-product of this solution, Q is the $\ell \times \ell$ matrix whose (i, j) th entry is $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{1}$ is an ℓ -dimensional vector with unitary components, \mathbf{y} is the vector of the labels of the training samples, and C is a parameter. Support vectors are associated with $\alpha_i > 0$.

It can be proven that equation 1.3 is equivalent to a linear discriminant function in the transformed space \mathcal{H} :

$$f(\mathbf{x}^*) = \langle w, \Phi(\mathbf{x}) \rangle + b \quad (1.5)$$

where $w \in \mathcal{H}$, the bias $b \in \mathbb{R}$, and the set of slack variables $\{\xi_i\}_{i=1}^{\ell}$, which determine the error of the training samples classification, solve the following minimization problem [36]:

$$\begin{cases} \min_{w, \boldsymbol{\xi}, b} \left(\frac{1}{2} \langle w, w \rangle + C \mathbf{1}^T \boldsymbol{\xi} \right) \\ y_i (\langle w, \Phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 2, \dots, \ell \end{cases} \quad (1.6)$$

The C parameter tunes the tradeoff between the terms $\boldsymbol{\xi}$, which favors fitting the discriminant function with the available training set, and $\langle w, w \rangle$, which conversely favors minimizing the error expectation on an unknown sample with a consequent overfitting minimization.

Generalization to $M > 2$ classes is usually achieved by decomposing the multiclass problem into a collection of binary subproblems [97, 152]. A popular approach is one-against-one (OAO), which usually represents a good

1.2. Unsupervised classification

tradeoff between accuracy and computational burden. The key idea is to determine a binary discriminant function for each couple of classes, apply each function to an unknown sample and assign it to the most voted class among the results obtained with each binary discriminant function. More details can be found in [152].

1.2 Unsupervised classification

Unsupervised classification is based on the identification of similarities between samples of a dataset without any prior information about the classes [115]. Indeed, sometimes the number and the characteristics of the classes are not clearly identifiable especially with highly heterogeneous data. The idea behind unsupervised classification is grouping samples into a number of natural classes by the discovery of hidden structures and similarities in the dataset. This approach is known as *clustering*. The classification process leads to identify the most representative clusters in the feature space and assigns a symbolic label to each cluster. At a later stage, an analyst may provide a suitable interpretation to each cluster.

The unsupervised classification process can be summarized in the following phases:

1. Classification parameter selection, in order to balance the need of selective variables for a correct grouping of samples and the avoidance of limited informative clusters;
2. Selection of a distance or similarity criterion, e.g. Euclidean, Minkowsky, or Chebychev distances, kernel-based similarity metrics, graphical models encoding similarities among pairs of samples, etc.;
3. Selection of the clustering algorithm, e.g. hierarchical models, in which data are grouped into nested subsets associated with different levels of similarities, or partitional models, in which the clusters are not arranged as a hierarchy and are often searched for altogether.

Unsupervised classification methods are sometimes more complex than those of supervised classification, although clustering algorithms are generally

suited to knowledge discovery tasks [51].

1.3 Contextual and non-contextual classification

1.3.1 Key ideas and terminology

When a classifier is meant to be applied to an image, in order to label each pixel with a certain class membership, an important role is represented by methods able to include in the decision process both the information of the feature vector of each pixel (pixelwise information) and the information given by the neighboring sample (spatial information). This kind of approach is known as contextual classification [146]. It can be seen as a special case of structured output learning, i.e., the area of machine learning that focuses on the estimation of unknown data with a certain dependence structure (spatial, temporal, multiscale, semantic, etc., dependences). The spatial-contextual approach plays a major role in the application to remote sensing data because it is usually effective especially with very high resolution (VHR) images in which it is appropriate to consider both the spatial heterogeneity and the geometric properties associated with the input data [95,98]. Indeed, although a non-contextual approach is more computationally tractable problem, it may produce a salt and pepper effect on this type of images.

On a preliminary note, we recall that, in this thesis, the phrase image classification will always be meant as the task of labeling each individual pixel with a class membership. This is consistent with the common terminology that has been used in the remote sensing literature for a few decades. In the recent computer vision literature, this task is usually named semantic segmentation (more precisely dense semantic segmentation), whereas image classification refers to the labeling of an entire image with a unique class membership (e.g., in data mining or face recognition applications). We also recall that, in the multispectral remote-sensing literature, spatial-spectral classification and purely spectral classification are often used instead of contextual classification and non-contextual classification to acknowledge that

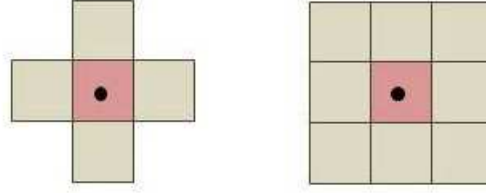


Figure 1.3: Examples of neighboring systems

pixelwise information in a multispectral image generally includes the spectral observations taken on each pixel (spectral channels or bands).

The idea behind the use of spatial information is that neighboring pixels in a real remotely sensed image are strongly correlated [110,115]. Contextual classification deals with spatial information by including pixels in a suitable neighboring system (Figure 1.3).

In the literature the main proposed approaches to contextual classification are represented by models based on *Markov random fields* (MRF), texture analysis and region-based methods.

MRFs are an extension of Markov chain to the two-dimensional case [76] and allow to equivalently express a global model of contextual information using only local relationships between neighboring pixels. Details about MRFs will be discussed in section 1.3.2. Given a region in an input image, texture analysis identifies a repetitive spatial distribution in pixel intensities. This kind of information can be included in the classification process through the extraction of features related to spatial relations between neighboring pixels. Among the approaches proposed in the literature many of them are based on moving window methods, in which texture features of a pixel are computed on samples included in a square or rectangular window [110]. A typical drawback is the creation of artifacts at the edges between different regions in an image with consequent effects on the effectiveness of the classification. The key idea of region-based approaches is the use of segmentation techniques in order to identify homogeneous areas [15] that are subsequently labeled as a

1.3. Contextual and non-contextual classification

whole by a suitable classifier.

1.3.2 Markov random fields

Markov random fields represent a family of effective tools in image processing because of their capability to integrate spatial information in the classification process [55].

Contextual image analysis has relevant application in different areas such as change detection, pixel labeling, edge detection, texture analysis, feature extraction. In these contexts, MRFs are able to formalize a model of the image in order to apply the Bayesian theory. In image classification the option of consider the pixels as independent and identically distributed entities is not a realistic hypothesis because pixel located in the same area may be highly correlated. Thus, the integration of pixelwise and spatial information allows to improve classification performances as compared to models that do not use contextual information. The performance improvement is due to error correction, removal of ambiguities and retrieval of incomplete information.

Potentially each pixel is dependent on all the pixels composing the image. However, a problem of Bayesian image analysis formalized with a global dependence would be computationally intractable. Furthermore, from a signal-processing perspective, the autocorrelation function of a real image generally decays with the increase in the spatial lag between pairs of pixels. Thanks to the Markovian property it is possible to limit the contextual information of a pixel to its neighborhood.

Let us define a neighborhood system N [84]. Each pixel of an image (*lattice*) is considered as an element (*site*) of a set S . Each site is associated with a neighborhood system expressed as a subset of S with the following properties:

- a site does not belong to its neighborhood;
- the neighborhood relation between two sites is mutual.

A *clique* is a subset of S whose sites are mutually neighbors. In Figure 1.4 first and second order neighborhoods are shown, and they are composed

1.3. Contextual and non-contextual classification

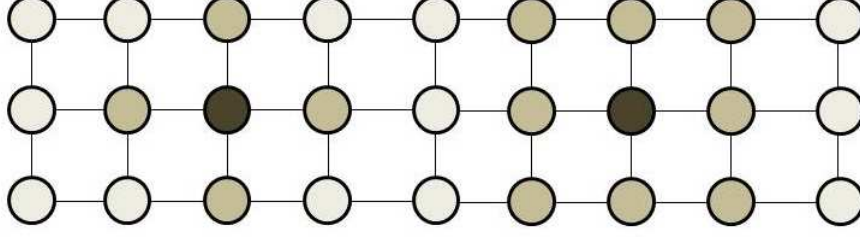


Figure 1.4: First and second order neighborhood systems (grey) of a site (black)

of 4 and 8 sites, respectively. The neighborhood can be extended to higher orders, allowing to model different levels of relationship between sites.

Let us now define a *Gibbs random field* [84]. Given S and N and a random field \mathbf{F} of a set of variables associated with the sites of the lattice, \mathbf{F} is a Gibbs random field if its $P(\mathbf{F})$ is:

$$P(\mathbf{F} = \mathbf{f}) = \frac{1}{Z} e^{-U(\mathbf{f})} \quad (1.7)$$

where Z is a normalization constant called *partition function* and $U(\mathbf{f})$ is an *energy function* expressed as follows:

$$U(\mathbf{f}) = \sum_{c \in C} V_c(\mathbf{f}) \quad (1.8)$$

and where $V_c(\mathbf{f})$ is the *potential function* associated with a clique belonging to the set C . It is important to stress that, in (1.8), the potential associated with each clique only depends on the sites belonging to that clique. Hence, although (1.8) provides a global model for the joint distribution of all random variables in the field, it also provides a decomposition in terms of local contributions on the cliques. The energy function represents the sum of the potential functions of all the clique associated with a neighborhood system.

A Gibbs random field considers the energy function in global terms giving a description of a whole lattice. The Markovian property allows to express

1.3. Contextual and non-contextual classification

the energy function in local terms related to a neighborhood system and can be written as follows:

$$P(f_i/f_{S-\{i\}}) = P(f_i/f_{\partial i}) \quad (1.9)$$

where ∂i is the neighborhood of the i -th site.

The Hammersley-Clifford theorem (1971) defines the equivalence between Gibbs and Markov random fields, as reported by Besag in [12]: if a Gibbs random field can be defined in terms of a clique in a neighborhood system then it is equivalent to a Markov random field and a unique MRF exists for the corresponding GRF and vice-versa. This result is of paramount importance because it relates a purely local Markovianity property with a global joint distribution model that can be plugged into a Bayesian analysis scheme.

Chapter 2

Spatial-contextual kernel-based classification of remote sensing images

2.1 Introduction

In this chapter the problem of spatial-contextual image classification is addressed in general terms through the development of two novel frameworks based on the integration of spatial-contextual information into kernel machines. Kernel-based SVMs have recently become very popular classifiers thanks to their accurate results in several applications, to their generalization capability, and to their robustness to the Hughes phenomenon that affects the application of classifiers to high-dimensional feature spaces [152]. Nevertheless, SVMs are intrinsically non-contextual and classify each pixel only on the basis of the corresponding spectral or radar observations and regardless of the neighboring pixels [152]. This issue represents a limitation (especially in the classification of VHR images) because it implies neglecting the spatial information associated with the input image. For this purpose, the attention in this chapter is focused on two alternative family of methods, namely MRFs probabilistic graphical models and region-based methods. As already mentioned in Chapter 1, MRFs have proven to be effective tools to describe and incorporate spatial information in the classification process [84].

2.1. Introduction

An MRF is a model for the spatial-contextual information and allows a Bayesian classification problem to be formulated in terms of the minimization of a suitable “energy function” [84]. Region-based methods rely on the combination of classification and segmentation algorithms [15] and are able to incorporate the spatial-geometrical information in the land cover estimation process. Many approaches have been proposed to address this problem, including region-growing algorithms, which progressively expand a collection of segments starting from a suitable set of seed points [110]; Markovian methods, which formulate segmentation as an unsupervised classification problem addressed through a Markov random field (MRF) model [41, 107, 126]; fuzzy-connectedness techniques, which formalize the uncertainty associated with image data through fuzzy theory [40, 149]; watershed methods, which represent the field of the pixel intensities as a topographic relief and search for the related catchment basins [123, 147]; and hierarchical algorithms, which generate a hierarchy of segmentation results associated with progressively increasing spatial detail [16, 39, 144]. A segmentation map intrinsically offers a characterization of the geometrical structure associated with the image. A simple intuitive region-based classifier could just label each segment as a whole rather than separately classifying each pixel. However, more sophisticated methods have been developed by properly combining segmentation, feature extraction, and possibly multiscale analysis [15, 140]. Especially appealing results can be obtained when the regions extracted through segmentation match semantic objects in the scene (hence, the popular name “object-based image analysis,” OBIA).

Specifically, the first proposed method is based on an extension of the approach in [38, 97] and provides a rigorous integration of SVMs, MRFs, and the last-generation energy minimization methods. Therefore, in this framework the two key aspects concern the minimization of the resulting Markovian energy function and the integration of SVMs and MRFs. The first one represents a challenging combinatorial problem which has been studied since the Eighties and is classically addressed using either time-expensive stochastic global minimizers or computationally cheap greedy local minimizers [13, 55]. Indeed, efficient energy minimization methods based on graph-theoretic concepts (graph cuts, GC, and belief propagation) have attracted increasing in-

2.1. Introduction

terest during the last decade. GC methods formulate energy minimization as a maximum flow problem on an appropriate graph associated with the image: in the case of binary classification, they allow converging to a global minimum in generally acceptable times; in the multiclass case, they are proven to reach a strong local minimum, which is characterized by good optimality properties [18, 19, 74]. Belief propagation-type methods formalize the intuitive idea of passing messages along the graph to decrease the energy. In particular, the max-product loopy belief propagation (LBP) technique operates on graphs with loops while the tree re-weighted message passing (TRW) method combines belief propagation with the construction of suitable spanning trees [48, 73, 156].

The second key aspect of merging SVMs and MRFs is not straightforward because the latter are framed within probabilistic Bayesian modeling, whereas the former are non-Bayesian learning machines. A common workaround is to postprocess the SVM discriminant function through the algorithms in [106, 162], which use parametric modeling, maximum-likelihood, and numerical analysis concepts to approximate pixelwise posteriors. The resulting probabilistic output is plugged into the unary energy. This approach is computationally efficient and has recently led to accurate results (e.g., [26, 57, 165]). However, it methodologically mixes i.i.d. and non-i.i.d. assumptions in the parameter estimation and MRF modeling stages, respectively. The method presented here aims at merging the analytical formulations of SVM and MRF and it is based on the approach proposed in [97]. Nevertheless, in this previous classifier the energy minimization task was performed through the greedy iterated conditional mode (ICM) algorithm that is proven to converge only to a local minimum and may exhibit critical dependence on the initialization [13]. In order to overcome these limitations through the application of an arbitrary energy minimization method, including GC, LBP, and TRW, the global posterior energy function requires a new explicit formulation and a general re-formalization of the integrated SVM-MRF framework becomes necessary.

This formulation has been developed, and specifically, under suitable assumptions, the global energy function related to the application of a Markovian classifier in the transformed space associated with a kernel has been

2.2. Markovian support vector classification

formalized in terms of a kernel expansion, similar to a traditional SVM, by using a Markovian kernel and an additional feature associated with the adopted MRF model. Parameters are automatically optimized through numerical techniques based on the minimum span bound and MSE criteria. The resulting classification is hereafter referred to as Markovian support vector classifier (MSVC).

Conversely, the second proposed method on an integration of region-based and kernel-based approaches. In particular, a novel region-based kernel function is proposed, that allows taking advantage of the correlation among neighboring pixels and of the geometrical structures associated with the image. This function can be incorporated into an arbitrary kernel machine and both single-channel and multispectral formulations can be formalized. A multiscale analysis can also be carried out using multiple segmentation results through the composite kernel method. Focusing on optical images, the methodological rationale behind the proposed kernel is to model the portion of image within each kernel as a set of samples drawn from a suitable Gaussian stochastic process, and to compute a kernel by combining the corresponding density functions through an integral inner product.

2.2 Markovian support vector classification

2.2.1 Proposed formulation of the energy function

Let \mathcal{I} be a regular pixel lattice associated with a remote-sensing image composed of d channels. Each pixel is characterized by a feature vector \mathbf{x}_i ($\mathbf{x}_i \in \mathbb{R}^d$) and a class label y_i . Let $\mathcal{S}^* \subset \mathcal{I}$ be the set of training pixels (i.e., y_i is *a-priori* known for $i \in \mathcal{S}^*$). We refer first to binary classification (the multiclass case will be addressed in Section 2.2.2), so two possible labels $y_i = 1$ or $y_i = -1$ can be assigned to the i -th pixel according to its class membership. Hence, a continuous-valued random field $\mathcal{X} = \{\mathbf{x}_i\}_{i \in \mathcal{I}}$ of feature vectors and a binary-valued random field $\mathcal{Y} = \{y_i\}_{i \in \mathcal{I}}$ of class labels are defined on the pixel lattice \mathcal{I} . In order to consider \mathcal{Y} as an MRF, a neighborhood system $\{\partial i\}_{i \in \mathcal{I}}$ is assumed to be defined on the lattice \mathcal{I} , where $\partial i \subset \mathcal{I}$ is the set of the neighbors of the i -th pixel [84]. In this MRF framework,

2.2. Markovian support vector classification

the “posterior global energy function” $U(\mathcal{Y}|\mathcal{X})$, associated with the adopted MRF model and neighborhood system, can be expressed as a sum of contributions associated with the pixels and conditioned to their feature vectors and neighboring labels:

$$U(\mathcal{Y}|\mathcal{X}) = \sum_{i \in \mathcal{I}} U_i(y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}), \quad (2.1)$$

where:

$$U_i(y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}) = g(\mathbf{x}_i|y_i) + \beta \mathcal{E}_i(y_i|\mathbf{y}_{\partial i}) \quad (2.2)$$

is composed of two contributions [19]. The first one is related to non-contextual pixelwise statistics, while the second one is a prior energy function \mathcal{E}_i that characterizes the adopted MRF model. The positive parameter β determines the influence of the spatial contribution on the overall energy function [19]. The model in (2.1) and (2.2) implicitly assumes that the MRF model assumed for the label field has non-zero potentials only for pairwise cliques, i.e.:

$$\mathcal{E}(y_i|\mathbf{y}_{\partial i}) = \sum_{j \in \partial i} V_{ij}(y_i, y_j), \quad (2.3)$$

where $V_{ij}(\cdot)$ is the potential associated with the pair (i, j) of pixels $(i, j \in \mathcal{I})$. Furthermore, the separation into a pixelwise term including features and labels and a spatial term including only labels reflects an implicit conditional independence assumption, i.e.:

$$p(\mathcal{X}|\mathcal{Y}) = \prod_{i \in \mathcal{I}} p(\mathbf{x}_i|y_i), \quad (2.4)$$

i.e., the global joint class-conditional density $p(\mathcal{X}|\mathcal{Y})$ of the feature vector field given the label field factors out in terms of the pixelwise class-conditional density $p(\mathbf{x}_i|y_i)$ on each site. This is a widely accepted assumption when MRF models are used for classification purposes and is used to favor analytical tractability [84]. As a common aspect in many Bayesian approaches, this assumption also results in decomposing the posterior energy (2.1) into a prior and likelihood contributions (see (2.2))

Several energy minimization techniques, including simulated annealing [55] and the aforementioned ICM, can be entirely expressed in terms of

2.2. Markovian support vector classification

the “local energy-difference function,” i.e. [84, 97]:

$$\Delta U_i(\mathbf{x}_i, \mathbf{y}_{\partial i}) = U_i(-1|\mathbf{x}_i, \mathbf{y}_{\partial i}) - U_i(1|\mathbf{x}_i, \mathbf{y}_{\partial i}). \quad (2.5)$$

The previous approach in [97] was based on a reformulation of ΔU_i in terms of an SVM kernel expansion. However, in order to make it possible to apply arbitrary and efficient energy minimization methods, including graph cuts, LBP, and TRW, a new explicit formulation is needed for the global posterior energy function. Specifically, this function can be written as:

$$\begin{aligned} U(\mathcal{Y}|\mathcal{X}) = & \sum_{i \in \mathcal{I}} \frac{U_i(y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}) + U_i(-y_i|\mathbf{x}_i, \mathbf{y}_{\partial i})}{2} + \\ & + \sum_{i \in \mathcal{I}} \frac{U_i(y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}) - U_i(-y_i|\mathbf{x}_i, \mathbf{y}_{\partial i})}{2}. \end{aligned} \quad (2.6)$$

Using $\Delta U_i(\mathbf{x}_i, \mathbf{y}_{\partial i})$, it is easily proven that:

$$U_i(-y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}) - U_i(y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}) = y_i \Delta U_i(\mathbf{x}_i, \mathbf{y}_{\partial i}). \quad (2.7)$$

Similarly:

$$\begin{aligned} & U_i(y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}) + U_i(-y_i|\mathbf{x}_i, \mathbf{y}_{\partial i}) = \\ & = [g(\mathbf{x}_i|1) + g(\mathbf{x}_i|-1)] + \beta[\mathcal{E}_i(1|\mathbf{y}_{\partial i}) + \mathcal{E}_i(-1|\mathbf{y}_{\partial i})], \end{aligned} \quad (2.8)$$

where the first bracketed term is obviously independent of \mathcal{Y} . For many popular choices of the function \mathcal{E}_i , such as, for instance, the well known Potts model, i.e. [84]:

$$\mathcal{E}_i(y_i|\mathbf{y}_{\partial i}) = - \sum_{j \in \partial i} \delta(y_i, y_j), \quad (2.9)$$

also the second bracketed term is a constant (i.e., it is independent of $\mathbf{y}_{\partial i}$). Assuming that the adopted MRF model satisfies this property, i.e., $\mathcal{E}_i(1|\mathbf{y}_{\partial i}) + \mathcal{E}_i(-1|\mathbf{y}_{\partial i})$ is a constant, and plugging Eqs. (2.7) and (2.8) into (2.6), we obtain:

$$U(\mathcal{Y}|\mathcal{X}) = -\frac{1}{2} \sum_{i \in \mathcal{I}} y_i \Delta U_i(\mathbf{x}_i, \mathbf{y}_{\partial i}) + G(\mathcal{X}), \quad (2.10)$$

where $G(\mathcal{X})$ depends on the feature vectors but not on the labels and has no effect on the minimization result.

2.2. Markovian support vector classification

Let K be a kernel function in \mathbb{R}^d . Then, there exists a mapping $\Phi : \mathbb{R}^d \rightarrow \mathcal{F}$ from the d -dimensional feature space to a transformed feature space \mathcal{F} , endowed with an inner product $\langle \cdot, \cdot \rangle$, such that $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ [152]. As proven in [97], under mild assumptions on the distribution of the transformed samples in \mathcal{F} , the local energy-difference function can be expressed as an SVM-like discriminant function associated with a special kernel, named Markovian kernel, i.e.:

$$\Delta U_i(\mathbf{x}_i, \mathbf{y}_{\partial i}) = \sum_{j \in \mathcal{S}} \alpha_j y_j K_{\text{MRF}}(\mathbf{x}_i, \varepsilon_i; \mathbf{x}_j, \varepsilon_j) + b, \quad (2.11)$$

where

$$K_{\text{MRF}}(\mathbf{x}_i, \varepsilon_i; \mathbf{x}_j, \varepsilon_j) = K(\mathbf{x}_i, \mathbf{x}_j) + \beta \varepsilon_i \varepsilon_j \quad (2.12)$$

is the Markovian kernel and ε_i is an additional feature defined through the adopted MRF model:

$$\varepsilon_i = \mathcal{E}_i(-1|\mathbf{y}_{\partial i}) - \mathcal{E}_i(1|\mathbf{y}_{\partial i}). \quad (2.13)$$

The set $\mathcal{S} \subset \mathcal{S}^*$ of support vectors and the weight coefficients α_j ($j \in \mathcal{S}$) are obtained through the training of an SVM with kernel K_{MRF} (further details can be found in [97]). Plugging Eqs. (2.11) and (2.12) into (2.10), we obtain the following expression for the global posterior energy function (up to additive and multiplicative terms constant with respect to \mathcal{Y}):

$$\begin{aligned} \tilde{U}(\mathcal{Y}|\mathcal{X}) &= - \sum_{i \in \mathcal{I}} y_i \left[\sum_{j \in \mathcal{S}} \alpha_j y_j K_{\text{MRF}}(\mathbf{x}_i, \varepsilon_i; \mathbf{x}_j, \varepsilon_j) + b \right] = \\ &= - \sum_{i \in \mathcal{I}} y_i \sum_{j \in \mathcal{S}} [\alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b] - \beta \sum_{i \in \mathcal{I}} y_i \varepsilon_i \sum_{j \in \mathcal{S}} \alpha_j y_j \varepsilon_j. \end{aligned} \quad (2.14)$$

Hence, the global posterior energy function corresponding to the application of a Markovian classifier in the transformed space \mathcal{F} associated with a kernel K can be expressed as a linear combination of two contributions related to either the pixelwise information conveyed by the feature vectors \mathbf{x}_i or the contextual information characterized through the additional feature ε_i ($i \in \mathcal{I}$). The MRF minimum energy rule in the space \mathcal{F} can be related to a kernel expansion, similar to a traditional SVM, by using an additional feature associated with the adopted MRF model and the Markovian kernel. This conclusion holds for an arbitrary choice of the kernel K , associated with an either finite- or infinite-dimensional space \mathcal{F} [97].

2.2.2 The proposed classification method

The proposed contextual classifier is iterative and is based on the application of an arbitrary energy minimization method to minimize the energy function in Eq. (2.14). In the initialization phase, a preliminary classification map is generated through a traditional non-contextual SVM, whose regularization and kernel parameters are collected in a vector $\boldsymbol{\theta}$. The optimization of $\boldsymbol{\theta}$ is obtained through the Powell-span-bound algorithm [97]. Then, since the set \mathcal{S} of support vectors and the related weights α_j ($j \in \mathcal{S}$) are obtained through the training of an SVM with kernel K_{MRF} , at each iteration of the method ε is updated on the basis of the current classification map, β is estimated through the Ho-Kashyap algorithm [97, 121], and $\tilde{U}(\mathcal{Y}|\mathcal{X})$ is minimized through the chosen energy minimization method. The method performs the following processing steps:

- Initialization phase:
 - a) Optimize $\boldsymbol{\theta}$ through the Powell-span-bound algorithm (see Section 2.2.3) and initialize β (e.g., $\beta = 1$).
 - b) Train a non-contextual SVM with the kernel K and the estimate of $\boldsymbol{\theta}$ obtained in step a), and generate a preliminary classification map by running the trained SVM.
- Iterative phase: for a fixed number of iterations, perform the following steps:
 1. Compute ε_i for each $i \in \mathcal{I}$ on the basis of the current classification map (see Eq. (2.13)).
 2. Estimate β by applying the Ho-Kashyap algorithm (see Section 2.2.3).
 3. Train an SVM with the Markovian kernel K_{MRF} , the contextual feature ε_i updated in step 1, and the estimates obtained for $\boldsymbol{\theta}$ and β in steps a) and 2, respectively.
 4. Compute $\tilde{U}(\mathcal{Y}|\mathcal{X})$ on the basis of the support vectors \mathcal{S} and the related weights α_j ($j \in \mathcal{S}$) obtained in step 3 (see Eq. (2.14)).

2.2. Markovian support vector classification

5. Minimize $\tilde{U}(\mathcal{Y}|\mathcal{X})$ with respect to \mathcal{Y} through the considered energy minimization method (see Section 2.2.3).

Depending on the selected minimization algorithm, convergence may not be guaranteed. For instance, adopting the GC approach, convergence is ensured for the application in step 5 [19]. However, convergence is not analytically guaranteed for the proposed iterative algorithm (although a good convergent behavior was remarked in all experiments). Thus, the iterative phase is run for a fixed number of iterations and the output classification map is the one with the minimum value of the global posterior energy within the maps generated in all iterations.

The proposed method, illustrated for the case of binary classification, can be extended to the multiclass case through a decomposition into multiple binary subproblems, as usual for SVM-type classifiers. Here, the “one-against-one” (OAO) approach is chosen because of its usual accuracy and limited sensitivity to unbalanced classes (especially as compared to alternate strategies, such as “one-against-all”). Details on the application of OAO can be found in [97].

2.2.3 Parameter estimation and energy minimization

The optimization of θ and β is automatic and does not require manual intervention. The smoothing parameter β regulates the tradeoff between the spatial and pixelwise energy components. Its optimization is automated by extending to the proposed approach the algorithm proposed in [122] and [97] and based on the Ho-Kashyap numerical procedure. For each $i \in \mathcal{S}^*$ (i.e., each support vector), the following inequality expresses a condition of correct classification of the i -th training sample:

$$\beta_1 y_i [U_i^P(-1) - U_i^P(1)] + \beta_2 y_i [U_i^C(-1) - U_i^C(1)] \geq 0, \quad (2.15)$$

where $\beta = \beta_1/\beta_2$ ($\beta_1 > 0, \beta_2 > 0$), and $U_i^P(\pm 1)$ and $U_i^C(\pm 1)$ are shorthand notation for the pixelwise and contextual energy terms associated in Eq. (2.14) with the i -th pixel and the labels $y_i = \pm 1$. This algorithm is run within the each iteration of the proposed method, so $U_i^C(\pm 1)$ is computed according to the values of the neighboring labels in the current classification

2.2. Markovian support vector classification

map. Stacking together all such inequalities, a system of linear inequalities is obtained:

$$\begin{bmatrix} y_1[U_1^P(-1) - U_1^P(1)] & y_1[U_1^C(-1) - U_1^C(1)] \\ y_2[U_2^P(-1) - U_2^P(1)] & y_2[U_2^C(-1) - U_2^C(1)] \\ \vdots & \vdots \\ y_\ell[U_\ell^P(-1) - U_\ell^P(1)] & y_\ell[U_\ell^C(-1) - U_\ell^C(1)] \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \geq \mathbf{0} \quad (2.16)$$

where ℓ indicates here the number of support vectors. The numerical solution of this system is addressed through the iterative Ho-Kashyap procedure that is a well known approach in pattern recognition to calculate linear discriminant functions [145]. If the solution exists, the convergence in a finite number of iterations is guaranteed [145].

The optimization of $\boldsymbol{\theta}$ is automatically performed through the Powell-span-bound algorithm that minimizes an upper bound (named “span bound”) on the leave-one-out error rate [97]. The span bound is computed by using the training samples in the initialization phase of the proposed method (details can be found in [97]) and is a non-differentiable function of $\boldsymbol{\theta}$. The Powell algorithm is used to numerically minimize the span bound. It ensures convergence to a local minimum through a procedure that emulates the conjugate-gradient technique without involving derivatives [111]. Additional details about the algorithm can be found in Section 6.4.1.

The proposed reformulation of the global posterior energy allows to address the minimization task in step 5 through computationally efficient methods, which rely on graph cut and belief propagation concepts and have attracted increasing interest during the last decade. In the case of binary classification, *graph cuts* make use of a reformulation based on the min-flow/max-cut theorem to reach a global energy minimum, provided the pairwise potential satisfies a suitable condition [59]. In the multiclass case, graph cut algorithms iteratively define a sequence of suitable binary problems, and under appropriate assumptions on the pairwise potential, converge to local minimum with strong optimality properties [18, 19, 74].

Belief propagation-type methods formalize the intuitive idea of passing messages along the graph to decrease the energy [67]. In particular, the *max-product loopy belief propagation (LBP)* technique operates on graphs

2.2. Markovian support vector classification

with loops, such as those that are usually associated with MRF neighborhoods. It may generally not converge, but when it does, it obtains a local minimum with good optimality properties [67, 139]. The complexity of efficient formulation of LBF is linear with respect to the numbers of pixels and classes [48]. The *tree re-weighted message passing (TRW)* method combines belief propagation with the construction of suitable spanning trees [156], and can be endowed with specific convergence properties by using an appropriate sequential formulation (TRW-S) [73]. The complexity of this formulation is linear with respect to the numbers of edges in the graph, of classes and of iterations [73].

2.2.4 Experimental validation

The MSVC framework is experimented with the data sets illustrated in the following paragraph in conjunction with three energy minimization algorithms, i.e., graph cuts, LBP, and TRW-S. Regarding LBP, for each iteration of the MSVC approach, both variants discussed in [137], which differ in the schedules for exchanging messages among the pixels, are used, and the solution with the lower energy is selected. In the case of TRW, the TRW-S formulation is used to favor a convergent behavior. The results of MSVC are compared to those obtained (i) by a benchmark non-contextual SVM; and (ii) by deriving approximate pixelwise posteriors from the SVM output through the method in [162] and by plugging them into the unary term. In all the experiments, the Gaussian radial basis function kernel is used, and the hyperparameters of the SVM and the parameter β are automatically optimized by using the method in 2.2.3. The Potts model (see Eq. (2.9)) was used for \mathcal{E}_i with a second order neighborhood system.

First, the experimental validation was carried out with three hyperspectral data sets. The first data set, named AVIRIS Indian Pine, was the well-known AVIRIS "Indian Pines" hyperspectral image, 145×145 pixels with a spatial resolution of 20m per pixel. This data set originally includes 220 spectral channels but 20 water absorption bands (104-108, 150-163, 220) have been removed, and the rest (200 bands) were taken into account for the experiments. The reference data contains 16 classes of interest, which

2.2. Markovian support vector classification

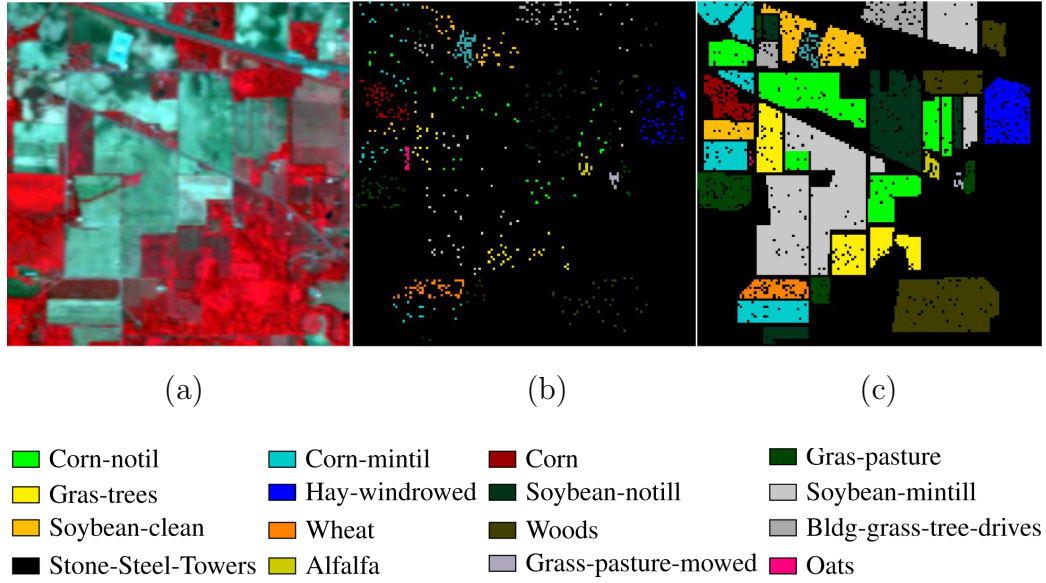


Figure 2.1: AVIRIS Indian Pines hyperspectral data. (a) Three band false color composite, (b) Reference data and (c) Color code.

represent mostly different types of crops and are detailed in Table 2.1. Figure 2.1 shows a three-band false color image and its corresponding reference samples.

The second data set, named "CASI Houston University," was acquired over the University of Houston campus and the neighboring urban area in June, 2012 by the Compact Airborne Spectrographic Imager (CASI). The size of the data is 349×1905 with the spatial resolution of 2.5 m. This data set is composed of 144 spectral bands ranging $0.38\text{-}1.05\mu\text{m}$. This data consists of 15 classes including: Grass Healthy, Grass Stressed, Grass Synthetic, Tree, Soil, Water, Residential, Commercial, Road, Highway, Railway, Parking Lot 1, Parking Lot 2, Tennis Court and Running Track. The "Parking Lot 1" includes parking garages at the ground level and in elevated areas, while "Parking Lot 2" corresponded to parked vehicles. Table 2.2 demonstrates different classes with the corresponding number of training and test samples. Figure 2.2 shows a three-band false color image and its corresponding training and test samples.

The third data set, named ROSIS-03 Pavia University, was composed of

2.2. Markovian support vector classification

Table 2.1: Indian Pines: Number of Training and Test Samples.

Class		Number of Samples	
No	Name	Training	Test
1	Corn-notill	50	1384
2	Corn-mintill	50	784
3	Corn	50	184
4	Grass-pasture	50	447
5	Grass-trees	50	697
6	Hay-windrowed	50	439
7	Soybean-notill	50	918
8	Soybean-mintill	50	2418
9	Soybean-clean	50	564
10	Wheat	50	162
11	Woods	50	1244
12	Bldg-grass-tree-drives	50	330
13	Stone-Steel-Towers	50	45
14	Alfalfa	50	39
15	Grass-pasture-mowed	50	11
16	Oats	50	5
Total		695	9671

2.2. Markovian support vector classification

Table 2.2: Houston: Number of Training and Test Samples.

Class		Number of Samples	
No	Name	Training	Test
1	Grass Healthy	198	1053
2	Grass Stressed	190	1064
3	Grass Synthetic	192	505
4	Tree	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking Lot 1	192	1041
13	Parking Lot 2	184	285
14	Tennis Court	181	247
15	Running Track	187	473
Total		2,832	12,197

2.2. Markovian support vector classification

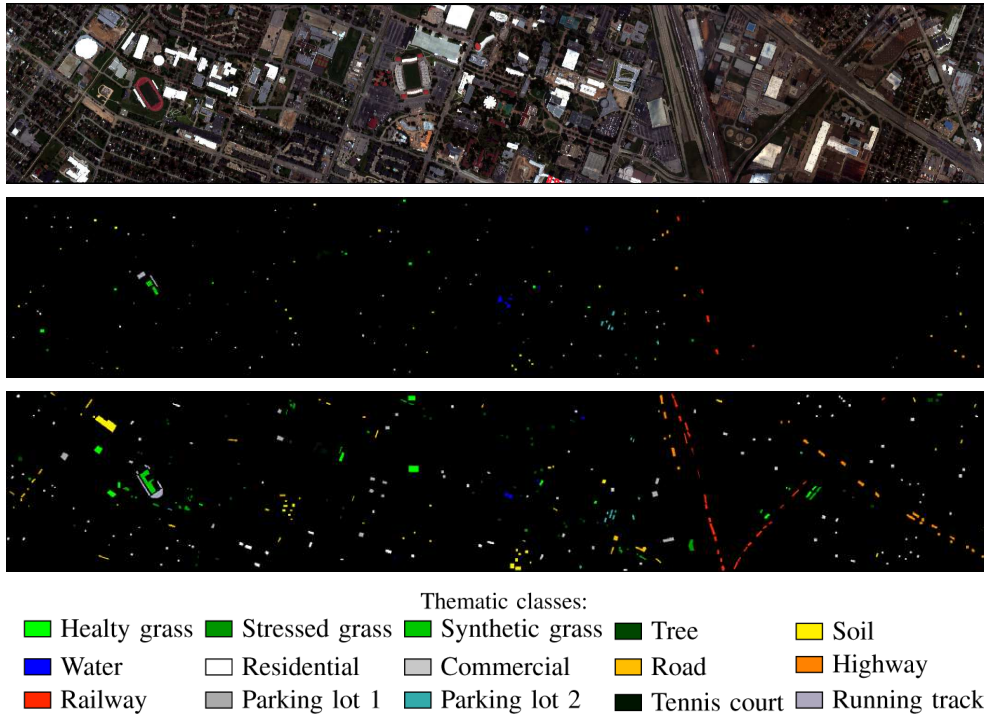


Figure 2.2: Houston - From top to bottom: A color composite representation of the hyperspectral data using bands 70, 50, and 20, as R, G, and B, respectively; Training samples; Test samples; and legend of different classes.

2.2. Markovian support vector classification

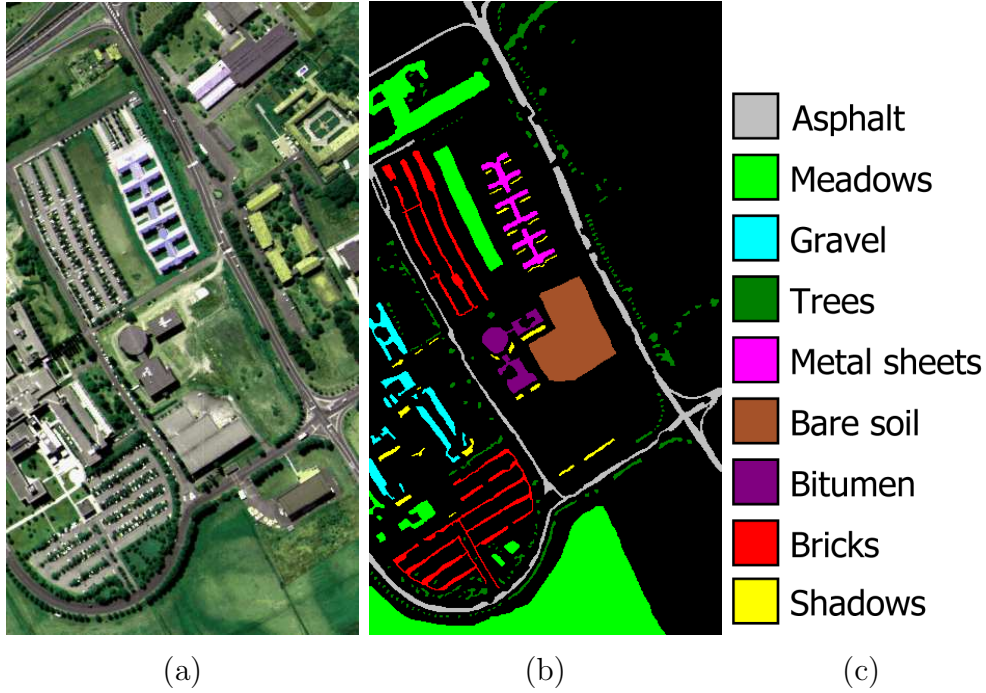


Figure 2.3: ROSIS-03 Pavia University hyperspectral data. (a) Three band false color composite, (b) Reference data and (c) Color code.

115 data channels with a spectral coverage ranging from 0.43 to $0.86 \mu m$. Twelve channels have been removed due to noise. The image (640×340 pixels) was acquired over the city of Pavia, Italy by the ROSIS-03 (Reflective Optics Spectrographic Imaging System) with a spatial resolution of 1.3 m. It presents 9 classes including trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow and soil. Figure 2.3 presents a false color image of ROSIS-03 Pavia University data and its corresponding reference samples.

The focus on hyperspectral data comes from the opportunity to jointly take benefit from the effectiveness of kernel methods and MRFs in the application to high-dimensional feature space and to VHR imagery, respectively. Furthermore, an additional experimental phase is conducted with the aim to investigate the capability of the proposed integrated framework to exploit the different sources of information provided by a multisensor (optical and synthetic aperture radar, SAR) data set. Specifically the data set is composed of two co-registered images (920×820 pixels) acquired over the port

2.2. Markovian support vector classification

Table 2.3: Pavia University: Number of Training and Test Samples.

Class		Number of Samples	
No	Name	Training	Test
1	Asphalt	548	6304
2	Meadow	540	18146
3	Gravel	392	1815
4	Tree	524	2912
5	Metal Sheet	256	1113
6	Bare Soil	532	4572
7	Bitumen	375	981
8	Brick	514	3364
9	Shadow	231	795
Total		3921	40002

area of Port-au-Prince (Haiti) by two different types of sensor: the first one is a 3-channels, 2.5-m resolution GeoEye image, while the second one consists of a 2.5-m resolution, HH-polarized COSMO-SkyMed SAR acquisition (Figures 2.7) The data set presents five classes: water, (buildings with) red roofs, urban and port areas, vegetation, and bare soil. Training and test fields are chosen to be spatially disjoint to minimize the correlation between training and test samples and the resulting possible bias in accuracy assessment. The experimental setup is the same as with the hyperspectral data sets.

The classification accuracies obtained by the aforementioned methods on the test samples of the three data sets are collected in Tables 2.4, 2.5, and 2.6. The MSVC framework provides overall accuracies (OA) around 91-92%, 82-87%, and 85-87% in the cases of Indian Pines, Pavia, and Houston, respectively. Remarkable improvements in accuracy are obtained by MSVC as compared to the non-contextual SVM benchmark. Furthermore, the Markovian classifier that uses, within its unary term, approximate pixel-wise posteriors derived from the SVM output also provides accurate results, although it obtains lower values of OA than MSVC with all considered energy minimization methods. On one hand, the improvement of MSVC over SVM is an expected result because the former is a spatio-spectral approach while

2.2. Markovian support vector classification

the latter is a purely spectral classifier. On the other hand, the aforementioned results confirm the effectiveness of MRF modeling for spectral-spatial classification. They also point out the capability of the integrated MSVC framework to benefit from both SVM and MRF in order to incorporate spatial information into a classification task in a high-dimensional feature space. This comment is also confirmed by a visual analysis of the classification maps (e.g., Figures 2.4(b), 2.5(b), and 2.6(b)), which points out the capability of MSVC to favor the desired spatially smooth behavior, especially as compared to purely spectral methods (e.g., Figures 2.4(a), 2.5(a), and 2.6(a)).

The three considered energy minimization algorithms overall exhibit similar behaviors. They obtain very similar accuracies in the cases of Indian Pines and Houston, while in the case of Pavia, TRW-S reaches 3-4% higher OA than graph cuts and LBP. On one hand, the high accuracies achieved confirm the effectiveness of current advanced graph cut and message passing techniques for MRF energy minimization in a HSI classification task, a conclusion that has been drawn in numerous image processing and computer vision applications [137]. On the other hand, the performances obtained using all three methods also suggest the flexibility of the MSVC framework in incorporating arbitrary energy minimization algorithms. This flexibility also comes together with the opportunity to fully automate the resulting classifiers through the aforementioned parameter optimization methods in [97, 121].

Given the similar behavior of the different energy minimization method adopted, the experimental validation using multisensor optical/SAR data has focused on the MSVC framework applied in conjunction with the GC approach. The results of the proposed MSVC-GC method on the multisensor optical/SAR dataset were compared to those obtained with the same approach applied separately 1) only to the optical image and 2) only to the SAR image. Here, the goal was mostly to ascertain the multisensor fusion capabilities of the proposed method, as they are expected from the fully non-parametric kernel-based formulation.

The high classification accuracies on the test set reached by the proposed method when applied to the optical/SAR image (see Table 2.7), showed the capability of the proposed MSCV-GC approach to exploit the information provided by multisensor data. Both the overall and average accuracies high-

2.2. Markovian support vector classification

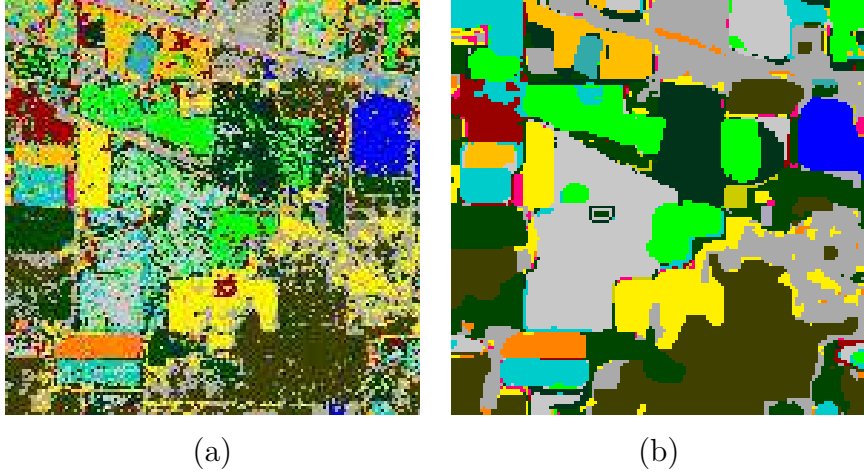


Figure 2.4: Classification maps for Indian Pines: (a) SVM, (b) MSVC with graph cuts.

light the improvement related to the application of the proposed classifier to multisensor data with respect to the use of single-sensor images. In particular, accurate classification maps were generated by the proposed method with an improvement of the values of OA and AA of 0.31% and 1.15%, respectively. In particular the proposed approach applied to multisensor data yielded a significant enhancement in the discrimination between the red roof and urban and port area classes which resulted in an improvement of the accuracy for the red roof class (+2,33%). The same consideration holds for the discrimination between the bare soil and urban and port area classes, where the joint use of SAR and optical images allowed an accuracy increase (+3,25%) for bare soil. All these results overall confirm the flexibility of the proposed integrated MRF / SVM / min-energy framework for the spatial-contextual classification of VHR imagery from single and multiple EO sensors. Furthermore, it should be noted that a previous MSVC formulation using ICM was originally developed in [97] and experimentally validated with various data modalities, including multispectral VHR imagery. Possible future extension of the proposed framework could consist in the integration with multiscale or multiresolution MRFs, such as those discussed in the next chapter.

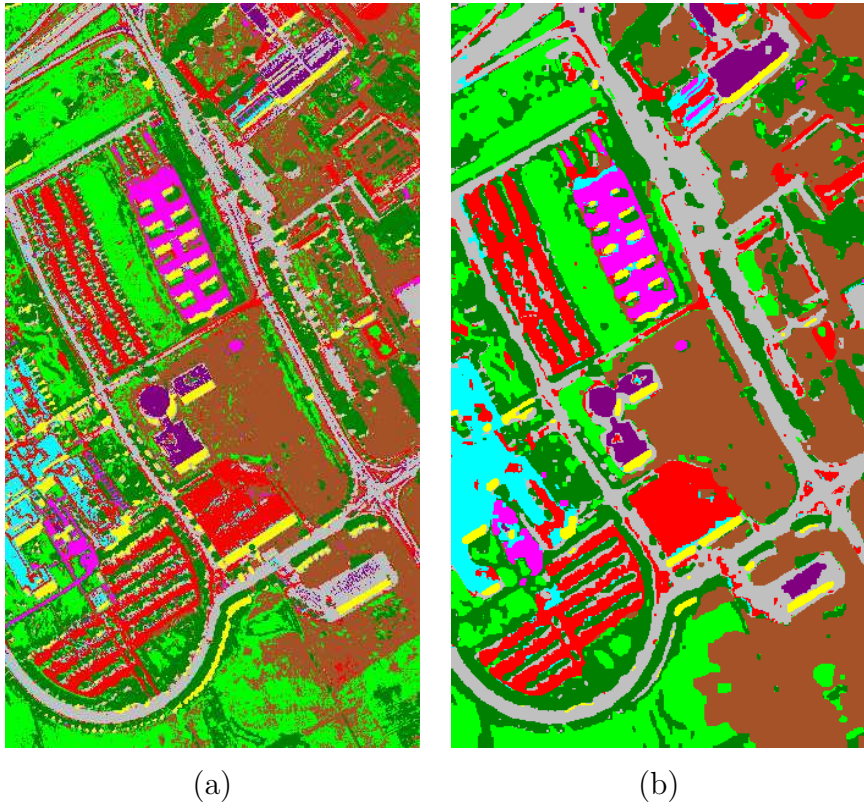


Figure 2.5: Classification maps for Pavia University: (a) SVM, (b) MSVC with LBP.

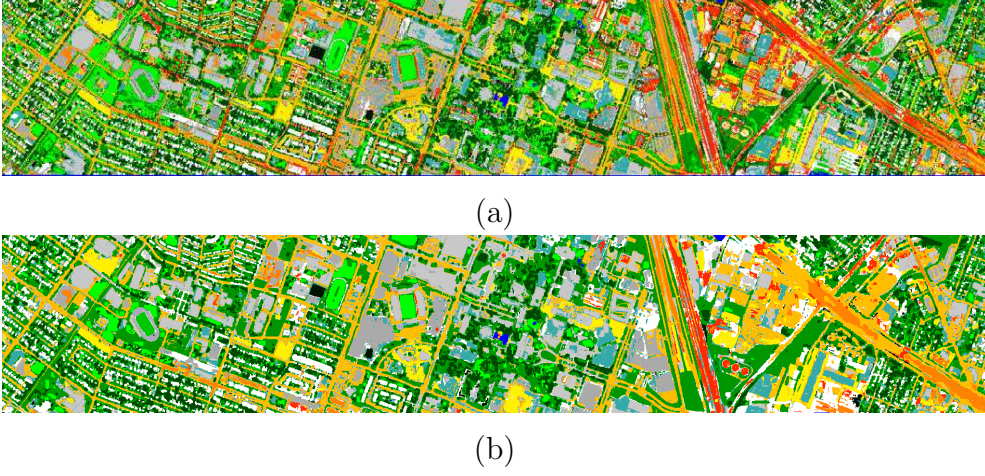


Figure 2.6: Classification maps for Houston:(a) SVM, (b) MSVC with TRW-S.

2.3 Region-based kernel for optical data

2.3.1 Methodology

Let \mathcal{I} be a regular pixel lattice associated with a remote sensing image composed of a finite set of homogeneous regions $\mathcal{R} : \bigcup_{r \in \mathcal{R}} r = \mathcal{I}$ and $r \cap s = \emptyset$ for $r \neq s$. No restrictions are placed by the proposed kernel frameworks on the characteristics of these regions or on the segmentation algorithm to be used for generating them. The pixels belonging to each individual region are modeled as samples drawn from a two-dimensional stationary Gaussian random process, which is a usually accurate assumption for pixel intensities in optical images [81]. In the simplest formulation of the proposed kernel, this Gaussian process may be modeled as white, a scenario that corresponds to neglecting information associated with intra-region correlations. In this case, owing to the stationarity assumption, each region $r \in \mathcal{R}$ is associated with a feature vector $(\mu_r, \sigma_r) \in \mathbb{R}^2$, where μ_r and σ_r are the mean and standard deviation of the intensities of the pixels in the region. The variance is actually known to be an indicator of texture (namely, a first-order statistical texture feature). Here, its region-wise computation makes it spatially adaptive to the segmentation map that identifies the set \mathcal{R} of regions. Given

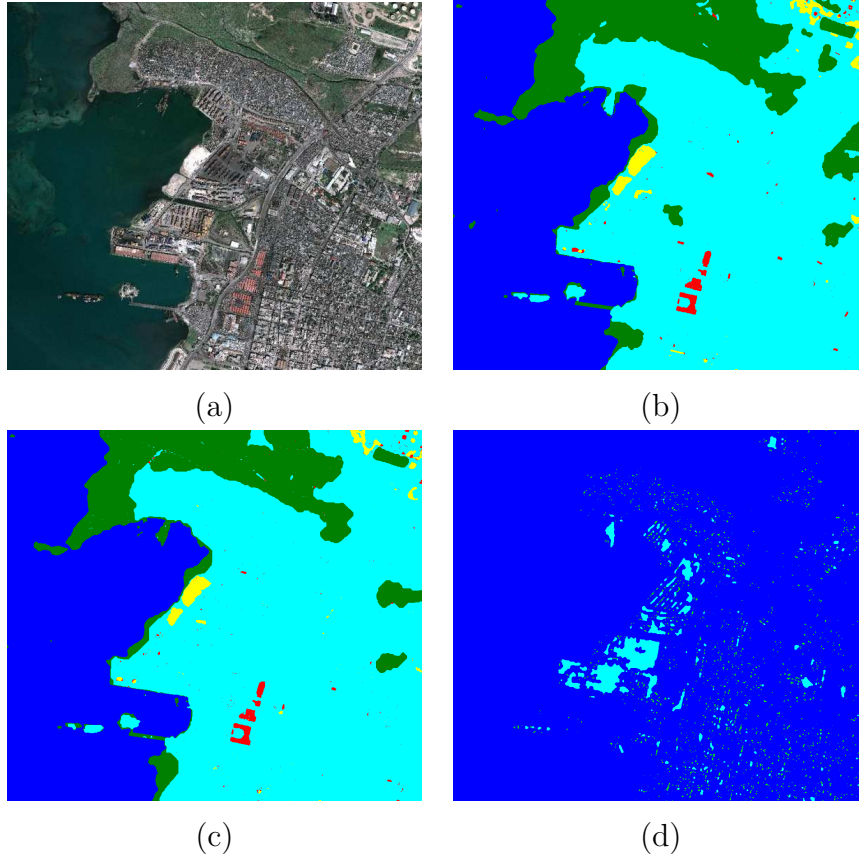


Figure 2.7: Port-au-Prince: (a) RGB composition of the optical channels; classification maps generated by the application of proposed method to the multisensor optical/SAR (b) and by same approach to only optical image (c) and only SAR image (d). Legend: water (blue), buildings with red roofs (red), urban and port areas (cyan), vegetation (green), bare soil (yellow).

2.3. Region-based kernel for optical data

two different regions r and s , a mapping function $\varphi: \mathbb{R} \times (0, +\infty) \rightarrow L^2$ and a kernel function K_{RB} can be defined such as $\phi(\mu, \sigma)$ is the $\mathcal{N}(\mu, \sigma^2)$ density function and:

$$K_{RB}(r, s) = \langle p_r, p_s \rangle_{L^2} = \int_{-\infty}^{+\infty} p_r(\xi) p_s(\xi) d\xi = \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_r\sigma_s} \sqrt{\exp\left[-\frac{(\xi - \mu_r)^2}{\sigma_r^2} - \frac{(\xi - \mu_s)^2}{\sigma_s^2}\right]} d\xi \quad (2.17)$$

where $p_r(\cdot)$ is the Gaussian first-order density associated with region r and $p_s(\cdot)$ has a similar meaning ($r, s \in \mathcal{R}$). The function K_{RB} is an admissible kernel by definition, as its value is equivalent to evaluating an inner product in the Hilbert space L^2 . Algebraic manipulations also allow $K_{RB}(r, s)$ to be related to the Bhattacharyya coefficient between the first-order Gaussian distributions associated with segments r and s . In particular, through the definition of Bhattacharyya coefficient for Gaussian distributions, up to a multiplicative factor, the proposed region-based kernel function is:

$$K_{RB}(r, s) = \frac{1}{\sqrt{\sigma_r^2 + \sigma_s^2}} \exp\left[-\frac{(\mu_r - \mu_s)^2}{2(\sigma_r^2 + \sigma_s^2)}\right]. \quad (2.18)$$

This kernel formulation can also be readily extended to the application to multispectral data, using either multivariate Gaussian distributions or composite kernels [24], and to model ling intra-region correlations through bivariate distributions taking into account the joint statistics of pairs of samples inside each region (i.e., second-order statistical texture information). Furthermore, multiscale analysis can also be carried out using multiple segmentation results, associated with distinct spatial scales, again through composite kernels. For example, if an input multispectral image composed of B bands and a set $\{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_L\}$ of segmentation maps corresponding to L spatial scales are considered, then a separate region-based kernel $K_{b\ell}(\cdot)$ can be computed using each map \mathcal{S}_ℓ and the mean and variance of each b -th band ($b = 1, 2, \dots, B; \ell = 1, 2, \dots, L$). Denoting as $r_{i\ell}$ the segment label of pixel $i \in \mathcal{I}$ in \mathcal{S}_ℓ , and as $\mathbf{r}_i = (r_{i1}, r_{i2}, \dots, r_{iL})$ the vector of all segmentation labels

2.3. Region-based kernel for optical data

of pixel i , then the following composite kernel:

$$K_{RB}(\mathbf{r}_i, \mathbf{r}_j) = \sum_{\ell=1}^L \sum_{b=1}^B K_{b\ell}(r_{i\ell}, r_{j\ell}) \quad (2.19)$$

can be used to apply the proposed region-based kernel in a multiscale and multispectral configuration. More generally, the proposed kernel can be plugged into an arbitrary kernel machine. For classification purposes, plugging this kernel into an SVM is clearly a natural choice and is the choice used here to experimentally validate it (see the next section). However, it can be combined as well with kernel-based feature extraction methods, such as the kernel principal component analysis of the kernel Fisher discriminant, or with unsupervised kernel machines including spectral clustering or kernel k means.

2.3.2 Experimental validation

Experiments have been carried out with three synthetic datasets, one real dataset, and two segmentation methods. First, the graph-based region-merging segmentation method in [48] was used. In the first synthetic dataset, the data belonging to two different classes were generated according to a Gaussian distribution in a two-dimensional feature space. The chosen covariance matrices and mean vectors lead to a simple case study with two well separated classes (Figure 2.8(a)). The second synthetic dataset differs from the first one in the Gaussian parameter setting which determines two strongly overlapping classes (Figure 2.9(a)). In the third dataset, the two considered classes were generated according to a uniform distribution (Figure 2.10(a)).

Figures 2.8(c), 2.9(c), and 2.10(c) show the results achieved by the proposed method in terms of classification maps. The proposed classifier achieves high accuracies in the application to the first two datasets where the method assumption of Gaussian distribution is satisfied. Nevertheless, a visual analysis of the classification maps generated by the proposed classifier makes it possible to also appreciate the behavior of the considered method with the third image characterized by a mixture of two uniform distributions,

2.3. Region-based kernel for optical data

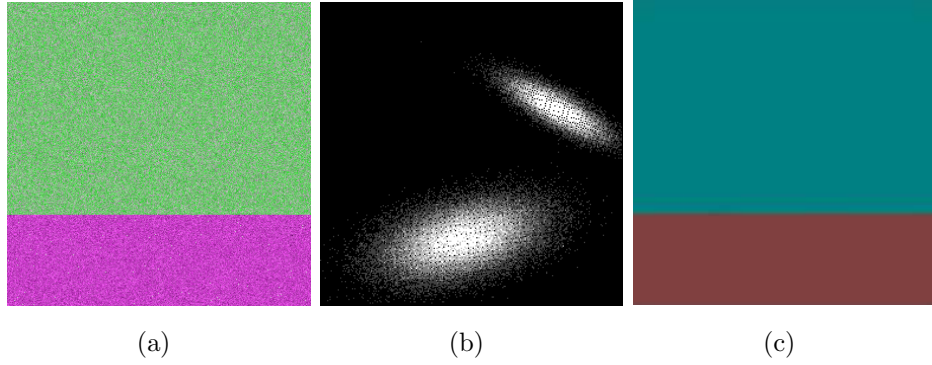


Figure 2.8: First synthetic dataset. RGB false-color composition (a), two-dimensional histogram (b), and classification map generated by the proposed method (c)

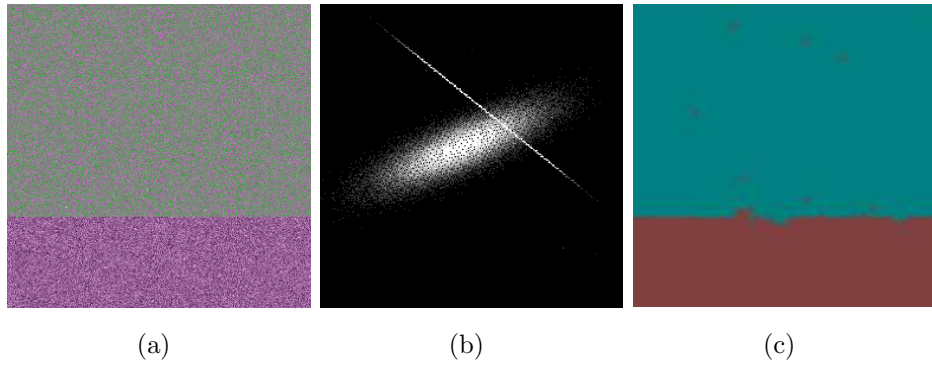


Figure 2.9: Second synthetic dataset. RGB false-color composition (a), two-dimensional histogram (b), and classification map generated by the proposed method (c)

2.3. Region-based kernel for optical data

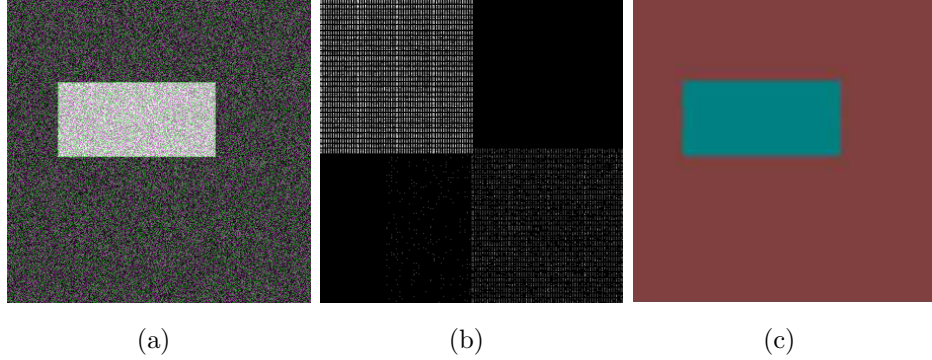


Figure 2.10: Third synthetic dataset. RGB false-color composition (a), two-dimensional histogram (b), and classification map generated by the proposed method (c)

thus showing remarkable robustness even when working assumptions are not fulfilled.

The real data set used for the experimental validation is an IKONOS image acquired around Alessandria, Italy, and composed of 1250×1250 pixels. The spatial resolution is 4 m, and four channels corresponding to blue, green, red, and near-infrared radiation are available (Figure 2.11). Seven main classes can be visually noted, i.e., “urban,” “forest,” “water,” “wet soil,” “bare soil,” and two agricultural covers (Table 2.8).

The proposed region-based kernel has been applied in its multispectral and multiscale formulation using an input multiscale segmentation result and white Gaussian processes. Specifically, the segmentation method in [48], which is a graph-theoretic region-merging technique parameterized by a scale parameter, has been applied with five different values of this parameter. The resulting stack of five segmentation results ranges from a map with a large number of small segments, corresponding to small-scale structures and objects, to a map with a few large segments, corresponding to large image regions and structures. An SVM with the proposed region-based kernel has been applied with this multiscale segmentation result. The proposed kernel has no internal parameters. Therefore, here, the only SVM parameter is the regularization coefficient C [152], which has been automatically optimized by minimizing the span bound [150] using Brent’s algorithm [22]. As already

2.3. Region-based kernel for optical data

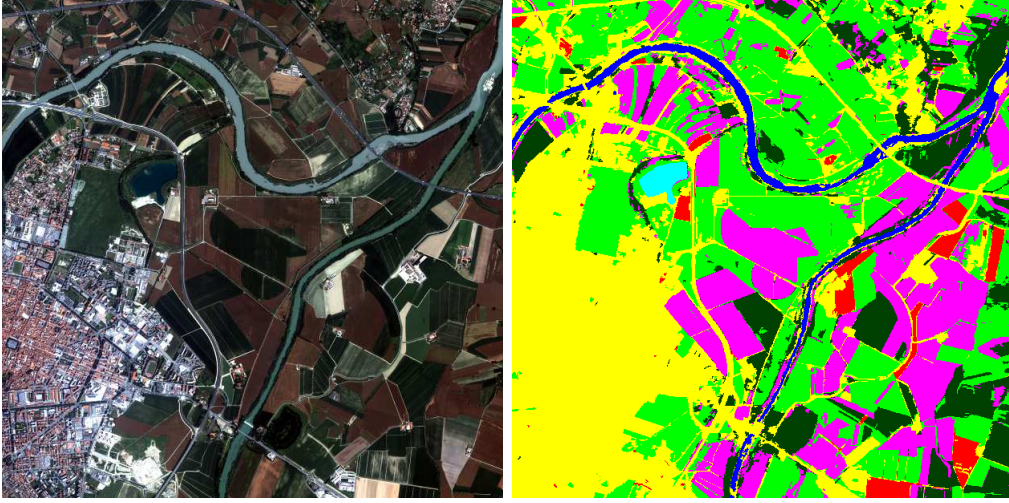


Figure 2.11: Alessandria data set: RGB true-color composition of the IKONOS channels (left) and classification map (right).

mentioned, the span bound is a tight upper bound on the generalization error of the SVM, and has been found effective for the optimization of the parameters of SVM classification and regression methods [28], [27], [96], [97]. Here, Brent’s algorithm has been applied because it is feasible for non-differentiable univariate functions [111] such as the considered span bound. Spatially disjoint training and test sets have been used to minimize the correlation between training and test samples and the corresponding possible biases in accuracy assessment.

The classification map in Figure 2.11 has been obtained, which corresponds to the confusion matrix in Table 2.8 with respect to the test set. Remarkable performances have been obtained in terms of both overall accuracy ($OA \approx 95\%$) and average accuracy ($AA \approx 93\%$), which confirms the effectiveness of the proposed approach for the classification of VHR optical imagery. In particular, high accuracies have been obtained for all classes except “forest.” On one hand, this confirms the effectiveness of the developed approach at discriminating classes with very diverse spatial behaviors, including spatially smooth (e.g., “water”) and geometrical or textured behaviors (e.g., “urban” or the agricultural covers). On the other hand, a more accurate discrimination of a strongly textured class such as “forest” could

2.3. Region-based kernel for optical data

be accomplished by the aforementioned formulation of the proposed kernel based on spatially correlated Gaussian processes and second-order textures. A visual inspection of the classification map further confirms the effectiveness of the proposed approach and especially its capability to identify the urban area as a whole, which is a challenging task because of the strong spatial heterogeneity of this class when observed through VHR sensors.

To ascertain the flexibility of the proposed kernel in combination with different segmentation methods, as a second phase of the experimental validation, watershed-type technique has been used. In this case, the data set is composed of the multispectral IKONOS image acquired over Itaipu area and used in Chapter 3 (see Tables 2.9, 2.10, 2.11). The first pre-processing step consists in the application of a segmentation algorithm to the considered dataset. For this purpose, a segmentation result provided by colleagues of the University of Naples, Italy, and obtained using the watershed-based segmentation method in [52] was used. In this method, a distance map is obtained from an edge map of the input image and used as digital elevation model. Two different markers are exploited in order to take into account local geometrical and spectral homogeneities and the watershed is applied to resulting digital elevation model. The adopted watershed-based method provides object-level segmentations of adaptive scale and, thanks to the morphological and spectral markers, showed excellent performance in reducing the fragmentation, which typically constitutes the main drawback of this kind of approach. Two different experiments were conducted. The first one shares the same settings as in the previous experimental phase, i.e denoting as \mathcal{I} the pixel lattice and as r_i the segment label of pixel $i \in \mathcal{I}$ an SVM classifier using the region-based kernel $K_{RB}(r_i, r_j)$ was applied to classify the aforementioned image. The optimization of the regularization coefficient C [151] was again performed by minimizing the span bound [150] using Brent's method [22].

Then, a composite kernel combining the aforementioned region-based kernel with a pixelwise kernel $K_P(\mathbf{x}_i, \mathbf{x}_j)$ was considered, i.e. $(i, j \in \mathcal{I})$ [25]:

$$K_{RBP}(\mathbf{x}_i, r_i, \mathbf{x}_j, r_j) = \lambda K_{RB}(r_i, r_j) + (1 - \lambda) K_P(\mathbf{x}_i, \mathbf{x}_j), \quad (2.20)$$

where \mathbf{x}_i is the pixelwise feature vector of pixel $i \in \mathcal{I}$, a Gaussian radial basis function (RBF) is used for $K_P(\cdot)$, and the parameter $\lambda \in [0, 1]$ weighs the

2.3. Region-based kernel for optical data

region-based and pixelwise contributions. The pixelwise features included the IKONOS spectral channels and the normalized difference vegetation index (NDVI) to favor the discrimination of the vegetated classes in the data set. An SVM with this composite kernel overall includes three parameters, i.e., C , λ , and the standard deviation σ of the RBF kernel. Their values were again automatically optimized through the minimization of the span bound. In this case, Powell’s algorithm was used to address the numerical minimization task because it is applicable to non-differentiable multivariate functions [109] [111], for which gradient-like methods are obviously unfeasible. The result of a purely pixelwise SVM using the RBF kernel was also considered for the sake of comparison. Powell’s algorithm and the span bound were used to optimize the parameters in this case as well.

The confusion matrices of the considered approaches are shown in Tables 2.9, 2.10, 2.11. The proposed region-based classifier in his simplest formulation showed high classification accuracy with an overall accuracy (OA) of around 98% and an average accuracy (AA) of 89%. This result confirmed the effectiveness of the proposed kernel function and the capability of the proposed approach to take advantage of the contextual information made available by the segmented image. In particular, the producer and user accuracies (PA and UA) were high for all the classes except “herbaceous” and “bare soil.” The former was partially confused with “urban” because of the highly heterogeneous spatial behavior of the urban class, which also included vegetated patches in the training and test sets of this data set. “Bare soil” was confused in part with “urban” and “built-up” because of the similar spectral responses in the three available IKONOS channels.

The use of the composite kernel allowed the classification result to be further improved. In particular, the discrimination of “herbaceous” improved substantially (with an increase of over 33% in PA). The “built-up” class was also discriminated with higher accuracy (increase of around 7% in PA), while the accuracies of “urban” and “bare soil” were slightly lower (decrease of around 3% in PA). OA and AA were approximately 98.5% and 94%, respectively. On the contrary, the application of a purely pixelwise SVM with the RBF kernel led to poorer performances, as expected. In particular, although OA was rather high (around 94%), poor discrimination was especially noted

2.3. Region-based kernel for optical data

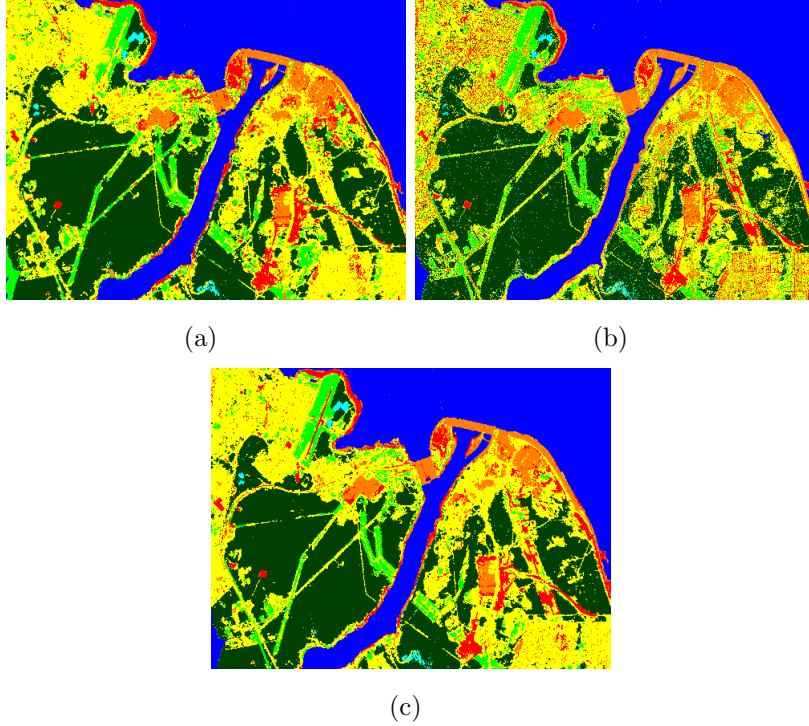


Figure 2.12: Classification maps obtained, for the Itaipu data set, by SVM classifiers using the region-based kernel (a), a composite kernel combining the region-based and the pixelwise RBF kernels (b), and the purely pixelwise RBF kernel (c). Color legend: see Table 2.9.

for “urban” ($PA \approx 58\%$). Indeed, this is consistent with the fact that, in this case, the classification task included identifying the urban area as a whole, which was a challenging task because of the aforementioned spatial behavior of this class. Accordingly, pixelwise processing was quite ineffective, while the segmentation and contextual classification approaches developed proved to be successful.

A visual analysis of the corresponding classification maps (Figure 2.12 and 2.13) further confirms these comments. In particular, the sharp improvement in the discrimination of the urban area obtained by the two region-based results as compared to the purely pixelwise result, and the improved discrimination of “herbaceous” achieved using $K_{RBP}(\cdot)$ as compared to $K_{RB}(\cdot)$ are especially evident in the detail in Figure 2.13.

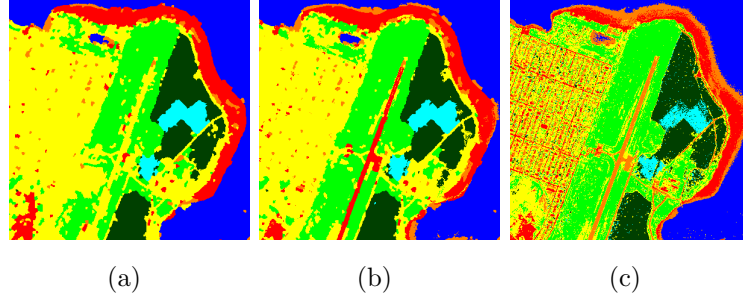


Figure 2.13: Classification maps obtained, for the Itaipu data set, by SVM classifiers using the region-based kernel (a), a composite kernel combining the region-based and the pixelwise RBF kernels (b), and the purely pixelwise RBF kernel (c). Color legend: see Table 2.9.

2.4 Conclusion

This chapter tackled the problem of the integration of spatial-contextual information into kernel machines through the development of two novel integrated framework based on Markov random fields approach and region-based methods. Specifically, the first proposed approach integrates SVMs, MRFs, and efficient energy minimization techniques. The method is based on a previous approach in which the energy minimization task was performed thorough ICM algorithm, able to reach only local minimum and characterized by critical dependence in the initialization. In order to overcome these limitations the integrated SVM-MRF framework is re-formalized and a reformulation of the global posterior energy function corresponding to the application of an MRF-based classifier in the transformed space associated with a kernel function is proposed. Thus, the last-generation energy minimization method, including graph cut and believe propagation-type approaches, may be exploited for this task. The experimental validation was carried out with diverse remote sensing data. Using three hyperspectral data sets, the capability of the proposed framework to classify at very high spatial resolution in very high-dimensional spaces was tested in conjunction with GC, LBP, and TRW-S algorithm. The remarkable classification accuracies pointed out the capability of the proposed method to benefit from both SVM and MRF in order to incorporate spatial information into the classification task. Further-

2.4. Conclusion

more, the GC-based classifier formalized in the proposed SVM-MRF integrated framework, was tested on a multisensor image classification problem confirming the flexibility of the proposed approach.

With regards to the second proposed spatial-contextual method, a novel kernel function has been proposed. Specifically, this approach allows exploiting the spatial information associated with the input image through the combination of the SVM kernel theory and the region-based approach in a unique classification process. The pixels belonging to each individual region are modeled as samples drawn from a two-dimensional Gaussian random process, which is a reasonable assumption for pixel intensities in VHR optical imagery. Based on this assumption, a kernel function whose entries are two segments rather than two pixelwise samples is defined as the inner product between the Gaussian densities of the two corresponding processes in the L^2 Hilbert space. This function is an admissible Mercer kernel by construction, and can be incorporated into an arbitrary kernel machine. Both single-channel and multispectral formulations can be formalized, and a multiscale analysis can also be carried out using multiple segmentation results, associated with distinct spatial scales, through the composite kernel method. Experiments on both simulated and real dataset were conducted. The high classification accuracies reached confirms the capability of the proposed approach for VHR image classification purpose. An SVM classifier using the proposed region-based kernel was applied to classify the adopted dataset, and then a further improvement in the overall and average accuracies was pointed out by the application of a composite kernel combining the aforementioned region-based kernel with a Gaussian radial basis function pixelwise kernel. The comparison with a purely pixelwise SVM with a RBF kernel highlights the ability of the two region-based classifiers in the discrimination of the urban area as a whole, which represents a challenging problem. Indeed in a VHR image this spatially heterogeneous class can be considered as composed of a collection of subclasses related to different objects or materials and the discrimination of “urban” class as a unique land cover may not be straightforward. The formulation of the proposed kernel validated through these experiments used a white model for the data within each region. This assumption is obviously violated in real data, which usually exhibits strong correlations

2.4. Conclusion

among neighboring pixels. On one hand, the method proved able to provide high accuracies in the application to challenging data sets nonetheless. On the other hand, the extension to a non-white model is straightforward, could allow textured classes to be better discriminated, and will be pursued as an extension of the proposed approach. The two approaches provided different and somewhat complementary insights on the capability to integrate spatial information into kernel-based learning and classification. It is worth noting that the two approaches are not mutually exclusive: the region-based kernel of the second proposed approach can be incorporated into arbitrary kernel machines, including the Markovian SVM of the first proposed approach.

2.4. Conclusion

Table 2.4: Indian Pines - Classification accuracies [%] obtained by the MSVC approach, by a non-contextual SVM, and by an MRF-based classifier whose unary term is derived through the method in [162]. Graph cuts (GC), TRW-S, and LBP are used to minimize the energies of both Markovian approaches.

Classes	SVM	MSVC			MRF with unary from [162]		
		GC	TWR-S	LBP	GC	TRW-S	LBP
1	77.17	88.01	91.26	93.50	86.34	85.69	85.19
2	78.06	94.64	96.81	96.68	83.80	83.80	83.29
3	91.85	97.28	97.83	97.83	100	100	100
4	95.53	95.30	93.51	93.74	96.20	96.20	96.2
5	92.40	96.56	97.42	97.99	97.85	98.28	97.85
6	96.36	91.12	94.31	92.03	99.32	99.32	99.32
7	71.90	87.15	81.70	87.80	86.93	87.04	86.93
8	67.16	92.47	90.74	89.04	88.50	87.26	84.45
9	83.33	86.35	84.04	85.11	96.45	94.68	93.97
10	99.38	100	99.38	99.38	99.38	99.38	99.38
11	89.55	91.80	92.93	92.36	88.59	89.31	90.68
12	77.27	92.73	92.42	92.12	100	100	99.39
13	100	100	93.33	97.78	100	100	100
14	89.74	82.05	76.92	74.36	92.31	94.87	92.31
15	90.91	100	81.82	81.82	100	100	100
15	100	100	100	100	100	100	100
OA	79.85	91.66	91.40	91.80	90.54	90.18	89.42
AA	87.54	93.47	91.53	91.97	94.73	94.74	94.31
kappa	0.7710	0.9044	0.9014	0.9062	0.8919	0.8878	0.8793

2.4. Conclusion

Table 2.5: Pavia University - Classification accuracies [%] obtained by the MSVC approach, by a non-contextual SVM, and by an MRF-based classifier whose unary term is derived through the method in [162]. Graph cuts, TRW-S, and LBP are used to minimize the energies of both Markovian approaches.

Classes	SVM	MSVC			MRF with unary from [162]		
		GC	TWR-S	LBP	GC	TRW-S	LBP
1	81.77	95.29	96.73	96.89	93.97	93.88	93.99
2	65.76	67.45	77.47	69.58	64.47	67.34	67.73
3	69.42	80.72	81.93	82.59	68.43	66.39	70.80
4	94.37	95.19	95.36	96.91	96.67	96.81	96.53
5	99.46	100	98.65	100	99.91	99.91	99.91
6	92.26	98.25	96.85	97.86	97.92	97.51	97.44
7	86.24	95.51	81.45	85.22	92.35	92.46	92.46
8	89.68	95.07	97.68	97.35	98.42	98.37	98.10
9	91.95	90.19	93.46	86.79	99.87	99.75	99.50
OA	77.53	82.35	86.93	83.60	80.70	81.85	82.19
AA	85.66	90.85	91.06	90.35	90.22	90.27	90.72
kappa	0.7166	0.7769	0.8312	0.7917	0.7573	0.7704	0.7745

2.4. Conclusion

Table 2.6: Houston - Classification accuracies [%] obtained by the MSVC approach, by a non-contextual SVM, and by an MRF-based classifier whose unary term is derived through the method in [162]. Graph cuts, TRW-S, and LBP are used to minimize the energies of both Markovian approaches.

Classes	SVM	MSVC			MRF with unary from [162]		
		GC	TWR-S	LBP	GC	TRW-S	LBP
1	82.15	82.91	83.10	82.81	82.72	82.62	82.24
2	98.03	100	100	100	97.74	97.74	98.31
3	99.80	99.21	99.80	99.80	99.80	99.80	99.80
4	98.48	97.35	97.35	98.3	98.86	99.05	98.86
5	98.48	99.81	99.91	99.91	98.39	98.48	98.39
6	99.30	99.30	97.90	98.60	98.60	98.60	98.60
7	89.18	91.70	91.79	92.26	88.53	88.62	88.62
8	49.67	53.85	57.08	55.84	46.72	45.68	48.34
9	82.91	84.70	86.59	89.24	83.00	83.29	83.19
10	70.27	76.64	73.65	77.41	75.58	74.32	74.61
11	82.73	74.48	72.01	74.76	86.62	86.53	86.81
12	72.53	83.09	79.25	81.27	75.98	76.08	76.27
13	67.37	83.51	80.70	82.11	73.68	73.33	71.93
14	100	100	100	100	99.60	99.60	99.60
15	97.04	97.25	91.75	94.93	97.46	97.25	97.46
OA	83.99	86.07	85.48	86.60	84.96	84.80	85.05
AA	85.86	88.25	87.39	88.48	86.89	86.73	86.87
kappa	0.8264	0.8489	0.8424	0.8546	0.8370	0.8352	0.8379

2.4. Conclusion

Table 2.7: Port-au-Prince - Classification accuracies [%] on the test set obtained by MSVC-GC method applied to optical/SAR, only optical and only SAR image.

Class	Optical/SAR	Optical only	SAR only
Water	98.08	98.35	100
Building with red roofs	84.67	82.34	0
Urban and port areas	99.88	99.86	7.67
Vegetation	99.79	99.35	0
Bare soil	98.48	95.23	0
OA	99.18	98.87	19.75
AA	96.18	95.03	21.53
kappa	0.9877	0.9832	0.0258

Table 2.8: Alessandria data set: confusion matrix and classification accuracies of an SVM classifier based on the proposed region-based kernel applied using input multiscale segmentation.

Test map	Classification map						
	urban	agriculture1	agriculture2	forest	water	wet soil	bare soil
urban	7223	0	0	0	0	0	0
agriculture1	0	19030	37	0	0	0	0
agriculture2	0	0	8817	0	0	0	0
forest	101	16	1840	2676	0	0	0
water	7	0	0	0	2643	0	0
wet soil	0	0	0	0	0	808	0
bare soil	272	0	0	0	0	0	4770
PA	100%	99.81%	100%	57.76%	99.74%	100%	94.61%
UA	95.00%	99.92%	82.45%	100%	100%	100%	100%
OA	95.29%						
AA	93.13%						
kappa	0.9381						

2.4. Conclusion

Table 2.9: Confusion matrix obtained on the test set of the Itaipu data set by SVM classifiers using the region-based kernel.

Region-based kernel	Test map	Classification map						
		urban	herbaceous	shrub-brush	forest	bare soil	built-up	water
	urban	18646	13	0	0	76	247	0
	herbaceous	1840	3706	0	0	0	0	0
	shrub-brush	8	0	48911	0	0	0	0
	forest	0	0	0	674	0	0	0
	bare soil	357	59	0	0	2932	507	0
	built-up	496	0	0	0	446	4325	0
	water	0	0	0	0	0	25	91521
	PA	98.23%	66.82%	99.98%	100%	76.06%	82.12%	99.97%
	UA	87.35%	98.09%	100%	100%	84.89%	84.74%	100%
	OA	97.67%						
	AA	89.03%						
	kappa	0.9631						

Table 2.10: Confusion matrix obtained on the test set of the Itaipu data set by a composite kernel combining the region-based and the pixelwise RBF kernels.

Composite kernel (region-based and pixelwise RBF)	Test map	Classification map						
		urban	herbaceous	shrub-brush	forest	bare soil	built-up	water
	urban	18043	12	0	0	242	685	0
	herbaceous	42	5504	0	0	0	0	0
	shrub-brush	0	0	48919	0	0	0	0
	forest	0	0	0	674	0	0	0
	bare soil	482	4	0	0	2834	535	0
	built-up	23	0	0	0	561	4683	0
	water	0	0	0	0	0	25	91521
	PA	95.05%	99.24%	100%	100%	73.51%	88.91%	99.97%
	UA	97.06%	99.71%	100%	100%	77.92%	79.00%	100%
	OA	98.51%						
	AA	93.81%						
	kappa	0.9764						

2.4. Conclusion

Table 2.11: Confusion matrix obtained on the test set of the Itaipu data set by the purely pixelwise RBF kernel.

	Test map	Classification map						
		urban	herbaceous	shrub-brush	forest	bare soil	built-up	water
Pixelwise RBF kernel	urban	11061	1232	524	208	2250	3703	4
	herbaceous	428	5118	0	0	0	0	0
	shrub-brush	243	0	48310	366	0	0	0
	forest	1	0	57	616	0	0	0
	bare soil	305	15	0	0	2634	901	0
	built-up	258	0	1	0	78	4922	8
	water	1	0	0	0	0	41	91504
	PA	58.27%	92.28%	98.76%	91.39%	68.33%	93.45%	99.95%
	UA	89.95%	80.41%	98.81%	51.76%	53.08%	51.45%	100%
	OA	93.92%						
	AA	86.06%						
	kappa	0.9045						

Chapter 3

Multiresolution classification of panchromatic and multispectral images

3.1 Introduction

Current Earth-observation (EO) missions offer the opportunity to collect data at very diverse spatial resolutions, ranging from a few kilometers to less than one meter, and suitably match the requirements of the desired applications in terms of spatial coverage and geometrical detail [70]. Several spaceborne passive EO systems, such as IKONOS, QuickBird, WorldView-2 and -3, GeoEye-1, SPOT 5 HRG, SPOT 6 and 7, Pleiades, Landsat 7 ETM+, Landsat 8 OLI, and EO-1 ALI or the forthcoming PRISMA mission, jointly acquire both a set of multispectral (MS) channels and an additional panchromatic (PAN) channel, the latter of which exhibits a much coarser spectral resolution than the MS bands, but with a finer spatial resolution [70]. This intrinsic complementarity between PAN and MS data conveys an important potential for image classification (e.g., in land cover mapping applications). However, to exploit this potential, novel multiresolution image classifiers are needed that allow the spectral information associated with the MS channels and the spatial information conveyed by the PAN channel to be fused to generate accurate classification results at the finest available resolution [70, 135].

3.1. Introduction

A popular approach to this multiresolution classification problem is to: (i) generate a common-resolution data set by resorting to down/up-sampling [29,50], pansharpening [8,32,102], or unmixing [112,113,168] procedures; and (ii) apply a conventional supervised classifier to the resulting single-resolution data set. Such strategies are conceptually and computationally straightforward, but the classification results may be affected by artifacts or distortions introduced by resampling, pansharpening, or unmixing processes. In particular, upsampling MS data to the PAN lattice generates a multispectral image with a smaller pixel size but obviously with no improvement in spatial resolution, i.e., in the size of the smallest discernible spatial detail. Its joint use with a PAN image for classification purposes may favor blocky artifacts in the output classification map, especially when nearest-neighbor resampling is used, or possible oversmoothing of small-scale spatial details that are visible in the PAN channel but cannot be appreciated in the upsampled MS channels. In contrast, downsampling PAN data generally causes aliasing issues that require a suitable anti-aliasing low-pass filtering to be applied. The application of unmixing techniques aims at estimating the abundance of each spectrally pure component (e.g., ground materials) in each pixel of an MS image [88]. Approaches have been developed especially in the case of hyperspectral images that allow one to benefit from rich spectral information. These approaches include linear unmixing through convex geometry, statistical Bayesian, and sparse regression or sparse coding methods [14,88], and nonlinear unmixing through radiative transfer theory, bilinear models, ray-tracing and radiosity methods, and neural or kernel techniques [44,63]. The application of unmixing to the joint classification of PAN-MS imagery is expected to be more accurate when the MS image is a hyperspectral image with many spectral channels than when it includes 4 to 8 spectral bands as images from current very high resolution MS-PAN spaceborne cameras (e.g., QuickBird, WorldView-2 and -3) do. Pansharpening techniques use MS and PAN data to synthesize an image that exhibits the spectral content of the MS data with enhanced spatial resolution [6,143]. Several approaches have been proposed including component substitution methods (e.g., IHS color space projection, principal component analysis, and Gram-Schmidt orthogonalization), multiresolution analysis techniques (e.g., using lowpass filters,

3.1. Introduction

Laplacian pyramids, wavelets, and contourlets), Bayesian approaches, compressive sensing, and superresolution algorithms. A recent comprehensive comparison among pansharpening methods can be found in [154]. In general, the application of a supervised single-resolution classifier to a pansharpened product may be affected by the quality of the product in terms of the injection of spatial details from the PAN image to the MS image, and possible artifacts.

Experimental analyses of the impacts of such issues on the classification accuracy can be found in [7, 32, 62, 102, 141, 142]. Similar analyses for multitemporal classification or change detection are described in [17, 86, 166]. Indeed, only a few techniques have been proposed for multiresolution classification *per se*. In [159], the problem is addressed by an object-oriented approach. In [29], two methods are proposed, based on the fusion of posterior probabilities and on a hierarchical top-down algorithm, respectively. In [10], the joint use of Landsat and other resolution data is discussed, and a technique based on a spectral rule is proposed. In [116], a linear mixture model and a Bayesian approach are combined to generate a high-resolution classification map from high-resolution structural information and a coarse-resolution time series. In [135], a Bayesian approach is developed that adapts Markov random fields (MRFs), the expectation-maximization (EM) [114] and Besag's [13] parameter-estimation algorithms, and the iterated conditional mode (ICM) energy-minimization method [84] to address supervised or unsupervised multiresolution classification. A hierarchical MRF is combined in [155] with wavelet transforms, finite mixture models, and copula functions to address the joint classification of images collected by multiple sensors at different resolutions over the same area. In [78] and [79], a graph-theoretic multiresolution technique to jointly segment a set of images acquired at multiple resolutions over the same area is proposed and adapted to the applicative frameworks of urban area classification and landslide extraction. While focusing not on classification but on geophysical parameter regression, a multiresolution Markovian approach is proposed in [71] to spatially downscale land surface temperatures retrieved from coarse resolution geostationary sensors using finer-resolution land cover information.

Here a novel technique based on MRFs and graph cuts is proposed for

3.1. Introduction

multiresolution supervised classification. The method is formulated with regard to the customary case of PAN-MS data, but its extension to more general configurations of input multiresolution data is straightforward. As already introduced in Chapter 1, MRFs allow both spatial context [84] and multiple information sources [98, 128] to be modeled for image-analysis purposes by formulating the “maximum *a-posteriori*” (MAP) criterion as the minimization of a suitable energy function. Several algorithms have been proposed for this minimization task [84]. Simulated annealing (SA) is a stochastic method converging to a global minimum; however, it typically has long computation times [72, 84]. ICM is a deterministic method that requires much shorter times than SA but allows reaching only a local minimum, which may be significantly suboptimal and affected by the initialization [13, 84, 128]. Graph cut approaches have become increasingly popular in the last decade and are based on the reformulation of the minimum-energy problem as a maximum-flow problem over a suitable graph [59]. In the case of binary classification, they allow a global minimum to be reached in polynomial time [59]. With more than two classes, graph cuts converge to a local minimum characterized by strong analytical optimality properties [18, 19, 74].

The proposed method combines a graph cut approach with the linear-mixture model formalized in [135] for the relationships between data at different resolutions to iteratively generate a contextual classification map at the highest spatial resolution available in the input data set (i.e., the resolution of the PAN channel). The estimation of the parameters of the method is performed by extending the iterative techniques proposed in [135] and [121], and based on the EM and Ho-Kashyap’s algorithms, respectively. In particular, in the Appendix of this Chapter an analytical proof is provided which demonstrate that, when all PAN pixels corresponding to the same training MS pixel share the same class label (i.e., when there is no mixed MS pixel in the training set), a simple and intuitive closed-form expression can be derived for the convergence point of EM. This analytical insight operatively allows the iterative EM procedure to be skipped and replaced by a simple one-shot estimator, when the aforementioned condition on mixed MS training pixels holds.

The main novelties consist in the development of a multiresolution super-

3.2. Methodology

vised image classification method that strengthen the multiresolution MRF model and the EM formulation introduced in [135] through the integration of graph cuts (i.e., instead of ICM) and of the MRF parameter-estimation algorithm in [121] (i.e., instead of the Besag’s unsupervised method). Experimental results with both semi-simulated and real (IKONOS and Landsat ETM+) data are presented, with comparisons to previous techniques.

3.2 Methodology

3.2.1 Multiresolution model and assumptions

Let a multiresolution image composed of a PAN channel and of d coarser-resolution MS channels be given. We denote by \mathcal{P} , \mathcal{S} , and ρ the pixel lattices associated with the PAN and MS data and the resolution ratio (i.e., the ratio between the spatial resolutions in meters or centimeters) of the MS and PAN channels, respectively. Such channels are assumed to be properly coregistered and for each $s \in \mathcal{S}$ and $p \in \mathcal{P}$, we write $p \prec s$ (or $s \succ p$) to indicate that the ground area of the s th MS pixel includes the ground area of the p th PAN pixel. Hence, for each MS pixel $s \in \mathcal{S}$, there are exactly ρ^2 PAN pixels $p \in \mathcal{P}$ such that $p \prec s$. Let $\Omega = \{\omega_1, \omega_2, \dots, \omega_M\}$ be the set of thematic classes in the imaged scene, x_s ($x_s \in \mathbb{R}^d$) be the vector of the d channels of the s th MS pixel ($s \in \mathcal{S}$), y_p and ℓ_p ($y_p \in \mathbb{R}, \ell_p \in \Omega$) be the intensity and class label of the p th PAN pixel ($p \in \mathcal{P}$), respectively. Hence, $\mathcal{L} = \{\ell_p\}_{p \in \mathcal{P}}$, $\mathcal{X} = \{x_s\}_{s \in \mathcal{S}}$ and $\mathcal{Y} = \{y_p\}_{p \in \mathcal{P}}$ represent the random fields of the class labels and of the MS and PAN observations. As in [135], we model the relationship between the observations at different resolutions in terms of linear mixtures and the spatial distribution of the class labels as an MRF. Specifically, we state the following assumptions:

1. A random field $\mathcal{V} = \{v_p\}_{p \in \mathcal{P}}$ of “virtual” feature vectors ($v_p \in \mathbb{R}^d$) exists over the PAN lattice, such that the MS random field \mathcal{X} can be obtained from \mathcal{V} through mosaic averaging, i.e.:

$$x_s = \rho^{-2} \sum_{p \prec s} v_p \quad \forall s \in \mathcal{S}. \quad (3.1)$$

3.2. Methodology

2. The joint probability density function (PDF) of the PAN and virtual fields $(\mathcal{Y}, \mathcal{V})$, when conditioned to the label field \mathcal{L} , is Gaussian. Moreover, the virtual vectors, when conditioned to the labels, are independent and identically distributed (i.i.d.).
3. \mathcal{L} is an MRF with respect to a given neighborhood system defined over the PAN lattice. In particular, a second-order Potts MRF model is used, i.e. [84]:

$$P(\ell_p | \ell_q, q \neq p) = P(\ell_p | \ell_q, p \sim q) \propto \exp \left[-\lambda \sum_{q: p \sim q} \delta(\ell_p, \ell_q) \right], \quad (3.2)$$

where $p \sim q$ indicates that p and q are neighbors with respect to the second-order neighborhood $(p, q \in \mathcal{P})$ and λ is a positive parameter.

4. A set $\mathcal{S}^* \subset \mathcal{S}$ of training pixels is available over the MS lattice such that the true class label ℓ_p^* is known for all $p \prec s$ and all $s \in \mathcal{S}^*$.

The virtual random field \mathcal{V} is an image with the same size and resolution of the PAN image (i.e., the finest resolution) and the same number of channels of the MS image. Assumption 1 allows the (actual) data acquired at the MS resolution to be expressed in terms of the (virtual) pixel intensities that are collected in \mathcal{V} at the PAN resolution. A more precise formulation could replace the mosaic average by a more advanced filter, such as the approximate modulation transfer function (MTF) filter [4, 154], to model the spatial behavior of the adopted MS sensor. Here, the mosaic-average formalization is accepted for analytical convenience and to keep the proposed method as general as possible without customizing it to any specific sensor and MTF.

Because of the i.i.d. statement in Assumption 2, the PDF of v_p , given $\ell_p = \omega_i$, is Gaussian $\mathcal{N}(\mu_i, \Sigma_i)$, the parameters μ_i and Σ_i being independent of the pixel location p ($p \in \mathcal{P}; i = 1, 2, \dots, M$). The linearity of the model in (3.1) implies that the joint PDF of \mathcal{X} and \mathcal{Y} , conditioned to \mathcal{L} , is also Gaussian. In particular, the PDF of the feature vector x_s of the s th MS pixel, conditioned to the labels $L_s = \{\ell_p\}_{p \prec s}$ of all the corresponding PAN pixels,

3.2. Methodology

is Gaussian with the following mean and covariance matrix ($s \in \mathcal{S}$) [135]:

$$\begin{aligned}\phi(L_s, \theta) &= E\{x_s | L_s, \theta\} = \rho^{-2} \sum_{p \prec s} \sum_{i=1}^M \delta(\ell_p, \omega_i) \mu_i \\ \Phi(L_s, \theta) &= \text{Cov}\{x_s | L_s, \theta\} = \rho^{-4} \sum_{p \prec s} \sum_{i=1}^M \delta(\ell_p, \omega_i) \Sigma_i,\end{aligned}\tag{3.3}$$

where θ is a vector collecting all μ_i and Σ_i parameters ($i = 1, 2, \dots, M$). It is recalled that Gaussianity is an often accepted model for the class-conditional statistics of multispectral data [70].

Assumption 3 is accepted to exploit MRF modeling to incorporate spatial-contextual information in the classification process, thus favoring the spatial regularity of the classification map and gaining robustness against noise. The Potts model is used because of its analytically simple formulation, to focus on the multiresolution-fusion properties of the proposed approach. However, the method can be readily generalized to include more sophisticated MRF models for \mathcal{L} . As usual with Potts-like models λ plays the role of a spatial smoothing parameter [84].

According to Assumption 4, the true class labels are meant to be known for training samples at the PAN resolution. This is consistent with the goal of generating a classification result at that resolution. Assumption 4 also implies that whenever an MS pixel is in the training set, the class memberships are known for all corresponding PAN pixels. This choice also supports the case of a training set including mixed MS pixels (i.e., MS pixels such that the corresponding PAN pixels have different class labels), provided that these labels are known.

3.2.2 Overview of the proposed method

The goal of the proposed method is to fuse the multiresolution information associated with the PAN-MS data set and the spatial-contextual information modeled through an MRF to generate a classification result at the PAN resolution. For this purpose, the key idea of the method is to iteratively perform a MAP estimation of the virtual channels based on the linear mixture model in (3.1) and on the current classification map to collect the estimated

3.2. Methodology

virtual channels together with the PAN channel in a unique data set. The resulting data set is classified at the finest resolution through graph cuts. The method is iterative and is initialized with a preliminary map generated by applying a basic upsampling procedure to the MS channels. The estimation of θ is integrated in the initialization. More precisely, the proposed method performs the following processing steps, where the superscript “ t ” denotes the quantities computed at the t th iteration.

- Initialization phase – The following operations are performed:
 - i) Estimate θ by applying EM to the training samples (see Section 3.2.3 for details);
 - ii) Generate a preliminary map \mathcal{L}^0 by upsampling the MS channels to the PAN resolution and by performing a conventional graph-cut-based joint classification of the PAN and upsampled MS channels with a Gaussian pixelwise class-conditional model and with $\lambda = 1$ in the Potts MRF [19, 84].
- Iterative phase – The following processing steps are performed at the t th iteration ($t = 0, 1, 2, \dots$):
 1. For each PAN pixel $p \in \mathcal{P}$, compute a MAP estimate v_p^t of v_p , given the feature vector x_s of the MS pixel $s \succ p$ and the current labels in L_s^t (see Section 3.2.4 for details);
 2. For each PAN pixel $p \in \mathcal{P}$, collect v_p^t and the PAN observation y_p in a stacked vector $z_p^t = (v_p^t, y_p)$;
 3. For each class ω_i , compute a sample-mean estimate \hat{m}_i^t of the class-conditional mean $m_i^t = E\{z_p^t | \ell_p = \omega_i\}$ and a sample-covariance estimate \hat{C}_i^t of the class-conditional covariance matrix $C_i^t = \text{Cov}\{z_p^t | \ell_p = \omega_i\}$ by using all stacked vectors z_p^t in the training set of ω_i (i.e., such that $p \prec s$ for some $s \in \mathcal{S}^*$ and $\ell_p^* = \omega_i$);
 4. Compute an estimate λ^t of λ by applying the technique in [121] (see Section 3.2.5 for details);

3.2. Methodology

5. Update the classification map (i.e., compute \mathcal{L}^{t+1}) by minimizing the following energy function:

$$U^t(\mathcal{L}|\mathcal{Z}^t) = - \sum_{p \in \mathcal{P}} \sum_{i=1}^M \delta(\ell_p, \omega_i) \ln g(z_p^t | \hat{m}_i^t, \hat{C}_i^t) + \\ - \lambda^t \sum_{p \sim q} \delta(\ell_p, \ell_q), \quad (3.4)$$

where $\mathcal{Z}^t = \{z_p^t\}_{p \in \mathcal{P}}$ is the random field of the stacked vectors and $g(\cdot|m, C)$ is the Gaussian density with mean m and covariance matrix C (see Section 3.2.5 for details);

6. If the percentage of pixels whose labels are different in \mathcal{L}^{t+1} and \mathcal{L}^t is smaller than a predefined threshold (which is equal to 0.01% in the performed experiments), stop; otherwise, go to step 1.

A flowchart of the method is shown in Figure 3.1. The energy in (3.4) is implicitly related to an i.i.d. class-conditional model for the stacked vectors. The validity of this condition will be discussed in Section 3.2.4.

3.2.3 PDF parameter estimation

The EM-based approach developed in [135] is applied to the training set to estimate θ in step i) of the initialization phase of the proposed method. EM is an iterative parameter-estimation technique proposed for problems characterized by data incompleteness and converging, under suitable assumptions, to maximum-likelihood estimates [114].

This approach is necessary because θ collects the parameters of the Gaussian class-conditional distributions of the virtual feature vectors that are defined in Assumption 1 but cannot be directly observed, i.e., the existence of v_p is postulated in the linear mixture model given in (3.1), while the MS camera provides measurements of the multispectral data x_s ($s \in \mathcal{S}$) and not directly of v_p ($p \in \mathcal{P}$). Therefore, well-known closed-form estimators of the vector mean and the covariance matrix of a multivariate Gaussian distribution from a set of i.i.d. samples cannot be used because input samples of the virtual features are not available.

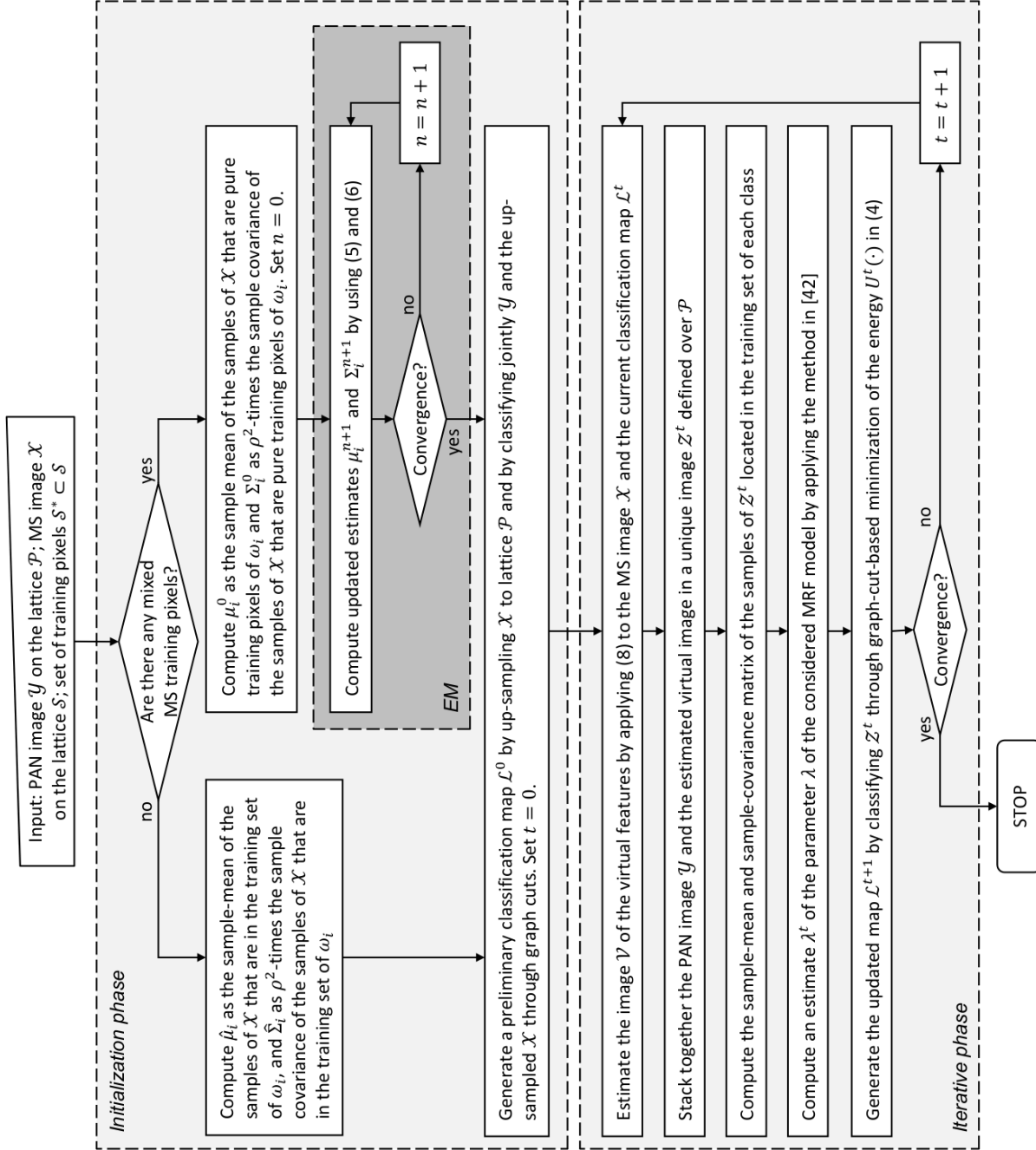


Figure 3.1: Flowchart of the proposed method.

3.2. Methodology

In the initialization phase of the proposed method, EM is initialized with preliminary estimates μ_i^0 and Σ_i^0 ($i = 1, 2, \dots, M$; comments on the choice of this starting point are reported below). Then, denoting the estimates computed at the n th EM iteration by a superscript n , the following update equations hold ($i = 1, 2, \dots, M$) [135]:

$$\begin{aligned}\mu_i^{n+1} &= \frac{\sum_{s \in \mathcal{S}^*} \nu_{si} \eta_{si}^n}{\sum_{s \in \mathcal{S}^*} \nu_{si}}, \\ \Sigma_i^{n+1} &= \frac{\sum_{s \in \mathcal{S}^*} \nu_{si} [\Delta_{si}^n + (\eta_{si}^n - \mu_i^{n+1})(\eta_{si}^n - \mu_i^{n+1})^\top]}{\sum_{s \in \mathcal{S}^*} \nu_{si}},\end{aligned}\tag{3.5}$$

where, for each $s \in \mathcal{S}^*$:

$$\begin{aligned}\eta_{si}^n &= \mu_i^n + \rho^{-2} \Sigma_i^n \Phi(L_s^*, \theta^n)^{-1} [x_s - \phi(L_s^*, \theta^n)] \\ \Delta_{si}^n &= \Sigma_i^n - \rho^{-4} \Sigma_i^n \Phi(L_s^*, \theta^n)^{-1} \Sigma_i^n \\ \nu_{si} &= \sum_{p \prec s} \delta(\ell_p^*, \omega_i),\end{aligned}\tag{3.6}$$

and $L_s^* = \{\ell_p^*\}_{p \prec s}$ is the vector of the (known) class labels of the PAN pixels corresponding to the MS training pixel s . The iteration is stopped when $\|\theta^{n+1} - \theta^n\|$ goes below a predefined threshold (which is set to 0.001). This formulation is feasible when the training set includes either mixed or non-mixed MS pixels. In particular, when there are no mixed training MS pixels, a simple and intuitive closed-form expression holds for the convergence point of EM.

As proven in the Appendix, if the training set includes no mixed MS pixels and if $\bar{\mu}_i$ and $\bar{\Sigma}_i$ are the sample-mean and sample-covariance of the MS training samples of ω_i , then: (i) the sequence $\{\mu_i^n\}_{n=1}^\infty$ is constant and coincides with $\bar{\mu}_i$; and (ii) the sequence $\{\Sigma_i^n\}_{n=1}^\infty$ asymptotically converges to $\rho^2 \bar{\Sigma}_i$ ($i = 1, 2, \dots, M$). This analytical result is intuitively quite appealing. Indeed, if there is no mixed training MS pixel, for each $s \in \mathcal{S}^*$, all PAN pixels $p \prec s$ share the same class label, say ω_i , and therefore, due to (3.3), the PDF of x_s given L_s simplifies to $\mathcal{N}(\mu_i, \rho^{-2} \Sigma_i)$. This would suggest that μ_i could be estimated as $\bar{\mu}_i$ and Σ_i as $\rho^2 \bar{\Sigma}_i$; the result proven in the Appendix states that EM actually converges to this intuitive solution.

Therefore, when the training set includes no mixed MS pixels, no actual iteration is needed, and the one-shot estimates $\bar{\mu}_i$ and $\rho^2 \bar{\Sigma}_i$ ($i = 1, 2, \dots, M$)

3.2. Methodology

can be used. No such simple solution is devised when there are mixed MS pixels in the training set. In this general case, a full EM cycle is needed and the aforementioned theorem suggests a reasonable initialization for EM, i.e., to compute μ_i^0 as the sample mean of the subset of non-mixed (i.e., pure) MS training pixels of ω_i and Σ_i^0 as ρ^2 -times the related sample-covariance.

The formulation of EM in [135] supported both supervised and unsupervised classification and involved averaging in (3.5) over not the training set but all image pixels with their labels iteratively assigned by the classifier. This approach is necessary in the unsupervised case or in a supervised case with a very small training set, but may provide biased estimates [84]. Here, the focus is on the supervised classification and only use a training set in the estimation of θ to minimize the possible estimation bias.

3.2.4 MAP estimation of the virtual feature vectors

As proven in [135], for each $s \in \mathcal{S}$ and each $p \prec s$, the PDF of v_p , given L_s and x_s , is Gaussian with the following mean:

$$E\{v_p|L_s, x_s, \theta\} = \sum_{i=1}^M \delta(\ell_p, \omega_i) \cdot \left\{ \mu_i + \rho^{-2} \Sigma_i \Phi(L_s, \theta)^{-1} [x_s - \phi(L_s, \theta)] \right\}, \quad (3.7)$$

and with a suitable covariance matrix. Due to this Gaussianity property, the MAP estimate of v_p , given x_s and L_s coincides with the conditional expectation in (3.7). Therefore, given the estimate $\hat{\theta}$ computed by EM during the initialization phase, the estimate of the virtual feature vectors computed at the t th iteration of the proposed method is obtained by applying (3.7) as follows:

$$\begin{aligned} v_p^t &= E\{v_p|L_s^t, x_s, \hat{\theta}\} = \\ &= \sum_{i=1}^M \delta(\ell_p^t, \omega_i) \left\{ \hat{\mu}_i + \rho^{-2} \hat{\Sigma}_i \Phi(L_s^t, \hat{\theta})^{-1} [x_s - \phi(L_s^t, \hat{\theta})] \right\}, \\ &\quad \forall s \in \mathcal{S} \quad \forall p \prec s. \end{aligned} \quad (3.8)$$

3.2. Methodology

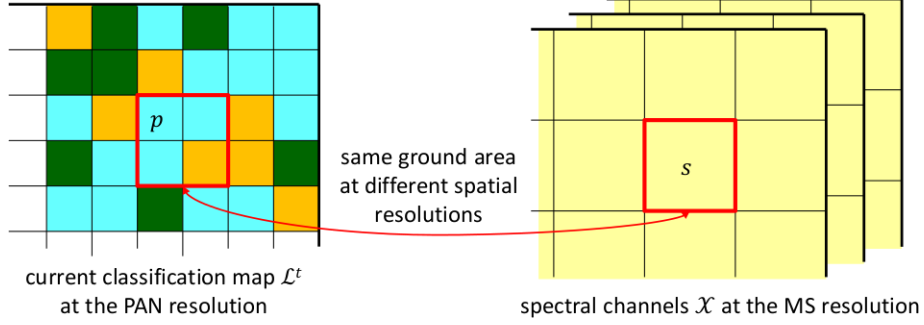


Figure 3.2: An example classification map on a PAN lattice and multispectral observations on an MS lattice in the MAP estimation of the virtual features. In this example, $\rho = 2$. The colors in the classification map denote different class labels.

Therefore, in each iteration of the proposed method, the virtual feature vector of each pixel in the PAN lattice is estimated as a function of the MS observations of the corresponding pixel on the MS lattice and of the current classification labels on the PAN lattice (see Figure 3.2). Because the estimate $\hat{\theta}$ of the parameter vector θ and, consequently, the estimates $\hat{\mu}_i$ of μ_i and $\hat{\Sigma}_i$ of Σ_i ($i = 1, 2, \dots, M$) are fixed after the initialization phase of the proposed method, the field of the estimated virtual features v_p^t ($p \in \mathcal{P}$) evolves due to the evolution of the field \mathcal{L}^t of class labels. Conversely, \mathcal{L}^t is updated in each iteration based on the classification of the stacked image \mathcal{Z}^t using graph cuts, i.e., by operating in the joint domain of the estimated virtual features and the PAN observations. Therefore, as the proposed method iterates, the evolution of the estimates of the virtual features is determined by the evolution of the classification map and vice versa.

The virtual features represent the components of the linear mixture model of the MS channels (see (3.1)). Therefore, when the method converges, the estimated virtual features may be interpreted as estimates of how the MS data would be if they had been acquired at the PAN spatial resolution, given the classification map. This result is formally similar to a pansharpening result, except for one major difference: pansharpening methods are typically unsupervised and synthesize a fused multiresolution image using signal processing concepts, whereas the virtual features involved in the proposed

3.2. Methodology

method (and in [135]) are components of a supervised classification procedure, and their estimates are conditioned to a classification result and the corresponding training samples.

Furthermore, v_p^t is a linear function of x_s and consequently $z_p^t = (v_p^t, y_p)$ is a linear function of (x_s, y_p) . Since the joint PDF of $(\mathcal{X}, \mathcal{Y})$, given \mathcal{L}^t , is Gaussian (see Section 3.2.1), also \mathcal{Z}^t given \mathcal{L}^t and $\hat{\theta}$ is Gaussian-distributed. This confirms the Gaussian class-conditional model involved in the energy function of (3.4). Conversely, the related i.i.d. condition on the stacked vectors, given the labels, is not rigorously fulfilled. For instance, for each $s \in \mathcal{S}$, all vectors z_p^t such that $p \prec s$ depend on x_s and L_s^t and therefore are generally not independent. Similarly, for each $s \in \mathcal{S}$ and $p \prec s$, the PDF of z_p^t , given ℓ_p^t may not be independent of the location p , since z_p^t depends on the labels ℓ_q^t of all other pixels $q \prec s, q \neq p$. Therefore, the i.i.d. model implicit in (3.4) and the related characterization of the PDF of z_p^t given $\ell_p^t = \omega_i$ by a unique mean m_i^t and a unique covariance matrix C_i^t (independent of the location p) play the role of simplifying conditions, which are accepted to favor analytical tractability.

Under these conditions (and not those discussed above for the estimation of θ in the initialization phase of the proposed method), sample-mean \hat{m}_i^t and sample-covariance \hat{C}_i^t estimators of the parameters m_i^t and C_i^t of the class-conditional Gaussian distribution of the stacked vectors z_p^t ($p \in \mathcal{P}$) can be used in step 3 of the proposed method, because, in this step, samples of the stacked vectors are available from the input PAN observations and the MAP estimation of the virtual features. The estimates \hat{m}_i^t and \hat{C}_i^t are updated in each iteration of the proposed method because the samples z_p^t ($p \in \mathcal{P}$) of the corresponding Gaussian distributions vary during the iterations because they include the estimated virtual features v_p^t ($p \in \mathcal{P}$) resulting from step 1.

3.2.5 MRF parameter estimation and energy minimization

The technique in [121] is applied at each iteration of the proposed method to estimate λ . This technique is feasible for MRF models whose energy functions depend linearly on the parameters to be estimated, which is a condition

3.3. Experimental validation

that holds for the relationship between U^t and λ in (3.4). The algorithm in [121] expresses a condition of correct classification of the training set as an overconditioned system of linear inequalities, which is numerically solved through the Ho-Kashyap’s algorithm. Details can be found in [121]. The Ho-Kashyap’s algorithm is iterative and, in each iteration of the proposed method, runs separately until it converges to an updated estimate λ^t of the parameter λ . Convergence is analytically guaranteed and the convergence properties of the Ho-Kashyap numerical procedure are well-known from the literature on supervised linear classifiers [145].

To minimize the energy function in (3.4), the swap-move graph cut algorithm is applied [19]. The graph cut approach is chosen because it represents an effective tradeoff between computational burden and classification accuracy. Compared to ICM, graph cuts are known to significantly improve the energy-minimization solution, without strongly affecting computation time. Compared to SA, they often provide comparable accuracies in remarkably shorter times [136]. The swap-move technique expresses a multiclass energy-minimization problem as a sequence of two-class subproblems, for which global-energy minimum are reached through the graph cut approach [19, 59]. This iterative procedure usually exhibits short convergence times, and the resulting solution can be proven to be a “strong” local minimum with respect to suitable optimality criteria [19].

In [135], the Besag’s and ICM algorithms were used to estimate the MRF parameters and minimize the energy, respectively. Here, the technique in [121] was preferred over the Besag’s method because the latter is an unsupervised algorithm and was found to sometimes provide significantly biased estimates and to underestimate the strengths of spatial-contextual interactions when applied to the supervised case [84, 121].

3.3 Experimental validation

3.3.1 Data sets and experimental setup

Three data sets were used to experimentally validate the proposed technique. The first one is a semi-simulated data set obtained from an originally single-

3.3. Experimental validation

resolution 3-channel 1998×1500 pixel-sized IKONOS image (with spatial resolution of 4 m) of the area of Itaipu (Brazil/Paraguay border; Figure 3.3(a)), including seven classes (Table 3.1). Simulated PAN and lower-resolution MS channels were generated by averaging the IKONOS visible channels and by mosaic-subsampling of the IKONOS spectral channels by a factor ρ , respectively. The cases $\rho = 2, 3$, and 4 were considered because these resolution ratios are used by the recent spaceborne sensors mentioned in Section 3.1. They correspond to 8, 12, and 16-m resolution for the simulated MS channels. Therefore, compared to the original single-resolution IKONOS image, this semi-simulated data set is composed of one PAN channel with a degraded spectral resolution and three MS channels with degraded spatial resolutions. The training set for the seven classes comprises 176688 pixels in the PAN lattice (Table 3.1), which correspond to 44172, 19632, and 11043 pixels on the MS lattice (i.e., an average of 6310.3, 2804.6, or 1577.6 training samples per class on the MS lattice) for the cases $\rho = 2, 3$, and 4, respectively.

Using this semi-simulated data set, a comparison between the results of the proposed multiresolution method and those of a single-resolution classifier applied to the original (non-subsampled) 4 m resolution multispectral image is feasible. The latter plays the role of a benchmark result because it was generated by classifying the original data at its full (spectral and spatial) resolution. This comparison assesses the possible loss in accuracy due to the classification of MS channels with coarser spatial resolutions and a PAN channel with coarser spectral resolution. Indeed, the aforementioned simple approach used to generate synthetic PAN-MS data was employed to make the comparison with the single-resolution benchmark possible without the need for customizing to a specific sensor, and was not meant as an accurate simulation of a PAN acquisition, which should take into account the MTF of the sensor to be simulated, geometrical resampling issues, and the contributions to PAN data resulting from near infrared radiation [154]. To perform a fair comparison, a conventional single-resolution graph cut classifier, based on the Potts MRF, Gaussian class-conditional models for pixelwise statistics of the original multispectral channels, and the swap-move energy minimization technique, was applied to the original IKONOS image. For this benchmark classifier, the spatial-smoothing parameter of the Potts model

3.3. Experimental validation

was estimated using the technique described in [121]. Furthermore, the results of the proposed method were compared to those of the multiresolution technique described in [135].

The other two data sets are real. The first one was acquired by IKONOS over the area of Alessandria (Italy) in 2004, is composed of a 1 m resolution 3000×3000 pixel-sized PAN channel and of four 4 m resolution 750×750 pixel-sized MS channels (Figures 3.5(a) and 3.5(b)), and presents seven classes. The number of training pixels in the PAN lattice was 615632, which corresponds to 38477 pixels in the MS lattice and an average of 5496.7 training samples per class in the MS lattice (Table 3.3). The second one is a Landsat ETM+ data set, collected over Pavia (Italy) in 2000, that comprises a 15 m 3000×3000 pixel-sized PAN channel and six 1500×1500 pixel-sized MS channels (Figures 3.6(a) and 3.6(b)), including five classes (Table 3.4). The size of the training set on the PAN lattice was 67124, which correspond to 16781 MS training pixels and an average of 3356.2 MS training pixels per class.

The results of the proposed technique were compared experimentally with those of the technique described in [135]. For the real data sets, the aforementioned single-resolution benchmark is unfeasible because the image composed of MS channels at the PAN spatial resolution, which was available for the semi-simulated data set, does not exist. On the contrary, in this case, a feasible comparison with a single-resolution classification result was made using a traditional two-step procedure: first, a pansharpening algorithm was used, and then, a single-resolution classifier was applied to the resulting pansharpened image. The well-known Gram-Schmidt (GS) pansharpening algorithm [80, 154] was used in the first step. The aforementioned single-resolution classifier based on graph cuts, the Potts model, Gaussian class-conditional statistics, and the parameter estimation method described in [121] was used in the second step. In particular, an MRF-based classifier was used in this step to ensure a fair comparison with the results of the proposed contextual classifier.

For each data set, the classification results were quantitatively evaluated by computing accuracy parameters with respect to a test set, which was spatially disjoint from the training set to minimize correlation between training

3.3. Experimental validation

and test samples and prevent an optimistic bias in performance assessment (see Figures 3.3(c), 3.5(d), and 3.6(d)). McNemar’s test was applied separately to each pair of classification maps generated for each data set using the aforementioned methods to assess whether the differences in the accuracy of these maps were statistically significant. Given the classification maps generated using the i th and j th methods, the test computes the following standardized normal statistics:

$$Z_{ij} = \frac{f_{ij} - f_{ji}}{\sqrt{f_{ij} + f_{ji}}}, \quad (3.9)$$

where f_{ij} is the number of test samples that are labeled erroneously by the i th method and correctly by the j th method [43, 49]. Using the commonly accepted 5% level of significance, the difference between two results is statistically significant if $|Z_{ij}| > 1.96$ [43, 49]. If this condition is met, a negative or positive value of Z_{ij} indicates that the i th or j th method, respectively, has a higher overall accuracy on the test set.

From an application-oriented viewpoint, the experiments with all three data sets were focused on land cover and not land use mapping problems. Hence, training and test samples referred to land cover classes. The “Itaipu” data set included spatially homogeneous classes (e.g., “water”), textured classes (e.g., “shrub and brush rangeland”), and classes involving geometrical structures (e.g., “urban”). This was also true of the “Alessandria” and “Pavia” data sets. In addition, several groups of classes (e.g., “urban,” “built-up land,” and “barren land” for “Itaipu”; the vegetated classes for “Alessandria”; and “barren land” and “urban” for “Pavia”) were spectrally overlapping in the feature space. Therefore, the overall collection of data sets used for the experimental validation represented a complex testbed for the multiresolution classification of PAN-MS imagery. In the cases of the “Itaipu” and “Alessandria” data sets, “urban” was meant to be discriminated as a whole regardless of the numerous subclasses (e.g., asphalt, concrete, metal, glass, water, grass) that were associated with ground materials and could be appreciated in the input very high resolution images. Accordingly, the training and test areas of “urban” covered these subclasses without distinguishing among them.

For each data set, five iterations sufficed for convergence in all runs of

3.3. Experimental validation

both the proposed method and the technique described in [135].

The next subsections are organized as follows. Experimental results and comparisons using the semi-simulated and real data sets are reported in Sections 3.3.2 and 3.3.3, respectively. The behavior of the proposed method and the algorithm described in [135] as functions of the number of input training samples is discussed in Section 3.3.4. Specifically, in this subsection, two aspects are analyzed experimentally: first, classification performance is evaluated while progressively reducing the number of training samples to appreciate the possible sensitivity to the total training set size. Second, as the numbers of available training samples of the various classes in the three data sets are rather imbalanced (see Tables 3.1, 3.3, and 3.4), a dedicated experiment is performed to investigate possible sensitivity to imbalanced classes.

3.3.2 Classification results and experimental comparisons with the semi-simulated data set

Very accurate classification maps were generated by the proposed method when applied to “Itaipu” in all three subsampled versions (Table 3.1). The overall (OA) and average (AA) accuracies and κ on the test samples were in all cases approximately 99%, 96% and 0.98. Note that, as usual in remote sensing, test samples were not exhaustive and were located inside image regions associated with each class and not at the spatial borders between distinct classes (Figure 3.3(c)). Erroneously classified pixels can be noted in the classification maps especially in such border areas (Figures 3.3(d), 3.4(b), and 3.4(c)). Furthermore, the choice to consider “urban” as a whole class, without distinguishing among its subclasses corresponding to ground materials and structures, implied a strong overlapping, in the feature space of the MS and PAN channels, between “urban” and the classes associated with vegetation and bare soil. Pixels misclassified due to this spectral overlapping can be visually noted in Figures 3.3(d), 3.4(b), and 3.4(c), although they are not included in the test set. When ρ was varied in $[2, 4]$, the results were overall very stable, even when considering individual classes. Slight improvements in the test-set accuracy were noted as ρ was increased. This is also explained by the fact that increasing ρ reduces the noise in the simulated MS channels,

3.3. Experimental validation

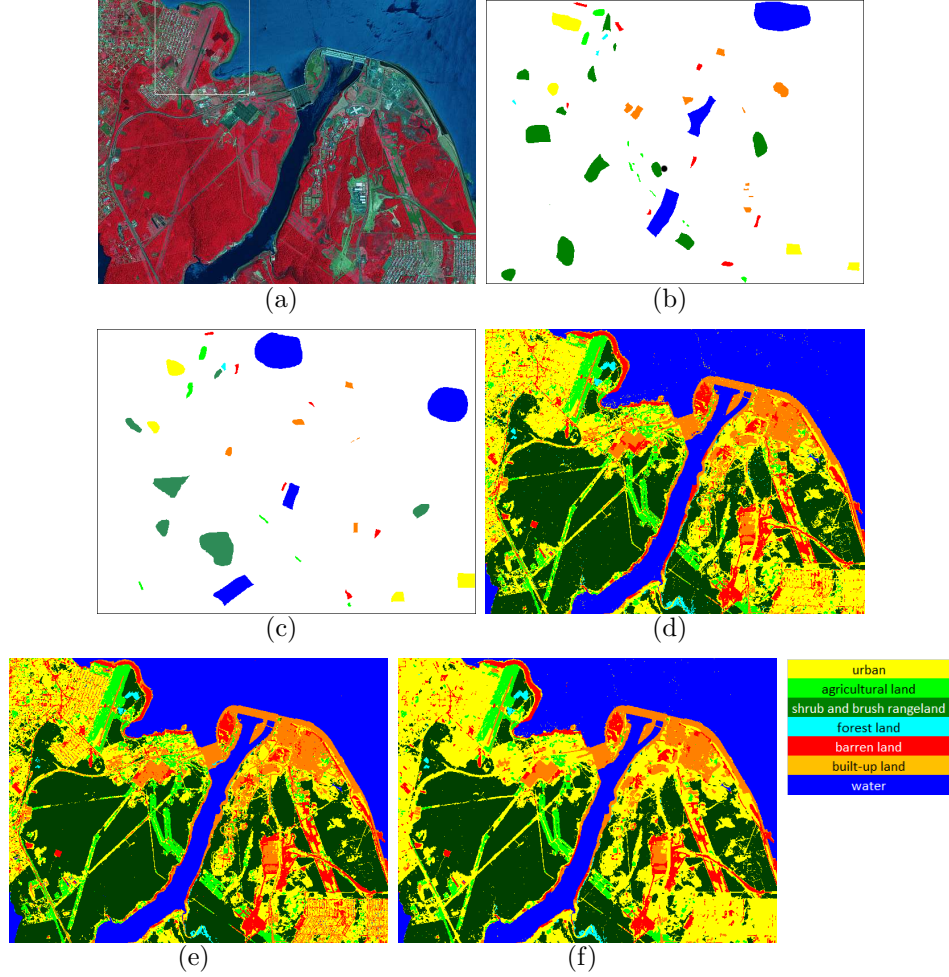


Figure 3.3: “Itaipu”: (a) RGB false-color composition of the IKONOS channels; (b) training map; (c) test map; and classification maps generated by the proposed method (d) and by the technique in [135] (e) (both applied with $\rho = 2$) as well as by a single-resolution Markovian classifier (f). The white box in (a) highlights the detail zoomed in Figure 3.4.

3.3. Experimental validation

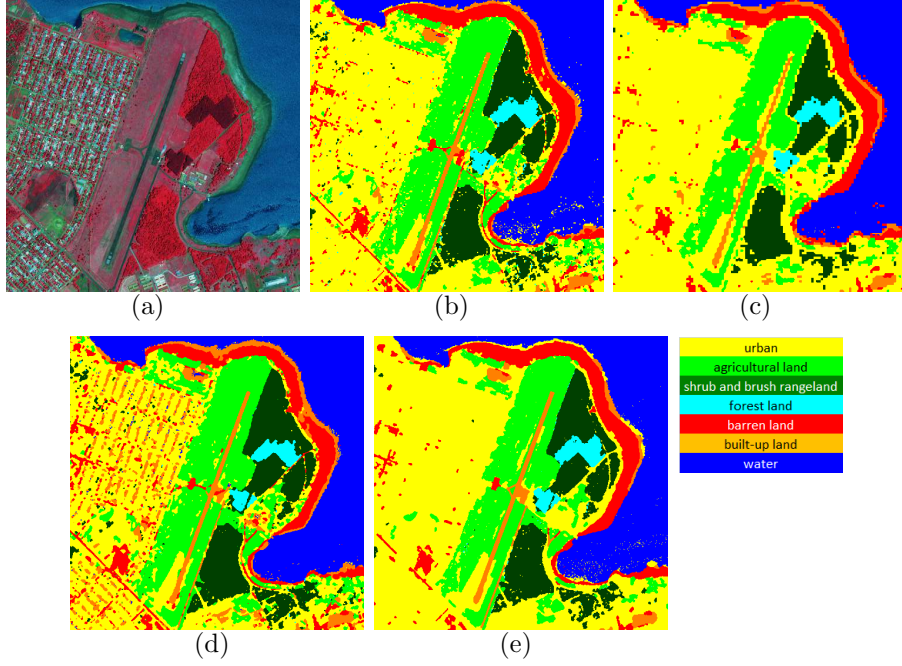


Figure 3.4: “Itaipu”: detail of the RGB false-color composition of the IKONOS channels (a) and of the classification maps generated by the proposed method with $\rho = 2$ (b) and $\rho = 4$ (c), by the technique in [135] with $\rho = 2$ (d), and by a single-resolution Markovian classifier (e).

even at the expense of a possible accuracy loss in the classification of small details and of border areas. A visual comparison of the zoomed image areas in Figures 3.4(b) and 3.4(c) notes a slight blocky effect with $\rho = 4$, but not with $\rho = 2$, which is consistent with the expected spatial degradation for larger resolution ratios.

The result of the benchmark single-resolution classification of the original IKONOS image was very similar to that of the proposed technique (Figures 3.3(d), 3.3(f), 3.4(b), 3.4(c), and 3.4(e)) but had a slightly higher accuracy. The differences between the values of the OA, AA, and κ obtained in the single-resolution run and by the proposed multiresolution method were less than 0.7%, 0.5%, and 0.01 for all $\rho \in [2, 4]$ (Table 3.1). The higher accuracy of this result than of that of the proposed method was expected because the former was from operations on the original IKONOS image at full spectral and spatial resolutions while the latter was obtained after degrading

3.3. Experimental validation

the spatial and spectral resolutions to simulate the MS and PAN data, respectively. Nevertheless, the similarity between the results of the benchmark case and the proposed method suggests that the latter effectively fuses the multiresolution information associated with PAN-MS imagery and generates a classification map with the finest resolution observable in the input data set. The accuracy of the proposed method is similar to that of the ideal situation, in which full spectral information is available at the panchromatic resolution.

When applied to “Itaipu,” the method in [135] also generated accurate maps that were slightly worse than the proposed technique in terms of accuracy on the test set. As ρ was varied in $[2, 4]$, the differences between the values of OA, AA, and κ of the two approaches ranged approximately in $[1\%, 3\%]$, $[2\%, 5\%]$, and $[0.01, 0.05]$, respectively. Compared to the previous technique, the proposed method yielded a large improvement in the accuracy for “urban” (+30%, 8% and 12% when $\rho = 2, 3$, and 4, respectively) and very small ($\leq 0.5\%$) reductions in the accuracies for “shrub and brush rangeland,” “forest land” for $\rho = 3$ and 4, “built-up land,” and “water.” A visual analysis of the maps in Figures 3.3(d) and 3.3(e) and of the detail in Figure 3.4(d) confirms these comments and notes a stronger confusion between “urban,” “built-up,” and “barren land” and a more noisy classification result for the method in [135] than for the proposed algorithm. These results are likely caused by the improved effectiveness of the proposed formulation, which uses graph cuts, the MRF parameter estimation algorithm in [121], and the MAP estimation of the virtual features, compared to the technique in [135], which is based on the ICM and Besag’s algorithms. The times taken by both methods, when run on a 2.33-GHz 4-GB RAM hardware configuration with a CPU-based non-parallelized implementation, ranged between 13 and 24 min and were overall comparable: a shorter time was required by the proposed method when $\rho = 2$ and by the method in [135] when $\rho = 4$, while both techniques took almost the same time when $\rho = 3$.

For all of the values of ρ considered, the differences among the accuracies of the results of the proposed method, the technique described in [135], and the single-resolution benchmark were all statistically significant according to McNemar’s test. In particular, values of Z_{ij} larger than 35 were obtained for

3.3. Experimental validation

Table 3.1: “Itaipu”: numbers of training and test samples on the PAN lattice; classification accuracies on the test set and computation times of the proposed method, of the technique in [135] and of a single-resolution graph-cut-based Markovian classifier.

class	training samples	test samples	$\rho = 2$		$\rho = 3$		$\rho = 4$		single- resolution
			proposed	[135]	proposed	[135]	proposed	[135]	
urban	19584	18864	95.11%	65.51%	97.65%	89.51%	98.40%	86.42%	98.79%
herbaceous rangeland	6048	5472	99.01%	98.76%	98.52%	96.16%	97.62%	93.38%	99.03%
shrub and brush rangeland	61056	48816	99.69%	99.97%	99.78%	99.97%	99.90%	99.95%	99.81%
forest land	1008	576	100%	96.14%	99.55%	100%	99.70%	100%	100%
barren land	6336	3744	80.86%	76.37%	81.76%	76.89%	81.43%	77.85%	79.43%
built-up land	12960	5184	97.15%	97.68%	96.98%	96.03%	97.85%	96.11%	100%
water	69696	91440	99.89%	99.98%	99.94%	99.98%	99.96%	99.97%	99.94%
overall accuracy			98.78%	95.59%	99.11%	98.09%	99.23%	97.68%	99.30%
average accuracy			95.96%	90.63%	96.31%	94.08%	96.41%	93.38%	96.71%
κ			0.9808	0.9306	0.9859	0.9699	0.9878	0.9634	0.9890
time [min]			16	24	19	20	24	13	4

3.3. Experimental validation

the difference between the proposed and previous multiresolution methods for all of the resolution ratios considered (see Table 3.2).

3.3.3 Classification results and experimental comparisons with the real data sets

When applied to “Alessandria” and “Pavia,” the proposed method obtained very high accuracies (i.e., above 97%) on the test samples of all classes (see Tables 3.3 and 3.4). The aforementioned comments about the location of test samples not at the spatial borders between different classes hold in these cases as well. This further suggests the effectiveness of the method in fusing multiresolution information for classification purposes in the application to diverse sensors (ETM+ and IKONOS) and resolutions (15-30 m and 1-4 m). The computation times, on the same hardware-software platform mentioned above, of the proposed method in the applications to “Pavia” and “Alessandria” (i.e., around 1 hour, and 1 hour and 10 min, respectively) were longer than in the experiments with the semi-simulated data sets due to the larger sizes of the two real data sets compared to the “Itaipu” images. However, these times are considered acceptable in the framework of laboratory experiments on land cover mapping applications. Note also that even though the resolution ratio for “Alessandria” is 4, no blocky artifacts resulted from the application of the proposed method (Figure 3.7(c)), possibly because of the intrinsic spatial smoothing introduced by the multispectral sensor.

The technique described in [135] resulted in a classification map for the “Pavia” data set (Figures 3.6(e) and 3.6(f)) that was similar to the map of the proposed method except for slightly lower values of the OA, AA, and κ and lower accuracies for all classes except “water.” However, a considerably longer execution time (more than two hours; Table 3.4) was required by this method than by the proposed technique. This computational difference may be explained by the fact that a full cycle of EM involving all image pixels is run by the method in [135] at each ICM iteration, whereas the proposed algorithm performs only one EM cycle, which is incorporated in the initialization phase and involves only the training samples.

In the case of “Alessandria,” very similar values of OA and κ were reached

Table 3.2: The results of McNemar’s significance test comparing the classification results in Tables 3.1, 3.3, and 3.4. “SR” denotes the single-resolution classification of the “Itaipu” data set. “PS + SR” denotes the pansharpening and single-resolution classification of the “Alessandria” and “Pavia” data sets. For each data set, the (i, j) entry of the corresponding table denotes the value of McNemar’s statistic Z_{ij} for method i and method j .

data set		Itaipu ($\rho = 2$)			Itaipu ($\rho = 3$)			Itaipu ($\rho = 4$)		
method		proposed	[135]	SR	proposed	[135]	SR	proposed	[135]	SR
proposed		–	–68.83	25.44	–	–35.36	11.12	–	–49.49	4.28
[135]		68.83	–	78.36	35.36	–	40.18	49.49	–	47.4
SR		–25.44	–78.36	–	–11.12	–40.18	–	–4.28	–47.4	–
data set		Alessandria			Pavia					
method		proposed	[135]	PS + SR	proposed	[135]	PS + SR			
proposed		–	4.69	–54.47	–	–12.92	–48.19			
[135]		–4.69	–	–41.95	12.92	–	–55.83			
PS + SR		54.47	41.95	–	48.19	55.83	–			

3.3. Experimental validation

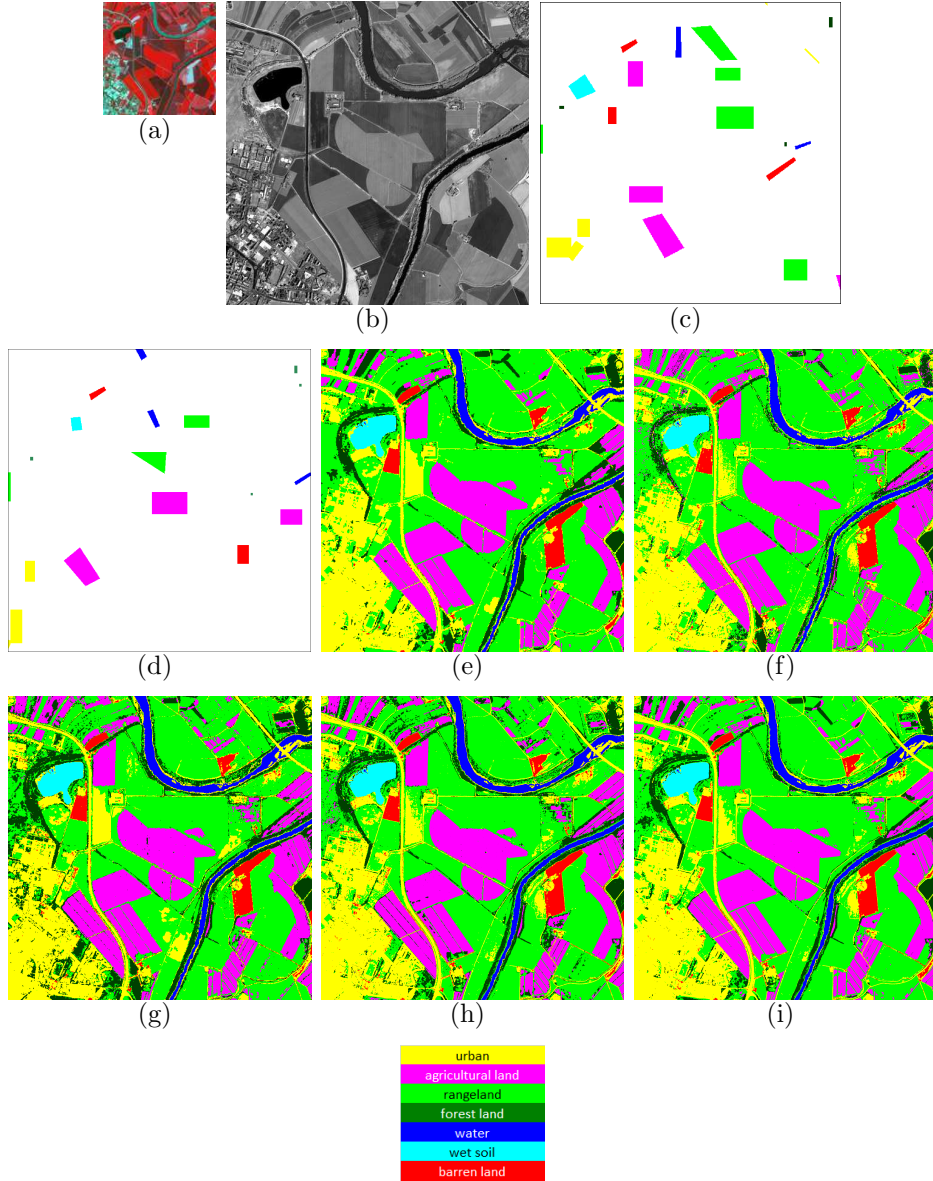


Figure 3.5: “Alessandria”: (a) RGB false-color composition of the MS channels, (b) PAN channel, (c) training map, (d) test map, classification maps generated by the proposed method (e), by the technique in [135] (f), by a single-resolution classification of the GS pansharpening result (g), by a single-resolution classification of the BSD pansharpening result (h), and by a single-resolution classification of the GSA pansharpening result (i).

3.3. Experimental validation

Table 3.3: “Alessandria”: number of training and test samples on the PAN lattice; classification accuracies on the test set and computation times of the proposed method, of the previous technique in [135], and of the application of a single-resolution graph cut classifier after pansharpening.

class	training samples	test samples	proposed method	method in [135]	pansharpening and single-res classification		
					Gram-Schmidt	BDS	GSA
urban	89248	61008	98.86%	99.05%	96.67%	98.69%	99.18%
agricultural land	186736	174352	97.54%	98.10%	97.48%	96.95%	97.41%
rangeland	235440	77328	99.55%	99.42%	95.09%	98.88%	98.97%
forest land	6160	4288	100%	81.90%	99.60%	97.97%	97.69%
water	21616	22656	99.51%	99.84%	99.72%	99.82%	99.84%
wet soil	36880	12928	100%	100%	100%	100%	100%
barren land	39552	29936	100%	100%	100%	100%	100%
overall accuracy			98.58%	98.65%	97.31%	98.14%	98.45%
average accuracy			99.35%	96.90%	98.37%	98.90%	99.01%
κ			0.9802	0.9812	0.9626	0.9742	0.9784
time [min]			70	51	14	16	16

by both methods with slightly higher values for the technique in [135]. Slightly better discrimination was obtained using this technique for the “urban,” “agricultural,” and “water” classes; the differences in accuracy were less than 0.2% compared to the proposed method. In contrast, the value of the AA was 2.5% higher for the proposed algorithm than for the algorithm described in [135]. This result is due to a sharp improvement (approximately +18%) in the accuracy of the proposed method for “forest” and is interpreted as a consequence of the improved energy-minimization capabilities granted by the graph cut approach, compared to ICM. The increased accuracy for “forest” is also visually confirmed by the maps in Figures 3.5(e) and 3.5(f) and in the zoomed area in Figure 3.7. The execution time of the algorithm described in [135] was less than it was for the proposed algorithm but both times were, overall, approximately one hour.

The application of a single-resolution graph cut classifier after Gram-Schmidt pansharpening led to accurate classification results even though the values of the OA, AA, and κ were less than those of the proposed method for the “Pavia” and “Alessandria” data sets (see Figures 3.5(g) and 3.6(g)). In the former case, the differences between the OAs, AAs, and κ ’s of the two

3.3. Experimental validation

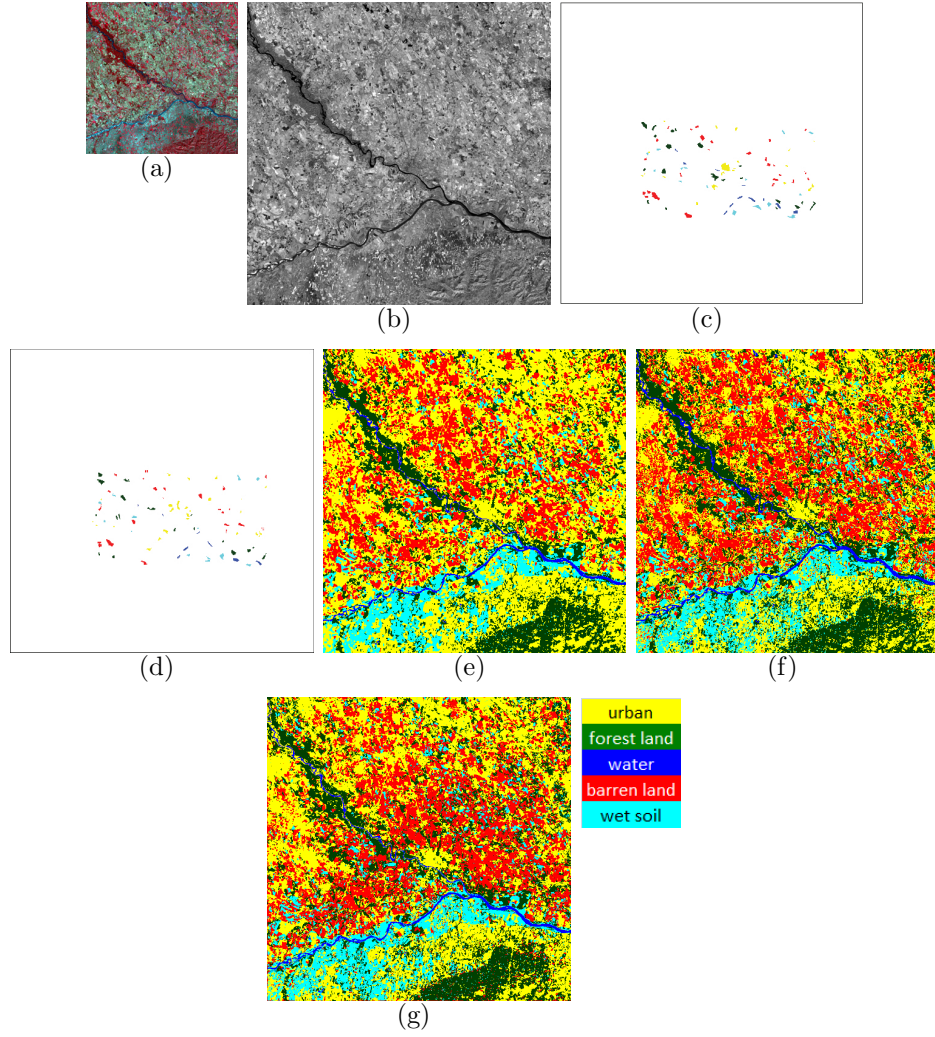


Figure 3.6: “Pavia”: (a) RGB false-color composition of the MS channels, (b) PAN channel, (c) training map, (d) test map, and classification maps generated by the proposed method (e), by the technique in [135] (f), and by a single-resolution classification of the pansharpening result (g).

3.3. Experimental validation

methods were approximately 0.05-0.06, and the proposed technique resulted in improved accuracies for all of the classes, especially “barren land” (around +7%) and “wet soil” (approximately +10%). This result is due to a reduction in the amount of confusion between “barren land” and “urban” and between “wet soil” and “water.” For the “Alessandria” data set, the improvement in the OA, AA, and κ obtained using the proposed method compared to the classification of the pansharpening result was approximately 0.01-0.02. Larger improvements were obtained in discriminating “urban” (approximately +2%) and “rangeland” (approximately +4.5%). This is confirmed by a visual comparison of the corresponding classification maps in Figures 3.5(e) and 3.5(g), which show more confusion and less regularity in the urban area detected by single-resolution classification after pansharpening. The improved accuracy of the proposed method compared to the single-resolution classification of the pansharpening result is interpreted as being because the multiresolution model used by the former is optimized using class information from the MAP estimate of the virtual features. In contrast, pansharpening generates a fused image on the basis of signal processing concepts without explicitly targeting a given supervised classification problem. This improvement in accuracy was obtained at the expense of an increased computational burden, because the pansharpening and single-resolution classification procedures together took approximately 14 and 9 minutes for the “Alessandria” and “Pavia” data sets, respectively. These results suggest a tradeoff between the applications of a single-resolution classifier after pansharpening and of the proposed multiresolution classifier. For the considered data sets, the latter option consistently gave longer execution times and higher accuracies. On one hand, the aforementioned differences in accuracy were significant, especially for certain classes, although not dramatic. On the other hand, as previously mentioned, the longer time required by the proposed approach, which was due to its iterative stage, was compatible with the usual timeline of land cover mapping applications.

For both of the real data sets, the accuracy differences among all of the aforementioned classification results were statistically significant based on McNemar’s test (see Table 3.2).

Specifically for the “Alessandria” data set, an additional experimental

3.3. Experimental validation

session was performed with the aim to further investigate the comparison between the performances obtained by the proposed method and the performances resulting from the single-resolution classification of a pansharpened image. Indeed, even though the application of the Gram-Schmidt algorithm allows the achievement of accurate results, two different techniques based on the component-substitution (CS) approach, namely band-dependent spatial detail (BDSD) and adaptive Gram-Schmidt (GSA), represent the most powerful approaches for pansharpening in the specific application to IKONOS imagery [154]. The former consists in a linear injection model in which for each multispectral band an optimal detail image is extracted from the panchromatic channel [54]. The latter is make use of a linear regression algorithm for the adaptive estimation of the coefficients for the generation of the equivalent panchromatic image [5]. The adoption of BSDS and GSA pansharpening techniques allowed an accuracy improvement with respect to the application of the same single-resolution classifier to the Gram-Schmidt pansharpened image with differences in OA, AA, and k of approximately 1% in both BSDS and GSA cases. In particular better discrimination was obtained for the “urban” (approximately +2%) and “rangeland” classes (approximately +4% for BSDS and +3% for GSA), while a decrease of around 2% was observed for the forest land class. Very similar accuracy values were reached by the proposed method and by the single-resolution classifier applied to the pansharpened image through BSDS and GSA. The differences between the OAs, AAs and ks of the aforementioned approaches were less than 1%, with slightly higher values obtained by the proposed method. A larger improvement was reached in discriminating the “forest land” class (more than 2%). This result can be also appreciated through a visual analysis of the corresponding classification map (see Figures 3.5(h) and 3.5(i)) which shows a more homogeneous area related to this land cover class. Similarly, the classification maps obtained by the proposed approach led to more regularity and less confusion in the detection of the “urban” area.

3.3. Experimental validation

Table 3.4: “Pavia”: number of training samples on the PAN lattice; classification accuracies on the test set and computation times of the proposed method, of the previous technique in [135], and of the application of a single-resolution graph cut classifier after pansharpening.

class	training samples	test samples	proposed method	method in [135]	pansharpening and single- resolution classification
urban	12880	11100	99.59%	97.61%	96.72%
forest land	19080	13128	99.92%	99.70%	97.88%
water	6832	4624	98.14%	99.44%	95.48%
barren land	18152	15008	99.02%	98.37%	92.10%
wet soil	10180	7840	99.87%	99.66%	89.58%
overall accuracy			99.42%	98.84%	94.48%
average accuracy			99.31%	98.96%	94.35%
κ			0.9925	0.985	0.9287
time [min]			62	139	9

3.3.4 Dependence on training set size

To assess the dependence of the performances of the proposed approach and the method described in [135] on the total number of training samples, both methods were applied to each of the aforementioned data sets while progressively subsampling the corresponding training sets. For each data set, the total number of training samples was iteratively reduced, first, by removing a randomly selected 10% of the original training set for each run until only 10% of the training samples were left. Then, denoting as N the average number of MS training pixels per class, each training set was further iteratively reduced to 8%, 6%, 4%, 2%, 1%, and 0.5% of its original size by stopping when N was approximately 30. For both the proposed technique and the algorithm described in [135], Figure 3.8 shows the behavior of the OA and AA as functions of N for the “Itaipu” with $\rho = 2$, “Pavia,” and “Alessandria” data sets. The plots for κ and for the “Itaipu” data set with $\rho = 3$ and 4 are similar to those shown and omitted for brevity.

Both of the multiresolution classification methods exhibited remarkable robustness to the reduction of the training set size. Specifically, for the

3.3. Experimental validation

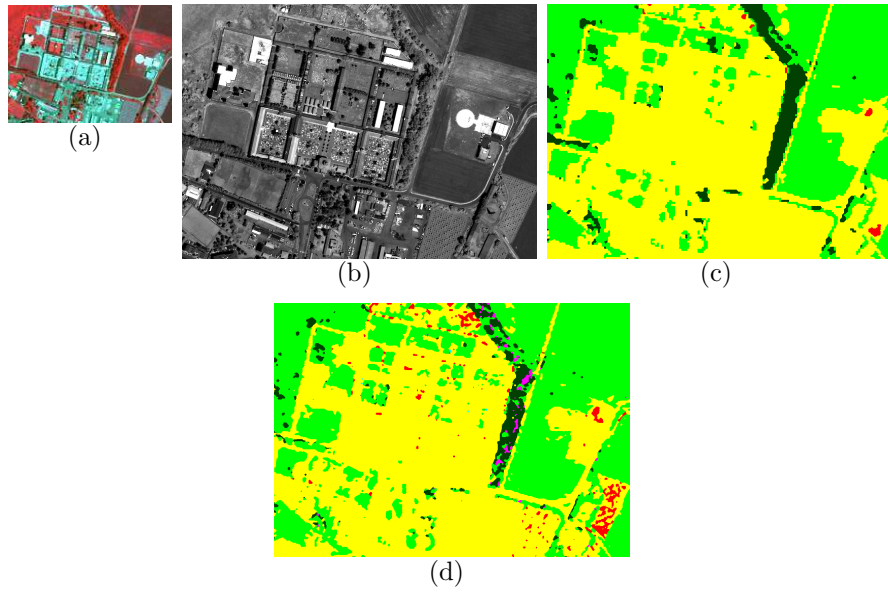


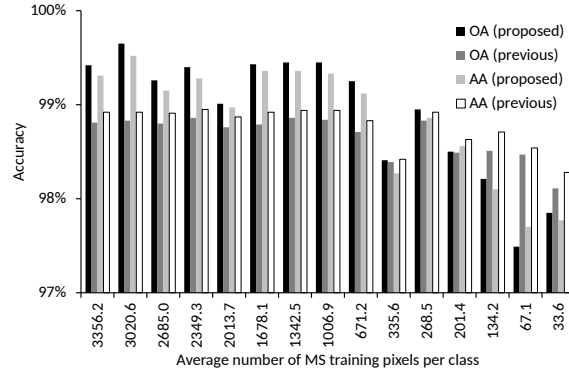
Figure 3.7: “Alessandria”: detail of the RGB false-color composition of the MS channels (a), of the PAN channel (b), and of the classification maps generated by the proposed method (c) and by the technique in [135] (d). The color legend is the same as in Fig 3.5.

3.3. Experimental validation

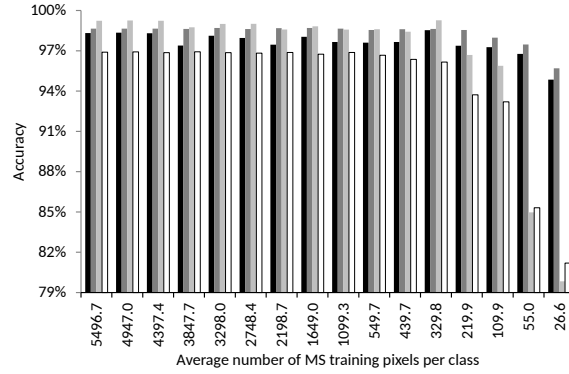
“Itaipu” data set with $\rho = 2$, the values of the OA and the AA for both methods were almost constant as long as N was at least approximately 500. The OA remained larger than 96% and 94.5% for the proposed and previous methods, respectively, for all values of N . Smaller values of the AA were obtained when $N \leq 126.1$ because of the low accuracy of “barren land,” which was due to its very small number of training pixels (there were, e.g., only 28 PAN training pixels of this class when $N = 29.1$) and its spectral overlapping with “urban” and “built-up land.” Nevertheless, for all values of N , the proposed method was more accurate than the method described in [135]. The averages over all runs of the differences between the OAs and AAs of the two algorithms were 2.7% and 4.6%, respectively. These results are consistent with those described in Section 3.3.2 and show the improvement in the effectiveness of the energy minimization and parameter optimization stages of the method developed here compared to the previous algorithm. Furthermore, the stability of the accuracies over a large range of values of N suggests that the proposed approach, which is based on parametric modeling through Gaussian processes and spatial-contextual labeling through MRFs and graph cuts, is remarkably robust for the application with rather small sample sizes.

Similar stability was observed in the results obtained for the “Pavia” and “Alessandria” data sets. In particular, for the “Pavia” data set, the values of the OA and AA of the two methods ranged from 97.5% to 99.7% as N was reduced to 33.6. Slightly higher accuracies were obtained using the proposed method for $N \geq 671.6$ and using the technique in [135] for $N \leq 134.2$, but the average over all runs of the differences between the OAs and AAs of the two algorithms were approximately 0.2% and 0.1%. For the “Alessandria” data set, the accuracies of both methods were quite stable for $N \geq 109.9$. For such values of N , the proposed method resulted in larger values of the AA than the previous method, which resulted in larger values of the OA. For smaller values of N , both methods resulted in $AA < 85\%$. The averages over all runs of the differences between the OAs and AAs of the proposed and previous methods were (-0.7%) and 1.9% , respectively. Indeed, as in Section 3.3.3 and Table 3.3, the proposed algorithm yielded a large increase in accuracy for “forest land” (which is endowed with the smallest number of test samples of this data set), compared to the previous technique at the

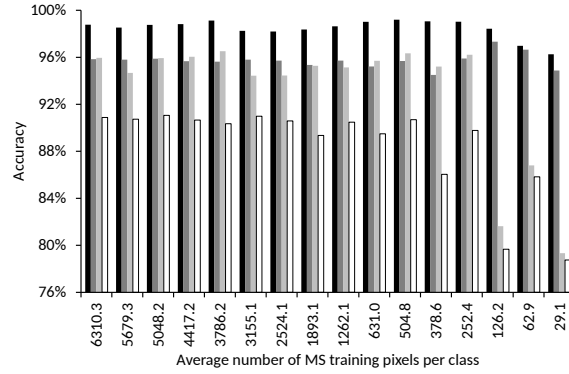
3.3. Experimental validation



(a) Pavia



(b) Alessandria



(c) Itaipu ($\rho = 2$)

Figure 3.8: The behavior of the overall (OA) and average (AA) accuracies of the proposed method and the method described in [135], as functions of the number of training samples for (a) the “Pavia” data set; (b) the “Alessandria” data set; and (c) the “Itaipu” data set with $\rho = 2$. The same legend is used for all the plots.

3.3. Experimental validation

Table 3.5: Behaviors of the OA, AA, and κ of the proposed method and of the technique in [135] after subsampling, for each data set, the size of the set of training samples of each class down to approximately the size of the smallest class. Means and standard deviations over 10 runs are shown.

data set	method	OA		AA		κ	
		mean	std. dev.	mean	std. dev.	mean	std. dev.
Itaipu ($\rho = 2$)	proposed	98.86%	0.29%	95.85%	0.67%	0.9819	0.0045
	[135]	94.00%	0.34%	86.60%	1.42%	0.9058	0.0053
Itaipu ($\rho = 3$)	proposed	98.64%	0.20%	95.17%	0.48%	0.9786	0.0031
	[135]	97.36%	0.36%	92.25%	0.88%	0.9583	0.0057
Itaipu ($\rho = 4$)	proposed	98.56%	0.16%	94.86%	0.34%	0.9773	0.0025
	[135]	96.48%	0.21%	90.11%	0.80%	0.9446	0.0034
Alessandria	proposed	97.29%	0.64%	98.59%	0.26%	0.9626	0.0087
	[135]	97.92%	0.39%	98.59%	0.14%	0.9713	0.0056
Pavia	proposed	99.61%	0.02%	99.52%	0.03%	0.9949	0.0003
	[135]	98.51%	0.11%	98.66%	0.09%	0.9807	0.0014

expense of small reductions in the accuracies for other classes, for which larger sets of test pixels were available.

Furthermore, the sizes of the sets of training samples of the classes were quite imbalanced for all three data sets (see Tables 3.1, 3.3, and 3.4). On one hand, the use of parametric Gaussian models for parameter estimation and training purposes is expected to make the proposed method not critically sensitive to imbalanced classes, especially as compared to non-parametric classifiers such as support vector machines. On the other hand, to quantitatively appreciate the possible sensitivity, a dedicated experiment was performed with each data set by reducing the size of the set of training samples of each class to approximately the size of the smallest of such sets, i.e., by balancing all the classes to the training sample size of the smallest one. This subsampling was performed randomly and repeated 10 times. Table 3.5 shows the means and the standard deviations of the values of OA, AA, and κ over these 10 runs for all data sets and for both the proposed method and the technique described in [135]. In the case of the proposed algorithm, the mean OA, AA, and κ were very similar to the values obtained using the original training set size, and the corresponding standard deviations were smaller than 0.01,

3.4. Conclusion

thus confirming the limited sensitivity of the method to imbalanced classes. Similar comments hold for the algorithm described in [135] as well, although it obtained lower mean values of OA, AA, and κ for all data sets except “Alessandria”. In the case of “Alessandria,” the two methods obtained the same AA, while the technique in [135] achieved slightly higher values of OA and κ . McNemar’s test can also be generally biased by the use of imbalanced classes. The results (not detailed for brevity) of the application of McNemar’s procedure after also balancing the test sample sizes of the classes to the size of the smallest class confirmed again that the differences in accuracy between the developed technique and the method in [135] were significant.

3.4 Conclusion

A novel method has been proposed for the supervised classification of multiresolution images made of a higher-resolution PAN channel and of several coarser-resolution MS channels. The method extends and improves the approach developed in [135] by integrating the multiresolution linear-mixture model and the formulation of EM in [135], graph-cut-based energy minimization, MAP estimation, and the supervised MRF parameter optimization algorithm in [121]. The properties of the adopted EM estimator have also been analytically investigated by proving that in a special case, a simple closed-form expression for the convergence point of EM can be obtained.

Experiments on semi-simulated and real data collected by two different spaceborne multiresolution sensors showed the capability of the method to generate accurate classification maps at the PAN resolution, thus confirming the effectiveness of the technique in fusing the spatial detail associated with the PAN channel and the spectral information conveyed by the MS data. When applied to a semi-simulated data set, whose PAN and MS channels were generated by spatially or spectrally downsampling a given multispectral image, the proposed method obtained results very similar to those given by a benchmark contextual classification of the original non-downsampled data. Moreover, the proposed method was experimentally compared with the previous multiresolution classification algorithm described in [135]. This algorithm, which generally supports supervised and unsupervised classifica-

3.4. Conclusion

tion and was applied in a supervised formulation here, also generated accurate classification maps when applied to either semi-simulated or real data. However, the proposed method exhibited higher average accuracies and/or shorter computation times than the previous method for all of the data sets considered. The overall accuracy was higher for all of the results of the proposed method than for those of the method described in [135] except for the “Alessandria” data set, for which the difference in the OA was less than 0.1%. These results are consistent with the capabilities of graph cut techniques to converge to either global energy minimum or local minimum satisfying strong optimality conditions, compared to the ICM strategy adopted in [135], which ensures convergence to a local minimum with no prescribed optimality property.

The results of applying the proposed method to real data sets were also compared to those obtained applying a single-resolution graph cut classifier after pansharpening. On one hand, higher accuracies, especially for certain classes, were obtained by the proposed method than by proceeding through pansharpening. This result suggests the capability of the method to optimize the fusion of data acquired at different spatial resolutions with respect to the specific problem of discriminating between a given collection of thematic classes characterized through training samples. On the other hand, the combination of single-resolution classification and pansharpening was consistently faster than the proposed multiresolution classification technique, which performs energy minimization through graph cuts and Bayesian estimation on every iteration. However, all execution times, evaluated using a CPU-based non-parallelized implementation, were compatible with the usual timeline of land cover mapping applications. Substantial reductions in execution time, especially in applications to very large input images, could be obtained by using GPU-based parallelized implementations, thanks to the well-known parallelizability properties of MRF-based classifiers [138].

The experimental results also remarked that increasing the ratio between the resolutions of the PAN and MS data had a minor impact on the classification accuracy, at least when this ratio was between 2 and 4. Such ratios are usual for current PAN-MS sensors. When dealing with the semi-simulated data set, a larger ratio yielded slightly improved numerical accuracies due to

3.4. Conclusion

the reduced impact of noise in the MS channels; however, a small number of blocky artifacts were observed at the spatial boundaries between different classes. Indeed, no such artifacts were remarked in the application to a real IKONOS data set with the same resolution ratio.

When applied after reducing the training set size until only approximately an average of 30 training samples per class were left on the MS lattice, the accuracy of the proposed method exhibited remarkable stability. This suggests that the proposed formulation based on Gaussian random fields, MRFs and graph cuts can effectively benefit from parametric and spatial-contextual modeling to minimize the requirements for training samples. Indeed, compared to the previous technique described in [135], similar behaviors were noted when using full and reduced training sets. For almost all of the subsampled training sets, improved accuracy was obtained by the proposed method for the semi-simulated data set, and very similar accuracies were obtained by the two approaches when they were applied to the “Pavia” real data set. For almost all of the training set sizes using the “Alessandria” real data set, the proposed method significantly improved the discrimination of a spatially textured class compared to the previous method at the expense of minor accuracy reductions for other classes. The proposed method also exhibited robustness to possible issues associated with imbalanced classes. For all data sets, the sizes of the sets of training samples of the classes were quite imbalanced. However, when applied after balancing all the classes to approximately the training sample size of the smallest one, the proposed technique obtained accuracies very similar to those achieved using the original training set.

The proposed classifier is iterative, and the experiments have shown good convergent behavior in a small number of iterations. As for the method in [135], case-specific convergence properties have not been analytically proven so far and represent an interesting issue worth being investigated. Furthermore, the method was formalized and tested in the framework of the classification of PAN-MS images, as this data typology represents a very typical case of multiresolution remote-sensing imagery. However, generalizing the method to the cases in which either more than two resolutions are involved or the finest-resolution data are multichannel is straightforward. Experimen-

3.4. Conclusion

tally validating such generalizations will be an interesting extension of this research as well. Furthermore, the proposed method is aimed at generating a classification map at the finest observed resolution, i.e., on the PAN lattice, but it would be interesting to extend it to generating classification results not only at the PAN resolution but also at the MS resolution. A simple way to accomplish this is to perform, for each pixel in the MS lattice, majority voting using the class labels determined for the corresponding pixels in the PAN lattice. However, more sophisticated decision fusion strategies [75] that consider, for example, the energy contributions associated with these PAN pixels in the adopted MRF model, are also worth investigating. A combination of the proposed method, which is based on a linear mixture model for the multiresolution input data, and hierarchical MRF models for class labels at multiple spatial resolutions [155] could also be an interesting extension. Other important generalizations aimed at further increasing accuracy, especially in the discrimination of urban areas, will be to adaptively incorporate edge (e.g., line processes [84]), segmentation [95], or texture information in the adopted MRF model [98].

Appendix

The aforementioned analytical proof about the convergence of the considered EM formulation is provided. This proof demonstrates that, under the assumptions and notations introduced in Section 3.2, if the training set includes no mixed MS pixels, then the following properties hold for the sequences $\{\mu_i^n\}_{n=1}^\infty$ and $\{\Sigma_i^n\}_{n=1}^\infty$ of parameter estimates computed by EM during the initialization phase of the proposed method:

- $\mu_i^n = \bar{\mu}_i$ for all $n = 1, 2, \dots$
- $\Sigma_i^n \longrightarrow \rho^2 \bar{\Sigma}_i$ for $n \longrightarrow +\infty$,

where $\bar{\mu}_i$ and $\bar{\Sigma}_i$ are the sample-mean and sample-covariance matrix of the MS training pixels of ω_i ($i = 1, 2, \dots, M$), respectively.

Because there are no mixed MS pixels in the training set, each training MS sample can be assigned a unique class label. This means that \mathcal{S}^* can be

3.4. Conclusion

partitioned into M subsets $\mathcal{S}_1^*, \mathcal{S}_2^*, \dots, \mathcal{S}_M^* \subset \mathcal{S}$, such that, for all $s \in \mathcal{S}_i^*$, the condition $\ell_p^* = \omega_i$ holds for all $p \prec s$ ($i = 1, 2, \dots, M$). Therefore, focusing on the n th EM iteration and the i th class ω_i ($i = 1, 2, \dots, M; n = 1, 2, \dots$), (3.3) and (3.6) imply that, for each $s \in \mathcal{S}_i^*$:

$$\begin{aligned} \phi(L_s^*, \theta^n) &= \mu_i^n, & \Phi(L_s^*, \theta^n) &= \rho^{-2} \Sigma_i^n, \\ \eta_{si}^n &= x_s, & \Delta_{si}^n &= (1 - \rho^{-2}) \Sigma_i^n. \end{aligned} \quad (3.10)$$

Similarly, we have:

$$\nu_{si} = \begin{cases} \rho^2 & \text{for } s \in \mathcal{S}_i^* \\ 0 & \text{for } s \in \mathcal{S}^* - \mathcal{S}_i^*. \end{cases} \quad (3.11)$$

Plugging these relationships into (3.5), we obtain:

$$\mu_i^{n+1} = \bar{\mu}_i, \quad \Sigma_i^{n+1} = (1 - \rho^{-2}) \Sigma_i^n + \bar{\Sigma}_i, \quad (3.12)$$

where:

$$\bar{\mu}_i = \frac{1}{N_i} \sum_{s \in \mathcal{S}_i^*} x_s, \quad \bar{\Sigma}_i = \frac{1}{N_i} \sum_{s \in \mathcal{S}_i^*} (x_s - \bar{\mu}_i)(x_s - \bar{\mu}_i)^\top \quad (3.13)$$

are the sample-mean and sample-covariance matrix of the training MS pixels of ω_i , respectively, and N_i is the number of such pixels (i.e., the cardinality of \mathcal{S}_i^*). Eq. (3.12) implies that the sequence $\{\mu_i^n\}_{n=1}^\infty$ is constant with the value $\bar{\mu}_i$ and that the sequence $\{\Sigma_i^n\}_{n=1}^\infty$ satisfies a first-order linear difference equation. It is easily proven by induction that the solution of this equation is ($n = 1, 2, \dots$):

$$\Sigma_i^n = (1 - \rho^{-2})^n \Sigma_i^0 + \bar{\Sigma}_i \sum_{k=0}^{n-1} (1 - \rho^{-2})^k, \quad (3.14)$$

where Σ_i^0 is the initialization matrix. By the definition of the resolution ratio, we have $\rho > 1$, which implies $1 - \rho^{-2} \in (0, 1)$. Therefore, for all initialization matrices, the EM sequence of covariance-matrix estimates converges to:

$$\lim_{n \rightarrow +\infty} \Sigma_i^n = \bar{\Sigma}_i \sum_{k=0}^{\infty} (1 - \rho^{-2})^k = \bar{\Sigma}_i \cdot \frac{1}{1 - (1 - \rho^{-2})} = \rho^2 \bar{\Sigma}_i. \quad (3.15)$$

This concludes the proof.

Chapter 4

Supervised classification of thermal infrared hyperspectral images

4.1 Introduction

The work presented in this chapter differs with respect to the previously introduced innovative classification approach. Indeed, here the aim is to explore the potential of pattern recognition approaches in the application to long-wave infrared (LWIR) hyperspectral image classification. Although in literature this innovative kind of remote sensing data has rarely been taken into account for the generation of land covers maps, the interest for thermal infrared hyperspectral images is considerably growing in the remote sensing community. The reasons of the increasing attentions lies in the appealing properties characterizing LWIR data, i.e. insensitivity to Sun illumination (day and night operations), reduced sensitivity to smoke and mist as compared to visible sensors, and complementarity with respect to visible and near infrared (VNIR) imagery. While VNIR images has been intensely studied for long in the literature [81], only a few recent works address the classification problem related to LWIR hyperspectral data. Specifically, in [125] a LWIR image composed of 32 bands was used jointly with emissivity based information in order to develop a clustering method for man-made object.

4.1. Introduction

The halo effect, which characterizes very hot or cold objects in the image, is exploited for the development of a segmentation algorithm. Then, considering foreground pixels containing only man-made objects, vegetation in the background is clustered through statistics of emissivity vectors and stochastic expectation maximization is trained on the obtained segmented data. In [124] the aim is to perform change detection taking advantage of a novel "Normalized Difference Thermal Index". The index proposed in the paper is developed in order to favor the discrimination between man-made objects and other structures. In this work, four methodological approaches (three classification and one feature reduction methods) have been considered and validated through experiments with a real high-resolution LWIR hyperspectral data set acquired over a complex urban and vegetated scene. The main interest of this work lies in the validation of powerful image classification approaches [98] [81] [122], which were previously found accurate in the case of VNIR hyperspectral images, in the application to this innovative type of remote sensing data. Here, a pattern recognition and image processing perspective is taken to experimentally discuss the potential and limitations in the application to LWIR hyperspectral data. As discussed in [125] [124], the alternate approach of physically-based methods, which incorporate prior information on LWIR image formation and may involve temperature-emissivity separation [37], usually demonstrate effective for the classification of LWIR hyperspectral data. Specifically, the considered classifiers are:

- a well-known non-contextual Bayesian classifier with Gaussian class-conditional distributions [81];
- a contextual classifier based on a Markov random field (MRF) model for the spatial neighborhood information [98] [72];
- a Markovian multiscale region-based classifier [98].

Since LWIR hyperspectral sensors are normally airborne and the resulting spatial resolutions can be very high, special focus has been put to integrating spatial information into the classification process by including MRF models (methods (ii) and (iii)) and region-based processing (method (iii)). Furthermore, the application of Bayesian classifiers to hyperspectral images

is generally sensitive to the Hughes' phenomenon, which occurs when the training set size is not large enough to ensure a reliable estimation of the classifier parameters [81]. Therefore, methods (i) and (ii) have also been applied in conjunction with a feature reduction technique [81] [122]. The considered approach is the sequential forward band extraction (SFBE) [122], which projects the original feature space onto a lower dimensional subspace by computing, from the original hyperspectral bands, a collection of synthetic multispectral bands optimized with respect to the considered classification problem.

4.2 Methodology

4.2.1 Sequential forward band extraction

Let a LWIR hyperspectral image containing n channels (h-bands) be available along with a training map associated with a set Ω of C information classes. SFBE was proposed and validated in the case of VNIR hyperspectral images [122]. Here, its possible relevance for feature reduction from LWIR hyperspectral data is experimentally discussed. SFBE computes a set \mathfrak{X} of m transformed features ($m \leq n$), that emulate the spectral bands (s-bands) of a synthetic multispectral sensor and are optimized in order to maximize an interclass distance measure $\mathcal{J}(\mathfrak{X})$ [122]. Specifically, the Jeffries-Matusita distance is used because of its relationship to the error probability of Bayesian non-contextual classifiers [23]. The distance is computed using a Gaussian model for each class-conditional distribution, a well-known model for the statistics of image data collected by passive sensors [81]. The class parameters are estimated as sample means and sample covariance matrices on the training samples. The solution \mathfrak{X} obtained by SFBE is sub-optimal with respect to the maximization of $\mathcal{J}(\cdot)$ but generally identifies a set of s-bands that well discriminate the considered classes. Further details can be found in [122].

4.2.2 Contextual and non-contextual classifiers

Let \mathcal{I} be the pixel lattice of a LWIR hyperspectral image, \mathfrak{X} be the set of m s-bands extracted by SFBE, $x_i \in \mathbb{R}^m$ be the vector of the s-bands (feature vector) of the i th pixel ($i \in \mathcal{I}$), $y_i \in \Omega$ be its class label, $\mathcal{X} = \{x_i\}_{i \in \mathcal{I}}$ and $\mathcal{Y} = \{y_i\}_{i \in \mathcal{I}}$ be the random fields of the feature vectors and class labels, respectively.

The first considered classification method is a pixelwise Bayesian maximum *a-posteriori* (MAP) classifier with Gaussian class-conditional distributions, chosen as a well-known and consolidated non-contextual classifier for hyperspectral data in conjunction with feature reduction [81]. It is equivalent to minimizing the following energy function:

$$U(\mathcal{Y} | \mathcal{X}) = - \sum_{i \in \mathcal{I}} \ln p(x_i | y_i) - \sum_{i \in \mathcal{I}} \ln P(y_i), \quad (4.1)$$

where the prior probability $P(y_i = \omega)$ of each class ω and the parameters of the Gaussian ω -conditional density $p(x_i | y_i = \omega)$ ($\omega \in \Omega$) are estimated as the related relative frequency, sample mean, and sample covariance matrix on the training samples.

The second classifier incorporates local contextual information through an MRF model. MRFs characterize, in a mathematically rigorous way, the relationships between spatially local and global properties of the image statistics [72]. Thanks to the Hammersley-Clifford theorem, it is possible to prove that, if \mathcal{Y} is an MRF, then the global (image-wise) MAP rule is equivalent to minimizing an energy function defined locally according to a neighborhood system [72]. Here, the Potts MRF is used to characterize spatial information, and the following energy is minimized:

$$U(\mathcal{Y} | \mathcal{X}) = - \sum_{i \in \mathcal{I}} \ln p(x_i | y_i) - \beta \sum_{i \sim j} \delta(y_i, y_j), \quad (4.2)$$

where $\delta(\cdot)$ is the Kronecker symbol, the pixelwise (unary) term is associated with the same Gaussian multivariate class-conditional distributions as in (4.1), $i \sim j$ indicates that the i th and j th pixels are neighbors ($i, j \in \mathcal{I}$), and β is a positive parameter weighting the contextual (binary) and unary terms.

4.3. Experimental validation

The third considered classifier is the MRF- and region-based method in [98] [95]. Given a set of K segmentation maps, generated on the basis of the input image and associated with different spatial scales, the method minimizes the following Markovian energy function:

$$U(\mathcal{Y} | \mathcal{X}) = - \sum_{i \in \mathcal{I}} \sum_{k=1}^K \alpha_k \ln P(s_{ik} | y_i) - \beta \sum_{i \sim j} \delta(y_i, y_j), \quad (4.3)$$

where $P(s_{ik} | y_i)$ is the class-conditional distribution of the segment labels s_{ik} of the pixels in the segmentation map at the k th scale ($k = 1, 2, \dots, K; i \in \mathcal{I}$), the last additive term corresponds again to the Potts model, α_k and β are positive weight parameters. The method is aimed at fusing spatial information related to both the local neighborhood of each pixel and to its membership to homogeneous regions. A multiscale approach is adopted, i.e., multiple segmentation maps, ranging from a few coarse-scale regions to many fine-scale regions, are used jointly. The graph-based region-growing method in [48] is used to generate these maps. Additional details can be found in [98] [95].

Both energies (4.2) and (4.3) are minimized using iterated conditional mode (ICM), a deterministic method that usually exhibits short execution times and converges to a local minimum. The weight parameters (i.e., $\alpha_k, k = 1, 2, \dots, K$, and β) are optimized using the technique in [121], which is based on the Ho-Kashyap numerical algorithm.

4.3 Experimental validation

The data used for experiments are obtained from the `grss_dfc_2014` data set of the 2014 IEEE GRSS Data Fusion Contest [85]. It includes an 84-channel LWIR hyperspectral image with 1-m spatial resolution (see Figure 4.1), acquired using the Hyper-Cam imager, which is based on a Fourier-transform spectrometer. The `grss_dfc_2014` data set was composed of both this image and of a visible RGB image at finer spatial resolution: the latter was not used in the present experimental analysis to focus entirely on the LWIR hyperspectral component. This experimental protocol made the land cover mapping problem especially challenging because the 7 classes of this data set

4.3. Experimental validation

(see Table 4.1), which included three roof classes (“red,” “grey,” and “concrete roof”) and two vegetated classes (“low vegetation” and “trees”), were drawn mostly on the basis of visible data and can hardly be discriminated when only thermal data are used. Accordingly, experiments with both this set of 7 classes and a set of 4 classes obtained by merging “low vegetation” and “trees” into one “vegetation” class and the three roof types into one “roof” class, were performed.

SFBE was applied to extract m s-bands with $2 \leq m \leq 82$. Figure 4.2 shows the behavior, as a function of m , of the overall accuracy (OA) of the Gaussian Bayesian classifier of (4.1) and of the MRF-Based Gaussian classifier of (4.2) on the test set, when applied to classify within the aforementioned 4 classes in the resulting m -dimensional feature space. As m increased, first a sharp increase in OA was obtained by both classifiers, and then, a saturating behavior was remarked. This is consistent with the results that are usually obtained when feature reduction and Bayesian classifiers are applied to VNIR hyperspectral images [72]. For this data set, both classifiers obtained their maximum values of OA on the test set when they were applied to the original 84-dimensional space. This result may not hold in general and suggests that the number of training samples of this specific data set was enough to compute accurate estimates of the class parameters without a severe impact of the Hughes’ phenomenon. Higher values of OA were obtained for all values of m by the MRF-based than by the non-contextual classifier. This is consistent with the capability of the former to incorporate spatial neighborhood information in the classification of a high spatial resolution image. The behavior of OA as a function of m in the case of 7 classes was similar (although with lower values of OA; see also below), and is omitted for brevity.

The confusion matrices obtained on the test set by the MRF-based Gaussian classifier in the cases of 7 and 4 classes are reported in Tables 4.1 and 4.2, respectively. In either case, the confusion matrix obtained using the number m of s-bands corresponding to the maximum value of OA is shown. When all the 7 classes in the `grss_dfc_2014` data set were used, substantial confusion among the three classes describing different kinds of roof was remarked (see Table 4.1). Indeed, these classes correspond to similarly shaped objects that differ in VNIR reflectance but may generally not exhibit large

4.3. Experimental validation

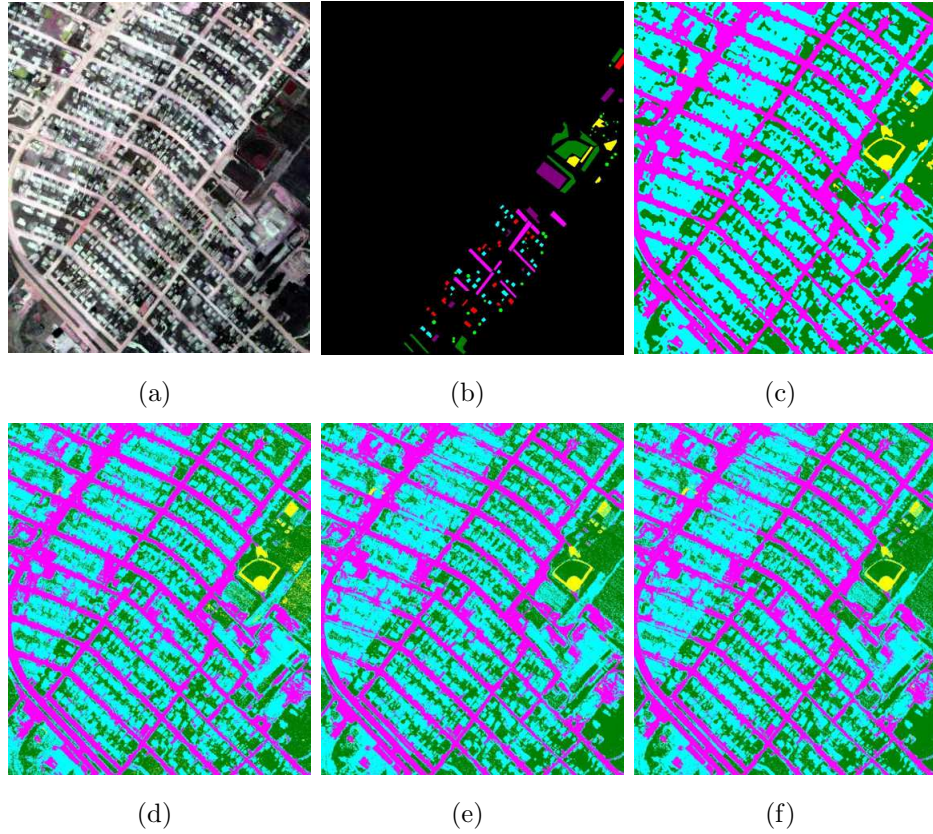


Figure 4.1: False color composite of bands no. 1, 40, and 75 of the LWIR hyperspectral image (a), training map drawn mostly on the basis of visible data (b), classification map obtained by the MRF-based Gaussian classifier (c), non-contextual Gaussian Bayesian classifier (d), contextual SVM classifier (e), and non-contextual SVM classifier (f). See Tables 4.2 and 4.3 for the color legend.

4.3. Experimental validation

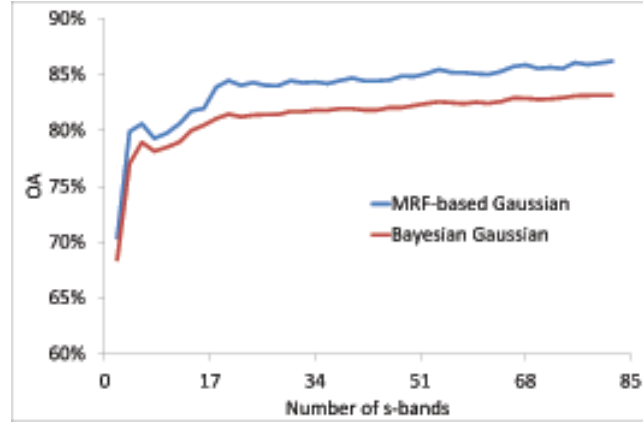


Figure 4.2: Behavior of the OA of the non-contextual Gaussian Bayesian and of the MRF-based Gaussian classifiers as a function of the number of s-bands extracted by SFBE (4 classes).

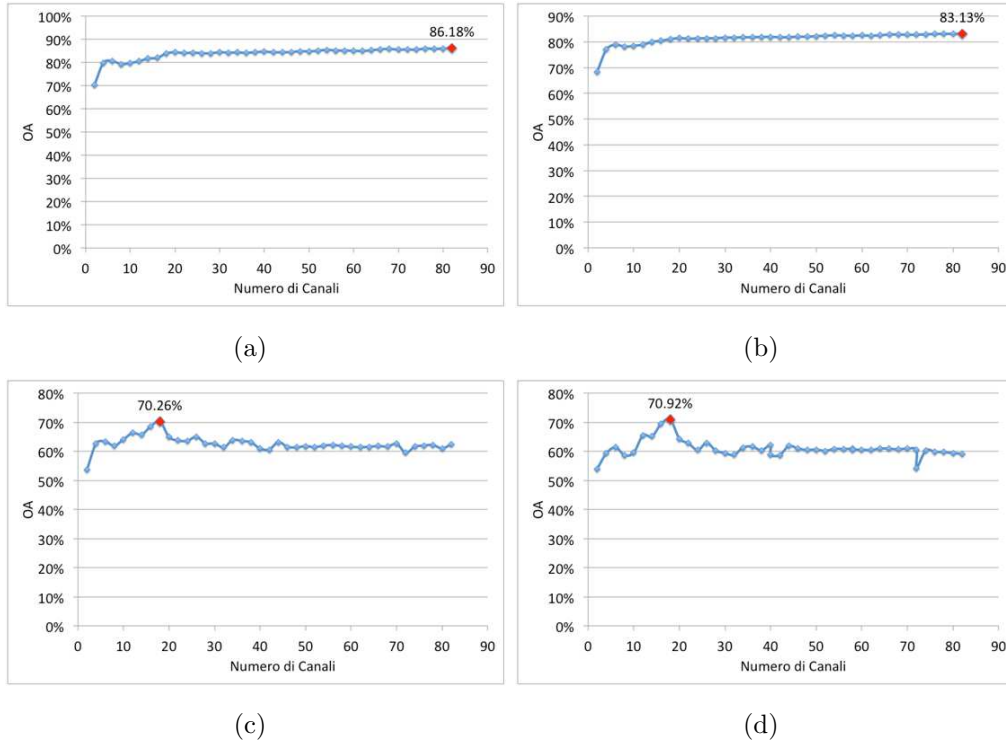


Figure 4.3: Feature reduction effects on the MRF-based Gaussian classifier (c), non-contextual Gaussian Bayesian classifier (d), contextual SVM classifier (e), and non-contextual SVM classifier (f).

4.3. Experimental validation

differences in emissivity or temperature, a scenario that makes it hard to distinguish among them, even though local spatial context is used through an MRF model. Similar comments hold with regard to the two vegetated classes. Discrimination between “bare soil” and “low vegetation” was also poor in the case of the pixel-based classifiers of (4.1) and (4.2). The comparison between the confusion matrices in the cases of 7 and 4 classes (see Tables 4.1 and 4.2) confirms that, when operating with 7 classes, a substantial amount of labeling errors was due to confusions within roof and vegetated cover types, whereas a unique roof class and a unique vegetated class could be more accurately discriminated using the considered LWIR hyperspectral image although no VNIR data were employed (see also Figure (4.1)).

In Table 4.3, the accuracies of all the three considered classifiers, when applied to the 4-class case, are shown (the results with 7 classes are again omitted for brevity). The segmentation method in [48] was applied to the original 84-band LWIR image to generate 5 segmentation maps corresponding to different spatial scales. As discussed in [95], the Markovian region-based classifier of (4.3) is usually not critically sensitive to the number of scales considered, provided that at least one very fine scale and one coarse scale segmentation results are employed. Increased values of OA and of the average accuracy (AA) were obtained by the MRF-based classifier of (4.2) as compared to the non-contextual method of (4.1) and by the multiscale region-based technique of (4.3) as compared to both pixel-based methods. This is consistent with the capability of the method in [98] [95] to integrate spatial information associated with both the local neighborhood of each pixel and homogeneous regions in the input image. When region-based processing was used, the accuracies in the discrimination of “bare soil” and “vegetation” increased of 37.4% and 8.4%, respectively, as compared to the use of the pixelwise MRF-based classifier, thus suggesting that confusion between bare and vegetated covers was reduced. However, less precise discrimination was obtained for “roof” by the classifier in (4.3) than by the one in (4.2), although still maintaining a satisfactory accuracy.

4.3. Experimental validation

Table 4.1: Confusion matrix of the MRF-based Gaussian classifier on the test set, in the application to 7 classes.

Test map	Classification map						
	road	red roof	grey roof	concrete roof	low vegetation	trees	bare soil
road	10220	6	8	26	5	5	0
red roof	347	898	612	37	31	0	0
grey roof	173	905	1191	58	25	0	0
concrete roof	191	157	365	96	0	0	0
low vegetation	164	49	22	179	1311	8	0
trees	65	92	272	219	613	2	0
bare soil	16	0	0	45	971	0	918

Table 4.2: Confusion matrix of the MRF-based Gaussian classifier on the test set, in the application to 4 classes.

Test map	Classification map			
	urban	vegetation	roof	bare soil
road	10137	3	130	0
vegetation	179	1836	981	0
roof	423	36	4627	0
bare soil	15	956	83	896

4.4. Conclusion

Table 4.3: Classification accuracies of the considered classifiers on the test set.

Test map	Classification map		
	Non-contextual Gaussian Bayesian	MRF-based Gaussian	Markovian region- based multiscale
road	97.9%	98.7%	98.7%
vegetation	53.8%	61.3%	69.7%
roof	85.0%	91.0%	80.0%
bare soil	45.5%	46.0%	83.3%
OA	83.1%	86.2%	88.3%
AA	70.6%	74.2%	82.9%

4.4 Conclusion

The potential and limitations of three approaches to supervised classification, including pixel-based, region-based, Markovian, and multiscale processing, have been experimentally analyzed in the application to thermal infrared hyperspectral data. Experimental results with a challenging data set pointed out that rather high accuracy was obtained when considering “high level” thematic classes associated with roads, vegetation, bare soil, and roofs. On one hand, it is worth noting that the input image was collected during day-time [85], thus making temperature relevant in the discrimination of thematic classes whose emissivities could be hardly distinguished otherwise. On the other hand, it was not possible to effectively discriminate more detailed classes, such as different types of vegetation or roofs made of different materials, due to the strong spectral overlapping in the feature space of the thermal hyperspectral channels. Significantly higher accuracies could be obtained through the fusion of LWIR and visible higher resolution imagery [85], although the latter was not used in the present study. Furthermore, in the case of the considered data set, the use of feature reduction, which is generally a major tool in the analysis of hyperspectral data, did not allow gains in

4.4. Conclusion

labeling performance to be achieved as compared to classifying in the original feature space directly. Further experiments, based on support vector classifiers and not reported for brevity, did not provide improved classification as compared to the aforementioned methods.

On one hand, the conducted experiments overall suggested the effectiveness of the pattern recognition approach in the application to LWIR hyperspectral image classification. On the other hand, the combination of this approach and of physically-based models describing the behaviors of temperature and emissivity [37] could allow the information brought about by LWIR hyperspectral data to be further and more accurately exploited.

Chapter 5

Parameter optimization for Markov random field models for remote sensing image classification

5.1 Introduction

As mentioned in the Introduction of this thesis, the most proposed method here for remote sensing application are rooted in the theory of Markov random fields models (MRFs). However, the objective of the research presented in this chapter is not to propose a classification approach but rather to deal with a key aspect related to the application of MRFs in the classification process, namely the energy minimization task. Indeed MRFs, owing to the Hammersley-Clifford theorem, allow formalizing Bayesian decision rules as the minimization of suitable energy functions [72]. On one hand, the application of the maximum *a-posteriori* decision rule to a global probabilistic model associated with an entire image would be computationally intractable due to the huge number of unknown variables. On the other hand, under Markovian assumptions, the minimum-energy reformulation, which is linked to a spatial neighborhood system defined on the pixel lattice, turns out to be computationally affordable thanks to efficient algorithms such as

5.1. Introduction

graph cuts or iterated conditional mode (ICM) [72] [98]. Indeed, the energy function of an MRF model generally includes parameters, whose values may critically affect the performances of the resulting classification schemes in terms of accuracy, spatial smoothing and anisotropy, and sensitivity to small-scale details and thin structures [66] [121]. The choice of appropriate values for these parameters is often accomplished through manual trial-and-error procedures that are usually time-consuming and possibly affected by empirical choices performed by the user/operator. Therefore, automatic techniques to optimize the parameters of MRF models represent important tools for image classification. This problem has been addressed mostly in the case of unsupervised classification and more scarcely in the case of supervised image classification. Direct maximum likelihood estimation of MRF model parameters is usually computationally unfeasible except for very specific MRF models [66]. Alternate approaches for unsupervised classification include pseudo-likelihood functions, which represent computationally feasible approximations of the log-likelihood but often lead to underestimating the strengths of spatial inter-pixel interactions [66] [121]; stochastic gradient and Monte Carlo techniques, which incorporate stochastic sampling processing steps [66]; and mean square error (MSE) approaches, which use least squares fitting to address suitable sets of linear algebraic conditions on the unknown parameters [42]. In the case of supervised classification, MSE, genetic, and case-specific heuristic algorithms have been proposed [72] [121]. Here, a novel parameter optimization method is proposed for MRF models for supervised classification. The method is focused on MRFs such that the energy functions can be expressed as linear combinations of different contributions (e.g., resulting from multiple information sources in a multisource fusion scheme) and the unknown parameters are the weights of these linear combinations. This methodological assumption is quite general and encompasses many well-known models developed for spatial-contextual [72], multisensor [121], and multitemporal image classification [91]. The proposed technique is based on the MSE approach and on the sequential minimal optimization (SMO) algorithm for the resolution of a suitable quadratic programming (QP) problem. The original SMO algorithm in [105] and its variants (e.g., [47]) have gained substantial popularity in recent years due to their successful applications to

the QP problems involved in the training of support vector machines (SVMs). An approach based on MSE and on the Ho-Kashyap algorithm was proposed in [121] for the same supervised classification task. Drawbacks of that approach included long computation time, unpredictable number of iterations to reach convergence, and non-guaranteed positivity of the solution. The method proposed here overcomes these limitations taking advantage of the combination with the SMO algorithm and of an improved MSE formulation.

5.2 Methodology

An image data set is assumed to be available together with a collection of training samples to be used for learning and parameter optimization. The data set may include single-date or multitemporal, single-sensor or multisensor imagery [72] [121] [98]. Given a pixel lattice \mathcal{I} and a neighborhood system on \mathcal{I} , an MRF model is postulated for the joint posterior distribution of the class labels $\{y_i\}_{i \in \mathcal{I}}$ of the pixels, conditioned to the corresponding satellite observations $\{\mathbf{x}_i\}_{i \in \mathcal{I}}$, i.e.:

$$P(y_i | y_j, j \neq i, \mathbf{x}_i) = P(y_i | y_j, j \sim i, \mathbf{x}_i) \propto e^{-U(y_i | \mathbf{y}_{\partial i}, \mathbf{x}_i)}, \quad (5.1)$$

where $i \sim j$ denotes that pixels i and j are neighbors, $\mathbf{y}_{\partial i}$ is the vector of the labels of the neighbors of pixel i ($i, j \in \mathcal{I}$), and $U(\cdot)$ is the (local posterior) energy function of the considered MRF model [72] [121] [98]. Real-valued parameters $\lambda_1, \lambda_2, \dots, \lambda_L$ are generally included in this function. Here, a linear relationship between $U(\cdot)$ and the vector $\boldsymbol{\lambda} \in \mathbb{R}^L$ of these parameters [121] is assumed to hold. In this case, λ_ℓ is usually the weight of the ℓ th term in a linear combination of L energy terms ($\ell = 1, 2, \dots, L$). As proven in [121], if this linear relationship holds, then the condition of correct classification of the training samples can be written as an overconditioned system of linear inequalities, i.e.:

$$E\boldsymbol{\lambda} \geq \mathbf{0}, \quad (5.2)$$

where the $R \times L$ matrix E ($R \gg L$) is derived from $U(\cdot)$ using an initial classification map and the training map. Details on the computation of E

5.2. Methodology

can be found in [121].

The approximate solution of this overconditioned linear system is addressed through the MSE criterion and reformulated as a constrained QP problem to take benefit from the computationally efficient SMO algorithms that have been recently developed in the framework of SVM. SMO belongs to the family of active-set methods that decompose a QP problem involving many variables into an appropriate sequence of subproblems, each focused on a subset of the variables. In particular, SMO generates a sequence of subproblems, which operate with pairs of variables and for which closed-form solutions can be computed [105] [47].

In the proposed method, the combination of the MSE approach to MRF parameter estimation and of SMO is accomplished by proving that a case-specific QP formulation exists that satisfies the following two conditions:

- it is consistent with the aforementioned system of linear inequalities associated with the correct classification of the training set;
- it belongs to the family of QP problems to which the SMO formulation in [47] is applicable.

On one hand, the second condition is operatively relevant because it allows taking benefit from the computational efficiency of the QP algorithm in [47], which was originally developed for SVM learning and is incorporated here within the proposed MRF parameter optimization scheme. On the other hand, it is a non-trivial condition because the family of QPs addressed by the algorithm in [47] does not encompass the numerical solution of an arbitrary MSE problem.

The system (5.2) of linear inequalities is reformulated as the following system of linear equalities involving a vector $\mathbf{b} \in \mathbb{R}^R$ of non-negative “margin” variables, i.e.:

$$E\boldsymbol{\lambda} = \mathbf{b}, \quad \mathbf{b} \geq \mathbf{0}. \quad (5.3)$$

If $E^\# = (E^T E)^{-1} E^T$ denotes the (left) pseudo-inverse of E , then a necessary condition for this system to be satisfied is:

$$E^\# (E\boldsymbol{\lambda} - \mathbf{b}) = \mathbf{0} \Rightarrow \boldsymbol{\lambda} = E^\# \mathbf{b}. \quad (5.4)$$

According to their meaning as weights of energy terms, it is generally desirable that the λ_ℓ parameters ($\ell = 1, 2, \dots, L$) are strictly positive. The following MSE formulation is proposed to integrate the aforementioned linear condition and this positivity constraint:

$$\begin{cases} \min_{\mathbf{a} \in \mathbb{R}^L, \mathbf{b} \in \mathbb{R}^R} \|\mathbf{a} + \epsilon \mathbf{1}_L - E^\# \mathbf{b}\|^2 \\ \mathbf{a} \geq \mathbf{0}, \mathbf{b} \geq \mathbf{0} \end{cases} \quad (5.5)$$

where \mathbf{a} is an auxiliary unknown vector such that $\boldsymbol{\lambda} = \mathbf{a} + \epsilon \mathbf{1}_L$, ϵ is a positive tolerance representing the smallest admissible value of each unknown parameter, and $\mathbf{1}_L$ denotes the vector in \mathbb{R}^L comprising L unitary components. After some algebraic manipulations, (5.5) can be written as:

$$\begin{cases} \min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \left(\frac{1}{2} \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} + \mathbf{p}^T \boldsymbol{\alpha} \right) \\ \boldsymbol{\alpha} \geq \mathbf{0} \\ \mathbf{1}_N^T \boldsymbol{\alpha} = 1 - L\epsilon \end{cases} \quad (5.6)$$

where $N = R + L$, $\mathbf{p} \in \mathbb{R}^N$, Q is an $N \times N$ matrix,

$$\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \quad Q_{hk} = \mathbf{e}_h^T \mathbf{e}_k, \quad p_k = \epsilon \mathbf{e}_k^T \mathbf{1}_L, \quad (5.7)$$

I_L denotes the $L \times L$ identity matrix, and $\mathbf{e}_k \in \mathbb{R}^L$ is the k th column of the $L \times N$ block matrix $[I_L | -E^\#]$ ($k = 1, 2, \dots, N$). Indeed this problem belongs to the family of QPs that can be addressed using the SMO formulation in [47]. The constraint $\mathbf{1}_N^T \boldsymbol{\alpha} = 1 - L\epsilon$ is acceptable because parameters and margins are determined only up to some arbitrary positive multiplicative coefficient. So the condition $\mathbf{1}_L^T \boldsymbol{\lambda} + \mathbf{1}_R^T \mathbf{b} = 1$ yields no loss of generality. It implies:

$$\mathbf{1}_L^T \mathbf{a} + \epsilon \mathbf{1}_L^T \mathbf{1}_L + \mathbf{1}_R^T \mathbf{b} = \mathbf{1}_N^T \boldsymbol{\alpha} + L\epsilon = 1, \quad (5.8)$$

which results in the equality constraint of (5.6), provided that $1 - L\epsilon > 0$, i.e., $\epsilon < 1/L$. Furthermore, it is also possible to prove that this constraint also implies that ($\ell, h = 1, 2, \dots, L$):

$$\frac{1}{\rho} \leq \frac{\lambda_\ell}{\lambda_h} \leq \rho = \frac{1}{\epsilon} - (L - 1), \quad (5.9)$$

i.e., ϵ is related not only to the minimum admissible value of each parameter but also to the maximum admissible ratio ρ between two distinct parameters.

5.3 Experimental validation

The proposed method was experimentally validated with three distinct MRF models, associated with both single-date and multitemporal image classification, and with five data sets. The data sets included:

- a multipolarization and multifrequency SIR-C/XSAR image (700×280 pixels) acquired over an agricultural area near Pavia (Italy) and including the “dry soil” and “wet soil” thematic classes;
- a Landsat-5 TM data set (494×882 pixels) acquired over an agricultural area near Alessandria (Italy) and presenting four classes (“bare soil”, “wet soil”, “wood”, and “cereals”);
- an airborne data set (250×350 pixels) composed of 6 optical bands and 9 multipolarization and multifrequency SAR channels, acquired over Feltwell (UK) and presenting five agricultural classes;
- a 4 m resolution IKONOS data set (1999×1501 pixels) acquired over Itaipu (Brazil/Paraguay border) and presenting seven classes, including vegetated and urban/built-up land covers;
- a multitemporal pair of 10-m resolution SPOT-5 HRG images (1500×1160 pixels) acquired near Beijing (China) in 2003 and 2005 and presenting five urban and vegetated classes.

The three considered MRFs included:

- a single-date model based on the well-known Potts spatial MRF and on pixelwise Gaussian class-conditional statistics [72];

5.3. Experimental validation

- a single-date model based on the Potts MRF and on the characterization of pixelwise statistics through the k -nearest neighbors (k -NN) algorithm [72];
- the multitemporal model in [91] that fuses the information associated with pixelwise statistics, spatial context, and temporal context through temporal transition probabilities.

In all experiments, energy minimization was addressed using ICM, which is usually an acceptable tradeoff between classification accuracy and computational burden. The results of the proposed method were compared with those given by:

- the previous method in [121], named HK-MSE in the following, which formulates the correct classification of the training pixels as the linear system (5.2) and, as compared to the proposed technique, uses a simpler quadratic formulation and the Ho-Kashyap numerical procedure;
- the application of the Goldfarb-Idnani (GI) method, which is a general-purpose active-set technique for convex QPs [56], to a QP problem resulting from the same linear system.

Preliminary experiments, not reported for brevity, confirmed limited sensitivity to the choice of the maximum ratio ρ between distinct parameters. In the following $\rho \simeq 25$ was used.

As a preliminary analysis, Table 5.1 shows, for each data set, the numbers of violated inequalities in the approximate solutions provided by the proposed method and HK-MSE for the system (5.2). Fewer inequalities turned out to be violated by the proposed method than by the previous one, which suggests an improved capability to identify parameter configurations that favor reduced errors on the training set. Table 5.1 also summarizes the accuracies on the test sets and the execution times of the proposed method and of the previous one in [121]. The proposed algorithm identified parameter configurations that ensured rather high classification accuracies on the test samples of all considered data sets. Especially accurate results were obtained with the Pavia, Alessandria, and Itaipu data sets, thus suggesting the effectiveness of the method when applied to diverse spatial resolutions and MRF

5.3. Experimental validation

Table 5.1: Classification accuracy on the test sets (OA = overall accuracy; AA = average accuracy), execution time, and number of violated inequalities (NVI) obtained by applying the proposed method and the previous HK-MSE method in [121]

Dataset	MRF model	HK-MSE				Proposed			
		OA	AA	Tune [s]	NVI	OA	AA	Tune [s]	NVI
Pavia	Potts/k-NN	94.96%	95.13%	3	0	94.96%	95.13%	3	0
Alessandria	Potts/Gauss	96.99%	86.49%	76	410	94.99%	86.20%	11	341
Fetwell	Potts/k-NN	83.02%	80.44%	2	199	83.80%	81.01%	1	115
Itaipu	Potts/Gauss	96.85%	92.62%	67	3411	97.42%	92.34%	67	2457
Beijing (2003)	Multitemporal [91]	84.19%	85.74%	810	41	84.33%	85.93%	570	41
Beijing (2005)	Multitemporal [91]	63.22%	71.75%	830		63.29%	71.79%	574	
	mean	86.21%	85.36%	298	812.2	86.47%	85.40%	204	591.4

5.3. Experimental validation

Table 5.2: Classification accuracies and comparison of the execution times of the proposed method and the Goldfarb-Idnani approach (OA = overall accuracy; AA = average accuracy)

Dataset	Goldfarb-Idnani			Proposed		
	OA	AA	Time: GI/HK-MSE	OA	AA	Time: Proposed/HK-MSE
Alessandria	94.20%	86.49%	>>1000	94.99%	86.20%	0.14
Fetwell	82.87%	80.33%	422	83.80%	81.01%	0.50

models. Lower accuracies were obtained for Feltwell due to the small equivalent number of looks of the related SAR channels and to the substantial overlapping among the related agricultural classes in the feature space. In the case of the multitemporal Beijing data set, more accurate maps were obtained for the image collected in 2003 than for that acquired in 2005, due to the stronger spectral overlapping among the classes in the latter than in the former. The comparison between the results of the proposed and HK-MSE algorithms pointed out that very similar accuracies were obtained by the two methods, which was an expected result due to the similarity between the corresponding MSE criteria. However, significantly shorter execution times were necessary to the proposed algorithm to identify appropriate parameter configurations for most data sets. An additional advantage of the developed method over HK-MSE is that the former analytically guarantees positivity of the parameters while the latter may converge to negative solutions. In the three MRFs considered for experiments, the parameters represent the weights of various energy contributions, so negative solutions are not analytically forbidden but are strongly undesired as they can affect classification accuracy. It is also worth recalling that, in [121], it was experimentally demonstrated that more accurate classification results were generally obtained by using the parameter values provided by HK-MSE than by the pseudo-likelihood approach. A second experimental comparison, focused on the Alessandria and Feltwell data sets, was performed between the developed technique and the GI method. In both cases, the two methods exhibited a strong similarity in the accuracy values on the test set with slightly higher OA values for the proposed method which allowed an increase in OA of less than 1%. This was an expected result because both approaches adopt the same energy

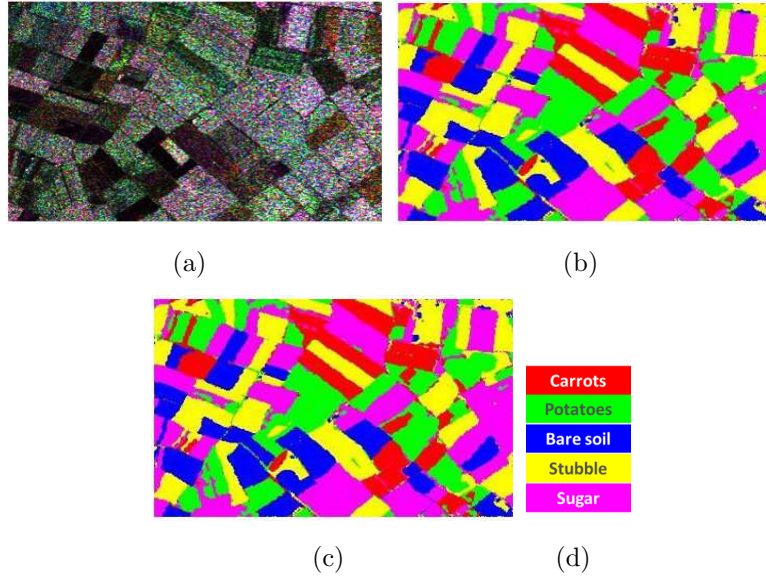


Figure 5.1: “Feltwell”: color composition of the image channels (a) and classification maps generated by the proposed method (b) and by the technique in [121] (c).

minimization algorithm (ICM) and the same MSE criterion for parameter estimation. Nevertheless, the computational time was significantly shorter for the proposed method with differences of several orders of magnitude compared to the GI one (Table 5.2). The experimental results and comparisons overall suggested that the proposed method is effective in identifying parameter configurations that favor high classification accuracy on test samples with reduced computational burden as compared to alternate and previous approaches. This is interpreted as due to the effective combination of SMO and of training information to appropriately characterize the correct labeling of the training data and derive effective parameter values.

5.4 Conclusion

In this chapter, a novel parameter optimization method has been proposed for MRF models for contextual supervised classification problems. The proposed technique generalizes a quadratic formulation and maps the parameter opti-

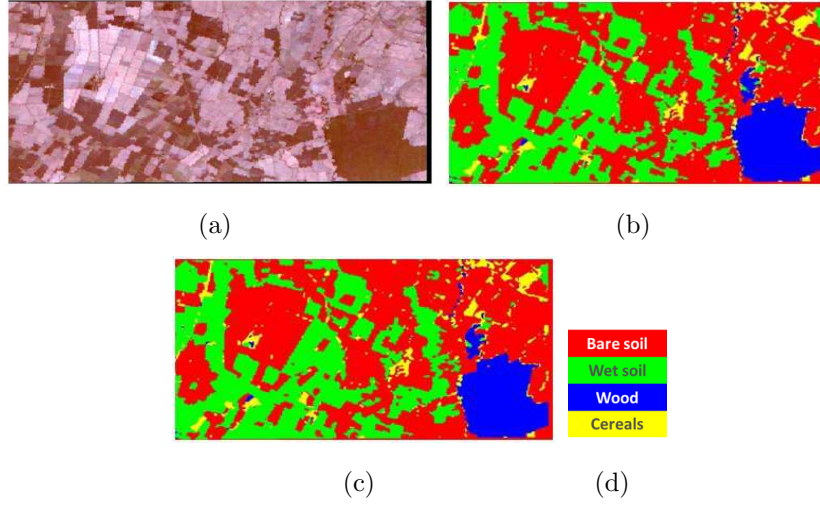


Figure 5.2: “Alessandria”: RGB false-color composition of the Landsat-5 TM channels (a) and classification maps generated by the proposed method (b) and by the technique in [121] (c).

mization problem to an appropriate quadratic programming problem feasible for the resolution through the SMO algorithm. The method has been tested with five different data sets which differ in type of sensor, spectral and spatial resolution, number of classes and number of channels, and has been compared with two previous techniques, the first one based on a similar QP formulation and on the Ho-Kashyap numerical method, and the second one based on the application of the Goldfard-Idnani method. As expected, the proposed approach reached the same or slightly higher overall accuracy values than these approaches because all considered methods share similar objective functions and use the same ICM algorithm for the classification task. However, the proposed method overcomes the limitations of the previous ones showing a remarkable reduction of computation time and allowing the positivity of the MRF parameters to be encoded as a constraint. This result validates the effectiveness of the proposed approach and the capability of Platts algorithm in terms of both computational time and memory occupation. Possible future extensions could be the validation of the proposed approach with other MRF models, including, for instance, multiscale or multiresolution modeling, texture components, or edge-preserving processing [98].

Chapter 6

Fault diagnosis for power generation plants based on solid oxide fuel cell

6.1 Introduction

Fuel cells (FCs) are electrochemical reactors that convert hydrogen energy into electrical energy through a reaction that uses oxygen [132] [11]. They are similar to flow batteries but can continuously produce electricity as long as the reaction is fed with fuel and oxygen. Power generation plants based on FCs represent one of the most promising and strategic technologies developed in recent decades because of their high energy conversion efficiency and environmental compatibility.

Proton exchange membrane FCs (PEMFCs) and solid oxide FCs (SOFCs) are the most widely used types of FCs [132] [11], and they have distinctive features, potential applications, and specific advantages and disadvantages. These two types of FCs are actually different electrochemical systems: PEMFCs are developed primarily for vehicular applications because they are lightweight systems that operate at approximately 80 °C, whereas SOFCs are preferred for distributed electricity generation because they operate at approximately 850 °C, providing high-quality waste heat to recover in cogeneration and bottoming cycles.

6.1. Introduction

Unfortunately, FC-based power generation plants suffer from low reliability and limited lifetimes; thus, the development of specific methods for automatic on-line fault diagnosis is of paramount importance for their commercial diffusion and constitutes an active field of research [104] [130]. The differences between PEMFCs and SOFCs, as well as the differences between the related power generation plants, result in significantly different malfunctioning mechanisms and fault typologies. Although numerous fault detection and isolation (FDI) methods have been developed for PEMFCs and related plants, as reviewed in [104], SOFCs and related plants have received insufficient attention until now [131] [9] [134] [103] [108]. The aforementioned differences make it impossible to directly apply the FDI methods developed for PEMFC plants to SOFC plants.

In this chapter, we focus on SOFC-based power generation plants. Although these plants are not expected to perform load following, they should work both at the design point and in several off-design operating conditions, i.e., in numerous steady-state operating conditions. In addition, the faults that affect these plants are not binary events; rather, their size or extent can take on values in wide continuous intervals. Thus, to avoid efficiency losses and irreversible damages in SOFC plants, it is particularly important to rapidly detect and identify possible faults as soon as they occur irrespective of their size or extent and of the current operating condition.

Despite these needs, the vast majority of the FDI methods proposed for SOFC-based power generation plants [131] [9] [134] [108] are designed to function under a single steady-state operating condition (or under a very restricted number of operating conditions) and for a small set of discrete sizes or extents of the considered faults. The reason for this limitation is that these FDI methods implement a conventional model-based FDI scheme in which the residuals (i.e., the differences between the measurements performed on the plant and the model predictions) are used for fault identification through an inference approach [68] [153] [69], namely, the fault signature matrix (FSM). This matrix associates each considered fault with a binary vector that has a number of elements equal to the number of residuals. If a given residual exceeds its threshold, then the corresponding vector element is equal to one; otherwise, it is equal to zero. As discussed in [46]

6.1. Introduction

for PEMFC plants and in [108] for SOFC plants, the use of binary codification of the residuals loses valuable information, and the strict association of each fault with a single vector does not allow all the possible combinations of fault sizes and operating conditions to be encompassed. Nevertheless, if the SOFC plant is operated in many operating conditions and different fault sizes can occur (relevant assumptions in the application to real-world SOFC systems), then each fault can generate multiple binary vectors. Vice versa, there are specific binary vectors that can be obtained as the result of two or more faults. Therefore, the relationship between binary vectors and faults is many-to-many, which implies an intrinsic ambiguity, i.e., one cannot univocally determine the fault from the observation of the binary vector. Furthermore, from a pattern-recognition perspective, residuals are used as features for the detection and classification of faults, and thresholding a feature implies partitioning the feature space with a hyperplane orthogonal to one of its axes. Therefore, the decision regions corresponding to any decision rule that operates only on binarized features are constrained to be aligned with these hyperplanes. This is severely restrictive, especially as compared to a classifier that can operate with the original features and their joint statistics, and is allowed to determine arbitrarily shaped decision regions to maximize classification accuracy. Finally, the design of the threshold values is a critical task that significantly affects FDI performance [108].

During the Ph.D. program, two different contribution to FDI methodologies has been developed with the aim to overcome the aforementioned limitations in SOFC plants. In the proposed approaches the dataset is collected through an SOFC plant model in which the capability of simulating faulty conditions has been implanted and the residuals are processed through a statistical classifier that detects and identifies the faults in place of the thresholds and FSM. The introduction of a classifier and of a wide dataset to train such a classifier is typical of the data-driven FDI strategy (also called history- or knowledge-based strategy) [153] [89] [164] [158] [53]. This combination between the conventional formulations of model-based and data-driven FDI strategies constitutes another FDI approach, named hybrid or integrated strategy [164] [53] [127] [87]. Specifically, in the first proposed FDI solution the statistical classifier is implemented through a support vector

6.2. Fault detection and isolation

machine (SVM). Furthermore, the problem of determine, among the variables that can be measured in the SOFC system, those which play the most critical role for the fault detection task is also addressed. For this purpose, a feature selection technique is integrated with the optimization of the parameter of the SVM classifier in a unique novel method. In the second proposed FDI solution, for the first time in FC field, random forest (RFs), a powerful supervised non-parametric classification approach, has been chosen as the pattern recognition technique to be used to develop an hybrid FDI approach for SOFC-based generation plant. Moreover, a data-driven strategy based again on RFs is developed and an experimental comparison between the proposed hybrid approach, the purely data driven strategy and a model-based FDI schemes adopting the FSM is carried out.

6.2 Fault detection and isolation

6.2.1 Model-based approach

Although the model-based FDI theory includes several schemes for the interaction between the real plant and the plant model [68] [153], the parity equation scheme with output errors [69] is the preferred option for SOFC plants [131] [9] [134], [108]. As shown in Figure 6.1, the plant and model receive the same inputs (which determine the plant operating conditions), and when no fault occurs, they are expected to produce the same outputs (which represent the monitored physical variables). In this condition, the residuals are zeros for less than the modeling and measurement errors. In the strategy commonly adopted for SOFC plants, the residuals are used to detect any possible fault through (absolute or relative) thresholds. Subsequently, the binarized residuals are used to identify the fault through an FSM, arranged based on a fault tree analysis, which is a deductive top-down tool and is typically used in safety and reliability engineering [9] [134] [108]. Because fault identification is performed using an inference approach, pattern recognition techniques are not involved in this task, and no dataset is necessary for training purposes. In this classical model-based FDI strategy for SOFC plants, the only possible use of pattern recognition techniques is to arrange a

6.2. Fault detection and isolation

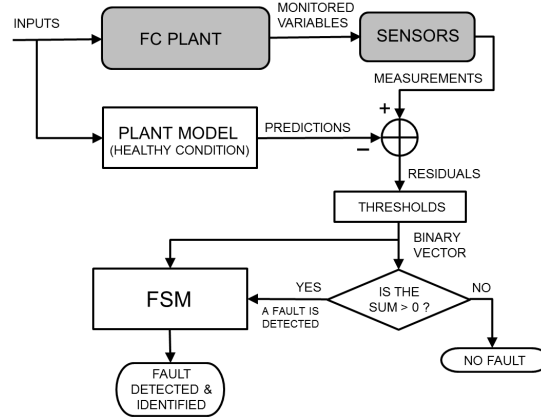


Figure 6.1: Schematic of the classical model-based FDI strategy for FC plants based on parity equations with output errors [69], residual binarization, and FSM.

qualitative model for the plant, in case a quantitative model is not available or viable, as reviewed in [157] [90].

6.2.2 Data-driven approach

To the best of the author knowledge, a purely data-driven FDI strategy [153] [89] [164] [53] has not yet been proposed for SOFC plants. However, some data-driven schemes have been considered for PEMFCs and related plants, as reviewed in [167]. As shown in Figure 6.2, in the general data-driven scheme, both plant inputs and outputs feed the FDI system. The latter is implemented by pattern recognition or artificial intelligence techniques and should be adequately trained through statistically representative datasets or *a priori* knowledge. SVMs, neural networks and expert systems are the preferred tools for the fault diagnosis [53] [117] [101] [118]. Unlike in the previous FDI strategy, a plant model is not required, and the fault detection and fault identification are potentially two undistinguished tasks (i.e., the plant healthy status represents a given class, such as each considered fault). Here, a data-driven FDI system is developed and tested by using the RFs to assemble a non-parametric supervised classifier.

6.2. Fault detection and isolation

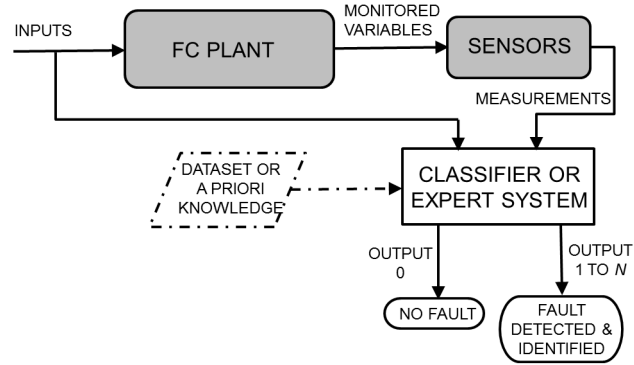


Figure 6.2: Schematic of the data-driven FDI strategy applied to FC plants.

6.2.3 Hybrid approach

A possible combination of the two previous strategies leads to a hybrid FDI strategy in which, as illustrated in Figure 6.3, a plant model that receives the same inputs of the real SOFC plant is used to predict the monitored variables. The residuals generated by the parity equations with output errors [69] are used to detect and identify faults through a statistical classifier, such as that used in data-driven approaches, adequately trained offline using statistically representative datasets. Additionally, in this case, the fault detection and fault identification are two undistinguished tasks, preventing the need to set detection thresholds. Moreover, the classifier exploits the original values of the residuals without any binarization. The strategies proposed in this chapter fall within this FDI family. In the former a quantitative model for an SOFC plant is used together an SVM classifier and a joint model and feature selection technique is developed. In the latter the introduction of RFs as a powerful, flexible and time inexpensive technique for assembling the statistical classifier allows the adoption of a rapid training and testing procedure for the FDI system in many different configurations.

6.2.4 Plant models simulating healthy and faulty conditions

The requirement of a large amount of data for training the statistical classifier used in the data-driven and hybrid FDI strategies (i.e., a collection of

6.2. Fault detection and isolation

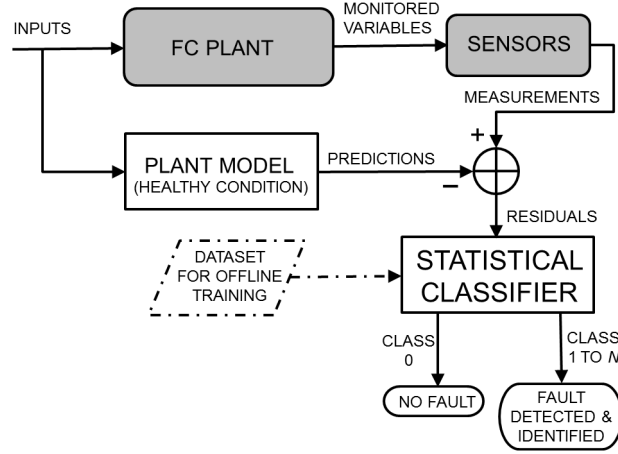


Figure 6.3: Schematic of the hybrid FDI strategy for FC plants, where residuals generated by the parity equations are processed by an adequately trained statistical classifier.

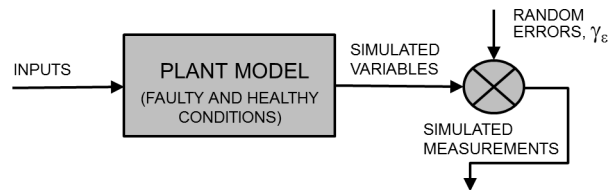


Figure 6.4: Replacement of the real FC plant and related sensors (see Figures 1 to 3) by a plant model, which simulates healthy and faulty conditions and whose output variables are subject to random errors.

6.2. Fault detection and isolation

the monitored variables or their residuals obtained under different healthy and faulty operating conditions and with various fault sizes) represents a practical problem. Not only is the collection of these data exceptionally time-consuming but the implanting of real faults in real systems also often produces irreparable damage with related economic loss. A similar problem occurs when the functioning of a conventional model-based FDI system should be assessed. Model-based FDI potentially offers a solution to this problem: if the model enables fictitious faults to be implanted inside it, such a model can be used to simulate a real plant operating under faulty conditions [103] [108] [94] [30] [99]. However, to also introduce model uncertainty and measurement tolerance, it is necessary to include random errors in the values of the physical variables simulated for the real plant. Each simulated variable can be multiplied by a random independent variable γ_ϵ , uniformly distributed in $[1 - \epsilon, 1 + \epsilon]$, where 100ϵ represents the maximum percentage error. In other words, the two grey blocks in Figures 6.1 6.2 6.3 can be replaced by the blocks depicted in Figure 6.4, devoted to generating a realistic simulation of the monitored variables also in faulty conditions. The data obtained due to this replacement can be used to train the classifier (if any) and evaluate the performance of the FDI system. Therefore, the availability of a plant model that simulates healthy and faulty conditions is important not only for model-based but also for data-driven strategies, particularly during offline training and testing operations.

The replacement of the real plant by a model is effective if the model is reliable and accurate. This requirement is typically satisfied when a quantitative mathematical model is deployed and has been validated with experimental trials that encompass several operating conditions. In addition, the inclusion of the physical equations that govern the real plant makes the quantitative model well suited for implanting fictitious faults with the required reliability and accuracy [130] [58].

6.3. Measurements and faults in solid oxide fuel cell plant

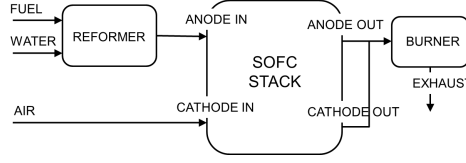


Figure 6.5: Schematic of the SOFC plant.

6.3 Measurements and faults in solid oxide fuel cell plant

6.3.1 Plant structure and operating conditions

The considered laboratory-scale power generation plant is composed of a reformer, an SOFC stack, and a post burner, as schematically illustrated in Figure 6.5. The reformer is fed with a mixture of methane and steam and converts the methane into hydrogen and carbon monoxide. These compounds are fed into the anode compartment of an SOFC stack while air is fed into the cathode compartment. The SOFC stack is composed of a number of rectangular planar cells superimposed onto each other. The electrochemical reaction produces steam and carbon dioxide along with electrical power and heat. The anode exhaust is mixed with the cathode exhaust and burned in the off-gas burner to reduce the release of pollutants and increase the temperature of the flue gas for further utilization.

The considered power generation plant is typically operated under two control strategies (i.e., constant-current and constant-voltage) and is equipped with an inverter. In addition to its primary task of converting the DC electric power produced by the plant into AC electric power, the inverter controls the power supplied by the plant and the operational mode of the stack. The flow rates of the fuel, air and water entering the power generation plant are regulated by the controller to maintain the average temperature of the stack and the fuel utilization factor as close as possible to the desired values. These values determine which voltage-current characteristic curve for the SOFC stack is used. When the stack temperature and fuel utilization factor are maintained constant, different operating conditions can be defined moving along the characteristic curve. A potential fault affecting the SOFC

6.3. Measurements and faults in solid oxide fuel cell plant

stack is expected to distort such a curve, whereas faults affecting other plant components can modify the parameters that identify the characteristic curve. In both cases, the characteristic curve changes, and the inverter should control the plant in such a way to keep the generated current or the generated voltage constant.

In the considered experiments, the inverter controls the plant to maintain the average temperature of the SOFC stack equal to 850 °C and the fuel utilization factor equal to 0.75. Under the design operating condition, the plant generates a voltage of 42.5 V and a current of 26 A (i.e., an electrical power of approximately 1.1 kW). Starting from this reference point ten different operating condition are defined by tuning the values imposed for constant voltage (or constant current) Additional details regarding the structure and control of the considered SOFC generation plant are described in [130].

6.3.2 Monitored variables and measurement difficulties

The monitored variables considered in this chapter for the fault diagnosis are the physical quantities listed in Table 6.1. The first five physiochemical variables are initially selected due to the ease with which they can be measured by a standard sensory suite deployed in the real plant. In addition to these variables, the adoption of a quantitative model in place of a real plant allows to consider five additional variables related to internal parameters of the SOFC (numbered from 6 to 10). Although they can potentially play an important role in understanding the plant health condition, they are typically difficult to measure in a real plant because of the difficulty in developing sensors that can operate in the harsh conditions encountered inside the FC (in the case of SOFC, the high temperature) and without significantly affecting the chemical reactions that occur. Even though the experimental phase in the next sections is mainly focused on the usage of the first five variable, the intended goal in considering the additional five variable is to analyze their potentiality for improving FDI performance. Moreover, despite these difficulties, some recent papers propose and test sensor technologies for measuring the internal parameters of FCs during their operation. In particular, the measurement of

6.3. Measurements and faults in solid oxide fuel cell plant

Table 6.1: Monitored variables used for the FDI in the considered SOFC plant.

N.	Physical quantity	Measurement difficulty
1	Generated electric power	Easy
2	Air flow rate entering the plant	Easy
3	Reformate fuel flow rate entering the SOFC stack	Easy
4	Air pressure loss between the inlet and outlet of the SOFC stack	Easy
5	Temperature at the burner outlet	Easy
6	Maximum temperature gradient	Moderate
7	Anodic activation losses	Difficult
8	Cathodic activation losses	Moderate
9	Ohmic losses	Difficult
10	Nerst voltage	Difficult

the temperature and temperature gradient in SOFCs [93] [60] [64] has been proposed by adopting thin thermocouples, the measurement of temperature and humidity in PEMFCs [77] [83] using multi-functional micro sensors, and the measurement of the losses due to the impedance increase inside the SOFC electrodes [3] [82] [33] [92] using the electrochemical impedance spectroscopy.

6.3.3 Fault classes

Concerning possible plant malfunctioning conditions, the following four fault classes are considered [130]:

1. SOFC stack degradation, i.e., a performance reduction due to the increase of the overall stack voltage loss due to internal problems, i.e., activation, ohmic and diffusion losses. It has been simulated by increasing the overall stack loss to between 105% and 160% of its nominal value.
2. Air leakage, i.e., a potential air leak between the air flow meter and the SOFC stack, reducing the flow that enters the stack to between 50% and 95% of the nominal rate.
3. Fuel leakage, i.e., a leak between the exit of the reformer and the en-

6.3. Measurements and faults in solid oxide fuel cell plant

trance of the SOFC stack, which reduces the flow to between 30% and 95% of the nominal rate.

4. Reformer degradation, i.e., a lower conversion of methane due to catalyst degradation or carbon deposition, simulated by reducing the reaction to between 30% and 95% of the nominal rate.

As shown in Figure 6.5, the SOFC plant under consideration includes an SOFC stack, a methane steam reformer, a burner, a fuel feeding system and an air feeding system. Because the burner is a well-tested and mature technology, its faults are not taken into account. The faults of the other four plant components are all considered, with one fault for each component. Because the different faults that can occur inside the SOFC stack provide similar effects [130], they have been combined into a single fault class. Similar considerations hold for the steam reforming reactor. Consequently, the FDI procedures discussed here cannot be used to distinguish among the different types of faults that occur inside the SOFC stack or in the methane steam reformer. If necessary, further investigations (for example, chemical or electrochemical tests) must be conducted to identify in detail the microscopic cause of the failure, as suggested in [130].

For each of the above faults, ten different sizes are considered by setting ten values that span the previously described ranges. Because each fault size is applied to each of the ten operating conditions, 100 combinations of operating conditions and fault sizes are identified for each fault. Overall, 400 combinations between the fault status (which includes the class and size of the fault) and operating conditions are considered in our investigation.

6.3.4 Quantitative plant model

The functioning of an SOFC-based power generation plant can be effectively simulated using a quantitative mathematical model, i.e., a model that is based on the physical equations ruling all the processes governing the plant, also called the “first-principle” or “white-box” model [157]. It can be obtained by coupling the models of the three main components (i.e., the FC stack, the reformer, and the burner). For the specific plant considered here,

6.3. Measurements and faults in solid oxide fuel cell plant

the adopted model is the one extensively described in [130]. The reliability of the model predictions is assured by the model validation that has been performed with the experimental data collected during the operation of a real SOFC plant identical to the assumed one [130]. The plant was manufactured by Staxera GmbH (D) [2] and was tested by EBZ GmbH (D) within the project Genius [1]. Five different steady-state operating conditions were considered, observing a maximum difference between the measured and predicted variables that was approximately 3% [130].

As previously mentioned, to preserve the experimental plant, the fault data can be obtained from a second version of the plant model that is modified to also simulate faulty conditions. To make the investigation more reliable, the model for the simulation of faulty conditions was also adequately validated. The validation exploited real data collected from the Staxera's SOFC plant in which a limited number of faulty conditions were experimentally mimicked, as described in [58] and similar to [99].

Finally, although the quantitative models are generally considered to be computationally heavy, the described plant model is perfectly suited for working online in the configurations shown in Figures 6.1 and 6.3. In fact, the considered FDI system is designed for steady-state operating conditions. This does not represent a limitation because transitions between operating conditions rarely occur in SOFC plants due to the slow transient response with characteristic times on the order of hours. SOFC systems are currently being studied primarily for applications as distributed power generation plants, where they are expected to work optimally at a given design point or otherwise in off-design operating conditions, which are still steady-state operating conditions. When the operating conditions change, it is only required that the plant model generate the prediction of the monitored variables before the real plant reaches the new steady-state condition. The adopted quantitative model satisfies this requirement. This model can potentially extend the functionality of the FDI to include transient conditions because the model is inherently dynamic [130], and its running times are significantly faster than those of the real system. For example, an experimental transient that is $2 \cdot 10^4$ s in duration that occurs in various operating conditions is simulated in less than 10^3 s on an Intel Xeon CPU operating at 2.93 GHz with 6 GB of

RAM [34].

6.4 Support vector for fault diagnosis

6.4.1 Proposed hybrid approach, parameter optimization and feature selection

In order to overcome the limited generality of the inference approach for fault identification in model-based FDI schemes, the concept of combining a pattern recognition approach, specifically a SVM, with physiochemical modeling is developed in the context of SOFC plants. In the proposed hybrid approach the residuals between the measured and the predicted values of the variables taken on the SOFC are used as feature for the SVM classifier. As previously mentioned, the classify-before-detect paradigm is applied, namely detection and classification are not a sequence of distinct tasks but are performed jointly because the non-faulty state is a class of the problem. The fully non-parametric formulation of SVMs favor their effective application to physically heterogeneous input variables, such as those generally involved in FDI system. Recent examples of this formulation can be found in [1, 148]. Moreover, a review of SVM deployments for fault diagnosis in FDI field is reported in [161]. In the considered problem five classes are involved in the classification task (four faulty and the non-faulty class) and the one-against-one (OAO) approach is used because of its good property in term of tradeoff between accuracy and computational burden. First, a SVM binary discriminant function $f_{hk}(\cdot)$ is separately determined to discriminate between the h th and k th classes when hk using only the training samples of these two classes. Then, to label an unknown vector \mathbf{x} , each function $f_{hk}(\cdot)$ is applied to \mathbf{x} , and a vote is cast in favor of either the h th or k th class depending on the sign of $f_{hk}(\mathbf{x})$. Finally, \mathbf{x} is assigned to the class that received the most votes. Detail on SVM classification approach can be found in Chapter 1 and in [152]. A relevant issue in the FDI method is to determine which of the residuals are the most informative with respect to the discrimination of the considered classes to minimize both the number of measurements taken in the SOFC plant and the memory and computational requirements of the

6.4. Support vector for fault diagnosis

FDI system. At the same time, the SVM classifier involves parameters (i.e., the regularization parameter C [152] and possible additional parameters in the kernel), whose values generally affect the classification accuracy. For example, in the experimental validation of the proposed method, a Gaussian radial basis function kernel is used:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right),$$

which includes the positive parameter σ . Here, a novel method is developed and embedded into the proposed FDI procedure to jointly perform feature selection and parameter optimization with the aim to achieve a full automation of the FDI procedure. The key idea is to identify the feature subset and the parameter configuration that minimize an analytical error bound, i.e., the so-called span bound. Under mild assumptions, the span bound can be proven to be a tight upper bound on the leave-one-out error rate [150]. It also exhibits a usually high correlation with the error rate on test samples disjointed from the training samples, provided that they are drawn from the same distribution [96,97]. However, its computation remarkably involves only training samples and no additional validation data. To minimize the span bound, the proposed method combines the approaches introduced in [97] and [119] for SVM parameter optimization and feature selection, respectively.

Let \mathcal{R} , \mathcal{S} , and $\boldsymbol{\theta}$ be the set of all d features (i.e., the residuals), a subset of m features ($\mathcal{S} \subset \mathcal{R}$), and a vector collecting the input SVM parameters, respectively. Focusing first on a binary classification problem, l samples associated with faulty and non-faulty condition are generated through the quantitative model and a vector \mathbf{x} ($i = 1, 2, \dots, l$) is defined. The dependence on \mathcal{S} and $\boldsymbol{\theta}$ can be explicitly stressed by denoting as $\alpha_i(\mathcal{S}, \boldsymbol{\theta})$ ($i = 1, 2, \dots, l$) and $f(\cdot|\mathcal{S}, \boldsymbol{\theta})$ the solution of the quadratic programming (QP) problem in [152] and the discriminant function related to the SVM algorithm, respectively, when the SVM is trained using the m features in \mathcal{S} and the input parameter vector $\boldsymbol{\theta}$. The span bound is defined as the fraction $\mathcal{J}(\mathcal{S}, \boldsymbol{\theta})$ of the training samples such that $\alpha_i(\mathcal{S}, \boldsymbol{\theta}) > 0$ and:

$$\alpha_i(\mathcal{S}, \boldsymbol{\theta}) S_i^2(\mathcal{S}, \boldsymbol{\theta}) \geq y_i f(\mathbf{x}_i|\mathcal{S}, \boldsymbol{\theta}),$$

where y_i is a label that takes on the value $+1$ or value -1 depending on th

6.4. Support vector for fault diagnosis

membership of the i th sample to either one of the two classes, and the coefficient $S_i^2(\mathcal{S}, \boldsymbol{\theta})$ (named span) can be obtained as a by-product of the training phase by solving a further QP problem [150] or through a fast linear algebra argument [61]. In the multiclass case, $\mathcal{J}(\mathcal{S}, \boldsymbol{\theta})$ is computed through OAO as a weighted average of the span bound values obtained separately for each pair of distinct classes, the weights being proportional to the relative frequencies of the classes in the training set [97]. Further details on this point can be found in [61, 97]. Indeed, $\mathcal{J}(\mathcal{S}, \boldsymbol{\theta})$ is a non-differentiable function of $\boldsymbol{\theta}$ [150]. According to the definition recalled above for the binary case, $\mathcal{J}(\mathcal{S}, \boldsymbol{\theta})$ is integer-valued and $\mathcal{J}(\mathcal{S}, \boldsymbol{\theta})$ proves to be piecewise constant on the space of the admissible parameter vectors $\boldsymbol{\theta}$. Similar comments hold in the multiclass case as well. This prevents applying numerical minimization algorithms that make use of derivatives (e.g., the gradient or the Newton-Raphsons methods) [28]. In general, suitable numerical gradients and difference quotients might be used to replace gradients and derivatives, but ad hoc convergence theorems would be necessary for their specific application to the span bound. In [28], a regularized differentiable version of the span bound is introduced to allow gradient descent to be applied, but an additional regularization parameter, which has to be manually tuned, is necessary. Here, similar to [97], the Powell's algorithm is used to minimize $\mathcal{J}(\mathcal{S}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Powell's method is an unconstrained minimization technique that emulates the behavior of the conjugate gradient method without using derivatives and converges, under mild assumptions, to a local minimum [120]. To minimize the resulting functional:

$$\mathcal{J}_*(\mathcal{S}) = \min_{\boldsymbol{\theta}} \mathcal{J}(\mathcal{S}, \boldsymbol{\theta})$$

with respect to \mathcal{S} , the steepest ascent algorithm in [119] is adapted and extended. It is an iterative algorithm, initialized with a preliminary subset of m features, which has been demonstrated effective when applied to the maximization of Bayesian interclass distance measures in problems of remote sensing image classification [119, 120]. Here, it is extended to the minimization (steepest descent) of the span bound functional, combined with the Powell's algorithm, and integrated in the proposed FDI procedure. Specifically, given a subset \mathcal{S} of m features, the proposed feature selection and

6.4. Support vector for fault diagnosis

parameter optimization method evaluates each possible replacement of one of the m features in \mathcal{S} by one of the $(d-m)$ features outside \mathcal{S} (i.e., in $\mathcal{R} - \mathcal{S}$) by computing the corresponding value of $\mathcal{J}_*(\cdot)$, i.e., by running Powell's algorithm until convergence. Let \mathcal{J}'_* be the minimum span bound obtained across all these $m(d-m)$ possible replacements. If $\mathcal{J}'_* < \mathcal{J}_*(\mathcal{S})$, the replacement is performed, and \mathcal{S} is correspondingly updated. This procedure is iteratively repeated, while reductions in the span bound are feasible through some replacement of a feature inside by a feature outside the current subset. The collection of the subsets of m features is finite, so finite-time termination is guaranteed. As discussed in [119], convergence to a local minimum of $\mathcal{J}_*(\cdot)$ is also guaranteed, whereby the notion of the local minimum is interpreted by endowing the discrete space of the subsets of \mathcal{R} with a metric-space topology through the well-known Hamming distance. This property, together with the aforementioned convergence behavior of the Powell's algorithm, suggests that, at the least, local minimum of the span bound functional are identified by the proposed method in the searches for both a feature subset and a parameter vector.

Initialization of the method is performed through the sequential forward selection (SFS) algorithm, that is, a well-known suboptimal approach to feature selection [160]. First, SFS starts from an empty subset of features, separately computes the values of $\mathcal{J}_*(\cdot)$ associated with the d subsets composed of one feature each, and selects the feature corresponding to the smallest value of $\mathcal{J}_*(\cdot)$. Then, it evaluates $\mathcal{J}_*(\cdot)$ for all the $(d-1)$ subsets of two features, which are obtained by separately pairing the previously selected feature with each other feature. Again, the resulting feature pair with the smallest value of $\mathcal{J}_*(\cdot)$ is selected. Then, the procedure is repeated iteratively, progressively adding one feature at a time until the desired number m of features is reached. Further details on the SFS and steepest descent (ascent) algorithms can be found in [119, 120].

6.4.2 Experimental validation

During the experimental validation of the model, an accuracy of approximately 3% was reported in predicting the variables of interest. Provided

6.4. Support vector for fault diagnosis

that accurate sensors are used, the maximum measurement error can be assumed considerably smaller. Therefore, the random errors to be introduced when the measurement of the ten monitored variables is simulated, should have a magnitude of approximately 3%. Thus, in the performance evaluation of the proposed SVM-based hybrid FDI system three values for the maximum percentage error are considered: 2%, 4%, and 6%. As previously mentioned, after setting the error magnitude, a dataset is composed by considering the 100 possible combinations of the operating condition and fault size for each fault. If independent, random errors were introduced for each monitored variable of each combination, a dataset of approximately 500 feature vectors would be generated (i.e., 5 classes are considered: four faulty classes and the non-faulty class). The one hundred vectors for the non-faulty class are generated by repeating the random errors for each operating condition ten times. Successive independent random errors allow the generation of an arbitrary number of datasets. Thus, we have produced a pool of datasets for each control strategy (i.e., constant voltage and constant current) and for each maximum percentage error (i.e., 2%, 4%, and 6%). Each dataset is composed of approximately 500 feature vectors, and each feature vector is composed of ten features. All of the features in a given dataset have been preliminarily normalized to ensure that each has a zero mean and unitary variance. This normalization is necessary due to the significantly different ranges of the measured residuals. It also helps to prevent overflowing and favors numerical stability in the solution of the QP problems for SVM training.

Optimal performance versus error magnitude

The first goal is to assess the classification performance as a function of the error magnitude for each possible subset of features. In this first phase of the experimental validation ten features are considered and 1023 possible subsets should be tested; the empty subset is clearly irrelevant. To increase the confidence in the achieved performance for each control strategy and error magnitude, the related pool of independent datasets is used as follows: one dataset is used as a training set for the SVM classifier, and the remaining

6.4. Support vector for fault diagnosis

datasets are used as test sets of the trained SVM classifier. This operation is repeated for each possible feature subset. The parameters of the SVM classifier are first automatically optimized using the algorithm described in Section 6.4.1. Conversely, the choice of the kernel function is held constant within the experimental validation, and the isotropic Gaussian RBF kernel is used. This kernel includes only one real-valued parameter (i.e., σ), which makes parameter optimization more computationally advantageous than in the case of other well-known examples of kernels, including the polynomial, sigmoid tanh, and anisotropic Gaussian RBF kernels [152]. The adopted kernel was found to be effective in numerous applications [97, 152]; therefore, it is used here as a tradeoff between generality and computational burden. For each test set, the performance of the SVM classifier is evaluated by the overall accuracy (OA) (i.e., the fraction of the correctly classified test samples, which is an estimate of the probability of correct classification). Because the trained classifier is tested using numerous test sets, the average, best, and worst OA values have been recorded for each feature subset. For each control strategy and error magnitude, the subset of the features that provides the highest average OA is selected. Table 6.2 shows these feature subsets and the associated optimal performance. The processing chain, including parameter optimization and SVM classification, that uses this subset will be referred to as the optimal classifier. For both control strategies, the classification performance decreases as the error magnitude increases. The OA values obtained during constant-voltage control are higher than those obtained during constant-current control, and the difference between them is approximately 0.06. The OA decreases by approximately 0.12 during constant-voltage control and 0.15 during constant-current control when the error magnitude is increased from 2% to 6%. The classification results are stable with respect to the choice of input training samples, as the difference between the best and worst probabilities for a given error magnitude does not exceed 0.04; these results demonstrate desirable behavior that suggests that the proposed approach to the FDI process is robust to overfitting issues. The feature subsets that achieve the best performances under different conditions (i.e., voltage/current, error magnitudes) differ and include a number of residuals ranging from 4 to 8. Three residuals are present in all of the subsets re-

6.4. Support vector for fault diagnosis

Table 6.2: Performance of the optimal classifier for a given control strategy and error magnitude.

Control	Max. Error	Overall Accuracy (OA)			Residual Subset
		Average	Worst	Best	
Constant voltage	2%	0.979	0.974	0.981	2,3,4,7,8
	4%	0.925	0.916	0.938	1,2,3,4,6,7,8,10
	6%	0.858	0.837	0.869	1,2,3,4,5,6,7,8
Constant current	2%	0.919	0.913	0.926	1,2,3,4,7,8,9
	4%	0.868	0.850	0.886	1,2,3,4,9
	6%	0.769	0.761	0.778	2,3,4,9

ported in 6.2: the residuals numbered with 2, 3, and 4 (see Table 6.1 for their physical meaning). These residuals are easy to measure.

Automatic feature selection

For a given control strategy, the design of a classifier that works with a specific feature subset that is composed of easy-to-measure residuals and provides a good performance when tested using datasets generated with different error magnitudes has significant practical relevance. The aim of second phase of the experimental validation is to search for a robust classifier and related feature subset that can ensure satisfactory performance when the error magnitude is not known accurately and potentially varies over time. To investigate this topic, a training set is created for each control strategy by merging three datasets together: one dataset for each maximum percentage error (e.g., 2%, 4%, and 6%). This training set is used with the automatic feature selection and parameter optimization technique described in Section 6.4.1 to determine the optimal feature subset when the number of the adopted, easy-to-measure residuals is increased from 1 to 5. For each control strategy and optimal feature subset, the classifier trained using the aforementioned mixture of error magnitudes is tested with several datasets that are characterized by a given error magnitude. To find the feature subset that yields the optimal performance, an average OA, which is referred to as OA_{avg} , is in-

6.4. Support vector for fault diagnosis

troduced by averaging the OAs obtained for several datasets associated with the same error magnitude and then averaging the three resulting average values obtained for the 2%, 4%, and 6% error magnitudes. The OA_{avg} of the classifiers trained with the mixture of error magnitudes when the number of adopted features increases is reported in 6.3. The subset of features adopted by the feature selection technique and the OA for each error magnitude are also listed. For both control strategies, the highest OA_{avg} is achieved when the SVM trained with the mixture of error magnitudes uses the optimal feature subset composed of 4 residuals. The feature subset is the same for the two control strategies, including residuals no. 1, 2, 3, and 4 (see Table 6.1 for their physical meanings). OA_{avg} values of approximately 89% and 82% are obtained during constant-voltage control and constant-current control, respectively. Although the SVM trained with the mixture of error magnitudes performed worse than the classifiers optimized for a specific error magnitude using all residuals in Table 6.1, the decrease in OA does not exceed 0.05. Using the 4 optimal residuals, the classifier trained with the mixture of error magnitudes during constant-voltage control achieves OA values higher than those of the corresponding classifier during constant-current control. Compared to the case reported in Table 6.2, the difference is a marginal increase from approximately 0.06 to approximately 0.075. In Table 6.3, the addition of the last easy-to-measure residual (i.e., residual no. 5) reduces the performance marginally. Using only two features that are different for the two control strategies produces a high level of performance, with $OA_{avg} \cong 86\%$ during constant-voltage control and $OA_{avg} \cong 75\%$ during constant-current control.

Fault detection performance

The OA does not quantify the fault detection performance, as it represents an estimation of the probability of correct classification and the healthy status is just one of the five classes considered. To assess the fault detection performance, the probabilities of correct detection, P_d , and of false alarm, P_{fa} , (i.e., the probabilities of detecting a fault when the plant is faulty and when the plant is healthy, respectively) have been estimated through their

6.4. Support vector for fault diagnosis

Table 6.3: Feature selection results and performance of the classifier trained with a mixture of error magnitude.

Control	No. of features	Overall Accuracy (OA)				Residual Subset
		Test-set max. error			OA _{avg}	
		2%	4%	6%		
Constant voltage	1	0.688	0.678	0.654	0.673	2
	2	0.902	0.876	0.817	0.865	1,2
	3	0.940	0.871	0.779	0.863	1,2,3
	4	0.967	0.889	0.825	0.894	1,2,3,4
	5	0.961	0.866	0.783	0.870	1,2,3,4,5
Constant current	1	0.609	0.592	0.582	0.594	2
	2	0.763	0.759	0.724	0.749	2,3
	3	0.834	0.829	0.792	0.818	2,3,4
	4	0.896	0.817	0.743	0.819	1,2,3,4
	5	0.900	0.819	0.730	0.816	1,2,3,4,5

relative frequencies. Figure 6.6 shows such a probability pair for the three maximum errors, considering a classifier trained with a mixture of error magnitudes and using only the five easy-to-measure residuals. As expected, P_d decreases and P_{fa} increases when the error magnitude increases. P_{fa} is less than 0.01 for a maximum error of 2% and equal to approximately 0.1 for a maximum error of 4%. If P_{fa} must not exceed 0.1 in a given application, we can derive from Figure 6.6 that the model must predict the monitored plant variables with an accuracy of better than 4%. Correspondingly, the P_d will be higher than 0.95 for the constant-voltage control and higher than 0.85 for the constant-current control. The OA (i.e., the probability of correct identification of the fault class) will be higher than 0.86 for the constant-voltage control and higher than 0.81 for the constant-current control (as reported in Table 6.3).

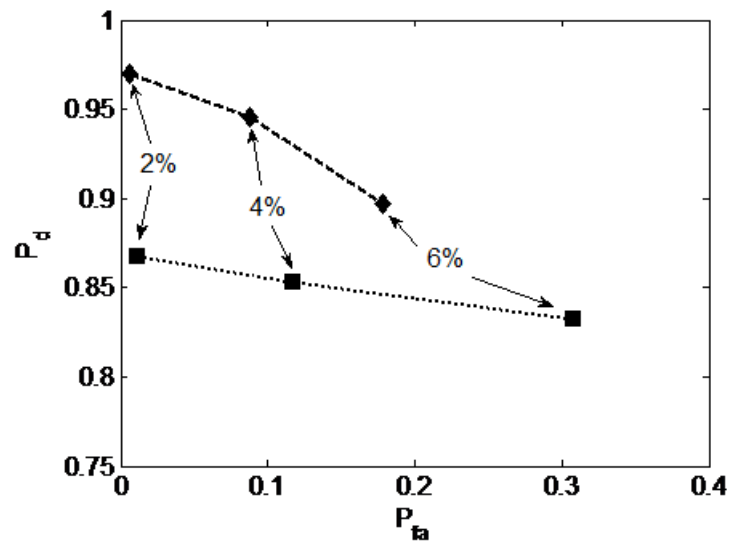


Figure 6.6: Probabilities of detection and false alarm of a classifier using the five easy-to-measure residuals and trained with a mixture of error magnitudes: constant-voltage case (diamonds and dashed line) and constant-current case (squares and dotted line). The probability pair is estimated for the 2%, 4%, and 6% error magnitudes.

6.5 Random forests for fault diagnosis

6.5.1 Proposed fault diagnosis approach

Following the same approach proposed in the previous section, the FDI task for a given SOFC-based plant is again formalized as a supervised classification problem in which the set \mathcal{C} of classes includes a finite number of distinct types of faults that can affect the plant and the non-faulty condition. Specifically, the potentiality of the hybrid approach that combines supervised pattern recognition and physicochemical modeling is demonstrated in Section 6.4 using a support vector classifier. Here, a different opportunity, based on the RF approach to classification, is investigated. RF is well known as a topical and effective approach to supervised classification and regression [20]. It is rooted in the theory of multiple classifiers, which make use of ensembles of individual classifiers to exploit the complementarity among several (possibly numerous, simple, and weak) learning machines to determine powerful discriminant functions through decision fusion [75]. In particular, here RF is chosen as the pattern recognition technique to be used within a hybrid FDI approach. In order to extensively test the proposed approach and to compare the different FDI strategies, a data-driven FDI method is developed by deploying RF as statistical classifier. Furthermore, the performance reached by a classic model-based scheme with fault signature matrix as inference approach is investigated. Concerning the proposed hybrid approaches, from a black-box perspective, SVM and RF, although they stem from different methodological approaches (kernel-based vs. ensemble learning), exhibit several common characteristics, including the fully non-parametric formulation and the applicability to data with arbitrary probability distributions. Their performances have often been compared [163] [129] [133] [100] with each other and with those of alternate non-parametric supervised methods, such as artificial neural networks, with opposite conclusions on different specific applications but with overall comparable classification accuracies. RFs are adopted here because, while they are expected to obtain accuracies similar to those of SVMs and other non-parametric techniques, they offer some advantages that are particularly relevant in operationally devising, develop-

6.5. Random forests for fault diagnosis

ing and testing FDI systems for FC-based plants. These advantages consist of the small number of input parameters, the possibility of generally tuning these parameters in an efficient way, the generally short computation times, and the possible scalability and parallelizability.

An RF is a suitable ensemble of tree classifiers, and a tree classifier is in turn defined by a set of decision rules sequentially interconnected by a tree topology [75] [21]. Each node of each tree in the forest implies a decision rule that “splits” the classification process into two or more branches. The rationale is that although the rule associated with each node is often very simple (e.g., thresholding one of the features), the combination of many individually simple rules within each tree and the fusion of the outputs of an ensemble of trees within the forest allows a flexible and possibly complex overall decision criterion to be determined in the feature space [75].

Specifically, in an RF classifier, T trees are parameterized and trained in an independent and identically distributed (i.i.d.) manner. Denoting as θ a vector of parameters that characterize a generic tree classifier, any set of T trees whose parameter configurations $\theta_1, \theta_2, \dots, \theta_T$ are i.i.d. random vectors generally defines an RF [20]. Consistent with the well-known results of ensemble learning [75], the accuracy of the forest generally benefits from maximizing the diversity among the corresponding trees. To randomly generate the trees of an RF, a typical strategy, which has been found effective in numerous applications, is to use random sampling of both training data and features. For each t th tree, a set \mathcal{L}_t of training data, including as many training samples as \mathcal{L} , is randomly drawn from \mathcal{L} uniformly and with replacement (bagging) and is used for training the tree itself ($t = 1, 2, \dots, T$). Denoting as \mathcal{N}_t the set of nodes of the t th tree, for each n th node ($n \in \mathcal{N}_t$), a subset $\mathcal{X}_{nt} \subset \mathcal{X}$ of F features ($1 \leq F \leq d; t = 1, 2, \dots, T$) is randomly drawn from \mathcal{X} with a uniform distribution and is used in the decision rule of that node (random input selection) [20] [75].

With regard to each individual tree classifier, the decomposition of the classification problem into the tree and the decision rule associated with each node may be defined, in principle, either heuristically, according to the user/operator’s prior knowledge about the addressed problem, or automatically on the basis of a training set. The latter approach is typically used in

6.5. Random forests for fault diagnosis

RF [21]. Here, binary trees are considered, the decision rule of each node consists of thresholding one of the features, and among the numerous supervised tree-based methods introduced in the literature (e.g., ID3, C4.5) [75], the CART (classification and regression tree) framework is used together with the aforementioned random sampling procedures. To construct the t th tree in the forest, CART is applied to the corresponding subset \mathcal{L}_t of training samples to hierarchically determine the set \mathcal{N}_t of nodes, the feature $x_{nt} \in \mathcal{X}_{nt}$ to be thresholded for each node $n \in \mathcal{N}_t$, and the corresponding threshold value ($t = 1, 2, \dots, T$). This is accomplished by minimizing an index of the impurity, in terms of class membership, of the two regions of the feature space determined by the decision rule on the node. More details can be found in [75] and [21].

Given an unknown d -dimensional feature vector \mathbf{x} to be classified, first, each t th tree is used to compute an estimate $y_t \in \mathcal{C}$ of the class membership of \mathbf{x} by applying the related sequence of decision rules to \mathbf{x} while the tree is traversed from its root to its leaves ($t = 1, 2, \dots, T$). Then, majority voting is used to fuse the decisions expressed by the trees in the forest, i.e., y_t is interpreted as a vote in favour of one of the classes, and \mathbf{x} is assigned to the class with the most votes out of the T trees.

The number F of features to be drawn on each node and the number T of trees are the parameters of the method. The accuracy of RF is generally not critically sensitive to their values. T clearly impacts the execution time [20] [75]. More details on the RF approach can be found in [20] along with the discussion of the generalization properties of the method.

6.5.2 Experimental validation

In this section the performance of the proposed RFs-based hybrid FDI strategy, the data-driven FDI with RF classifier and the model-based FDI with FSM are compared and assessed when the same operating conditions and faults are considered. The features used to discriminate between the four faulty classes and the non-faulty class may be either measurements of physical or chemical variables taken on the SOFC (data-driven approach) or residuals between the measured and predicted values of these variables (hybrid

6.5. Random forests for fault diagnosis

approach). As discussed in Sections 6.2 and 6.3, the set of training samples to be used to discriminate among the classes is generated through the physicochemical model in [130]. Here, the experimental validation is focused on the constant-voltage control and ten steady-state operating conditions are considered (ten design point and nine off-design points). Initially the first five variables in Table 6.1 are taken into account and the value for the maximum percentage error affecting each of the simulated variables (see Figure 6.4) is fixed to 4%. The procedure for the generation of the dataset is the same adopted for the previously described hybrid FDI approach based on SVM classifier, i.e., 400 possible combinations are identified between the ten operating conditions and the ten sizes for each of the four faults considered. If independent, random errors are introduced for each simulated variable of each combination; then, a set of 400 vectors of simulated measurements (see Figure 6.4) can be generated. Because each fault is represented by 100 vectors, 100 additional vectors can be added for the healthy status. They can be obtained by repeating the random errors for each operating condition ten times. Thus, a complete dataset is composed of 500 vectors of simulated measurements. Successive independent random errors allow for the generation of an arbitrary number of these datasets.

For the FDI strategies that use a vector of residuals for the online diagnosis, such as those depicted in Figures 6.1 and 6.3, the interaction between the model that replaces the real plant and the model that simulates the plant in healthy conditions enables the 500 vectors of simulated measurements to be transformed into just as many vectors of residuals. These residuals can be used to test the performance of a classical model-based FDI system, such as that depicted in Figure 1. However, the hybrid FDI system depicted in Figure 3 also needs an initial training phase. The residuals generated by some datasets can be used for this purpose, whereas the residuals generated by other datasets can be used for the performance test.

The residual generation is not necessary for the data-driven FDI system depicted in Figure 6.2: some datasets of simulated measurements can be used for the training phase (assuming that a supervised classifier is adopted), and other datasets can be used for testing the diagnostic performance.

Model-based strategy with fault signature matrix

The FSM expresses the association between each fault and a binary vector whose elements represent the residuals after the thresholding procedure. This matrix is typically arranged using a fault tree analysis, which is a deductive top-down tool used at the component level and based on the knowledge of the components' interactions at the system level [9] [134] [108]. Thus, in principle, the creation of the FSM does not require any collection of data or any plant model. However, when a statistically representative dataset is available, the FSM can be arranged from the analysis of the binarized residuals by associating to each fault the binary vector that is the most frequent when such a fault occurs. In this study, this option is followed.

To investigate the binary vectors produced by the considered faults, a relative threshold [104] [108] is set for each residual: the corresponding vector element is equal to one if the residual absolute value exceeds a percentage Ω of the predicted absolute value for that monitored variable; otherwise, it is equal to zero. Because the first five monitored variables listed in Table 6.1 are taken into account, 32 different binary vectors can be defined, ranging from [00000] to [11111]. In the considered vectors, the most significant bit is related to the residual of the variable $n. 1$ in Table 6.1, and so on. Figure 6.7 shows the binary vectors obtained for the healthy status and for every faulty status when the 500 vectors of a dataset are used, and the relative threshold Ω is initially set to be equal to 4.5%. This figure shows that each fault produces more than one binary vector; two or three primary vectors can be identified for each fault. Additionally, the figure shows that different faults can yield the same binary vector.

An accurate fault identification through the FSM is surely problematic in this situation, but this does not prevent achieving high performance in fault detection. If a fault is detected each time that the binary vector is different from [00000] (see Figure 6.1), Figure 6.7 clearly shows that missed alarms (i.e., the fault is not detected when a fault is actually present) are very rare events.

To maximize the fault identification performance, the FSM can be composed by associating each fault with the binary vector that has the maximum

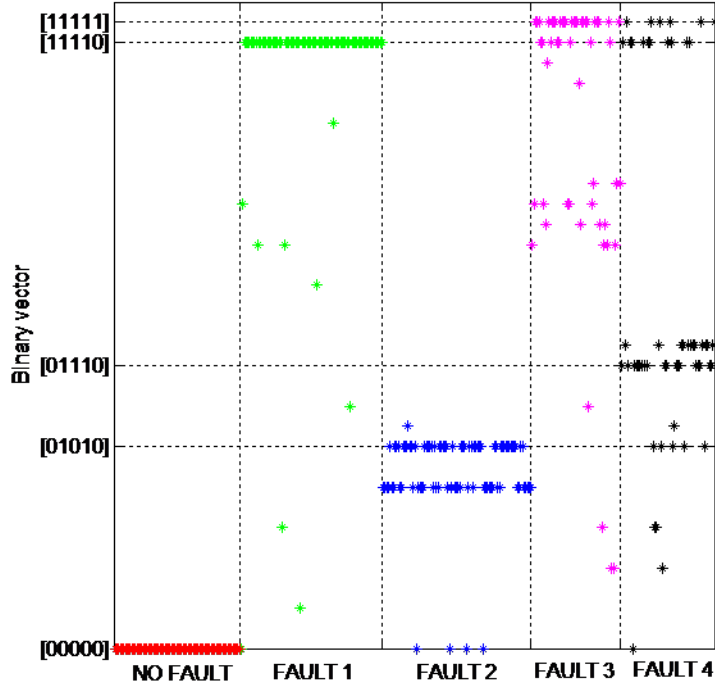


Figure 6.7: Binary vectors, from $[0\ 0\ 0\ 0\ 0]$ to $[1\ 1\ 1\ 1\ 1]$, produced by the model-based FDI system (see Figure 6.1) for the SOFC plant operating under healthy and faulty conditions. For a given fault, each asterisk indicates the binary vector produced by a given combination between the operating condition and fault size. The constant-voltage control for the SOFC plant is considered, the maximum percentage error is 4%, and the relative threshold Ω is 4.5% .

6.5. Random forests for fault diagnosis

Table 6.4: FSM of the model-based FDI system for the SOFC plant operating in constant-voltage conditions. The binary digits from R_1 to R_5 represent the residuals of the variables listed in Table 6.1 after the threshold operation.

Status	R_1	R_2	R_3	R_4	R_5
No fault	0	0	0	0	0
Fault n. 1	1	1	1	1	0
Fault n. 2	0	1	0	1	0
Fault n. 3	1	1	1	1	1
Fault n. 4	0	1	1	1	0

relative frequency for such a fault. Working in this way, the FSM presented in Table 6.4 is obtained.

The definition of the FSM enables the assessment of the FDI system performance. All the datasets that have not been exploited to arrange the FSM can be used. Note that if a vector of residuals generates a binary vector that does not correspond to any of the five vectors inside the FSM, the identification of the related status (i.e., no fault, fault n. 1, ..., fault n. 4) is not possible. Therefore, a fault is detected but is not classified. The relative threshold value is not necessarily equal to the value Ω used for arranging the FSM. Let Ξ be the percentage used as the relative threshold during the testing phase. Globally, the performance can be evaluated through the overall accuracy (OA), i.e., the fraction of test vectors whose status is correctly identified. We observed that the best OA is obtained when Ξ is between 4% and 5%. For Ξ belonging to this interval, an OA equal to 67% is obtained.

Clearly, the performance strictly depends on the FSM and in turn on the value Ω used to arrange such a matrix. If Ω ranges between 4% and 6%, then the association between faults and binary vectors is slightly different from that shown in Figure 6.7, but the maximum relative frequency occurs for the same binary vectors selected when $\Omega = 4.5\%$. Consequently, the FSM of Table 6.4 does not change and the OA is unaltered.

In contrast, if the relative threshold Ω is set significantly higher than the maximum percentage error due to modelling and measurement inaccuracies

6.5. Random forests for fault diagnosis

(e.g., $\Omega = 9\%$), then the number of missed alarms dramatically increases and the problem due to the association of different faults with the same binary vector worsens. Conversely, if Ω is set slightly lower than the maximum percentage error (e.g., $\Omega = 3\%$), then a considerable number of false alarms (i.e., a fault is detected when a fault is not actually present) occurs and the number of binary vectors produced by each fault dramatically increases. In both cases, an FSM different from that of Table 6.4 is obtained, and irrespective of the value of Ξ , the OA is definitely lower than that obtained for the case of $\Omega = 4.5\%$, $4\% \leq \Xi \leq 5\%$.

After many experiments are performed with different FSMs (i.e., different values for Ω) and using the value Ξ that maximizes the performance, the result is that the best OA that is attainable for the datasets described in Section 6.5.2 does not exceed 67%. The adoption of the fault tree analysis to arrange the FSM has no actual possibility of improving this performance because the FSM has been composed in such a way to maximize the OA for a given dataset. Although the latter is independent from the datasets used for the testing phase, it has the same statistical distribution of these datasets. Thus, the adopted FSM also represents the optimum choice for the test datasets. Instead, a performance improvement is potentially achievable by setting a specific threshold value for each residual. However, this optimization requires knowledge of the residuals' statistics in both normal and faulty states, encompassing all the operating conditions, which is difficult to reach.

Data-driven strategy with random forest

In this study, the data-driven FDI strategy depicted in Figure 6.2 was implemented by deploying a statistical supervised classifier assembled through RF. This classifier operates by receiving the measurements of the monitored variables as input features. The training of the RF classifier can be performed using a dataset composed as described in the first part of Section 6.5.2. The remaining independent datasets are used for testing of the data-driven FDI system.

The parameters used in the RF classifier are set as follows: $T = 10$, $F = 3$.

6.5. Random forests for fault diagnosis

Preliminary experiments, which are not reported for brevity, confirmed that these values allow for rather high accuracies to be obtained and that, most importantly, the accuracy of the classifier on the test set does not significantly change as T and F vary. For example, the accuracy remains remarkably stable as T varies from 10 to 100.

The obtained OA on the test set is equal to 72%. Although this value is 5% higher than that obtained with the FSM, this outcome cannot be considered a satisfactory result for the fault diagnosis task. Additionally, the performance for the basic fault detection task is poor: the relative frequency of correct fault detection is 91%, and the relative frequency of false alarms is approximately 27%. In particular, the last value is unacceptably high.

Hybrid strategy with plant model and random forest

In the FDI strategy depicted in Figure 6.3, the statistical classifier operates by receiving vectors of residuals as input features. The pool of independent datasets is used as follows: one dataset is used for training of the RF classifier, and the remaining datasets are used for testing of the hybrid FDI system.

The parameters used in the RF classifier are set as in the previous experiment with the data-driven approach (see Section 6.5.2).

In this case, the obtained OA is equal to 86%. The relative frequency of correct fault detection is 96%, and the relative frequency of false alarms is approximately 7%. Therefore, valuable performance is obtained for both fault detection and fault identification. Actually, the higher accuracy obtained by this hybrid approach than by the model-based and data-driven approaches in Sections 6.5.2 and 6.5.2 is an expected result because the former makes use of both supervised learning and physicochemical modeling whereas the other approaches can only benefit from one of these two contributions.

Because this strategy provides the highest OA (19% higher than the traditional model-based OA and 14% higher than the data-driven OA), an additional investigation is performed. To evaluate the FDI performance for each specific status of the SOFC plant (i.e., no fault, fault n. 1, ..., fault n. 4), the producer accuracy (PA) is introduced, i.e., the fraction of test vectors belonging to a given status that are correctly classified. Table 6.5 reports

6.5. Random forests for fault diagnosis

Table 6.5: OA and PA obtained by the hybrid FDI system when the five monitored variables of Table 6.1 are used, when the maximum temperature gradient (MTG) inside the SOFC stack is added, and when the cathodic activation losses (CAL) inside the SOFC stack are additionally introduced.

Status	Variables 1-5		Var. 1-5, MTG		Var. 1-5, MTG, CAL	
	PA	OA	PA	OA	PA	OA
No fault	93%		93%		96%	
Fault n. 1	87%		93%		95%	
Fault n. 2	94%	86%	94%	92%	96%	96%
Fault n. 3	82%		85%		99%	
Fault n. 4	75%		95%		96%	

the PA of each considered status, showing that no fault, faults n. 1 and 2 obtain a PA greater than or equal to the OA, whereas faults n. 3 and 4 obtain a PA lower than the OA. These faults are primarily confused with fault n. 1, which occurs with a relative frequency of 12% for fault n. 3 and a relative frequency of 19% for fault n. 4. This phenomenon is in accordance with Figure 6.7, where it can be observed that faults n. 3 and 4 occasionally produce the same binary vector associated with fault n. 1.

Hybrid strategy adding variables measured inside the fuel cell stack

To achieve a potential improvement in the PA of the hybrid FDI system, the introduction of an additional monitored variable, which is measured inside the SOFC stack, is proposed. As hypothesized in Section 6.3, it is assumed that the temperature gradient inside the SOFC stack can be measured and introduce the maximum temperature gradient as a monitored variable in addition to the first five variables listed in Table 6.1. Because here the real plant is replaced by a quantitative model (see Figure 6.4), this operation is easy. The generation of a pool of residual datasets is repeated (Section 6.5.2) and training of the RF classifier (Section 6.5.2) by considering six monitored variables in place of the previous five variables. The datasets not used for

6.6. Conclusion

the training are used for testing of the FDI system. The achieved OA and PAs are summarized in Table 6.5: the OA value reaches 92%, showing a 6% increase with respect to the previous case, and the PA values are equal to or greater than those of the previous case. In particular, only for fault n. 3 is the PA lower than the OA, being equal to 85%. The inclusion of the maximum temperature gradient significantly improved both the OA and PAs, pushing their values above 92%, except the PA for fault n. 3 that was only marginally increased, from 82% to 85%. This problem holds because fault n. 3 is still confused with fault n. 1.

To investigate whether the introduction of another physical quantity measured inside the FC stack can improve performance, particularly the PA of fault n. 3, the measurement of the cathodic activation losses is assumed to be possible and is included as an additional variable. As in the previous case, the quantitative model used in place of the real plant makes it possible to implement this measurement, the generation of a pool of residual datasets, and the training of the RF classifier by considering seven monitored variables. The OA and PAs obtained by also using the cathodic activation losses are summarized in Table 6.5: the OA value is now 96%, and all the PA values are greater than or equal to 95%. The high performance reached in this configuration for the fault diagnosis task is associated with the likewise high performance for the fault detection task: the relative frequency of correct fault detection is 97%, and the relative frequency of false alarms is approximately 4%.

6.6 Conclusion

In this chapter, two different hybrid FDI strategies have been developed for fault diagnosis in an SOFC power-generation plants. Specifically, RFS and SVM have been investigated as the pattern recognition techniques used in the FDI systems with the aim to find a solution that functions correctly in typical working conditions when several types of faults can occur with random sizes under several operating condition. Furthermore, an experimental comparison with other two classical FDI strategies (model-based with fault signature matrix and data-driven with RF classifier) has been carried out. The tests

6.6. Conclusion

performed point out that the model-based FDI schemes that adopt the FSM and purely data-driven FDI schemes do not provide satisfactory performance when numerous operating conditions for the SOFC plant and many sizes for each fault are considered. Rather, the hybrid FDI approaches in which supervised classifiers are used to analyze residuals generated through a quantitative plant model can reach a high rate of fault detection and identification and a low rate of false alarms. In particular, the SVM-based approach provides an OA that is generally higher than 0.8, even if the maximum percentage error reaches 6%. The SVM classifier performs better for constant-voltage control than for constant-current control, and the OA difference is approximately 0.075. The hypothetical inclusion of the residuals that are difficult to measure and the exact knowledge of the error magnitude do not significantly improve the performance: the increase in OA does not exceed 0.05. The experimental validation of the FDI scheme based on the RF classification approach was focused on the constant-voltage control strategy and even though this approach can obviously obtain different accuracies with respect to the SVM-based method on specific individual data sets on a case-by-case basis, the performance evaluation confirms that the adoption of supervised pattern recognition method in conjunction with physicochemical modeling turns out to be innovative and effective in the application of FDI in SOFC plant. The performance of the hybrid RF-based FDI strategy could be also further improved by including, among the monitored variables, two physical quantities measured inside the SOFC stack, namely, the maximum temperature gradient and cathodic activation losses. Although the practical measurement of these quantities during the FC functioning is currently extremely difficult, recent research results provide the opportunity to achieve future solutions. In the context of SOFC-based power generation, the following advantages of RFs have proved to be particularly important: (i) RF is a fully non-parametric classifier, a property that makes it possible to jointly exploit the multisensor information associated with measurements of highly heterogeneous physical and chemical variables; (ii) in addition to the aforementioned high classification accuracy, RF has demonstrated remarkable computational efficiency with execution times of a few seconds to complete the training and testing tasks in all experiments; (iii) the method includes only two param-

6.6. Conclusion

ters, which do not typically affect accuracy significantly and for which tuning is quite straightforward. Although this study primarily investigated experimental, laboratory-scale SOFC generation plants, no conceptual changes are required to apply the proposed FDI approach to plants of greater size and complexity. Although the level of complexity certainly increases, previous experiences with the modeling of plants and the development of classifiers suggest that the increased complexity does not prevent the possibility of successfully updating the FDI system.

Conclusion

Voice recognition, fingerprint identification, diagnostic support through biomedical images, optical character recognition, genomic or proteomic sequences analysis are just a few of the many examples that fall within pattern recognition applications with the aim to identify regularities (or pattern) in data belonging to heterogeneous sources. The pattern recognition field has experienced remarkable growth in the last decades and aims at realizing decision-making processes typically carried out by human brain, such as classification, by using computer algorithms. This thesis is focused on two thematic areas belonging to different application fields which have been tackled with a common pattern-recognition perspective. In particular in the first part of the manuscript several supervised classification methodologies have been developed in order to deal with land-use or land-cover mapping, forest inventory, urban planning and resource management. Within this framework the proposed approaches have been design in order to exploit the huge amount of information provided by last-generation VHR image sensors. First, the challenging problem of the integration of spatial-contextual information into kernel machines has been addressed by developing two novel integrated frameworks. The first one fuses together kernel-based SVM and Markov random fields along with efficient energy minimization techniques. The experimental validation has been carried out on hyperspectral images and multisensor optical/SAR datasets. The high values of classification accuracy pointed out the flexibility of the proposed methodology along with the effectiveness of graph cut and message passing technique for MRF energy minimization. In the second framework, spatial information is characterized using segmentation algorithms and a region-based kernel for VHR optical data is proposed using a Gaussian process model for each region of a seg-

mentation map and a suitable Hilbert space. Experiments on both simulated and real data were conducted showing high classification accuracy. The two approaches provided different and somewhat complementary insights on the capability to integrate spatial information into kernel-based learning and classification. It is worth noting that the two approaches are not mutually exclusive: the region-based kernel of the second proposed approach can be incorporated into arbitrary kernel machines, including the Markovian SVM of the first proposed approach.

Again with regard to the challenging task of extracting thematic information from VHR satellite imagery, a second addressed problem was the classification of data collected at multiple spatial resolutions. This is customary in the case of panchromatic and multispectral sensors associated with most current VHR satellite missions. The proposed methodological approach combines MRF, graph cuts, and Bayesian estimation into a multiresolution classifier that benefits from both panchromatic and multispectral observation to label at the panchromatic resolution. The experiments pointed out high accuracies and remarkable multiresolution fusion capabilities. The formalization and experimental validation were conducted with panchromatic/multispectral data. However, generalizing the method taking to take into account finest-resolution data and images with more than two resolutions is straightforward.

In supervised classification methods based on MRFs a key point also concerns the impact of MRF model parameter optimization on the classification performances. Therefore, this problem has been addressed in the thesis through the development of a parameter optimization technique that benefits of the effectiveness of the SMO algorithm in both memory occupation and computational time. The method evaluation was performed on a multiple datasets, which included multipolarization and multifrequency SAR, multispectral high-resolution, single date and multitemporal imagery. The method was compared with two previous techniques based on the Ho-Kashyap and Goldfard-Idnani numerical algorithms. Experiments showed a remarkable reduction in computational time along with high accuracy results.

As a last contribution on EO image analysis, an exploratory study was conducted on a rather new typology of data for the remote sensing field,

namely hyperspectral images in the thermal infrared range. Three supervised classification approaches and one feature reduction technique were tested on a 84-channel LWIR hyperspectral image with 1-m spatial resolution. In this case, the goal was essentially application-oriented: to investigate the capability to use this data modality, which has become of prominence only lately, for land cover mapping purposes. Experimental results pointed out that rather high classification accuracies could be reached only when the goal was to discriminate between "high-level" classes. The strong spectral overlapping of more detailed classes, clearly separable using data from the visible and near-infrared ranges but more similar in the thermal range feature space, does not allow a clear distinction among them. The results suggest that an improvement in the classification performance could be achieved by combining the pattern recognition approach with physically-based models for temperature and emissivity.

In the application to remote sensing image classification, the developed approaches have explored a variety of methodological and applicative problems with several common traits including the focus on VHR imagery and the role of Markov modeling. The future developments of these approaches, and more generally of the field of probabilistic graphical models, will face the challenging task of dealing with progressively further increasing spatial resolutions. Indeed, although the last generation satellite sensors is able to reach a spatial resolution of 30 cm, the adoption in this field of unmanned aerial vehicles, commonly known as drones, is already overcoming the current resolution limits of spaceborne sensors. The novel acquired imagery shows extremely high correlation between pixels belonging to extended neighborhoods, thus requiring for suitable methodological approaches able to exploit the consequent huge amount of information. In this context, MRF and MRF-type methods are expected to play a key role again. On one hand, they have proven to be effective tools to describe and include spatial information in the classification process. On the other hand, MRF-based classifiers are characterized by well-known parallelizability properties. Looking forward, this aspect is of paramount importance because, although the amount of information grows exponentially, a GPU-based parallelized implementation of the classification task is able to mitigate this challenging problem.

The second application field addressed in the manuscript was related to the development of fault detection and isolation techniques for power generation plants based on solid-oxide fuel cells. Automatic on-line diagnosis for FC power generation plants is of crucial importance for the progression of this technology. In particular, here this monitoring problem has been formalized with a pattern recognition approach and two different strategies have been proposed through the combination of statistical classifiers with a quantitative model of the SOFC plant. The dataset was collected through a physico-chemical considered model in which the capability of simulating faulty condition has been implanted. The high performances reached proved that the proposed FDI solutions are able to properly function in several operation condition of an SOFC plant and for many sizes of each possible fault. These results were of remarkable importance as compared to the current literature of FDI for fuel cells, which previously took benefit of the potential of pattern recognition approaches only marginally. Besides confirming the effectiveness of kernel machines and tree ensembles as flexible non-parametric classifiers, the results of the proposed FDI methods demonstrated the capability of current advanced learning methods in this topical application to sustainable energy as well. In this framework, a challenging task that could be tackled in the future will consist in the adoption of a new variable that has not been taken into account in the literature for the considered application, i.e the dynamics of the time evolution of the fuel cell. As a further connection point between the two macro-themes addressed in this Ph.D. thesis EO image classification and FDI in fuel cells, Markov modeling could be a valuable tool in this case as well. The theory of one-dimensional hidden Markov models could be properly applied in this field to model the time series of measurements taken on the fuel cell, further increase the correct fault detection probability, and reduce the relative frequency of false alarms.

Bibliography

- [1] Generic diagnosis instrument for sofc systems (genius), european union, collaborative project no. fch ju 245128. <http://www.fch-ju.eu/project/generic-diagnosis-instrument-sofc-systems>. Accessed: 2016-02-02.
- [2] Technical documentation integrated stack module, staxera gmbh, dresden, germany. http://www.fuelcellmarkets.com/content/images/articles/00864_AS_Dokumentation_ISM.pdf. Accessed: 2016-02-02.
- [3] AGLZIM, E.-H., ROUANE, A., AND EL-MOZNINE, R. An electronic measurement instrumentation of the impedance of a loaded fuel cell or battery. *Sensors* 7, 10 (2007), 2363–2377.
- [4] AIAZZI, B., ALPARONE, L., BARONTI, S., GARZELLI, A., AND SELVA, M. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogramm. Eng. Remote Sens.* 72, 5 (2006), 591–596.
- [5] AIAZZI, B., BARONTI, S., AND SELVA, M. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing* 45, 10 (2007), 3230–3239.
- [6] ALPARONE, L., AIAZZI, B., BARONTI, S., AND GARZELLI, A. *Remote Sensing Image Fusion*. CRC Press, 2015.
- [7] ARAI, E., SHIMABUKURO, Y. E., PEREIRA, G., AND VIJAYKUMAR, N. A multi-resolution multi-temporal technique for detecting and map-

- ping deforestation in the Brazilian Amazon rainforest. *Remote Sensing* 3, 9 (2011), 1943–1956.
- [8] ARENAS-CASTRO, S., JULIEN, Y., JIMENEZ-MUNOZ, J. C., SOBRINO, J. A., FERNANDEZ-HAEGER, J., AND JORDANO-BARBUDO, D. Mapping wild pear trees (*Pyrus bourgaeana*) in Mediterranean forest using high-resolution QuickBird satellite imagery. *Int. J. Remote Sensing* 34, 9–10 (2013), 3376–3396.
- [9] ARSIE, I., FILIPPI, A. D., MARRA, D., PIANESE, C., AND SORRENTINO, M. Fault tree analysis aimed to design and implement on-field fault detection and isolation schemes for sofc systems. In *Proceedings of the ASME 2010 Eighth International Fuel Cell Science, Engineering and Technology Conference* (2010), pp. 14–16.
- [10] BARALDI, A., DURIEUX, L., SIMONETTI, D., CONCHEDDA, G., HOLECZ, F., AND BLONDA, P. Automatic spectral-rule-based preliminary classification of radiometrically calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 imagery – Part I: System design and implementation. *IEEE Trans. Geosci. Remote Sensing* 48, 3 (2010), 1299–1325.
- [11] BEHLING, N. *Fuel Cells. Current Technology Challenges and Future Research Need*. Elsevier, Oxford, 2012.
- [12] BESAG, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* (1974), 192–236.
- [13] BESAG, J. On the statistical analysis of dirty pictures. *J. R. Statist. Soc.* 68 (1986), 259–302.
- [14] BIOUCAS-DIAS, J. M., PLAZA, A., DOBIGEON, N., PARENTE, M., DU, Q., GADER, P., AND CHANUSSOT, J. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing* 5 (2012), 354–379.

- [15] BLASCHKE, T., LANG, S., AND HAY, G. *Object-based image analysis: spatial concepts for knowledge-driven remote sensing applications*. Springer Science & Business Media, 2008.
- [16] BOUMAN, C. A., AND SHAPIRO, M. A multiscale random field model for bayesian image segmentation. *IEEE Transactions on image processing* 3, 2 (1994), 162–177.
- [17] BOVOLO, F., BRUZZONE, L., CAPOBIANCO, L., GARZELLI, A., MARCHESI, S., AND NENCINI, F. Analysis of the effects of pan-sharpening in change detection on VHR images. *IEEE Geosci. Remote Sensing Letters* 7, 1 (2010), 53–57.
- [18] BOYKOV, Y., AND KOLMOGOROV, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Machine Intell.* 26, 9 (2004), 1124–1137.
- [19] BOYKOV, Y., VEKSLER, O., AND ZABIH, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Machine Intell.* 23, 11 (2001), 1222–1239.
- [20] BREIMAN, L. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [21] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. *Classification and regression trees*. Chapman & Hall/CRC, 1984.
- [22] BRENT, R. P. *Algorithms for minimization without derivatives*. Prentice-Hall, 1973.
- [23] BRUZZONE, L., ROLI, F., AND SERPICO, S. An extension of the jeffreys-matusita distance to multiclass cases for feature selection. *IEEE Trans. Geosci. Remote Sensing* 33, 6 (1995), 1318–1321.
- [24] CAMPS-VALLS, G., GOMEZ-CHOVA, L., MUNOZ-MARI, J., VILA-FRANCES, J., AMOROS-LOPEZ, J., AND CALPE-MARAVILLA, J. Retrieval of oceanic chlorophyll concentration with relevance vector machines. *Remote Sens. Environ.* 105, 1 (2006), 23–33.

- [25] CAMPS-VALLS, G., GOMEZ-CHOVA, L., MUNOZ-MARI, J., VILA-FRANCES, J., AND CALPE-MARAVILLA, J. Composite kernels for hyperspectral image classification. *IEEE Geosci. Remote Sensing Letters* 3 (2006), 93–97.
- [26] CAO, X., XU, L., MENG, D., ZHAO, Q., AND XU, Z. Integration of 3-dimensional discrete wavelet transform and markov random field for hyperspectral image classification. *Neurocomputing* 226 (2017), 90–100.
- [27] CHANG, C.-C., AND LIN, C.-J. *LIBSVM: a library for support vector machines*, 2001. Online: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [28] CHAPELLE, O., VAPNIK, V., BOUSQUET, O., AND MUKHERJEE, S. Choosing multiple parameters for support vector machines. *Mach. Learn.* 46 (2002), 131–159.
- [29] CHEN, D., AND STOW, D. Strategies for integrating information from multiple spatial resolutions into land-use/land-cover classification routines. *Photogramm. Eng. Rem. Sensing* 69, 11 (2003), 1279–1287.
- [30] CHIANG, L. H., RUSSELL, E. L., AND BRAATZ, R. D. *Fault detection and diagnosis in industrial systems*. Springer Science & Business Media, 2000.
- [31] CHRISTOPHER, M. B. *PATTERN RECOGNITION AND MACHINE LEARNING*. Springer-Verlag New York, 2016.
- [32] COLDITZ, R., WEHRMANN, T., BACHMANN, M., STEINNOCHER, K., SCHMIDT, M., STRUNZ, G., AND DECH, S. Influence of image fusion approaches on classification accuracy: A case study. *Int. J. Remote Sensing* 27, 15 (2006), 3311–3335.
- [33] COMMINGES, C., FU, Q. X., ZAHID, M., STEINER, N. Y., AND BUCHELI, O. Monitoring the degradation of a solid oxide fuel cell stack during 10,000 h via electrochemical impedance spectroscopy. *Electrochimica Acta* 59 (2012), 367–375.

- [34] COSTAMAGNA, P., DE GIORGI, A., MAGISTRI, L., MOSER, G., PELLACO, L., AND TRUCCO, A. A classification approach for model-based fault diagnosis in power generation systems based on solid oxide fuel cells. *IEEE Transactions on Energy Conversion* 31, 2 (2016), 676–687.
- [35] COVER, T. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, 3 (1965), 326–334.
- [36] CRISTIANINI, N., AND SHAW-TEY, J. *An Introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.
- [37] CUBERO-CASTAN, M., CHANUSSOT, J., ACHARD, V., BRIOTTET, X., AND SHIMONI, M. A physics-based unmixing method to estimate subpixel temperatures on mixed pixels. *IEEE Trans. Geosci. Remote Sensing* 53, 1 (2015), 1894–1906.
- [38] DE GIORGI, A., MOSER, G., AND SERPICO, S. Contextual remote-sensing image classification through support vector machines, markov random fields and graph cuts. In *Proc. of the 2014 IGARSS Symposium, Quebec, Canada, 13-18 July* (2014), pp. 3722–3725.
- [39] D’ELIA, C., POGGI, G., AND SCARPA, G. A tree-structured markov random field model for bayesian image segmentation. *IEEE Transactions on Image Processing* 12, 10 (2003), 1259–1273.
- [40] DELLEPIANE, S., FONTANA, F., AND VERNAZZA, G. L. Nonlinear image labeling for multivalued segmentation. *IEEE Transactions on Image Processing* 5, 3 (1996), 429–446.
- [41] DENG, H., AND CLAUSI, D. Unsupervised segmentation of synthetic aperture radar sea ice imagery using a novel markov random field model. *IEEE Transactions on Geoscience and Remote Sensing* 43, 3 (2005), 528–538.

- [42] DERIN, H., AND ELLIOTT, H. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 1 (1987), 29–55.
- [43] DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comp.* 10 (1998), 1895–1923.
- [44] DOBIGEON, N., TOURNERET, J., RICHARD, C., BERMUDEZ, J. C. M., MCLAUGHLIN, S., AND HERO, A. O. Nonlinear unmixing of hyperspectral images. *IEEE Signal Processing Magazine* (2014), 82–94.
- [45] DUDA, R. O., HART, P. E., AND STORK, D. G. *Pattern classification*. Wiley Interscience, 2001.
- [46] ESCOBET, T., FEROLDI, D., LIRA, S. D., PUIG, V., QUEVEDO, J., RIERA, J., AND SERRA, M. Model-based fault diagnosis in pem fuel cell systems. *Journal of Power Sources* 192, 1 (2009), 216–223.
- [47] FAN, R.-E., CHEN, P.-H., AND LIN, C.-J. Working set selection using second order information for training SVM. *J. Mach. Learn. Res.* 6 (2005), 1889–1918.
- [48] FELZENSZWALB, P. F., AND HUTTENLOCHER, D. Efficient graph-based image segmentation. *International Journal Computer Vision* 59 (2004), 167–181.
- [49] FOODY, G. M. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogramm. Eng. Remote Sensing* 70, 5 (2004), 627–633.
- [50] FORZIERI, G., MOSER, G., VIVONI, E. R., CASTELLI, F., AND CANOVARO, F. Riparian vegetation mapping for hydraulic roughness estimation using very high resolution remote sensing data fusion. *J. Hydraul. Eng.* 136 (2010), 855–867.

- [51] FRAWLEY, W., PIATETSKY-SHAPIO, G., AND MATHEUS, C. Knowledge discovery in databases: An overview. *AI magazine* 13, 3 (1992), 57.
- [52] GAETANO, R., MASI, G., POGGI, G., VERDOLIVA, L., AND SCARPA, G. Marker-controlled watershed-based segmentation of multiresolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 53, 6 (2015), 2987–3004.
- [53] GAO, Z., CECATI, C., AND DING, S. X. A survey of fault diagnosis and fault-tolerant techniquespart i: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics* 62, 6 (2015), 3757–3767.
- [54] GARZELLI, A., NENCINI, F., AND CAPOBIANCO, L. Optimal mmse pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing* 46, 1 (2008), 228–236.
- [55] GEMAN, S., AND GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6 (1984), 721–741.
- [56] GOLDFARB, D., AND IDNANI, A. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* 27 (1983), 1–33.
- [57] GOLIPOUR, M., GHASSEMIAN, H., AND MIRZAPOUR, F. Integrating hierarchical segmentation maps with MRF prior for classification of hyperspectral images in a bayesian framework. *IEEE Transactions on Geoscience and Remote Sensing* 54, 2 (2016), 805–816.
- [58] GRECO, A., SORCE, A., LITWIN, R., COSTAMAGNA, P., AND MAGISTRI, L. Reformer faults in sofc systems: Experimental and modeling analysis, and simulated fault maps. *International Journal of Hydrogen Energy* 39, 36 (2014), 21700–21713.

- [59] GREIG, D. M., PORTEOUS, B. T., AND SEHEULT, A. H. Exact maximum a posteriori estimation for binary images. *J. R. Statist. Soc. B* 51, 2 (1989), 271–279.
- [60] GUAN, W. B., ZHAI, H. J., JIN, L., XU, C., AND WANG, W. G. Temperature measurement and distribution inside planar sofc stacks. *Fuel cells* 12, 1 (2012), 24–31.
- [61] GUO, J., TAKAHASHI, N., AND NISHI, T. An efficient method for simplifying decision functions of support vector machines. *IEICE Trans. Fundamentals E* 89-A, 10 (2006), 2795–2802.
- [62] HESTER, D. B., CAKIR, H. I., NELSON, A. C., AND KHORRAM, S. Per-pixel classification of high spatial resolution satellite imagery for urban land-cover mapping. *Photogramm. Eng. Rem. Sensing* 74, 4 (2008), 463–471.
- [63] HEYLEN, R., PARENTE, M., AND GADER, P. A review of nonlinear hyperspectral unmixing methods. *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing* 7 (2014), 1844–1868.
- [64] HUBER, T. M., OPITZ, A. K., KUBICEK, M., HUTTER, H., AND FLEIG, J. Temperature gradients in microelectrode measurements: Relevance and solutions for studies of sofc electrode materials. *Solid State Ionics* 268 (2014), 82–93.
- [65] HUGHES, G. F. On the mean accuracy of statistical pattern recognizers. *IEEE Tran. Inform. Theory* 14, 1 (1968), 55–63.
- [66] IBANEZ, M. V., AND SIMO', A. Parameter estimation in Markov random field image modeling with imperfect observations. A comparative study. *Pattern Rec. Lett.* 24, 14 (2003), 2377–2389.
- [67] IHLER, A., FISHER III, J., AND WILLSKY, A. Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research* 6 (2005).

- [68] ISERMANN, R. Supervision, fault-detection and fault-diagnosis methodsan introduction. *Control engineering practice* 5, 5 (1997), 639–652.
- [69] ISERMANN, R. Model-based fault-detection and diagnosis–status and applications. *Annual Reviews in control* 29, 1 (2005), 71–85.
- [70] JIA, X., AND RICHARDS, J. A. Cluster-space representation for hyperspectral data classification. *IEEE Trans. Geosci. Remote Sensing* 40, 3 (2002), 593–598.
- [71] KALLEL, A., OTTLE, C., HEGARAT-MASCLE, S. L., MAIGNAN, F., AND COURAULT, D. Surface temperature downscaling from multiresolution instruments based on Markov models. *IEEE Trans. Geosci. Remote Sensing* 51, 3 (2013), 1588–1612.
- [72] KATO, Z., AND ZERUBIA, J. *Markov Random Fields in Image Segmentation*. NOW publishers, 2012.
- [73] KOLMOGOROV, V. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 10 (2006), 1568–1583.
- [74] KOLMOGOROV, V., AND ZABIH, R. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Machine Intell.* 26, 2 (2004), 147–159.
- [75] KUNCHEVA, L. *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004.
- [76] KUNT, M. A statistical model for correlation functions of two-level digital facsimiles. *Proceedings of the IEEE* 63, 2 (1975), 327–329.
- [77] KUO, L.-S., HUANG, H.-H., YANG, C.-H., AND CHEN, P.-H. Real-time remote monitoring of temperature and humidity within a proton exchange membrane fuel cell using flexible sensors. *Sensors* 11, 9 (2011), 8674–8684.

- [78] KURTZ, C., PASSAT, N., GANCARSKI, P., AND PUISSANT, A. Extraction of complex patterns from multiresolution remote sensing images: A hierarchical top-down methodology. *Pattern Recognition* 45, 2 (2012), 685–706.
- [79] KURTZ, C., STUMPF, A., MALET, J.-P., GANCARSKI, P., PUISSANT, A., AND PASSAT, N. Hierarchical extraction of landslides from multiresolution remotely sensed optical images. *ISPRS J. Photogramm. Remote Sensing* 87 (2014), 122–136.
- [80] LABEN, C. A., AND BROWER, B. V. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening, 2000.
- [81] LANDGREBE, D. A. *Signal theory methods in multispectral remote sensing*. Wiley, 2003.
- [82] LANG, M., AUER, C., EISMANN, A., SZABO, P., AND WAGNER, N. Investigation of solid oxide fuel cell short stacks for mobile applications by electrochemical impedance spectroscopy. *Electrochimica Acta* 53, 25 (2008), 7509–7513.
- [83] LEE, C.-Y., FAN, W.-Y., AND CHANG, C.-P. A novel method for in-situ monitoring of local voltage, temperature and humidity distributions in fuel cells using flexible multi-functional micro sensors. *Sensors* 11, 2 (2011), 1418–1432.
- [84] LI, S. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.
- [85] LIAO, W., HUANG, X., COILLIE, F. V., GAUTAMA, S., PIZURICA, A., PHILIPS, W., LIU, H., ZHU, T., SHIMONI, M., MOSER, G., , AND TUIA, D. Processing of multiresolution thermal hyperspectral and digital color data: outcome of the 2014 ieee grss data fusion contest. *IEEE Journal Selected Topics in Applied Earth Observations and Remote Sensing* 8, 62 (2015), 2984–2996.
- [86] LU, D., BATISTELLA, M., AND MORAN, E. Integration of Landsat TM and SPOT HRG images for vegetation change detection in the

- Brazilian Amazon. *Photogramm. Eng. Rem. Sensing* 74, 4 (2008), 421–430.
- [87] LUO, J., NAMBURU, M., PATTIPATI, K. R., QIAO, L., AND CHIGUSA, S. Integrated model-based and data-driven diagnosis of automotive antilock braking systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40, 2 (2010), 321–336.
- [88] MA, W., BIOUCAS-DIAS, J. M., CHAN, T., GILLIS, N., GADER, P., PLAZA, A. J., AMBIKAPATHI, A., AND CHI, C. A signal processing perspective on hyperspectral unmixing. *IEEE Signal Processing Magazine* (2014), 67–81.
- [89] MACGREGOR, J., AND CINAR, A. Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods. *Computers & Chemical Engineering* 47 (2012), 111–120.
- [90] MARRA, D., SORRENTINO, M., PIANESE, C., AND IWANSCHITZ, B. A neural network estimator of solid oxide fuel cell performance for on-field diagnostics and prognostics applications. *Journal of Power Sources* 241 (2013), 320–329.
- [91] MELGANI, F., AND SERPICO, S. B. A Markov random field approach to spatio-temporal contextual image classification. *IEEE Trans. Geosci. Remote Sensing* 41, 11 (2003), 2478–2487.
- [92] MILLICHAMP, J., MASON, T. J., BRANDON, N. P., BROWN, R. J. C., MAHER, R. C., MANOS, G., NEVILLE, T. P., AND BRETT, D. J. L. A study of carbon deposition on solid oxide fuel cell anodes using electrochemical impedance spectroscopy in combination with a high temperature crystal microbalance. *Journal of Power Sources* 235 (2013), 14–19.
- [93] MOREL, B., ROBERGE, R., SAVOIE, S., NAPPORN, T. W., AND MEUNIER, M. An experimental evaluation of the temperature gradient in solid oxide fuel cells. *Electrochemical and solid-state letters* 10, 2 (2007), B31–B33.

- [94] MOSER, G., COSTAMAGNA, P., DE GIORGI, A., GRECO, A., MAGISTRI, L., PELLACO, L., AND TRUCCO, A. Joint feature and model selection for svm fault diagnosis in solid oxide fuel cell systems. *Mathematical Problems in Engineering 2015* (2015).
- [95] MOSER, G., AND SERPICO, S. B. Classification of high-resolution images based on MRF fusion and multiscale segmentation. In *Proc. of IGARSS-2008, Boston, USA, 7-11 July 2008* (2008).
- [96] MOSER, G., AND SERPICO, S. B. Automatic parameter optimization for support vector regression for land and sea surface temperature estimation from remote-sensing data. *IEEE Trans. Geosci. Remote Sensing* 47 (2009), 909–921.
- [97] MOSER, G., AND SERPICO, S. B. Combining support vector machines and markov random fields in an integrated framework for contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing* 51 (2013), 2734–2752.
- [98] MOSER, G., SERPICO, S. B., AND BENEDIKTSSON, J. A. Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE* 101 (2013), 631–651.
- [99] NAJAFI, B., MAMAGHANI, A. H., RINALDI, F., AND CASALEGNOA, A. Long-term performance analysis of an ht-pem fuel cell based micro-chp system: operational strategies. *Applied Energy* 147 (2015), 582–592.
- [100] NEF, T., URWYLER, P., BÜCHLER, M., TARNANAS, I., STUCKI, R., CAZZOLI, D., MÜRI, R., AND MOSIMANN, U. Evaluation of three state-of-the-art classifiers for recognition of activities of daily living from smart home ambient data. *Sensors* 15, 5 (2015), 11725–11740.
- [101] NG, S. S. Y., TSE, P. W., AND TSUI, K. L. A one-versus-all class binarization strategy for bearing diagnostics of concurrent defects. *Sensors* 14, 1 (2014), 1295–1321.

- [102] NIKOLAKOPOULOS, K. G. Comparison of nine fusion techniques for very high resolution data. *Photogramm. Eng. Rem. Sensing* 74, 5 (2008), 647–659.
- [103] PELLACO, L., COSTAMAGNA, P., DE GIORGI, A., GRECO, A., MAGISTRI, L., MOSER, G., AND TRUCCO, A. Fault diagnosis in fuel cell systems using quantitative models and support vector machines. *Electronics Letters* 50, 11 (2014), 824–826.
- [104] PETRONE, R., ZHENG, Z., HISSEL, D., PÉRA, M., PIANESE, C., SORRENTINO, M., BECHERIF, M., AND YOUSFI-STEINER, N. A review on model-based diagnosis methodologies for pemfcs. *International Journal of Hydrogen Energy* 38 (2013), 7077–7091.
- [105] PLATT, J. Sequential minimal optimization: a fast algorithm for training support vector machines. Technical report msr-tr-98-14, Microsoft Research, 1998.
- [106] PLATT, J. *Probabilistic outputs for support vector machines and comparison to regularized likelihood methods*. in Advances in large margin classifiers, ed.: A. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans, MIT Press, 2000.
- [107] POGGI, G., SCARPA, G., AND ZERUBIA, J. Supervised segmentation of remote sensing images based on a tree-structured mrf model. *IEEE Transactions on Geoscience and Remote Sensing* 43, 8 (2005), 1901–1911.
- [108] POLVERINO, P., PIANESE, C., SORRENTINO, M., AND MARRA, D. Model-based development of a fault signature matrix to improve solid oxide fuel cell systems on-site diagnosis. *Journal of Power Sources* 280 (2015), 320–338.
- [109] POWELL, M. J. D. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comp. J.* 7 (1964), 155–162.
- [110] PRATT, W. K. *Digital image processing*. Wiley Interscience, 2007.

- [111] PRESS, W. H., TEUKOLSKY, S. A., WETTERLING, W. T., AND FLANNERY, B. P. *Numerical recipes in C*. Cambridge University Press, Cambridge (UK), 2007.
- [112] PRICE, J. C. Combining multispectral data of differing spatial resolution. *IEEE Trans. Geosci. Remote Sens.* 37, 3 (1999), 1199–1203.
- [113] PUYOU-LASCASSIES, P., PODAIRE, A., AND GAY, M. Extracting crop radiometric responses from simulated low and high spatial resolution satellite data using a linear mixing model. *Int. J. Remote Sens.* 15, 18 (1994), 3767–3784.
- [114] REDNER, R. A., AND WALKER, H. F. Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review* 26, 2 (1984), 195–239.
- [115] RICHARDS, J., AND JIA, X. *Remote sensing digital image analysis*. Springer, 2006.
- [116] ROBIN, A., LE HEGARAT-MASCLE, S., AND MOISAN, L. Unsupervised subpixelic classification using coarse-resolution time series and structural information. *IEEE Trans. Geosci. Remote Sensing* 46, 5 (2014), 1359–1374.
- [117] RUIZ-GONZALEZ, R., GOMEZ-GIL, J., GOMEZ-GIL, F. J., AND MARTÍNEZ-MARTÍNEZ, V. An svm-based classifier for estimating the state of various rotating components in agro-industrial machinery with a vibration signal acquired from a single point on the machine chassis. *Sensors* 14, 11 (2014), 20713–20735.
- [118] SANTOS, P., VILLA, L. F., REÑONES, A., BUSTILLO, A., AND MAUDES, J. An svm-based solution for fault detection in wind turbines. *Sensors* 15, 3 (2015), 5627–5648.
- [119] SERPICO, S. B., AND BRUZZONE, L. A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE Trans. Geosci. Remote Sensing, Special Issue on Analysis of Hyperspectral Image Data* 39, 7 (1994), 1360–1367.

- [120] SERPICO, S. B., D'INCÀ, M., MELGANI, F., AND MOSER, G. A comparison of feature reduction techniques for classification of hyperspectral remote-sensing data. In *Proceedings of the SPIE Conference on Image and Signal Processing for Remote Sensing VIII, Crete, Greece, 22-27 September (2002)*, pp. 347–358.
- [121] SERPICO, S. B., AND MOSER, G. Weight parameter optimization by the Ho-Kashyap algorithm in MRF models for supervised image classification. *IEEE Trans. Geosci. Remote Sensing* 44 (2006), 3695–3705.
- [122] SERPICO, S. B., AND MOSER, G. Extraction of spectral channels from hyperspectral images for classification purposes. *IEEE Trans. Geosci. Remote Sensing* 45 (2007), 484–495.
- [123] SHAFARENKO, L., PETROU, M., AND KITTLER, J. Automatic watershed segmentation of randomly textured color images. *IEEE transactions on Image Processing* 6, 11 (1997), 1530–1544.
- [124] SHIMONI, M., HAEELTERMAN, R., AND PERNEEL, C. Short temporal change detection in complex urban area. In *Proceedings of the 2013 IGARSS Symposium, Melbourne, Australia, July (2013)*, pp. 2196–2199.
- [125] SHIMONI, M., AND PERNEEL, C. Dedicated classification method for thermal hyperspectral imaging. In *Proceedings of the 2012 IGARSS Symposium, Munich, Germany, July (2012)*, pp. 1397–1400.
- [126] SMITS, P., AND DELLEPIANE, S. Synthetic aperture radar image segmentation by a detail preserving Markov Random Field approach. *IEEE Trans. Geosci. Remote Sensing* 35, 4 (1997), 844–857.
- [127] SOBHANI-TEHRANI, E., AND KHORASANI, K. *Fault diagnosis of non-linear systems using a hybrid approach*, vol. 383. Springer Science & Business Media, 2009.

- [128] SOLBERG, A. H. S., TAXT, T., AND JAIN, A. K. A Markov random field model for classification of multisource satellite imagery. *IEEE Trans. Geosci. Remote Sensing* 34, 1 (1996), 100–113.
- [129] SON, J., NIU, G., YANG, B., HWANG, D., AND KANG, D. Development of smart sensors system for machine fault diagnosis. *Expert systems with applications* 36, 9 (2009), 11981–11991.
- [130] SORCE, A., GRECO, A., MAGISTRI, L., AND COSTAMAGNA, P. Fdi oriented modeling of an experimental sofc system, model validation and simulation of faulty states. *Applied Energy* 136 (2014), 894–908.
- [131] SORRENTINO, M., AND PIANESE, C. Grey-box modeling of sofc unit for design, control and diagnostics applications. In *In Proceedings of the European Fuel Cell Forum 2009, Lucerne, Switzerland, 29 June 2009* (2009).
- [132] SRINIVASAN, S. *Stochastic processes in engineering systems*. Springer-Verlag, New York, 2006.
- [133] STATNIKOV, A., WANG, L., AND ALIFERIS, C. F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics* 9, 1 (2008), 319.
- [134] STEINER, N. Y., HISSEL, D., MOÇOTÉGUY, P., CANDUSSO, D., MARRA, D., PIANESE, C., AND SORRENTINO, M. Application of fault tree analysis to fuel cell diagnosis. *Fuel Cells* 12, 2 (2012), 302–309.
- [135] STORVIK, G., FJORTOFT, R., AND SOLBERG, A. H. S. A Bayesian approach to classification of multiresolution remote sensing data. *IEEE Trans. Geosci. Remote Sensing* 43, 3 (2005), 539–547.
- [136] SZELISKI, R., ZABIH, R., SCHARSTEIN, D., VEKSLER, O., KOLMOGOROV, V., AGARWALA, A., TAPPEN, M., AND ROTHER, C. A comparative study of energy minimization methods for Markov random

- p>fields. In
- Proc. of the Ninth European Conference on Computer Vision (ECCV-2006), Graz, Austria, May 2006*
- (2006), vol. 2, pp. 16–29.
- [137] SZELISKI, R., ZABIH, R., SCHARSTEIN, D., VEKSLER, O., KOLMOGOROV, V., AGARWALA, A., TAPPEN, M., AND ROTHER, C. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 6 (2008), 1068–1080.
 - [138] SZIRANYI, T., ZERUBIA, J., CZUNI, L., GELDREICH, D., AND KATO, Z. Image segmentation using Markov random field model in fully parallel cellular network architectures. *Real-Time Imaging* 6, 3 (2000), 195–211.
 - [139] TAPPEN, M. F., AND FREEMAN, W. T. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Proceedings of the IEEE International Conference on Computer Vision* (2003), vol. 2, pp. 900–907.
 - [140] TARABALKA, Y., CHANUSSOT, J., AND BENEDIKTSSON, J. A. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. *IEEE Trans. Syst. Man Cybern. B* 40, 5 (2010), 1267–1279.
 - [141] TAYLOR, S., KUMAR, L., AND REID, N. Mapping lantana camara: Accuracy comparison of various fusion techniques. *Photogramm. Eng. Rem. Sensing* 79, 6 (2010), 691–700.
 - [142] TEGGI, S., CECCHI, R., AND SERAFINI, F. TM and IRS-1C-PAN data fusion using multiresolution decomposition methods based on the a-trous algorithm. *Int. J. Remote Sensing* 24, 6 (2006), 1287–1301.
 - [143] THOMAS, C., RANCHIN, T., WALD, L., AND CHANUSSOT, J. Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics. *IEEE Trans. Geosci. Remote Sensing* 46 (2008), 1301–1312.

- [144] TILTON, J. C. Analysis of hierarchically related image segmentations. In *Proc. of the IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, Greenbelt, MD, USA, 27-28 ottobre 2003* (2003), vol. 39, pp. 553–561.
- [145] TOU, J. T., AND GONZALEZ, R. C. *Pattern Recognition Principles*. Addison-Wesley, 1974.
- [146] TOUSSAINT, G. The use of context in pattern recognition. *Pattern Recognition* 10, 3 (1978), 189–204.
- [147] TROGLIO, G., LE MOIGNE, J., BENEDIKTSSON, J. A., MOSER, G., AND SERPICO, S. B. Automatic extraction of ellipsoidal features for planetary image registration. *IEEE Geoscience and Remote Sensing Letters* 9, 1 (2012), 95–99.
- [148] TSOUMAS, I., GEORGOULAS, G., MITRONIKAS, E., AND SAFACAS, A. N. Asynchronous machine rotor fault diagnosis technique using complex wavelets. *IEEE Transactions on Energy Conversion* 23, 2 (2008), 444–459.
- [149] UDUPA, J. K., AND SAHA, P. K. Fuzzy connectedness and image segmentation. *Proceedings of the IEEE* 91, 10 (2003), 1649–1669.
- [150] VAPNIK, V., AND CHAPPELLE, O. Bounds on error expectation for support vector machines. *Neural Comp.* 12 (2000), 2013–2036.
- [151] VAPNIK, V. N. *The nature of statistical learning theory*. Springer, 1995.
- [152] VAPNIK, V. N. *Statistical learning theory*. Wiley, 1998.
- [153] VENKATASUBRAMANIAN, V., RENGASWAMY, R., YIN, K., AND KAVURI, S. N. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & chemical engineering* 27, 3 (2003), 293–311.
- [154] VIVONE, G., ALPARONE, L., CHANUSSOT, J., DALLA MURA, M., GARZELLI, A., LICCIARDI, G. A., RESTAINO, R., AND WALD, L.

- A critical comparison among pansharpening algorithms. *IEEE Trans. Geosci. Remote Sensing* 53 (2015), 2565–2586.
- [155] VOISIN, A., KRYLOV, V. A., MOSER, G., SERPICO, S. B., AND ZERUBIA, J. Supervised classification of multisensor and multiresolution remote sensing images with a hierarchical copula-based approach. *IEEE Trans. Geosci. Remote Sensing* 52, 6 (2014), 3346–3358.
- [156] WAINWRIGHT, M. J., JAAKKOLA, T. S., AND WILLSKY, A. S. Map estimation via agreement on trees: Message-passing and linear programming. *IEEE Transactions on Information Theory* 51, 11 (2005), 3697–3717.
- [157] WANG, K., HISSEL, D., PÉRA, M. C., STEINER, N., MARRA, D., SORRENTINO, M., PIANESE, C., MONTEVERDE, M., CARDONE, P., AND SAARINEN, J. A review on solid oxide fuel cell models. *International Journal of Hydrogen Energy* 36, 12 (2011), 7212–7228.
- [158] WANG, L., WU, L., GUAN, Y., AND WANG, G. Online sensor fault detection based on an improved strong tracking filter. *Sensors* 15, 2 (2015), 4578–4591.
- [159] WEMMERT, C., PUISSANT, A., FORESTIER, F., AND GANCARSKI, P. Multiresolution remote sensing image clustering. *IEEE Geosci. Remote Sensing Letters* 6, 3 (2009), 533–537.
- [160] WHITNEY, A. W. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers* 100, 9 (1971), 1100–1103.
- [161] WIDODO, A., AND YANG, B.-S. Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical systems and signal processing* 21, 6 (2007), 2560–2574.
- [162] WU, T.-F., LIN, C.-J., AND WENG, R. C. Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.* 5 (2004), 975–1005.

- [163] YANG, B., DI, X., AND HAN, T. Random forests classifier for machine fault diagnosis. *Journal of mechanical science and technology* 22, 9 (2008), 1716–1725.
- [164] YIN, S., DING, S. X., XIE, X., AND LUO, H. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics* 61, 11 (2014), 6418–6428.
- [165] YU, H., GAO, L., LI, J., LI, S. S., ZHANG, B., AND BENEDIKTSSON, J. A. Spectral-spatial hyperspectral image classification using subspace-based support vector machines and adaptive Markov random fields. *Remote Sensing* 8, 4 (2016).
- [166] YUAN, F. Land-cover change and environmental impact analysis in the greater munkato area of minnesota using remote sensing and gis modelling. *Int. J. Remote Sensing* 29, 4 (2008), 1169–1184.
- [167] ZHENG, Z., PETRONE, R., PÉRA, M. C., HISSEL, D., BECHERIF, M., PIANESE, C., STEINER, N. Y., AND SORRENTINO, M. A review on non-model based diagnosis methodologies for pem fuel cell stacks and systems. *International Journal of Hydrogen Energy* 38, 21 (2013), 8914–8926.
- [168] ZHUKOV, B., OERTEL, D., LANZL, F., AND REINHACKEL, G. Unmixing-based multisensor multiresolution image fusion. *IEEE Trans. Geosci. Remote Sens.* 37, 3 (1999), 1212–1226.