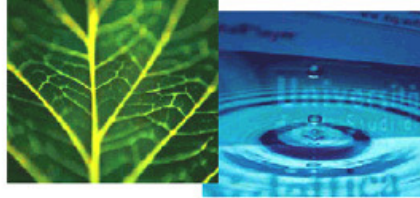


PhD Dissertation



**International Doctorate School in Information and
Communication Technologies**

University of Trento

**ADVANCED METHODS FOR THE ANALYSIS OF
MULTISPECTRAL AND MULTITEMPORAL REMOTE SENSING
IMAGES**

Massimo Zanetti

Advisor:

Prof. Lorenzo Bruzzone

Università degli Studi di Trento

April 2017

Abstract

The increasing availability of new generation remote sensing satellite multispectral images provides an unprecedented source of information for Earth observation and monitoring. Multispectral images can be now collected at high resolution covering (almost) all land surfaces with extremely short revisit time (up to a few days), making it possible the mapping of global changes. Extracting useful information from such huge amount of data requires a systematic use of automatic techniques in almost all applicative contexts. In some cases, the strict application requirements force the practitioner to develop strongly data-driven approaches in the development of the processing chain. As a consequence, the exact relationship between the theoretical models adopted and the physical meaning of the solutions is sometimes hidden in the data analysis techniques, or not clear at all. Although this is not a limitation for the success of the application itself, it makes however difficult to transfer the knowledge learned from one specific problem to another. In this thesis we mainly focus on this aspect and we propose a general mathematical framework for the representation and analysis of multispectral images. The proposed models are then used in the applicative context of change detection. Here, the generality of the proposed models allows us to both: (1) provide a mathematical explanation of already existing methodologies for change detection, and (2) extend them to more general cases for addressing problems of increasing complexity. Typical spatial/spectral properties of last generation multispectral images emphasize the need of having more flexible models to image representation. In fact, classical methods to change detection that have worked well on previous generations of multispectral images provide sub-optimal results due to their poor capability of modeling all the complex spectral/spatial detail available in last generation products. The theoretical models presented in this thesis are aimed at giving more degrees of freedom in the representation of the images. The effectiveness of the proposed novel approaches and related techniques is demonstrated on several experiments involving both synthetic datasets and real multispectral images. Here, the improved flexibility of the models adopted allows for a better representation of the data and is always followed by a substantial improvement of the change detection performance.

Keywords. Remote sensing, multispectral images, multitemporal images, free discontinuity models, change detection, Mumford-Shah, Blake-Zisserman, Rician distributions.

Contents

1	Introduction	1
1.1	Background on optical imagery in remote sensing	1
1.2	Motivation of the thesis	4
1.3	Objectives of the thesis	6
I	Novel techniques for numerical minimization of variational functionals for image approximation	11
2	Background	13
2.1	An optical image model	13
2.2	Mathematical methods to image approximation	14
2.3	Challenges and novel contributions	20
3	Numerical minimization of a variational functional to image approximation	23
3.1	Introduction	23
3.2	Numerical minimization of the Blake-Zisserman functional	29
3.3	Numerical results	37
3.4	Conclusions	54
4	Variational approximation of vector-valued images and curves	59
4.1	Introduction	59
4.2	The Blake-Zisserman model for the approximation of vector-valued images and curves	62
4.3	Experimental results: piecewise linear approximation of vector-valued images	69
4.4	Experimental results: polygonal approximation of planar closed curves . .	80
4.5	Conclusions	83

II	Novel statistical models for change detection in multispectral images	87
5	Background	89
5.1	Overview of the change detection problem	89
5.2	Change detection methods for multispectral images	91
5.3	Challenges and novel contributions	94
6	The Rayleigh-Rice mixture model for binary change detection	97
6.1	Introduction	98
6.2	A problem of binary change detection	100
6.3	The EM algorithm for parameter estimation of the Rayleigh-Rice mixture density	103
6.4	Experimental results on multispectral images	110
6.5	Conclusions	118
6.6	Acknowledgments	121
6.A	Appendices	121
7	A compound multi-class mixture model for change detection	127
7.1	Introduction	128
7.2	Statistical study of the difference image	130
7.3	Derivation of a method for binary decision based on the difference mixture model	135
7.4	Experimental results	143
7.5	Discussion and conclusion	149
7.A	Appendices	152
8	A class-wise spatial-contextual approach based on a free discontinuity model for change detection in multispectral images	155
8.1	Introduction	155
8.2	A spatial-contextual framework for class-wise statistical reduction of multispectral images	156
8.3	Binary change detection based on class-wise statistical reduction	160
8.4	Experimental results	161
8.5	Conclusion	165
9	Conclusions and future developments	169
9.1	Conclusions	169
9.2	Future developments	171

List of Tables

1.1	Relevant spaceborne multispectral sensors operating from the early eighties to nowadays. First group decommissioned. Second group ongoing missions.	3
3.1	Outer/inner iterations and execution time (seconds) observed in the run of the proposed algorithms.	41
3.2	Accuracy in the approximations given by the GS, BCDA and BCDA _c with respect to the <i>ideal</i> solution for the considered datasets.	43
4.1	Outer(k)/inner(iter) iterations and execution time observed in the minimization of the three functional models.	76
4.2	Estimation of additive Gaussian noise variance in color images using the MS and the BZred models.	80
6.1	Classification error in a two-class decision problem. ω_n unchanged pixels, ω_c changed pixels	103
6.2	Iteration details of the proposed EM algorithm on Rayleigh-Rice mixtures with parameters $\alpha = 0.4$, $b = 1$, $\sigma = 1$ and for different values of ν .	109
6.3	Iterations details of the proposed EM algorithm on Rayleigh-Rice mixtures with parameters $\alpha = 0.4$, $b = 1$, $\sigma = 2$ and for different values of ν .	109
6.4	Parameter estimation via EM algorithm and data fitting evaluation for the three considered mixture densities.	116
6.5	Performance of change detection based on the thresholding of the magnitude image for different values of the threshold.	119
7.1	EM algorithm iteration details and parameter estimations for both the considered mixtures.	148
7.2	Comparison of the change detection performance of the method in [1] and the proposed method with respect to optimal performance.	150

8.1	Class variance calculated on both the original (d) and the reduced (\bar{d}) difference images using channels 5 and 6. The change class contains the fire, the unchange class all the rest.	164
8.2	Change detection performance of the Rayleigh-Rice binary detector (rR) and optimal (opt) performance in the case of the original difference image d (see [1]) and the statistically reduced difference image \bar{d} . MA, FA and OE are missed alarms, false alarms and overall errors, respectively.	167

List of Figures

1.1	Spectral and spatial properties of Landsat, SPOT and Sentinel-2 multi-spectral imagery, [2].	4
1.2	A fire occurred in Sardinia Island (Italy) between August 7-9, 2013 (the total extention of the event is approximately 2400 hectares): (a) environmental image of the event, the village on the left is Laconi; (b) false color image of the area affected by the fire composed by three spectral bands of a Landsat 8 image of the scene, spatial resolution is 30 <i>m</i>	5
2.1	A typical optical imaging system, [3].	14
2.2	An example of scale-space representation via the Mumford-Shah piecewise smooth approximation. (a) The original image g is a portion of QuickBird image representing an urban area in Reggio Emilia, Italy. (b) Piecewise smooth approximation u of the image. Functional parameters are selected to represent the image at scale in which textural details are eliminated but basic shapes of the buildings are preserved. (c) map of the detected edges K	19
2.3	Example of a <i>cartooned</i> image. (a) is the original image g , (b) is the approximation u (the cartoon) obtained via Mumford-Shah piecewise constant approximation. (c) is the set of edges K	20
3.1	Limitation of the MS model of detecting second-order geometrical features. (a,b) Gray-scale image with second-order edges. (c) Edge-detection via Mumford-Shah functional compared to (d) a full theoretical exact detection of 2nd-order features.	26
3.2	Slice section of the discontinuity set S and its approximation via the recovering function σ_ϵ realizing the Γ -convergence.	28
3.3	Datasets of the experiment including three gray-scale images and a digital surface model obtained from airborne LiDAR points acquired over Trento, Italy.	39

3.4	Energy-versus-time at every outer iteration for all three datasets with GS, BCDA and BCDAc. Algorithms are stopped by criterion (3.22) with tolerance $TOL_F = 10^{-3}$	42
3.5	Results of the BCDA method for the considered datasets. First column is the smooth approximation \mathbf{u} , second column is the edge-map \mathbf{s} and third column is the edge/crease-map \mathbf{z}	44
3.6	Particulars of the segmentation for the dataset pearl	45
3.7	Particulars of the segmentation for the dataset airport	46
3.8	Particulars of the segmentation for the dataset barracks . Above, there are the 3D renderings of the surface model \mathbf{g} and its smooth approximation \mathbf{u} . In the bottom, there are the edge map \mathbf{s} and the edge+crease map \mathbf{z}	47
3.9	Smoothing on synthetic images with different noise levels, image size is 100×100 . Row 1: input noisy images \mathbf{g} . Row 2: difference $\mathbf{g} - \mathbf{u}$. Row 3: edge-detection functions \mathbf{s} . Row 4: edge/crease-detection functions \mathbf{z}	49
3.10	Performance details of BCDA by varying μ and for different variances of the noise. Test images are 1000×1000 pixels compositions of the pyramidal elements showed in Figure 3.9. For each value of μ we plot (a) the execution time, (b) the total number, and (c) the average number of iterations of the PCG solver related to \mathbf{u}	50
3.11	Performance of BCDA with Diagonal (D) and Block-Diagonal (BD) preconditioners for the PCG solver related to \mathbf{u} , versus the size of \mathbf{g} . We plot (a) the execution time, (b) the total number, and (c) the average number of iterations of the PCG solver related to \mathbf{u}	51
3.12	Test images. (a) Image with jump of variable height h . (b) Image with a crease of variable slope θ . The slices of functions \mathbf{s}, \mathbf{z} plotted in Figures 3.13 and 3.14 are located in correspondence of the red dashed lines.	53
3.13	Slices of functions \mathbf{s} and \mathbf{z} obtained by minimizing F_ϵ in the case of test image with a jump (Figure 3.12a) for different values of ϵ (a,b) and h (c,d).	55
3.14	Slices of functions \mathbf{s} and \mathbf{z} obtained by minimizing F_ϵ in the case of test image with a gradient discontinuity (Figure 3.12b) for different values of ϵ (a,b) and $\tan(\theta)$ (c,d).	56
4.1	Craquelure removal via MS and BZ approaches. (a) The input color image represents a portion of the oil painting <i>Girl with a pearl earring</i> . Its approximations are obtained by the three different models: (b) AT-MS , (c) AFM-BZap , (d) AFM-BZ	72

4.2	Discontinuity functions computed for the three functional models. (a) AT-MS , (b,c) AFM-BZap , (d) AFM-BZ . White corresponds to 1 and black to 0, gray values are in between.	73
4.3	Particular of the approximations by zooming the area in the red square in Figure 4.1. (a) Original image, and the results for: (b) AT-MS , (c) AFM-BZap , (d) AFM-BZ . Notice the over-segmentation effect in (b,c).	74
4.4	Pixel scatterplots of the image portions represented in Figure 4.3. (a) Original image, (b) AT-MS , (c) AFM-BZap , (d) AFM-BZ	75
4.5	Energy-versus-time at each outer iteration for the three functional minimization cases: (a) AT-MS , (b) AFM-BZap , (c) AFM-BZ . The plots illustrate the descent of each additive term in the functional models. The black dashed line is the total energy. Blue is the Hessian component and Cyan is the AT component associated to the Hessian (present only in AFM-BZap and AFM-BZ). Red is the gradient component and Magenta is the AT component associated to the gradient (present only in AT-MS and AFM-BZap). Green is the distance term.	77
4.6	Estimation of Gaussian additive noise in color image containing challenging geometries. (a) Synthetic noise-free generated image, (b) noisy image. Reconstructions of the noisy image are obtained by the (c) AT-MS model, (d) AFM-BZ model. Particulars zoomed at the crack-tip end for the (e) AT-MS solution (white traits emphasize the main directions of the discontinuity edges) and the (f) AFM-BZ solution.	79
4.7	Extraction of building edges from DSM. (a) 3d rendering of a DSM representing an old barrack. (b) Edge map of the main (U-shaped) building obtained by segmenting the DSM using the BZ model for gray-scale images [4]. The points correspond to the pixels where the function s (the edge detection function) is 0.	81
4.8	Curve approximation results obtained for different parameter choices of the CSSP model. The range of values used in the experiments allowed us to explore the behaviour of the solution from over- to under- fitting. No gradient discontinuity is allowed by the model.	83
4.9	Curve approximation results obtained for different parameter choices of the BZ model. The range of values used in the experiments allowed us to explore the behaviour of the solution from under- to over- fitting passing also through polygonal solutions. The best polygonal approximation is (i), i.e., for parameters $\lambda = 10^{-5}$ and $\nu = 10^{-4}$	84

5.1	General high-level scheme for the design and the implementation of a change detection procedure.	90
6.1	Geometrical interpretation of the posterior probability that x originated in the h -th component given Ψ' , as $p(\omega_h x, \Psi') = C_h/(C_1 + C_2)$ for $h = 1, 2$	105
6.2	Histograms of samples \mathbf{x} generated from $p(\rho \Psi)$ with $\alpha = 0.4, b = 1, \sigma = 1$ and for different values of ν	110
6.3	Histograms of samples \mathbf{x} generated from $p(\rho \Psi)$ with $\alpha = 0.4, b = 1, \sigma = 2$ and for different values of ν	110
6.4	Dataset A. Synthetic two-band difference image. (a) Magnitude of the difference image, (b) map of simulated changed pixels (black).	112
6.5	Dataset B: images of Lake Mulargia (Italy) acquired by the Thematic Mapper sensor of the Landsat 5 satellite: (a) channel 4 of the image acquired in September 1995; (b) channel 4 of the image acquired in July 1996; (c) change reference map indicating the enlargement of the lake (black) and an open quarry (red).	113
6.6	Dataset C: images of Lake Omodeo and surrounding area (Italy) acquired by the Operational Land Imager sensor of the Landsat 8 satellite: (a) channel 5 of the image acquired in July 2013; (b) channel 5 of the image acquired in August 2013; (c) change reference map indicating the burned area extension (black) and other minor changes related to clouds and water (red).	114
6.7	Histograms of the magnitude of the difference image and plot of the estimated densities.	117
6.8	Change detection maps obtained by thresholding the magnitude image using T^{RR} (a,d,g), T^{GG} (b,e,h) and T^{MOE} (c,f,i). In black are the estimated changed pixels W_c , in white the estimated unchanged pixels W_n	120
7.1	Example of statistical dependency of the difference image with respect to the input pair when the number of natural classes is $C = 2$. (a) The natural class ϕ_2 is only observable in image \mathbf{y}^2 , thus the resulting difference image has only one unchange behaviour ω_1 and one change behaviour ω_{12} (from class ϕ_1 to ϕ_2). (b) Both natural classes ϕ_1, ϕ_2 are observable in the two images. Therefore, they have their own unchange statistical behaviours ω_1, ω_2 , and mutual change behaviours ω_{12} (from class ϕ_1 to ϕ_2), and ω_{21} (from class ϕ_2 to ϕ_1).	134

7.2	Illustration of the datasets analyzed in the experiments. The pictures show pre and post images in one band and a reference map of the changes (red pixels are minor changes with respect to the main changes that are represented by black pixels) for (a,b,c) dataset A, (d,e,f) dataset B, (g,h,i) dataset C, and (j,k,l) dataset D.	146
7.3	Fitting capability of the considered methods. Estimated densities are superimposed to the histograms of the magnitude samples. In the legends also the distance metrics between the estimated densities and the histograms are given.	147
7.4	Change detection map obtained on (a,b) dataset A, (c,d) dataset B, (e,f) dataset C, and (g,h) dataset D. Blue pixels are missed alarms, red ones are false alarms and green ones are correctly detected changes.	151
8.1	The proposed tiling scheme with related size parameters.	159
8.2	Histograms of the differences between the solutions. x -axis (binned difference values) is restricted to a portion where more than 99,99% of pixels are counted. The central bars are cutted at the top.	162
8.3	A portion of the difference $u - u^t$ and of the function z^t in correspondence of a cross tile junction (indicated by the red traits). The scene represents some old barracks and surrounding area.	162
8.4	Landsat 8 dataset used in the experiments. (a) Original difference image (false color composition), see [1]. (b) Reference map of the changes. Black pixels belong to the fire, the red ones are minor changes unrelated to the fire. (c) Difference image after statistical reduction (false color composition), this image is used for computations in the experiments. (d) Map of the class boundaries extracted from the reduced image. The false color compositions are obtained by putting channels 3,5,6 of the Landsat image into R,G,B.	164
8.5	Histogram of the magnitude of the statistically reduced difference image and estimated density. Notice that the two classes are well separated. . . .	165
8.6	Results of the change detection. (a) CD map obtained with the Rayleigh-Rice binay detector (rR) applied to the reduced image, (b) CD map related to the optimal threshold (opt) applied to the reduced image. For comparison with the methods proposed in the two previous chapters: (c) CD map obtained with the rR binary detector (Chapter 6) applied to the image before reduction, and (d) CD map obtained with the rrR binary detector (Chapter 7) applied to the image before reduction. In all CD maps red, blue and green are false alarms, missed alarms and hits, respectively. . . .	166

Chapter 1

Introduction

This chapter gives an overview of the role played by multispectral imagery in remote sensing with its applications, it describes the main motivations and the objectives of the thesis and presents the whole structure and organization of the document. This dissertation is divided into two main parts; the content of each part is described more in detail in each corresponding introductory section.

1.1 Background on optical imagery in remote sensing

Remote Sensing (RS) encompasses a variety of technologies and techniques to continuously observe the Earth surface by means of sensors mounted on aircraft or spacecraft platforms. The peculiarity of RS is that certain properties of real objects can be measured remotely without relying on physical contact. This makes RS an essential tool for the study of our planet evolution. Indeed, it is already widely employed in different application domains such as forestry, agriculture, urban management, oceanography, natural disaster monitoring, etc.

RS sensors can be divided into two main groups: active and passive systems. In active systems (e.g., Synthetic Aperture Radar (SAR) and Light Detection and Ranging (LiDAR)) a signal is emitted from the sensor and its multiple returns scattered by the objects are then recorded. In passive systems, the sensor (e.g., scanner, radiometer) is able to measure the portion of sunlight radiation reflected (or emitted) by the objects in a specific spectral interval. The larger is the spectral interval, the higher is the geometric resolution. According to this principle, satellite RS optical images can be divided into four categories: panchromatic (PAN), multispectral at very high resolution (VHR), multispectral at moderate resolution (MS) and hyperspectral (HS) images. Panchromatic sensors record the total intensity of radiation falling on each pixel, therefore they can achieve the highest geometrical resolution (up to tens of centimeters). The multispectral

at VHR typically records the radiation in the red, green and blue (RGB) regions of the visible range (from 390 to 700 nm), in the near-infrared (NIR) region (from 700 to 900 nm) and possibly in the short-wave infrared (SWIR) region (from 1100 to 3000 nm). Geometrical resolution of satellite VHR images ranges between 2 and 5 meters. Multispectral sensors at moderate resolution (from 10 to hundreds of meters) are able to record the radiation in the visible and NIR ranges (in shorter intervals) but they can also measure farther portions of the spectrum covering the middle-infrared (MIR) and far-infrared (FIR or thermal) regions. Hyperspectral sensors sample over several (from tens to hundreds) narrow contiguous portions of the spectrum. These four types of optical images present complementary features, they have their own advantages and disadvantages and they can serve at different purposes in various applicative contexts. One particular feature that makes MS imagery very attractive for global studies is that MS sensors typically have large swaths (up to two or three hundreds of kilometers). Therefore, images can be collected at global scale with shorter revisit time, if compared to the others. We report in Table 1.1 specifications of some of the more relevant satellite missions that have been launched with MS sensors mounted on board since the early eighties.

Since the inception of the Landsat program, mid-resolution spaceborne sensors have provided the vast majority of multispectral datasets to image analysts studying land use/land cover change, vegetation and agricultural production trends and cycles, water and environmental quality, soils, geology, and other earth resource and science problems. Starting from Landsat 7, data were distributed for free and the same policy has been extended to all Landsat data. The French SPOT satellites have been another important source of high-quality, high-resolution multispectral data (SPOT can also collect stereo pairs). The brand new Sentinel-2 mission has been designed as a dependable multispectral EO system that will ensure the continuity of Landsat and SPOT observations and improve the availability of data for users. In comparison with the SPOT and Landsat precursor series of satellites, the Sentinel-2 mission offers an unprecedented combination of systematic global coverage of land surfaces, high revisit frequency, high spatial resolution (up to 10 m) and the largest swath. Spectral and spatial properties of Landsat, SPOT and Sentinel-2 imagery are compared in Figure 1.1. Optical imagery at very high spatial resolution (meter or submeter geometric resolution) such as QuickBird, WorldView, IKONOS, GeoEye and OrbView, is sold commercially. The trade-off for high spatial resolution is the limited geographic coverage. For vast areas, it is difficult to obtain seamless, cloud-free, high-resolution multispectral imagery within the single season or at the particular moment of the phenological cycle of interest to the researcher. In this thesis, we focus our attention on the study of MS images at moderate resolution. At moderate resolution, the passive response of objects over large spectral intervals allows natural classes in the image

Table 1.1: Relevant spaceborne multispectral sensors operating from the early eighties to nowadays. First group decommissioned. Second group ongoing missions.

Satellite (Sensor)	Spatial Resolution (m)	Swath (km)	Spectral bands	Quant.	Revisit time (days)
Terra (ASTER)	15, 30, 90	60	16 (VNIR, SWIR, TIR)	8 bit	16
Landsat 5 (TM)	30, 120	185	7 (VNIR, SWIR, TIR)	8 bit	16
Landsat 7 (ETM+)	15, 30, 60	185	8 (PAN, TM)	8 bit	16
SPOT 4 (HRVIR)	10, 20	60	5 (PAN, VNIR)	8 bit	26
SPOT 5 (HRG)	5 Pan, 10, 20	60	5 (SWIR, HRVIR)	8 bit	26
Landsat 8 (OLI, TIRS)	15, 30, 100	183	11 (ETM+, Coastal)	11 bit	16
SPOT 6, 7	1.5 Pan, 6	60	5 (PAN, VNIR)	12 bit	1 to 3
SENTINEL 2 (MSI)	10, 20, 60	290	13 (VNIR, SWIR, RedEdge)	16 bit	5

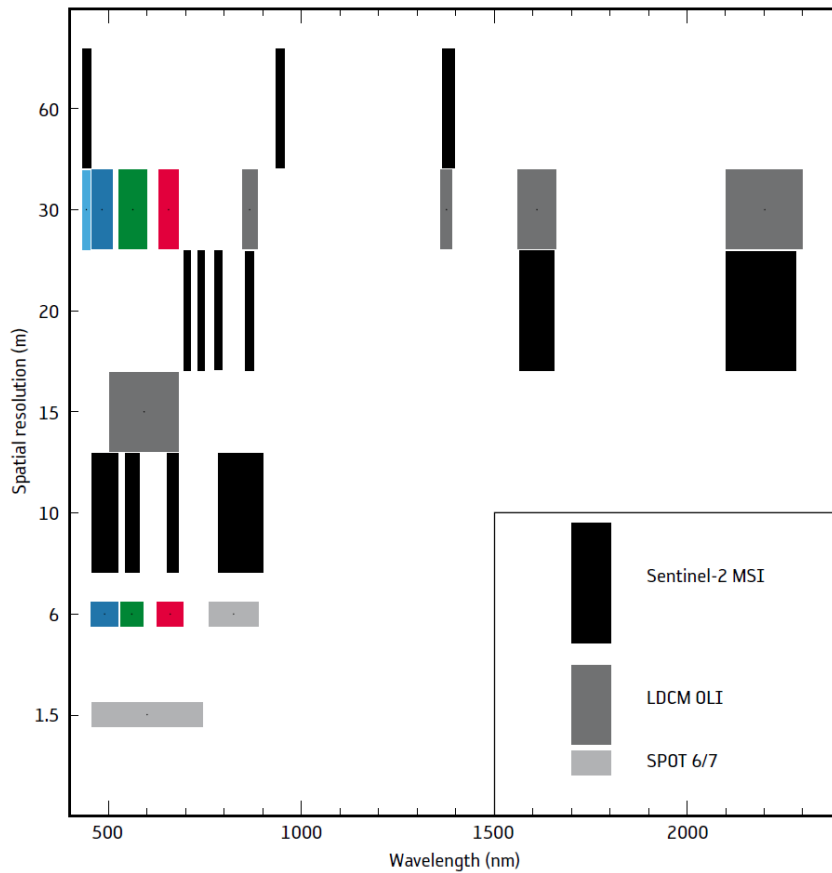


Figure 1.1: Spectral and spatial properties of Landsat, SPOT and Sentinel-2 multispectral imagery, [2].

to be essentially described in statistical terms. Indeed, they show spatial homogeneity and they can be accurately characterized by spectral vectors. In the sequel, these properties are investigated in greater detail. Novel results in the mathematical modeling of multispectral images are proposed and applied in the context of change detection. However, the character of the proposed theory is general and can be applied to other contexts as well.

1.2 Motivation of the thesis

For decades, the large number of launched EO satellites have provided a unique way to observe our living planet from space. Thanks to the revisiting properties of EO satellites, a huge amount of multitemporal images is now available in archives. The increasing demand coming from the different application domains has promoted new advances in technology and processing capability to guarantee operational continuity and provide observations



Figure 1.2: A fire occurred in Sardinia Island (Italy) between August 7-9, 2013 (the total extension of the event is approximately 2400 hectares): (a) environmental image of the event, the village on the left is Laconi; (b) false color image of the area affected by the fire composed by three spectral bands of a Landsat 8 image of the scene, spatial resolution is 30 *m*.

for the next generation of operational products such as land-cover maps, land-use change detection maps and geophysical variables. This allows for an accurate monitoring of the land surface changes in wide geographical areas according to both long term (e.g., yearly) and short term (e.g., daily) observations. The detection and understanding of changes occurred at the ground level is essential for studying the global change, the environmental evolution and the anthropic phenomena. For this reason, the development of change-detection techniques for the analysis of multitemporal remotely sensed images is now becoming one of the most important research topics in remote sensing.

Change Detection (CD) is the process of identifying changes occurred on the same geographical area between different observation times. According to the final applications there are different approaches of analysis and related methodological procedures. Fast analysis of changes occurring at the ground level is widely exploited in different applications, in particular in those activities for the monitoring of natural disasters. The new generation of satellite sensors has increased the availability of MS images and now provides data also with a significantly short revisiting time (e.g. up to daily observations). For example, in case of a natural disaster such as a fire, an earthquake or a tsunami, this allows for giving an almost real-time estimation of the area involved by the event (see e.g. Figure 1.2). Such estimations can greatly help the authorities in the planning of the first aid and related maneuvers. As a consequence, the development of effective methods for change detection is needed and the accuracy of the final results must be improved in order to guarantee a well balanced dislocation of resources. Moreover, CD is becoming one of the main tools to manage and monitor land usage and natural resources. For example, for

the purpose of environmental monitoring it is possible to detect changes in snow coverage, extension or reduction of thematic areas (e.g. forests, deserts, urban areas); in vegetation monitoring and agriculture surveys it is possible to detect vegetation indexes fluctuations, changes in vegetation life conditions, soil moisture variations.

Thanks to the spatial/spectral level of detail available in MS images, it is possible to specifically identify the spectral signature of certain classes and then perform detecting steps aimed at identifying variations in their spectral behaviour. Given its well-recognized importance in terms of applications, CD is a widely covered topic in literature. A huge vastity of models and methods have been proposed to solve as many different problems. However, in most of the cases the methodological development of CD procedures is driven by the problem and often the literature lacks in general models. This fact limits the possibility of passing and adapting the knowledge learned from certain problems to even slightly different applicative contexts and makes the general understanding of the underlying physical mechanism more difficult. Another limitation of current state-of-the-art methods is that, many empirical models that has worked well for the first generation of MS imagery show to be inefficient for addressing the new challenges introduced with the last generation products.

1.3 Objectives of the thesis

The aim of this thesis is to propose a general mathematical framework for the representation and analysis of RS images. If transposed to the applicative context of CD, the proposed models are able to provide both (1) a mathematical explanation of already existing methodologies, and (2) a framework to extend them to more general cases for addressing problems of increasing complexity. In our presentation, the mathematical representation of images as functions plays a central role. Certain assumptions can be made on the function models to let them accurately describe image properties that are of particular interest for the application. We will focus in particular on the description of geometrical and statistical properties of the images. It is a matter of fact that, the full information content present in an image can be completely modeled only in ideal cases. Rather, simplified models of the images are always (explicitly or implicitly) used as a basis for the rationale of most of the image processing techniques. Unfortunately, when these ideal models are not clear to the practitioner, it is difficult to interpret in a meaningful and reasonable way the results obtained after processing.

In this thesis we introduce and we take advantage of some novel image models to solve some typical problems of remote sensing image processing. We mainly focus our attention on: (1) variational approaches to image approximation (Part I of this thesis), and, (2)

statistical models for change detection in multispectral images (Part II of this thesis). In the following, we describe how the thesis presentation is structured, by focusing in particular on the main challenges and the innovative contributions introduced in the PhD activity.

1.3.1 Part I

The first part of the thesis is mainly concerned with image approximation based on variational methods. The variational approach provides a way to return an approximation of an image which is a composition of several pieces on which the image values are *homogeneous* in terms of derivatives. According to the derivative order that is penalized, we have first-order and second-order variational models to image approximation. Two well-known variational models to image approximation from computer vision due to Mumford and Shah [3] and Blake and Zisserman [5] are investigated. Given the large (almost unexplored) potential of variational methods to the processing of MS images, we have put a lot of effort in the study and the application of image approximation models in the remote sensing context. The RS highlights some well-known drawbacks of mathematical methods to image processing: the difficult scalability to large sized data and the extension of already existing models from their native scalar version (in which they are originally formulated) to the vector-valued case. An extensive study in this direction is proposed.

Chapter 2. In this chapter mathematical methods to image approximation are presented, with particular emphasis on variational methods and their applications. The chapter also introduces to the content of the first part of the thesis and describes the novel contributions.

Chapter 3. In this chapter we propose a novel approach to the numerical minimization of variational functionals with free-discontinuities that relies on a compact matricial formulation and an efficient iterative solver. In our exposition we take advantage of an extended formulation of the Mumford-Shah variational functional depending also on second-order derivatives due to Blake and Zisserman. Several experiments on both synthetic and real images are presented to demonstrate the various capabilities of variational models to provide simplified approximations of images.

Chapter 4. Applying free-discontinuity models to MS images is a non-trivial problem. Multispectral images are vector-valued and no available methods to address the minimization of variational functionals such as Mumford-Shah and Blake-Zisserman in the vector-valued case have chances to be successfully applied in a RS context due to their intrinsic computational complexity. In this chapter, our convenient

numerical approach to minimization is extended to the case of vector-valued inputs. The proposed method is then applied to difficult problems of image restoration and boundary approximation.

1.3.2 Part II

In the second part of this thesis we mainly investigate the statistical properties of multispectral images with special focus on the analysis of multitemporal images for CD. Very often RS images are modeled as realizations of a set of random variables following a specific statistical distribution. However, in some cases the statistical models associated to the images are not flexible enough to allow a precise description of them. As a consequence, some methodologies that rely on too simple statistical models may give inaccurate results. In the thesis we introduce some new statistical models to the description of the distribution of spectral difference-vectors and we derive from them novel methods to change detection based on image differencing. Differently to what is typically done in the literature, we do not rely on a-priori assumptions on the difference image. Instead, we weaken this constraint and propose a theoretical study by showing that the statistical description of the difference image can be derived from general assumptions on the single time images. By taking advantage of these general models, we further study the statistical distribution of the change vectors when coordinates are changed to magnitude and we devise efficient numerical algorithms for the binary detection of changes in bitemporal MS images that overcome the performance of typical empirical approaches.

Chapter 5. This chapter presents an overview of the change detection problem in remote sensing and the methods proposed in literature to address it on multispectral images. The chapter also includes a more detailed description of the contributions of this thesis with respect to the state-of-the-art.

Chapter 6. In this chapter, the standard two-class unchange/change model for binary CD is derived starting from the hypothesis of Gaussian distribution of natural classes in the difference image. When coordinates are changed to magnitude, we show that the two-class model can be described by a Rayleigh-Rice mixture. Parameter estimation of this mixture is a non-trivial task as this model is non-conventional (a typical conventional model which is often used for binary decision is based on a Gaussian mixture). Therefore, we devise a version of the Expectation-Maximization (EM) algorithm which is specifically tailored for the Rayleigh-Rice mixture.

Chapter 7. Further developments of the previous results are presented. Indeed, the typical two-class unchange/change model is sometimes not flexible enough to well model

the spectral behaviour of the difference image. Therefore, the intrinsic nature of the unchange/change macro classes is expected to be more complex. This is even more evident in last generation MS images, where the radiometric resolution is sensibly improved and more classes can be represented via their spectral characteristics. To have more chances to better represent last generation data, in this chapter we propose a new compound multiclass model for the CD problem that both extends what is previously done and also provides a general framework for many CD approaches.

Chapter 8. The study of large scenes poses some criticalities in the utilization of statistical models for MS image analysis due the larger statistical variability of data. In this chapter we investigate the possibility of reducing the statistical variability of images by exploiting an approximating approach based on variational methods. In order to do that, we first define a novel spatially related statistical model for MS images that localizes the multiclass distribution over homogeneous spatial regions of the image. The approximating framework allows us to reduce the statistical variability of class members by still preserving their spatial homogeneity. The proposed novel approach is then successfully applied to the purpose of change detection on MS images. Here, it shown that the discriminative capability of the statistical-based CD algorithms increases and large datasets can be addressed via a parallelizable tiling approach.

Part I

Novel techniques for numerical minimization of variational functionals for image approximation

Chapter 2

Background

This chapter introduces to the content of Part I. Firstly, a mathematical optical image model and the theory of mathematical methods to image approximation are briefly reviewed and variational models are introduced. In particular, some important properties of free discontinuity models are recalled. Then, the main challenges addressed in this part of the thesis and a description of related contributions are presented.

2.1 An optical image model

The typical function representation of an optical image can be explained in terms of the basic optical acquisition system [3]. The optical device is located in point P in the 3-D real world and it is pointed towards the subject of the scene, see Figure 2.1. The sensor is able to record the intensity of the light (restricted to a specific interval of the spectrum) coming from the reflective source in radial direction (with center P). This direction intersects the focal plane in one point, therefore a set of two-dimensional coordinates can be associated to the record of the light intensity. Thus, let us define $x \in \Omega_0 \subset \mathbb{R}^2$ the spatial coordinate (for convenience restricted to a bounded rectangular region Ω_0 of the plane) and $g(x)$ the recorded light intensity at point x . The value $g(x)$ can be either a scalar or a B -dimensional vector depending on the capability of the sensor of recording the spectral response of objects in different intervals of the spectrum (e.g., this is the typical case of multispectral scanners). The function $g(x) : \Omega_0 \rightarrow \mathbb{R}^B$ is called an image (this is a general notation that also includes the scalar case for $B = 1$). To better understand what kind of function is g like, one can consider the simple, yet effective, model depicted in Figure 2.1. Objects O_i that are homogenous in terms of the reflected light intensity and that can be seen from point P would be imaged to spatially homogenous regions $\Omega_i \subset \Omega_0$. Relative position of objects O_i (front, back, partial occlusion) would cause imaged edge boundaries to appear as image discontinuities. Therefore, at a first approximation the function g is

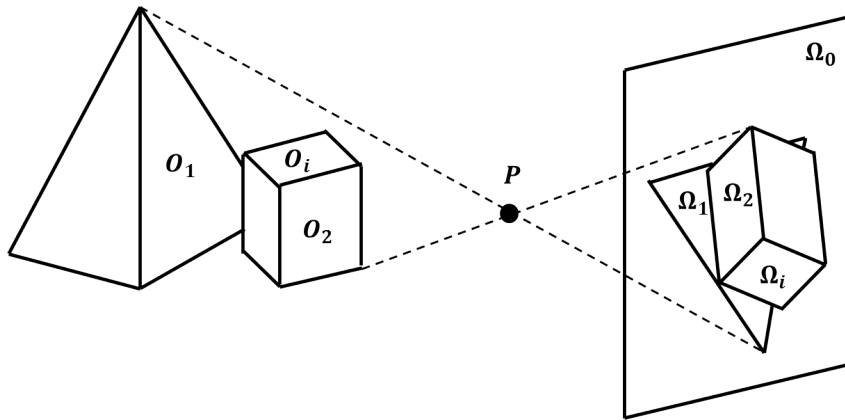


Figure 2.1: A typical optical imaging system, [3].

expected to be piecewise smooth, i.e., well modeled by a set of smooth functions g_i defined over disjoint sets Ω_i . Of course, this model is ideal as what is observed in real images is typically more complex: often objects are characterized by textured or fragmented patterns instead of homogeneous patches, shadows would typically result in shallow edges, etc. Noise measurement is another source of deviation from the ideal model. This process of *contamination* of the ideal image can be formalized by incorporating the physics of the vision mechanism, noise measurement, etc., in a functional transformation Λ that relates the ideal image u to the real recorded g as

$$g = \Lambda(u). \quad (2.1)$$

In general, the transformation Λ is not explicit, thus the problem of recovering u from g is an inverse ill-posed problem. The aim of mathematical methods to image approximation is that of recovering a *regular* approximation u of a real image g .

2.2 Mathematical methods to image approximation

According to the assumptions that are made on u (and on Λ), there are different approximation approaches. Mathematical methods to image approximation can be mainly divided into two groups: partial differential equation (PDE) methods and variational methods. In the following these two approaches are briefly reviewed.

2.2.1 Partial differential equation methods

In the early 80s Witkin [6] proposed a noise-reduction scale-space coarsening of signals via convolution with Gaussian kernels. Following a well-established axiomatization of

formal requirements the image processing transforms must satisfy [7], this model reflects causality, linearity and isometry invariance which are characterizing properties of the heat semi-group. Images at every scale form a family that depend on a temporal parameter t . Each image $u(t, \cdot)$ of the family satisfies the partial differential equation (PDE):

$$\partial_t u = \operatorname{div} [\nabla u] \quad \text{in } \mathbb{R} \times \Omega_0, \quad \text{with } u(0, \cdot) = g, \quad (2.2)$$

where g is the input image, a non-negative real-valued function defined over the (rectangular) domain $\Omega_0 \subset \mathbb{R}^2$. Here the concept of scale-space is strictly related to the parameter t that determines the width of the Gaussian kernel, i.e. to the amount of noise reduction. Anyway, the model is fully isotropic and does not take into account any structural information of the input image. It blurs out both noise and interesting features of the image such as edges or points of gradient discontinuity. For recovering at this drawback, some modifications of the previous formulation have been proposed. Perona and Malik [8] introduced in formulation (2.2) a term that makes the diffusion anisotropic. More specifically, the diffusion is inhibited according to local properties of the image that are detected by a function $c = c(|\nabla u|)$ that depends on the gradient. The result is a slightly changed concept of scale-space that can be formalized by the differential problem

$$\partial_t u = \operatorname{div} [c(|\nabla u|)\nabla u] \quad \text{in } \mathbb{R} \times \Omega_0, \quad \text{with } u(0, \cdot) = g. \quad (2.3)$$

Although this formulation is general, particularly effective results in enhancing edges have been obtained by using $c(s) = (1 + s^2/\lambda^2)^{-1}$. Here the positive parameter λ ensures that the diffusion is low when high gradients are detected. A variation on this model have been proposed independently by Rudin and Osher [9]. In particular in their formulation, which is equivalent to the Perona-Malik case where the function $c = -1/|\nabla u|$, the resulting PDE represents the flow generated by the minimization of the Total Variation, i.e. the quantity

$$\int_{\Omega_0} |\nabla u| \, dx, \quad (2.4)$$

with some constraints on u to be fulfilled. In this different view, the variational nature of the denoise-recovering problem becomes more evident.

2.2.2 Variational methods

Probably, the major limitation of the previous approaches is the limited flexibility in modeling all the components that play a relevant role in the recovering process. On the one hand, the action of the different components of the models is not even fully understood and only some effects are known, cfr. e.g., [10, 11]. As a consequence, the progressive

modification of the PDEs to obtain more and more meaningful solutions becomes a critical step. On the other hand, the real edge-detection step that locates the object boundaries at different scales is still considered as a subsequent and independent step. Diffusion-based methods only enhance edges and do not explicitly provide their detection. In this perspective, the variational approach seems to be more flexible and allows for having a proper and explicit modelization of all the components: smoothing, edge-detection, scale-space representation.

A first-order model: the Mumford-Shah functional

By fully exploiting a variational framework, Mumford and Shah [3] proposed a model for image approximation based on the minimization of the following functional

$$E(u, K) = \int_{\Omega_0 \setminus K} |\nabla u|^2 dx + \alpha \mathcal{H}^1(K) + \mu \int_{\Omega_0} |u - g|^2 dx \quad (2.5)$$

among all the functions continuously differentiable outside K , i.e. $u \in C^1(\Omega_0 \setminus K)$, where $K \subset \Omega$ is compact. Here, \mathcal{H}^1 is the 1-dimensional Hausdorff measure, and α, μ are positive parameters. This functional model has several interpretations and specializations that can be formulated depending on specific assumptions/constraints that can be made on u and the other parameters. Hereafter are some examples.

Piecewise smooth approximation. This is the most general case, as defined above.

The minimization of the first term in (3.1) forces u to be smooth outside K , which has to be a one-dimensional set with finite length because of the term $\mathcal{H}^1(K)$. The last distance term forces u to be close to (i.e., an approximation of) the original image g . The unknown set K can be easily understood as the set of the discontinuities of u , indeed the minimization of (3.1) is a prototype of *free discontinuities problem*, [12]. A solution of this formulation would explicitly provide: the set K of the edges, and, u a piecewise smooth version of the original image g . By changing the values of the parameters in (3.1) to augment the noise-reduction effect, different scales are reachable. This fact is enforced in [13], where a strict relationship between the Mumford-Shah and Perona-Malik approaches to segmentation is found. In particular, it is shown how the parameters of the Mumford-Shah functional can be interpreted as parameters regulating an anisotropic diffusion process applied to the image g . An example is shown in Figure 2.2.

Piecewise constant approximation. This formulation is obtained by restricting the minimization of (3.1) to the class of piecewise constant functions over partitions of Ω_0 . Let us define $\Omega := \{\Omega_i : i = 1, \dots, n_\Omega\}$ a *regular* partition of Ω_0 as:

- Ω_i are open connected sets with smooth boundary $\partial\Omega_i$
- let $K = \cup_{i=1}^{n_\Omega} \partial\Omega_i$ be the total partition boundary, then $\cup_{i=1}^{n_\Omega} \Omega_i = \Omega_0 \cup K$;
- $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$.

We define the space of piecewise constant functions over the partition Ω to be $PC(\Omega) := \{u : u(x) = u_i \in \mathbb{R} \text{ for all } x \in \Omega_i\}$. The functional (3.1) restricted to this class of functions is

$$E(u, K) = \sum_{i=1}^{n_\Omega} \int_{\Omega_i} |u_i - g|^2 dx + \nu_0 \mathcal{H}^1(K) \quad (2.6)$$

as the gradient term vanishes. Fixed Ω , it can be proved that the unique minimizer u^* of (2.6) is the piecewise function whose values are the integral means of g over the sets Ω_i , i.e., $u_i^* = \bar{g}_{\Omega_i} := \frac{1}{|\Omega_i|} \int_{\Omega_i} g dx$. Therefore, the minimization of (2.6) is equivalent to the minimization of

$$E_0(K) = \sum_{i=1}^{n_\Omega} \int_{\Omega_i} |g - \bar{g}_{\Omega_i}|^2 dx + \nu_0 \mathcal{H}^1(K) \quad (2.7)$$

which depends on the only variable K . It can be shown that the minimization of E_0 is a well-posed problem. If g is continuous there exists a minimizing K made up of a finite number of singular points joined by a finite set of C^2 -arcs. The resulting piecewise constant function u^* obtained by minimizing (2.7) is often called a *cartoon* of g , see Figure 2.3.

Markov Random Fields. By changing the representation of the image model from continuous to discrete, it can be shown that the piecewise constant approximation can be derived as a specialization of a Markov Random Field (MRF) model. Consider $\mathcal{G} = (V, E)$ to be an undirected graph with nodes $V = \{1, \dots, n\}$ and edges $E = \{(i, j) \in V \times V : \text{if } i, j \text{ are connected}\}$. To each node i we assign a random variable v_i that takes values on a discrete set $A = \{a_1, \dots, a_k\}$ and we call (\mathcal{G}, A) a random field. We denote a possible configuration of the random field by $v = (v_1, \dots, v_n)$ where each $v_i \in A$. If we measure the energy of a configuration v as

$$U(v) = \sum_{i \in V} G(v_i) + \nu_1 \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} F(v_i, v_j), \quad (2.8)$$

where $\mathcal{N}(i)$ is the set of i neighbors, then the Hammersley-Clifford theorem tells us that the random field has the Markov property and its configuration probability can be given in terms of Gibbs distribution as $P(v) = Z^{-1} e^{-1/TU(v)}$ where Z is a normalizing constant and T is a positive parameter. In (2.8), the function G encodes the

sensitivity of the random field to external forces (it is often called external interaction function), whereas F describes the energy generated by local inner interactions (it is often called internal interaction function). If the edges of the graph form a reticular structure, the random variables v_i can be associated to pixels values of an image u via the simple assignment $v_i = u(x_i)$, where x_i is a spatial discrete coordinate. Given an image g , by defining $G(v_i) := [u(x_i) - g(x_i)]^2$ and $F(v_i, v_j) := \delta(|u(x_i) - u(x_j)|)$, where δ is the kronecker function, we get that (2.8) is exactly the discrete version of (2.6) for a suitable choice of the parameter ν_1 . Notice that, via the kronecker function, the double summation in (2.8) returns a multiple of the length of the interface that separates pixels with different values, i.e., a multiple of the length of K in a discrete setting.

A second-order model: the Blake-Zisserman functional

The Mumford-Shah model assumes the ideal image u approximating g to be essentially piecewise constant. Indeed, the gradient term in the energy penalizes variations of intensity outside a set of finite one-dimensional length, i.e., the set of image discontinuities. This gradient penalization is often called 1st-order penalization as it affects 1st-order derivatives of the solution. In some cases, 1st-order penalization is a too strong assumption as some gradients of light intensity are proper characteristics of the image and the piecewise nearly-flat approximation of them might result too coarse. In order to give more flexibility to the approximation model, Blake and Zisserman [5] proposed a 2nd-order penalization variational model to image approximation with the aim of providing a piecewise linear approximation of the image. The model depends on 2nd-order derivatives and allows free discontinuities and free gradient discontinuities and it is based on the minimization of

$$\begin{aligned}
 E(u, K_0, K_1) = & \int_{\Omega_0 \setminus (K_0 \cup K_1)} |Hu|^2 dx + \mu \int_{\Omega_0} |u - g|^2 dx \\
 & + \alpha \mathcal{H}^1(K_0) + \beta \mathcal{H}^1(K_1 \setminus K_0),
 \end{aligned} \tag{2.9}$$

among all functions u that are twice differentiable outside $K_0 \cup K_1$ and at least differentiable outside K_0 . K_0 and K_1 vary among all the compact sets such that $K_0 \cup K_1$ is closed in Ω_0 . μ, α, β are positive parameters. Here Hu denotes the Hessian operator, i.e., the 2×2 square matrix containing all the second-order derivatives of u . It is important to note that, for an admissible solution u discontinuities are allowed on $K_0 \cup K_1$, whereas discontinuities of the gradient are allowed only on K_1 . α and β are contrast parameters regulating the total length of the discontinuity sets.



Figure 2.2: An example of scale-space representation via the Mumford-Shah piecewise smooth approximation. (a) The original image g is a portion of QuickBird image representing an urban area in Reggio Emilia, Italy. (b) Piecewise smooth approximation u of the image. Functional parameters are selected to represent the image at scale in which textural details are eliminated but basic shapes of the buildings are preserved. (c) map of the detected edges K .

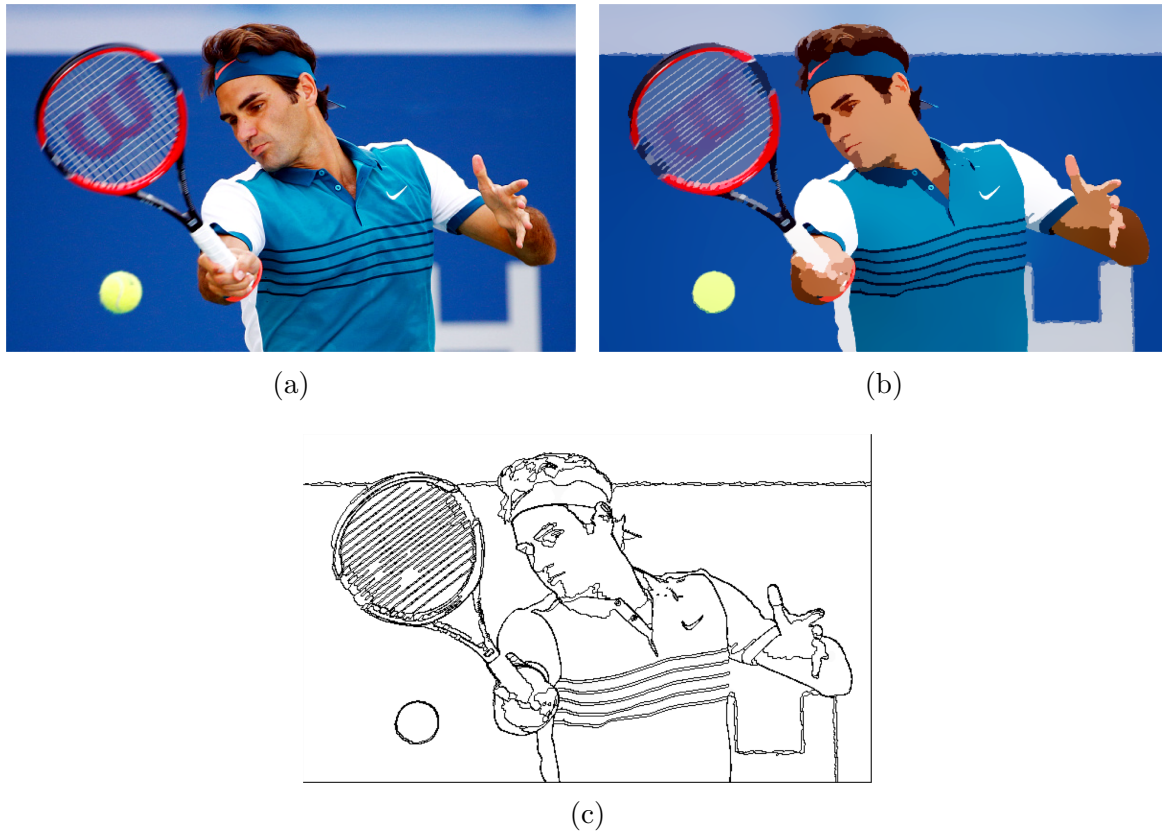


Figure 2.3: Example of a *cartooned* image. (a) is the original image g , (b) is the approximation u (the cartoon) obtained via Mumford-Shah piecewise constant approximation. (c) is the set of edges K .

2.3 Challenges and novel contributions

This part of the thesis presents novel results for the numerical minimization of variational functionals for image approximation due to Mumford-Shah and Blake-Zisserman. We propose in the following chapters an extensive and detailed presentation of the application of variational methods in different applicative contexts both in the remote sensing (statistical reduction of multispectral images, piecewise linear approximation of urban Digital Surface Models) and the image processing (image restoration, polygonal boundary recovering) domains. Our main requirements for the implementation of these methods were: (1) obtaining high computational efficiency, (2) dealing with vector-valued images, and, (3) developing highly parallelizable code.

Free discontinuity problems are difficult to implement because the unknown set variables (the discontinuity sets) introduce adverse conditions to the minimization problem. Among the solutions that have been proposed in literature to circumvent this problem,

the variational approximation by Ambrosio, Faina and March (AFM) [14] has some advantages: it is particularly suited for numerical implementation and it allows the explicit detection of image discontinuities. Chapter 3 is entirely devoted to the derivation of a novel numerical algorithm to the minimization of the AFM variational approximation of the Blake-Zisserman functional. The approximation model is general enough to include the Mumford-Shah model as a particular case. The numerical approach we propose exploits a compact matricial representation of the functional and a specifically tailored version of a block coordinate descent algorithm which is proved to converge to a stationary point of the objective energy.

Extending variational models to multispectral image analysis requires dealing with vector-valued inputs. Although in literature some approaches have been proposed to address the numerical minimization of the Mumford-Shah functional in the case of vector-valued inputs, most of them are not capable of explicitly detecting image discontinuities and also, they cannot deal with large size images (as typical in remote sensing). Therefore, in Chapter 4 we propose an extension of the proposed algorithm to the minimization of the functional model in Chapter 3 to the case of vector-valued inputs. Our application emphasizes the capability of the Blake-Zisserman model to provide piecewise linear approximation of the input both in the case of images and curves.

Chapter 3

Numerical minimization of a variational functional to image approximation

In this chapter¹ we address the numerical minimization of a variational approximation of the Blake-Zisserman functional given by Ambrosio, Faina and March. Our approach exploits a compact matrixial formulation of the objective functional and its decomposition into quadratic sparse convex sub-problems. This structure is well suited for using a block-coordinate descent method that cyclically determines a descent direction with respect to a block of variables by few iterations of a preconditioned conjugate gradient algorithm. We prove that the computed search directions are gradient related and, with convenient step-sizes, we obtain that any limit point of the generated sequence is a stationary point of the objective functional. An extensive experimentation on different datasets including real and synthetic images and digital surface models, enables us to conclude that: (1) the numerical method has satisfying performance in terms of accuracy and computational time; (2) a minimizer of the proposed discrete functional preserves the expected good geometrical properties of the Blake-Zisserman functional, i.e., it is able to detect first and second order edge-boundaries in images; (3) the method allows the segmentation of large images.

3.1 Introduction

Image approximation is a typical and widely investigated topic in image processing. It can be essentially defined as the process of finding a simplified version u of an image g

¹parts of this chapter appear in [4]

that, under certain conditions, it represents its ideal content prior to contamination due to acquisition mechanism, noise, etc (cfr. Section 2). By fully exploiting a variational framework, Mumford and Shah [3] proposed a model for image approximation based on the minimization of the following functional

$$\mathcal{MS}(u, K) = \int_{\Omega_0 \setminus K} |\nabla u|^2 dx + \alpha \mathcal{H}^1(K) + \mu \int_{\Omega_0} |u - g|^2 dx. \quad (3.1)$$

Here $\Omega_0 \subset \mathbb{R}^2$ and $g \in L^\infty(\Omega_0)$ is the input image. The minimization is among all the functions continuously differentiable outside K , i.e. $u \in C^1(\Omega_0 \setminus K)$, where $K \subset \Omega_0$ is compact. \mathcal{H}^1 is the 1-dimensional Hausdorff measure, and α, μ are positive parameters. The minimization of the first term forces u to be smooth (a piecewise constant behavior is expected) outside K . Because of the term $\mathcal{H}^1(K)$, K is a one-dimensional set with finite length. The last integral term is a distance term that forces u to be close to the original image g . The set K can be easily understood as the set of the discontinuities of u , indeed this is a typical problem belonging to a general class of problems called *free discontinuities problems*, [12].

From a practical point of view, the minimization of the MS functional (3.1) cannot be addressed because the measure term $\mathcal{H}^1(K)$ is not semi-continuous with respect to any reasonable topology. As suggested in [12], by relaxing the problem into the weaker space of Special Functions of Bounded Variation $SBV(\Omega_0)$, the methods of Calculus of Variations can be used to prove the existence of minima [15]. The advantage of this approach is that for every $u \in SBV(\Omega_0)$, the discontinuity set S_u is uniquely determined by geometrical properties of the function. This results in a functional formulation of the MS problem that uniquely depends on the function u :

$$\mathcal{G}(u) = \int_{\Omega_0} \left(\mu |u - g|^2 + |\nabla u|^2 \right) dx + \alpha \mathcal{H}^1(S_u), \quad (3.2)$$

where $u \in SBV(\Omega_0)$ and S_u is the complement set of Lebesgue points of u . Using compactness and lower semi-continuity theorems [16] it is showed that under mild conditions, there exists a solution such that $\mathcal{H}^1(S_u) < \infty$. Moreover, by regularity results one has that $\mathcal{H}^1(\overline{S_u} \setminus S_u) = 0$ and the couple $(u, \overline{S_u})$ can be identified with a minimizer of the strong formulation.

Based on this relaxed formulation, many techniques have been proposed to tackle the problem of numerically computing a minimizer. The free discontinuity term poses a serious problem. Ambrosio and Tortorelli [17], by exploiting a nice result of Modica and Mortola [18], proposed a Γ -convergence approximation via integral functionals defined on proper Sobolev spaces. In their approximation the discontinuity set is replaced by an auxiliary function that plays the role of indicator function. Numerical solutions based

on the Ambrosio-Tortorelli approximation are given in the framework of Finite Element Method (FEM) in [19], and via finite-difference discretization of Euler-Lagrange equations in [20]. In [21], a Γ -convergence approximation using local integral functionals defined on a discrete space is given. Numerical implementation of the method is presented in [22]. Another minimization technique is based on a convex relaxation of the functional [23]. A level set approach to minimization is presented in [24]. With no intent of being exhaustive, we refer the interested reader to the overview on the numerical approaches for solving the MS functional given in [25].

3.1.1 The Blake-Zisserman model for image segmentation

Being a first-order model, the MS variational segmentation suffers of some side effects [5, 26]. The minimization of the gradient norm forces the solution to be locally constant (zero gradient). In those regions where the gradient of g is too steep, this local approximation results in a step-wise function characterized by many fictitious discontinuities. This phenomenon is well-known as *over-segmentation* of steep gradients. Moreover, the minimization of the length term results in an approximation of complex edge junctions by triple-junctions where edges meet at $2/3\pi$ wide angles. This may lead to a degradation of the real geometry of boundaries. Lastly, properly because of its first-order nature, the MS model is unable to detect second-order geometrical features such as points of gradient discontinuity, see Figure 3.1. Since very often such points correspond to object boundaries, the MS model has the limitation that is not capable of detecting them.

With the specific intent to overcome such problems, Blake and Zisserman proposed a variational model based on second order derivatives, free discontinuities and free gradient discontinuities [5]. In their original formulation one has to minimize

$$\begin{aligned} \mathcal{BZ}(u, K_0, K_1) = & \int_{\Omega_0 \setminus (K_0 \cup K_1)} |Hu|^2 dx + \mu \int_{\Omega_0} |u - g|^2 dx \\ & + \alpha \mathcal{H}^1(K_0) + \beta \mathcal{H}^1(K_1 \setminus K_0), \end{aligned} \quad (3.3)$$

among all functions u that are twice differentiable outside $K_0 \cup K_1$ and at least differentiable outside K_0 . K_0 and K_1 vary among all the compact sets such that $K_0 \cup K_1$ is closed in Ω_0 . μ, α, β are positive parameters. Here Hu denotes the Hessian matrix of u . Notice that, for an admissible solution u , discontinuities are allowed on $K_0 \cup K_1$, whereas discontinuities of the gradient are allowed only on K_1 . α and β are contrast parameters regulating the total length of the discontinuity sets.

As in the MS case, also for the BZ functional minima existence and numerical issues can be addressed by considering a relaxation of the functional. Following [27], a relaxation in the space of Generalized Special Functions of Bounded Variation $GSBV(\Omega_0)$, is given by

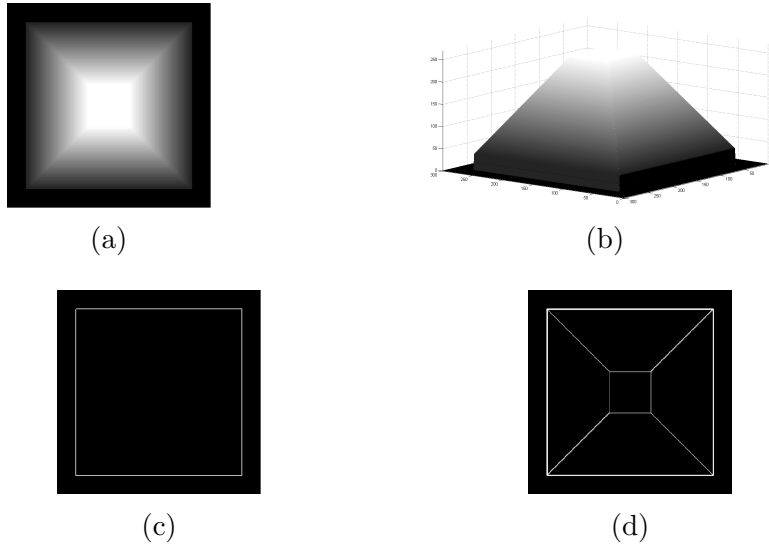


Figure 3.1: Limitation of the MS model of detecting second-order geometrical features. (a,b) Gray-scale image with second-order edges. (c) Edge-detection via Mumford-Shah functional compared to (d) a full theoretical exact detection of 2nd-order features.

the functional

$$\mathcal{F}(u) = \int_{\Omega_0} \left(\mu |u - g|^2 + |Hu|^2 \right) dx + (\alpha - \beta) \mathcal{H}^1(S_u) + \beta \mathcal{H}^1(S_{\nabla u} \cup S_u), \quad (3.4)$$

where $u \in GSBV^2(\Omega_0) := \{w \in GSBV(\Omega_0) : \nabla w \in [GSBV(\Omega_0)]^2\}$. In this weaker space, a proper definition of Hu and $S_{\nabla u}$ (the theoretic discontinuity set of ∇u) as geometrical property of the function u , is possible. By regularity arguments it can be proved [28] that a minimizer of (4.2) can be identified with a minimizing couple of the strong formulation, provided $\beta \leq \alpha \leq 2\beta$. Thus, the optimal set $K_0 \cup K_1$ is recovered via the discontinuity set S_u and the gradient discontinuity set $S_{\nabla u}$.

A vivid research interest is devoted to the Blake-Zisserman functional as it represents the generalization of the well-known and widely used Mumford-Shah. From a theoretical point of view it is a challenging topic, well-posedness of the problem and uniqueness of the solution [29] as well as regularity properties of minimizers [30–32] are still under investigation. Recently, a concise survey of the main results about the functional have been presented [33].

Segmentation based on the Blake-Zisserman model, because of its second-order nature, is specifically suitable for addressing problems such as: (1) image inpainting [34], where the functional minimization allows for predicting partially occluded regions in an image and their contours continuation, and (2) 3D data segmentation [35], where the unique capability of the functional of tracing second-order edges (creases) allows for precisely

locating planar objects (such as roof planes) in remote sensing 3D models of urban areas.

3.1.2 Variational approximation of the Blake-Zisserman relaxed functional

Implementing gradient descent of (3.3) with respect to the unknown free discontinuity sets is extremely difficult. Γ -convergence has shown to be fundamental to solve the problem of numerically computing a minimizer. This notion of convergence, suitable for functionals, has been introduced by [36]. For a deep treatment of this topic we refer to [37, 38]. The key point in Γ -convergence is that a specific functional, which may not have good properties for minimization, can be approximated by a sequence of regular functionals all admitting minimizers. The sequence of these approximate minimizers converges (in the classical sense) to a minimizer of the original objective functional. Besides its importance as mathematical tool, Γ -convergence is very attractive also from a numerical point of view as it allows for the solution of several difficult numerical problems in Computer Vision, Physics, and many other fields. See for instance [38, 39].

Following the idea of Ambrosio and Tortorelli, in [40] a Γ -convergence result is proved for the BZ functional in dimension 1. A full proof in dimension 2 and a partial result for any dimension n is given by Ambrosio, Faina and March [14]. The authors, by properly adapting the techniques of [40] and [17], have introduced two auxiliary functions $s, z : \Omega_0 \rightarrow [0, 1]$ (aimed at approximating the indicator functions of the discontinuity sets) to the model and proposed a Γ -convergence approximation of \mathcal{F} via the family of uniformly elliptic functionals

$$\begin{aligned} \mathcal{F}_\epsilon(s, z, u) &= \delta \int_{\Omega_0} z^2 |Hu|^2 dx + \xi_\epsilon \int_{\Omega_0} (s^2 + o_\epsilon) |\nabla u|^2 dx \\ &\quad + (\alpha - \beta) \int_{\Omega_0} \epsilon |\nabla s|^2 + \frac{1}{4\epsilon} (s - 1)^2 dx \\ &\quad + \beta \int_{\Omega_0} \epsilon |\nabla z|^2 + \frac{1}{4\epsilon} (z - 1)^2 dx \\ &\quad + \mu \int_{\Omega_0} |u - g|^2 dx, \end{aligned} \tag{3.5}$$

where $(s, z, u) \in [W^{1,2}(\Omega_0, [0, 1])]^2 \times W^{2,2}(\Omega_0) =: \mathcal{D}(\Omega_0)$. Here ϵ is the convergence continuous parameter, ξ_ϵ, o_ϵ are infinitesimals and the convergence is intended for $\epsilon \rightarrow 0$. To prove Γ -convergence, one has to show that for any $u \in GSBV^2(\Omega_0)$, $s \equiv 1, z \equiv 1$ the two following properties are verified:

Liminf inequality: for any sequence $\{(s_\epsilon, z_\epsilon, u_\epsilon)\}_{\epsilon>0} \subset \mathcal{D}(\Omega_0)$ that $[L^1(\Omega_0)]^3$ -converges to (s, z, u) it holds that $\mathcal{F}(u) \leq \liminf_{\epsilon \rightarrow 0} \mathcal{F}_\epsilon(s_\epsilon, z_\epsilon, u_\epsilon)$.

Limsup inequality: there exists a sequence $\{(s_\epsilon, z_\epsilon, u_\epsilon)\}_{\epsilon>0} \subset \mathcal{D}(\Omega_0)$ that $[L^1(\Omega_0)]^3$ -converges to (s, z, u) such that $\limsup_{\epsilon \rightarrow 0} \mathcal{F}_\epsilon(s_\epsilon, z_\epsilon, u_\epsilon) \leq \mathcal{F}(u)$.

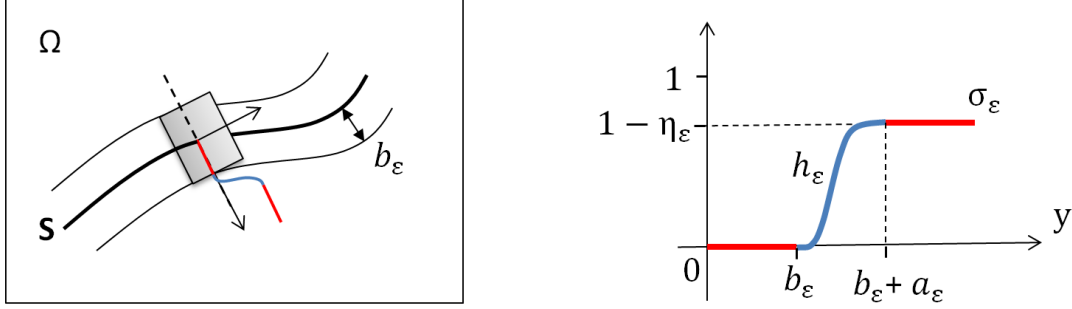


Figure 3.2: Slice section of the discontinuity set S and its approximation via the recovering function σ_ϵ realizing the Γ -convergence.

By standard arguments of functional analysis it is possible to prove that for any $\epsilon > 0$ the functional \mathcal{F}_ϵ always admits a minimizing triplet. Let us denote it by $(s_\epsilon, z_\epsilon, u_\epsilon)$. By sending $\epsilon \rightarrow 0$, thanks to the compactness properties of the Γ -convergence, the sequence $\{(s_\epsilon, z_\epsilon, u_\epsilon)\}_{\epsilon > 0}$ converges in the $[L^1(\Omega_0)]^3$ -norm to a triplet (s, z, u) where u is a minimizer of the limit functional \mathcal{F} and $s, z \equiv 1$ almost everywhere over Ω_0 .²

The constructive part of the Γ -convergence (Limsup inequality) provides us the tremendous advantage of keeping trace of the discontinuity sets S_u and $S_u \cup S_{\nabla u}$ via their regular function approximations. For a fixed $\epsilon > 0$, the two discontinuity sets, enjoying the regularity properties of $GSBV^2(\Omega_0)$ functions, are approximated by s_ϵ and z_ϵ (respectively) using a slicing argument and Coarea-formula for Lipschitz functions [14]. Let S be either S_u or $S_u \cup S_{\nabla u}$ and let us consider a 2-dimensional orthogonal slice of S (see Figure 3.2). The idea is to build a function σ_ϵ that is 0 in a tubular neighborhood of radius b_ϵ of the set S and that tends to 1 smoothly elsewhere. The tubular neighborhood shrinks as $\epsilon \rightarrow 0$. Formally the function σ_ϵ is defined as:

$$\sigma_\epsilon := \begin{cases} 0, & (S)_{b_\epsilon} \\ 1 - \eta_\epsilon, & \Omega_0 \setminus (S)_{b_\epsilon + a_\epsilon} \\ h_\epsilon \circ \tau, & \text{elsewhere} \end{cases} \quad (3.6)$$

where $a_\epsilon, b_\epsilon, \eta_\epsilon$ are infinitesimals as $\epsilon \rightarrow 0$, $\tau(y) := \text{dist}(y, \overline{S})$ and $(S)_r := \{y \in \mathbb{R}^2 : \text{dist}(y, S) < r\}$. The function h_ϵ (the blue piece of function in Figure 3.2) is obtained as the solution of the differential problem $h' = (1-h)/2\epsilon$, $h(b_\epsilon) = 0$, where $h(b_\epsilon + a_\epsilon) = 1 - \eta_\epsilon$. Exploiting the Schwartz inequality $a^2 + b^2 \geq 2ab$ it is possible to prove that such h_ϵ is energetically optimal in the class of the admissible functions (a general result is given in [18] and used for the approximation of discontinuity sets in [14, 17]). Because of the

²In practice it is assumed that $\mathcal{H}^2(\{\sigma = 0\}) = 0$ and $0 \leq \mathcal{H}^1(\{\sigma = 0\}) < \infty$, for either $\sigma = s$ and $\sigma = z$.

global minimization of \mathcal{F}_ϵ , the distance term $\mu|u_\epsilon - g|^2$ keeps the function u_ϵ close to g . High values of $|\nabla u_\epsilon|$ (associated to discontinuities of g) and high values of $|Hu_\epsilon|$ (associated to crease points of g) force the transition of the functions s_ϵ and z_ϵ from 1 to 0. Elsewhere, the minimization of the two terms containing the differential operators causes the smoothing of g . We remark here the importance of the parameters $\delta, \mu, \alpha, \beta$, that control the ratio at which the whole mechanism described before takes place.

From the discussion above it follows that, for small values of ϵ , the computation of a minimizing triplet of (4.3) provides u_ϵ , an approximation of a real minimizer u of \mathcal{F} , and s_ϵ, z_ϵ , the functions that map the tubular neighborhoods of the discontinuity sets S_u and $S_{\nabla u} \cup S_u$, respectively. The price to pay for having such nice outputs is computational complexity. In the remainder of the chapter we will show how the explicit minimization of (4.3) can be addressed in an efficient way by exploiting the nice properties of the functional and a compact formulation via finite-difference schemes enjoying good properties of convergence.

3.2 Numerical minimization of the Blake-Zisserman functional

In this section the numerical minimization of (4.3) is addressed. Firstly the functional is discretized and written in matricial form. Because of nice properties of the functional, the finite-difference discretization of the functional leads to a quadratic function with respect to each block variable when the others two are left fixed.

3.2.1 Discretization

A simple discretization technique, commonly used for computer vision problems (see for example [41, 42]), can be applied to the functional (4.3) in a straightforward way. The rectangular domain $\Omega_0 \subset \mathbb{R}^2$ is discretized by a lattice of points $\Lambda = \{(it_x, jt_y); i = 1, \dots, N, j = 1, \dots, M\}$ with step sizes t_x and t_y on the x and y directions respectively, giving rise to a point grid Λ of size $n := NM$. Using the standard representation of grey-scale images as matrices, the values of the image g on the grid points (it_x, jt_y) are denoted g_{ij} . Similarly, the approximate values of the functions s, z, u on the grid points are denoted s_{ij}, z_{ij}, u_{ij} . Furthermore, for any function $v \in \{g, s, z, u\}$, we denote by \mathbf{v} the column vector of dimension n obtained from the corresponding matrix rearranging the elements v_{ij} by a column-wise vectorization. The function $w(i, j) := (j - 1)N + i$ makes a bijective correspondence between the entry v_{ij} and its position in the vector \mathbf{v} . Shortly, $[\mathbf{v}]_{w(i,j)} = v_{ij}$. Given a vector \mathbf{v} , let us denote $\mathbf{R}_\mathbf{v}$ the diagonal matrix with diagonal entries equal to the elements of \mathbf{v} . Furthermore, we also denote \mathbf{v}^2 the vector

of the squared coefficients of \mathbf{v} , i.e., $[\mathbf{v}^2]_i = ([\mathbf{v}]_i)^2$ and $\mathbf{e} := (1, 1, \dots, 1)^T$. The maximum value of the entries of a vector is denoted by $\|\mathbf{v}\|_\infty := \max_i [\mathbf{v}]_i$.

The first and second order differential operators appearing in the functional can be approximated via finite difference-schemes as follows

$$\begin{aligned}
\partial_x v_{ij} &:= \frac{v_{i+1,j} - v_{i,j}}{t_x} = [\mathbf{D}_x \mathbf{v}]_{w(i,j)} \\
\partial_y v_{ij} &:= \frac{v_{i,j+1} - v_{i,j}}{t_y} = [\mathbf{D}_y \mathbf{v}]_{w(i,j)} \\
\partial_{xx} v_{ij} &:= \frac{v_{i+1,j} - 2v_{i,j} + v_{i-1,j}}{t_x^2} = [\mathbf{D}_{xx} \mathbf{v}]_{w(i,j)} \\
\partial_{yy} v_{ij} &:= \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{t_y^2} = [\mathbf{D}_{yy} \mathbf{v}]_{w(i,j)} \\
\partial_{xy} v_{ij} &:= \frac{1}{t_y} \left(\frac{v_{i+1,j+1} - v_{i,j+1}}{t_x} - \frac{v_{i+1,j} - v_{i,j}}{t_x} \right) = [\mathbf{D}_{xy} \mathbf{v}]_{w(i,j)}
\end{aligned} \tag{3.7}$$

for $i = 1, \dots, N$ and $j = 1, \dots, M$. By assuming zero boundary conditions ($v_{0,j} = v_{N+1,j} = v_{i,0} = v_{i,M+1} = 0$) as in [14], the above matrices \mathbf{D}_x , \mathbf{D}_y , \mathbf{D}_{xx} , \mathbf{D}_{yy} are given by

$$\begin{aligned}
\mathbf{D}_x &:= \frac{1}{t_x} \mathbf{I}_M \otimes \mathbf{A}_N^1 & \mathbf{D}_y &:= \frac{1}{t_y} \mathbf{A}_M^1 \otimes \mathbf{I}_N \\
\mathbf{D}_{xx} &:= \frac{1}{t_x^2} \mathbf{I}_M \otimes \mathbf{A}_N^2 & \mathbf{D}_{yy} &:= \frac{1}{t_y^2} \mathbf{A}_M^2 \otimes \mathbf{I}_N \\
\mathbf{D}_{xy} &:= \mathbf{D}_y \mathbf{D}_x = \mathbf{D}_x \mathbf{D}_y
\end{aligned}$$

where \otimes is the Kronecker product. Here \mathbf{I}_K denotes the identity matrix of dimension K and \mathbf{A}_K^1 , \mathbf{A}_K^2 are square matrices of order $K > 0$, representing a forward-scheme approximating first-derivative and a central-scheme approximating second-derivative respectively³:

$$\mathbf{A}_K^1 := \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & -1 \end{pmatrix}, \quad \mathbf{A}_K^2 := \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{pmatrix}. \tag{3.8}$$

By using the following approximations over each grid point

$$|Hv_{ij}|^2 = ([\mathbf{D}_{xx} \mathbf{v}]_{w(i,j)})^2 + ([\mathbf{D}_{yy} \mathbf{v}]_{w(i,j)})^2 + 2([\mathbf{D}_{xy} \mathbf{v}]_{w(i,j)})^2,$$

³The implementation of homogeneous Neumann boundary conditions follows straightforwardly by replacing the entries: $[\mathbf{A}_K^1]_{K,K} = 0$ and $[\mathbf{A}_K^2]_{1,1} = [\mathbf{A}_K^2]_{K,K} = -1$.

$$|\nabla v_{ij}|^2 = ([\mathbf{D}_x \mathbf{v}]_{w(i,j)})^2 + ([\mathbf{D}_y \mathbf{v}]_{w(i,j)})^2,$$

we can approximate the integral over Ω_0 with a simple 2-D composite rectangular rule, obtaining the following discrete form of the functional (4.3):

$$\begin{aligned} F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u}) &:= t_x t_y \left\{ \delta [\mathbf{u}^T \mathbf{D}_{xx}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xx} \mathbf{u} + \mathbf{u}^T \mathbf{D}_{yy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{yy} \mathbf{u} + 2\mathbf{u}^T \mathbf{D}_{xy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xy} \mathbf{u}] + \right. \\ &+ \xi_\epsilon [\mathbf{u}^T \mathbf{D}_x^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_x \mathbf{u} + \mathbf{u}^T \mathbf{D}_y^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_y \mathbf{u}] + \\ &+ (\alpha - \beta) [\epsilon (\mathbf{s}^T \mathbf{D}_x^T \mathbf{D}_x \mathbf{s} + \mathbf{s}^T \mathbf{D}_y^T \mathbf{D}_y \mathbf{s}) + \frac{1}{4\epsilon} (\mathbf{s} - \mathbf{e})^T (\mathbf{s} - \mathbf{e})] + \\ &+ \beta [\epsilon (\mathbf{z}^T \mathbf{D}_x^T \mathbf{D}_x \mathbf{z} + \mathbf{z}^T \mathbf{D}_y^T \mathbf{D}_y \mathbf{z}) + \frac{1}{4\epsilon} (\mathbf{z} - \mathbf{e})^T (\mathbf{z} - \mathbf{e})] + \\ &\left. + \mu (\mathbf{u} - \mathbf{g})^T (\mathbf{u} - \mathbf{g}) \right\}. \end{aligned} \quad (3.9)$$

Here, with abuse of notation, $\mathbf{R}_{\mathbf{s}^2}$ is the diagonal matrix composed by the elements of the vector $\mathbf{s}^2 + o_\epsilon$ (instead of \mathbf{s}^2).

Globally this functional is not convex, but it is quadratic with respect to each block of variables \mathbf{s} , \mathbf{z} , \mathbf{u} . The terms of F_ϵ containing \mathbf{s} or \mathbf{z} depend only on \mathbf{u} . On the other hand, the terms containing \mathbf{u} depend on \mathbf{s} and \mathbf{z} . Indeed, by fixing the variable \mathbf{u} or the other two variables \mathbf{s} and \mathbf{z} , we can write

$$\begin{aligned} F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u}) &= t_x t_y \left\{ \frac{1}{2} (\mathbf{s}^T \mathbf{z}^T) \begin{pmatrix} \mathbf{A}_s & 0 \\ 0 & \mathbf{A}_z \end{pmatrix} \begin{pmatrix} \mathbf{s} \\ \mathbf{z} \end{pmatrix} - (\mathbf{s}^T \mathbf{z}^T) \begin{pmatrix} \mathbf{b}_s \\ \mathbf{b}_z \end{pmatrix} + \mathbf{c}_{sz} \right\} \\ F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u}) &= t_x t_y \left\{ \frac{1}{2} \mathbf{u}^T \mathbf{A}_u \mathbf{u} - \mathbf{u}^T \mathbf{b}_u + \mathbf{c}_u \right\} \end{aligned} \quad (3.10)$$

where $\mathbf{A}_s = \mathbf{A}_s(\mathbf{u})$, $\mathbf{A}_z = \mathbf{A}_z(\mathbf{u})$, $\mathbf{A}_u = \mathbf{A}_u(\mathbf{s}, \mathbf{z})$ and $\mathbf{b}_s, \mathbf{b}_z, \mathbf{b}_u$ are given by

$$\begin{aligned} \mathbf{A}_s &= 2\xi_\epsilon \mathbf{R}_{|\nabla \mathbf{u}|^2} + 2\epsilon(\alpha - \beta)(\mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y) + \frac{\alpha - \beta}{2\epsilon} \mathbf{I} \\ \mathbf{b}_s &= \frac{\alpha - \beta}{2\epsilon} \mathbf{e} \\ \mathbf{A}_z &= 2\delta \mathbf{R}_{|\nabla^2 \mathbf{u}|^2} + 2\epsilon\beta(\mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y) + \frac{\beta}{2\epsilon} \mathbf{I} \\ \mathbf{b}_z &= \frac{\beta}{2\epsilon} \mathbf{e} \\ \mathbf{A}_u &= 2\delta(\mathbf{D}_{xx}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xx} + \mathbf{D}_{yy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{yy} + 2\mathbf{D}_{xy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xy}) + 2\xi_\epsilon(\mathbf{D}_x^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_x + \mathbf{D}_y^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_y) + 2\mu \mathbf{I} \\ \mathbf{b}_u &= 2\mu \mathbf{g} \end{aligned} \quad (3.11)$$

with $|\nabla \mathbf{u}|^2 := (\mathbf{D}_x \mathbf{u})^2 + (\mathbf{D}_y \mathbf{u})^2$ and $|\mathbf{H}\mathbf{u}|^2 := (\mathbf{D}_{xx} \mathbf{u})^2 + (\mathbf{D}_{yy} \mathbf{u})^2 + 2(\mathbf{D}_{xy} \mathbf{u})^2$. Vectors \mathbf{c}_{sz} and \mathbf{c}_u are constant, thus irrelevant for the minimization. In view of the terms $\frac{\alpha - \beta}{2\epsilon} \mathbf{I}$, $\frac{\beta}{2\epsilon} \mathbf{I}$ and $2\mu \mathbf{I}$, with $\epsilon, \mu, \beta, \alpha - \beta > 0$, the matrices $\mathbf{A}_s, \mathbf{A}_z, \mathbf{A}_u$ are symmetric and positive

definite. Furthermore, these matrices are very sparse and structured: \mathbf{A}_s and \mathbf{A}_z are block tridiagonal matrices where the diagonal blocks are tridiagonal and the off-diagonal blocks are diagonal. \mathbf{A}_u is a block five matrix, with at most 13 nonzero entries for each row.

In the following, for notation convenience, a generic point in \mathbb{R}^{3n} is represented by either \mathbf{y} or $(\mathbf{s}, \mathbf{z}, \mathbf{u})$. This makes a simple correspondence of the type: $\mathbf{y}_1 = \mathbf{s}$, $\mathbf{y}_2 = \mathbf{z}$ and $\mathbf{y}_3 = \mathbf{u}$. Accordingly, throughout the chapter a similar correspondence is used for denoting operators/vectors related to a specific block of variables. For example: $\mathbf{A}_s = \mathbf{A}_1$, $\mathbf{A}_z = \mathbf{A}_2$, $\mathbf{A}_u = \mathbf{A}_3$, and $\mathbf{b}_s = \mathbf{b}_1$, $\mathbf{b}_z = \mathbf{b}_2$, $\mathbf{b}_u = \mathbf{b}_3$ etc. Furthermore, we denote the gradient of F_ϵ with respect to the generic block of variables \mathbf{y}_i , computed at $\bar{\mathbf{y}}$, by $\nabla_i F_\epsilon(\bar{\mathbf{y}}) = \mathbf{A}_i \bar{\mathbf{y}}_i - \mathbf{b}_i$.

3.2.2 Minimization method

We address here the minimization of the function $F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u})$. Firstly we can observe that the objective function is continuously differentiable, and in view of the positive definiteness of matrices \mathbf{A}_i , $i = 1, 2, 3$, it is strictly convex with respect to each block component \mathbf{y}_i , when the others are left fixed. Let us prove that F_ϵ is also coercive.

Lemma 1. *The function $F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u})$ is coercive in \mathbb{R}^{3n} .*

Proof. Given a sequence $\mathbf{y}^k = (\mathbf{s}^k, \mathbf{z}^k, \mathbf{u}^k) \in \mathbb{R}^{3n}$ such that $\lim_{k \rightarrow \infty} \|\mathbf{y}^k\| = +\infty$ the lemma is proved if we show that $\lim_{k \rightarrow \infty} F_\epsilon(\mathbf{y}^k) = +\infty$. The hypothesis on \mathbf{y}^k implies that there exists a coordinate index $j \in \{1, 2, \dots, 3n\}$ such that $\lim_{k \rightarrow \infty} |\mathbf{y}_j^k| = +\infty$. If $1 \leq j \leq n$, then j corresponds to an index i in the \mathbf{s} block and we have that $\lim_{k \rightarrow \infty} |\mathbf{s}_i^k| = +\infty$. In particular $\lim_{k \rightarrow \infty} (\mathbf{s}_i^k - 1)^2 = +\infty$ and since $(\mathbf{s}_i^k - 1)^2 \leq (\mathbf{s}^k - \mathbf{e})^T (\mathbf{s}^k - \mathbf{e})$ we also have that $\lim_{k \rightarrow \infty} F_\epsilon(\mathbf{y}^k) = +\infty$. A similar argument works in the case of $n + 1 \leq j \leq 2n$, where j corresponds to an index in the \mathbf{z} block. If $2n + 1 \leq j \leq 3n$ then j corresponds to an index i in the \mathbf{u} block. From $\lim_{k \rightarrow \infty} |\mathbf{u}_i^k| = +\infty$ we have that $\lim_{k \rightarrow \infty} (\mathbf{u}_i^k - \mathbf{g}_i)^2 = +\infty$ and, since $(\mathbf{u}_i^k - \mathbf{g}_i)^2 \leq (\mathbf{u}^k - \mathbf{g})^T (\mathbf{u}^k - \mathbf{g})$, then we have again $\lim_{k \rightarrow \infty} F_\epsilon(\mathbf{y}^k) = +\infty$. \square

By using a truncation argument we have that the functions \mathbf{s} and \mathbf{z} that minimize the objective functional F_ϵ belong to a specific compact subset of \mathbb{R}^{3n} . In fact, given $\tau(\mathbf{v}) := 0 \vee \mathbf{v} \wedge 1$, i.e., the function that truncates \mathbf{v} at 0 and 1, one can see that for any triplet $(\mathbf{s}, \mathbf{z}, \mathbf{u}) \in \mathbb{R}^{3n}$

$$F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u}) \geq F_\epsilon(\tau(\mathbf{s}), \tau(\mathbf{z}), \mathbf{u}) \quad (3.12)$$

holds (in fact, the truncation of \mathbf{s} and \mathbf{z} does not increase their gradients). It follows that for a minimizer $(\mathbf{s}, \mathbf{z}, \mathbf{u})$ the functions $\mathbf{s}, \mathbf{z} \in [0, 1]^n$. A similar argument is used for the

Mumford-Shah functional in [17] to prove that the optimal u is such that $\|u\|_\infty \leq \|g\|_\infty$ (maximum principle). Unfortunately, the Hessian component of the Blake-Zisserman functional does not allow us to exploit the maximum principle and an explicit bound for the function \mathbf{u} cannot be calculated, [33].

The structure of the function $F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u})$ justifies the use of a block decomposition method, such as the **block non-linear Gauss-Seidel (GS)** method. Starting from $(\mathbf{s}^0, \mathbf{z}^0, \mathbf{u}^0)$, in view of (3.10), the method has the following form:

$$\begin{cases} \mathbf{s}^{k+1} = \arg \min_{\mathbf{s}} F_\epsilon(\mathbf{s}, \mathbf{z}^k, \mathbf{u}^k) \\ \mathbf{z}^{k+1} = \arg \min_{\mathbf{z}} F_\epsilon(\mathbf{s}^{k+1}, \mathbf{z}, \mathbf{u}^k) \\ \mathbf{u}^{k+1} = \arg \min_{\mathbf{u}} F_\epsilon(\mathbf{s}^{k+1}, \mathbf{z}^{k+1}, \mathbf{u}) \end{cases} . \quad (3.13)$$

Because of the the block diagonal structure of the matrix related to the quadratic functional obtained by fixing \mathbf{u} in the first subproblem in (3.10), \mathbf{s}^{k+1} and \mathbf{z}^{k+1} can be obtained by further subdividing this subproblem into two independent tasks.

Theorem 6.2 in [43] assures that the algorithm generates a sequence $\{\mathbf{s}^k, \mathbf{z}^k, \mathbf{u}^k\}$ such that every limit point is a stationary point of F_ϵ . Because of coercivity, the level sets $\mathcal{L}_\alpha = \{(\mathbf{s}, \mathbf{z}, \mathbf{u}) : F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u}) \leq \alpha\}$ are compact for every $\alpha > 0$. Since in particular \mathcal{L}_{α^0} is compact, where $\alpha^0 = F_\epsilon(\mathbf{s}^0, \mathbf{z}^0, \mathbf{u}^0)$, the theorem also guarantees that $\nabla F_\epsilon(\mathbf{s}^k, \mathbf{z}^k, \mathbf{u}^k) \rightarrow 0$ as $k \rightarrow \infty$ and there exists at least a limit point that is a stationary point of F_ϵ .

Nevertheless, any step of the non-linear Gauss-Seidel method requires the solution of three large and sparse systems. Although such systems can be efficiently solved by the Preconditioned Conjugate Gradient (PCG) algorithm, the whole method could be too expensive, above all for large images.

Therefore, we propose to solve our minimization problem with a **block coordinate descent algorithm (BCDA)**, based on the line search technique described in [43]. The basic idea of the method is to cyclically determine for each block variable a descent direction \mathbf{d}_i by few iterations of an iterative solver; then by an Armijo-type procedure a suitable step size is devised to assure a sufficient decrease of the objective function along this direction with respect to the i -th block variable, when the remaining variables are fixed.

In view of the special structure of $F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u})$, for each subproblem in (3.10) we can cyclically obtain a descent direction by few iterations of the PCG method applied to the linear system $\mathbf{A}_i^k \mathbf{d}_i = \mathbf{b}_i - \mathbf{A}_i^k \mathbf{y}_i^k$. In the first subproblem, \mathbf{d}_s^k and \mathbf{d}_z^k can be independently obtained. Furthermore, in view of the quadratic structure of the objective function with respect to each block of variables when the others are fixed, the step-lengths along the computed descent directions can be determined without having to use

Algorithm 1 BCDA

Step 0: Given $\mathbf{s}^0, \mathbf{z}^0, \mathbf{u}^0, \rho_{sz} > 0, \rho_u > 0, \gamma_s \in (0, 2), \gamma_z \in (0, 2), \gamma_u \in (0, 2)$;**Step 1:** $k = 0$;**Step 2:** Inexact minimization with respect to \mathbf{s} and \mathbf{z} :

- compute the search directions \mathbf{d}_s^k and \mathbf{d}_z^k ;
- compute $\alpha_s^k = \gamma_s \frac{-(A_s^k \mathbf{s}^k - \mathbf{b}_s)^T \mathbf{d}_s^k}{\mathbf{d}_s^{kT} A_s^k \mathbf{d}_s^k}$, $\alpha_z^k = \gamma_z \frac{-(A_z^k \mathbf{z}^k - \mathbf{b}_z)^T \mathbf{d}_z^k}{\mathbf{d}_z^{kT} A_z^k \mathbf{d}_z^k}$
- update $\mathbf{s}^{k+1} = \mathbf{s}^k + \alpha_s^k \mathbf{d}_s^k$; $\mathbf{z}^{k+1} = \mathbf{z}^k + \alpha_z^k \mathbf{d}_z^k$.

Step 3: Inexact minimization with respect to \mathbf{u} :

- compute the search directions \mathbf{d}_u^k ;
- compute $\alpha_u^k = \gamma_u \frac{-(A_u^k \mathbf{u}^k - \mathbf{b}_u)^T \mathbf{d}_u^k}{\mathbf{d}_u^{kT} A_u^k \mathbf{d}_u^k}$
- update $\mathbf{u}^{k+1} = \mathbf{u}^k + \alpha_u^k \mathbf{d}_u^k$.

Step 4: Set $k = k + 1$ and go to Step 2;

an Armijo-type procedure. Indeed, it is well known that, for a symmetric positive definite quadratic function, a sufficient decrease is assured when the step size α_i^k is chosen as $\gamma_i \frac{-(A_i^k \mathbf{y}_i^k - \mathbf{b}_i)^T \mathbf{d}_i^k}{\mathbf{d}_i^{kT} A_i^k \mathbf{d}_i^k} = \gamma_i \frac{-\nabla_i F_\epsilon(\mathbf{y}^k)^T \mathbf{d}_i^k}{\mathbf{d}_i^{kT} A_i^k \mathbf{d}_i^k}$, with $0 < \gamma_i < 2$; in particular, for $\gamma_i = 1$, we obtain the exact one-dimensional minimizer of the quadratic function along the direction \mathbf{d}_i^k . As consequence, we can devise a specialized version of the block-coordinate descent algorithm for $F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u})$; such scheme is outlined in Algorithm 1.

Gradient related search directions

In order to obtain convergence results for BCDA, the vectors \mathbf{d}_i^k , $i = 1, 2, 3$, have to be chosen so that they are *gradient related search directions*. Equivalently, they have to satisfy the following assumption:

- $\mathbf{d}_i^k = 0$ if and only if $\nabla_i F_\epsilon(\mathbf{y}^k) = 0$,
- there exists a forcing function $\sigma_i : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that:

$$\frac{\nabla_i F_\epsilon(\mathbf{y}^k)^T \mathbf{d}_i^k}{\|\mathbf{d}_i^k\|} \leq -\sigma_i(\|\nabla_i F_\epsilon(\mathbf{y}^k)\|) \quad (3.14)$$

for all k satisfying $\nabla_i F_\epsilon(\mathbf{y}^k) \neq 0$.

In order to determine a gradient related search direction when $\mathbf{A}_i^k \mathbf{y}_i^k - \mathbf{b}_i \neq 0$, we can execute several iterations of PCG method for the symmetric positive linear system $\mathbf{A}_i^k \mathbf{d}_i = \mathbf{b}_i - \mathbf{A}_i^k \mathbf{y}_i^k$ by stopping the algorithm when the residual $\mathbf{r}^\ell = \mathbf{b}_i - \mathbf{A}_i^k \mathbf{y}_i^k - \mathbf{A}_i^k \mathbf{d}_i^\ell$ of the system at the $\bar{\ell}$ -th iteration satisfies the rule

$$\|\mathbf{r}^\ell\| \leq \eta_i \|\mathbf{A}_i^k \mathbf{y}_i^k - \mathbf{b}_i\| \quad \eta_i \leq \frac{c}{\sqrt{K((\mathbf{A}_i^k)^{-1})}}, \quad (3.15)$$

where $K((\mathbf{A}_i^k)^{-1})$ is the spectral condition number of $(\mathbf{A}_i^k)^{-1}$ and $c < 1$. We observe that $K((A_i^k)^{-1}) = K(A_i^k)$ is bounded by a positive constant L in \mathcal{L}_{α^0} . Then, we set $\mathbf{d}_i^{\bar{\ell}} = \mathbf{d}_i^k$. We can prove that \mathbf{d}_i^k satisfies the assumption (3.14). Indeed, recalling that $\|\cdot\|_{A^{-1}}$ denotes the A^{-1} -norm (that is $\|\mathbf{x}\|_{A^{-1}} = \sqrt{\mathbf{x}^T A^{-1} \mathbf{x}}$), for $\mathbf{d}_i^{\bar{\ell}} = \mathbf{d}_i^k$ we have

$$\begin{aligned} \frac{\nabla_i F_\epsilon(\mathbf{y}^k)^T \mathbf{d}_i^k}{\|\mathbf{d}_i^k\|} &\leq \frac{(\mathbf{A}_i^k \mathbf{y}_i^k - \mathbf{b}_i)^T (A_i^k)^{-1} A_i^k \mathbf{d}_i^{\bar{\ell}}}{\|\mathbf{d}_i^{\bar{\ell}}\|} + \\ &\quad \frac{1}{2\|\mathbf{d}_i^{\bar{\ell}}\|} (\|\mathbf{r}^{\bar{\ell}} + \nabla_i F_\epsilon(\mathbf{y}^k)\|_{(A_i^k)^{-1}}^2 + \|\nabla_i F_\epsilon(\mathbf{y}^k)\|_{(A_i^k)^{-1}}^2 - \|\nabla_i F_\epsilon(\mathbf{y}^k)\|_{(A_i^k)^{-1}}^2) \\ &= \frac{1}{2\|\mathbf{d}_i^{\bar{\ell}}\|} \left(\|\mathbf{r}^{\bar{\ell}}\|_{(A_i^k)^{-1}}^2 - \|\nabla_i F_\epsilon(\mathbf{y}^k)\|_{(A_i^k)^{-1}}^2 \right) \\ &\leq \frac{1}{2\|\mathbf{d}_i^{\bar{\ell}}\|} \left(\lambda_{\max}((A_i^k)^{-1}) \|\mathbf{r}^{\bar{\ell}}\|^2 - \lambda_{\min}((A_i^k)^{-1}) \|\nabla_i F_\epsilon(\mathbf{y}^k)\|^2 \right) \quad (3.16) \\ &\leq \frac{1}{2\|\mathbf{d}_i^{\bar{\ell}}\|} (\lambda_{\max}((A_i^k)^{-1}) \eta_i^2 - \lambda_{\min}((A_i^k)^{-1})) \|\nabla_i F_\epsilon(\mathbf{y}^k)\|^2 \quad (3.17) \end{aligned}$$

where (3.16) follows from the well-known inequalities

$$\lambda_{\min}((A_i^k)^{-1}) \|\mathbf{x}\|^2 \leq \mathbf{x}^T (A_i^k)^{-1} \mathbf{x} \leq \lambda_{\max}((A_i^k)^{-1}) \|\mathbf{x}\|^2$$

for any \mathbf{x} , and the inequality (3.17) follows from (3.15). Furthermore, the bound on η_i implies $(\lambda_{\max}((A_i^k)^{-1}) \eta_i^2 - \lambda_{\min}((A_i^k)^{-1})) < 0$; then, since $\mathbf{d}_i^{\bar{\ell}} = (\mathbf{A}_i^k)^{-1} (-\nabla_i F_\epsilon(\mathbf{y}^k) - \mathbf{r}^{\bar{\ell}})$, we have

$$\begin{aligned} \|\mathbf{d}_i^{\bar{\ell}}\| &\leq \|(\mathbf{A}_i^k)^{-1}\| (\|\nabla_i F_\epsilon(\mathbf{y}^k)\| + \|\mathbf{r}^{\bar{\ell}}\|) \\ &\leq \lambda_{\max}((\mathbf{A}_i^k)^{-1}) (1 + \eta_i) \|\nabla_i F_\epsilon(\mathbf{y}^k)\| \end{aligned}$$

Using this inequality in (3.17), we can conclude that

$$\begin{aligned} \frac{\nabla_i F_\epsilon(\mathbf{y}^k)^T \mathbf{d}_i^k}{\|\mathbf{d}_i^k\|} &\leq \frac{1}{2(1 + \eta_i)} \left(\eta_i^2 - \frac{1}{K((\mathbf{A}_i^k)^{-1})} \right) \|\nabla_i F_\epsilon(\mathbf{y}^k)\| \\ &\leq \frac{c^2 - 1}{2L} \|\nabla_i F_\epsilon(\mathbf{y}^k)\|, \quad (3.18) \end{aligned}$$

where $c^2 - 1 < 0$. Consequently, in the level set \mathcal{L}_{α^0} , the search directions \mathbf{d}_s^k , \mathbf{d}_z^k and \mathbf{d}_u^k generated by a PCG method with stopping rule (3.15) are *gradient related search directions*. Therefore, by Theorem 7.1 in [43], we can affirm that for BCDA the same convergence results hold as for block non-linear Gauss Seidel method, that is $\nabla F_\epsilon(\mathbf{u}^k, \mathbf{z}^k, \mathbf{z}^k) \rightarrow 0$ as $k \rightarrow \infty$ and there exists at least a limit point in \mathcal{L}_{α^0} that is a stationary point of F_ϵ .

Algorithm parameters and preconditioning

We observe that from the practical point of view the computation of the condition number of $(\mathbf{A}_i^k)^{-1}$ (which equals that of \mathbf{A}_i^k), can be avoided. Indeed, in view of the condition (3.15) on η_i , it is sufficient to have an upper bound for $K(\mathbf{A}_i^k)$. Then, using the inequalities between the matrix norms, we have $\lambda_{max}(\mathbf{A}_i^k) \leq \|\mathbf{A}_i^k\|_\infty$. Since \mathbf{A}_s and \mathbf{A}_z are strictly diagonally dominant, the first Gerschgorin's theorem [44] enables us to determine as lower bound for the minimum eigenvalue the intersection between the union of the Gerschgorin circles and the x -axis of \mathbb{R}^2 , given by $\tilde{\lambda}_i^k = \min_t((\mathbf{A}_i^k)_{tt} - \sum_{v \neq t} |(\mathbf{A}_i^k)_{vt}|)$ for $i = 1, 2$. For the matrix \mathbf{A}_3 , a lower bound for the minimum eigenvalue is the value $\tilde{\lambda}_3^k = 2\mu$. Therefore, we have

$$K(\mathbf{A}_i^k) = \frac{\lambda_{max}(\mathbf{A}_i^k)}{\lambda_{min}(\mathbf{A}_i^k)} \leq \frac{\|\mathbf{A}_i^k\|_\infty}{\tilde{\lambda}_i^k}, \quad (3.19)$$

consequently, we can set

$$\eta_i = \sqrt{\frac{\tilde{\lambda}_i^k}{\|\mathbf{A}_i^k\|_\infty}}. \quad (3.20)$$

From the computational point of view, the quadratic structure of F_ϵ restricted to any block variable \mathbf{y}_i implies that the condition related to the Armijo rule that has to be verified to accept the step size $\alpha_i^k = \max_{j \geq 0} \{\delta_i^j \Delta_i^k\}$ can be

$$\alpha_i^k \leq 2(\gamma_i - 1) \frac{(A_i^k \mathbf{y}_i^k - \mathbf{b}_i)^T \mathbf{d}_i^k}{(\mathbf{d}_i^k)^T A_i^k \mathbf{d}_i^k}. \quad (3.21)$$

Furthermore, since \mathbf{A}_s and \mathbf{A}_z are block tridiagonal matrices, an inexpensive diagonal preconditioner enable us to satisfy the stopping rule (3.15) with a very few iterations, as shown by numerical experiments in the Section 3.3.

For the linear system related to matrix \mathbf{A}_u we can use a diagonal preconditioner or a block diagonal preconditioner. In this last case, each diagonal block is a tridiagonal matrix that can be easily factorized by the Cholesky algorithm. Although the factorization can be calculated in advance, PCG requires the factorization of the preconditioner and, at each iteration, the solution of bidiagonal lower and upper systems; thus PCG coupled

with block preconditioner becomes effective with respect to the version with diagonal preconditioner for large order of the system and only when high accuracy is required. Furthermore, unless for the first outer iteration, we use as starting vector of the PCG algorithm in Step 2 and 3 of BCDA the direction computed at the previous outer iteration. For the first iteration, the starting vector is the null vector.

In the numerical experiments, standard values equal to 1 are set for γ_s and γ_z ; we set $\gamma_u = 1.5$, since for this value we obtain a slightly better performance.

Initialization and stopping criteria

The objective functional to minimize is non-convex; thus, the significance of the solution returned by the iterative method (a stationary point) strongly depends on the choice of the first iterates. Using prior knowledge on the properties of the theoretical solution, an effective choice of the initial values can be made [14]. Being the functions $s, z \equiv 1$ almost everywhere over Ω_0 (in the limit case), we set the corresponding variables to $\mathbf{s}^0, \mathbf{z}^0 \equiv \mathbf{e}$. Since the function u is an approximation of the input image g , we assume that an energetically convenient initialization of the corresponding variable is $\mathbf{u}^0 = \mathbf{g}$.

In the experiments described in the next section, the algorithms are stopped at the iteration k such that the relative variation of the energy satisfies the condition

$$\left| \frac{F_\epsilon(\mathbf{u}^k, \mathbf{s}^k, \mathbf{z}^k) - F_\epsilon(\mathbf{u}^{k-1}, \mathbf{s}^{k-1}, \mathbf{z}^{k-1})}{F_\epsilon(\mathbf{u}^k, \mathbf{s}^k, \mathbf{z}^k)} \right| < TOL_F, \quad (3.22)$$

where TOL_F is a fixed tolerance.

3.3 Numerical results

In this section we present the results of an extensive numerical experimentation aimed at assessing different properties of the proposed block-coordinate descent method applied to (3.9). In Section 3.3.1 the proposed BCDA is compared with the GS in order to evaluate its performance both in terms of efficiency and accuracy. In Section 3.3.2, by focusing the attention on the noise reduction properties of the model, we compare the performance of the BCDA when a Point-Diagonal and a Block-Diagonal preconditioner is used for the solution of the PCG related to the linear systems involving \mathbf{A}_u . In Section 3.3.3 we make a discussion on the choice of the Γ -convergence parameter ϵ , which may critically affect the quality of the detection of first and second order discontinuities.

Very different datasets are considered in the tests, including both real and synthetic images and also Digital Surface Models (DSMs) obtained from remote sensing LiDAR (Light Detection and Ranging) data [45]. DSMs are obtained from airborne LiDAR point

clouds by interpolation over a regular planimetric grid. The value of the DSM on each pixel (grid point) corresponds to the height of the object hit by the laser pulse. In particular DSMs are very attractive as they represent the *real geometry* of the objects instead of the *light geometry* provided by gray-scale images.

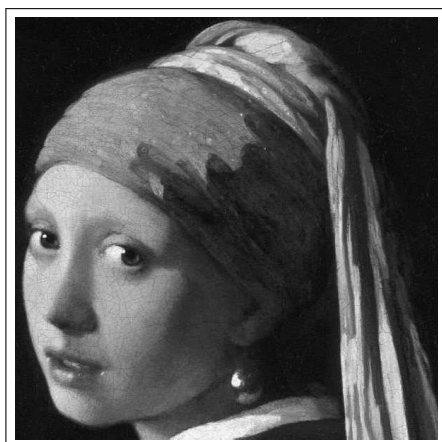
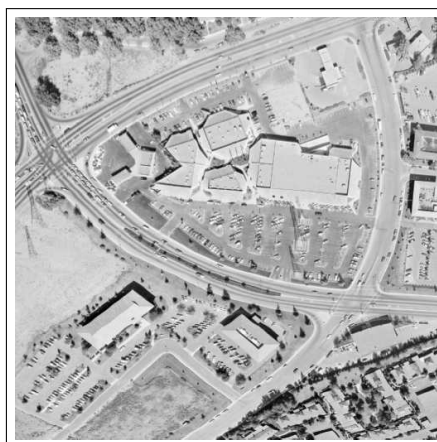
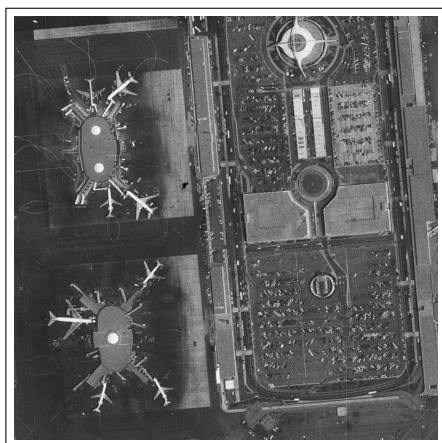
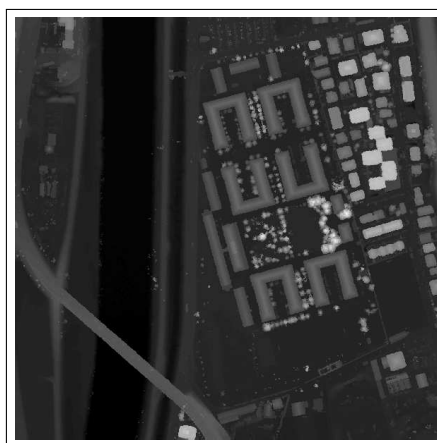
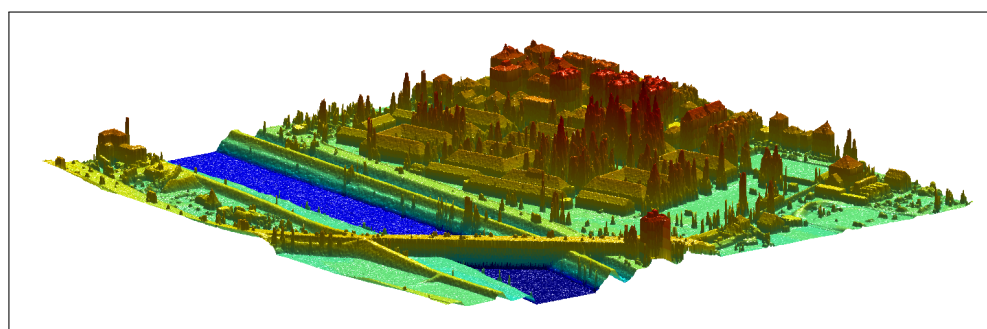
All tests are performed using MATLAB[®] on a standard workstation. Hardware is Intel(R) Core(TM) i5-4750 CPU @ 3.20 GHz, 8.00 GB Ram. For all numerical tests that follow, some common parameters to control convergence of outer/inner iterations are used. The algorithms are always stopped at the iteration k such that the corresponding relative variation of the energy (3.22) is less than $TOL_F = 10^{-3}$. A maximum number of outer iterations is also fixed as stopping criterion to 30. It is worth noting that in all computations this bound has never been reached. Regarding the solution of internal PCGs, we fixed a maximum number of iterations to 1000. It has been observed in a very large number of tests that, the linear systems involving $\mathbf{A}_s, \mathbf{A}_z$ are solved within an inner tolerance TOL_{PCG} in no more than 3 iterations in the case of GS, and in only 1 iteration in the case of BCDA. Other parameters and tolerances are explicitly specified in the tests.

3.3.1 Comparison of GS and BCDA performance

In this section we show how the proposed BCDA produces accurate solutions by also significantly reducing computational time if compared with a GS method. In order to compare the effectiveness of the two methods, we compute an *ideal* solution $\mathbf{s}^*, \mathbf{z}^*, \mathbf{u}^*$ by performing a lot of iterations of the GS method, i.e., by running GS until it reaches stagnation since all PCGs (with a very strong relative tolerance $TOL_{PCG} = 10^{-10}$) do not make any progress.

For the inner PCGs in the GS method, a strong relative tolerance $TOL_{PCG} = 10^{-8}$ is required; smaller tolerances have never resulted in lower minimizers. Tolerances for the solution of inner iterations of the PCGs in the BCDA are theoretically defined by (3.15) and (3.20). Furthermore, we propose also an hybrid version of the BCDA, to which we will refer to as BCDA_c, where the number of iterations for solving the inner PCGs is capped at 10 (in view of a previous remark this affects only the solution of the system involving \mathbf{A}_u). The main idea behind this choice it to show that actually just few steps of the inner solvers are needed to reach satisfying results at lower computational cost, even though a small (negligible) amount of accuracy is payed. Since the performance of the method when a block-diagonal preconditioner is used is analyzed in detail in Section 3.3.2, in all tests conducted in this section a diagonal preconditioner is used for the PCGs.

We consider as test problems the four datasets represented in Figure 3.3. The first image is a 600×600 portion of the oil painting “Girl with a Pearl Earring” by Johannes

(a) pearl (600×600)(b) aerial (512×512)(c) airport (1024×1024)(d) barracks (600×600)

(e) barracks (3D rendering)

Figure 3.3: Datasets of the experiment including three gray-scale images and a digital surface model obtained from airborne LiDAR points acquired over Trento, Italy.

van der Meer. We refer this dataset to as **pearl**. Then, two aerial gray-scale images⁴ are considered. The first image is called **aerial** and it has size of 512×512 pixels. The second image is called **airport** and it has size of 1024×1024 pixels. The last dataset is a subset of a DSM of Trento⁵, in Trentino Alto-Adige, Italy. The considered scene presents some old barracks and surrounding area. The size of the grid is 600×600 , spatial resolution is 1mt. We refer this dataset to as **barracks**. In Figure 3.3e a 3D rendering of the surface model is shown. Functional parameters for the minimization on these datasets are set as follows:

- **pearl**: $\epsilon = 0.01, \delta = 3, \alpha = 2, \beta = 1, \mu = 0.07, t = 1$;
- **aerial**: $\epsilon = 0.01, \delta = 1, \alpha = 2, \beta = 1, \mu = 0.05, t = 1$;
- **airport**: $\epsilon = 0.01, \delta = 1, \alpha = 2, \beta = 1, \mu = 0.05, t = 1$;
- **barracks**: $\epsilon = 0.01, \delta = 30, \alpha = 2, \beta = 1, \mu = 1, t = 1$.

In Table 3.1 we report the total number of outer (k) and inner (totiter) iterations of GS, BCDA and BCDA_c together with the total time in seconds (the mean of ten runs) required to compute an approximate solution, with $TOL_F = 10^{-3}$ in (3.22). Computational performance of GS, BCDA, and BCDA_c is better illustrated in Figure 3.4, where the value of the objective function at each outer iteration is plotted against the cumulative execution time.

The solutions obtained with all the proposed methods are very similar each other, in such a way that they cannot be distinguished only visually. The accuracy (and the similarity) of the solutions obtained with GS, BCDA and BCDA_c can be measured in terms of their distance to the *ideal* solutions. To this aim we defined a *normalized* distance based on the L^1 norm as follows. In view of (3.12), the solutions of the algorithms are such that $\mathbf{s}, \mathbf{z} \in [0, 1]^n$. Given $H := [0, h]^n$, with $h > 0$, we easily have that

$$d(h) := \max_{\mathbf{x}, \mathbf{y} \in H} \|\mathbf{x} - \mathbf{y}\|_1 = hn. \quad (3.23)$$

Therefore, we can define a normalized distance function $d_h : H \times H \rightarrow [0, 1]$ by setting $d_h(\mathbf{x}, \mathbf{y}) := \frac{1}{d(h)} \|\mathbf{x} - \mathbf{y}\|_1$. As a consequence, the maximum possible distance in $H := [0, 1]^n$ is 1, and the value $100 \cdot d_h(\mathbf{x}, \mathbf{y})$ can be interpreted as the percentage of image content which is changed between \mathbf{x} and \mathbf{y} . Thus, the distances for the functions \mathbf{s} and \mathbf{z} are given in terms of d_1 . As already mentioned in Section 3.2.2, it is not possible to define an explicit bound for the values of the function \mathbf{u} . Although our method produces a

⁴Images are downloadable at <http://sipi.usc.edu/database/database.php?volume=misc#top>

⁵DSMs are downloadable at <http://www.territorio.provincia.tn.it/portal/server.pt/community/lidar/847/lidar/23954>

Table 3.1: Outer/inner iterations and execution time (seconds) observed in the run of the proposed algorithms.

dataset	method	k	totiter (s-z-u)	time
pearl 600×600	GS	18	43-49-6435	190.3
	BCDA	14	14-14-690	38.9
	BCDA _c	14	14-14-106	25.0
aerial 512×512	GS	19	49-55-4592	105.9
	BCDA	15	15-15-492	25.8
	BCDA _c	11	11-11-62	14.4
airport 1024×1024	GS	20	51-55-4937	460.1
	BCDA	16	16-16-493	107.9
	BCDA _c	12	12-12-72	62.1
barracks 600×600	GS	14	31-39-3906	119.1
	BCDA	12	12-12-463	29.3
	BCDA _c	10	10-10-69	17.3

sequence of iterates $(\mathbf{s}^k, \mathbf{z}^k, \mathbf{u}^k) \in \mathcal{L}_{\alpha^0}$, the inequality $F_\epsilon(\mathbf{s}^k, \mathbf{z}^k, \mathbf{u}^k) \leq F_\epsilon(\mathbf{e}, \mathbf{e}, \mathbf{g})$ does not necessarily imply that $\|\mathbf{u}^k\|_\infty \leq \|\mathbf{g}\|_\infty$. However, in all numerical experiments presented in this chapter (and in many other experiments performed by the authors) it is observed that the optimal \mathbf{u} satisfies $\|\mathbf{u}^k\|_\infty \leq \|\mathbf{g}\|_\infty$, thus it is meaningful to compute the distances of the solutions \mathbf{u} by means of d_G , where $G := \|\mathbf{g}\|_\infty$.

The distances between the solutions of GS, BCDA, BCDA_c with respect to the *ideal* solutions are given in Table 3.2. By analyzing the results we see that the BCDA algorithm, if compared to GS, significantly decreases the time of computation of approximately 75% and returns solutions that do not differ more than 1% with respect to the *ideal* solution. BCDA and GS always resulted in the same accuracy. In the case of BCDA_c, the time of computation further decreases of 10% and the difference of the solutions with respect to the *ideal* solution are never greater than 1%. Notice that the obtained solutions have a lower accuracy as the normalized distances are about 10 times those obtained by the other two methods. In Figure 3.5, for each dataset we plot the smooth approximation \mathbf{u} , the edge-map \mathbf{s} and the edge/crease-map \mathbf{z} obtained by the BCDA.

Second-order segmentation provides a piecewise linear approximation of the input image. Therefore, if compared to first order models, it avoids the problem of over-segmentation and the real geometry of objects is followed properly. For instance, in the

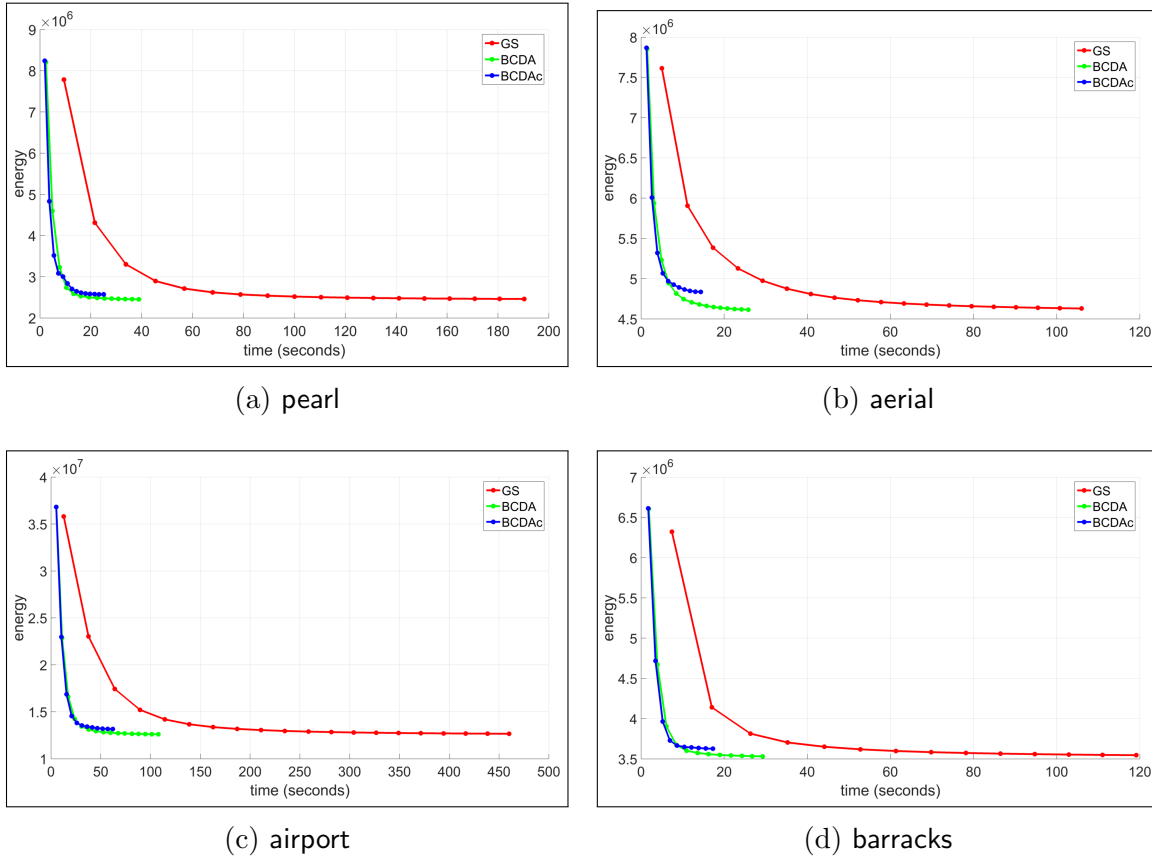


Figure 3.4: Energy-versus-time at every outer iteration for all three datasets with GS, BCDA and BCDAc. Algorithms are stopped by criterion (3.22) with tolerance $TOL_F = 10^{-3}$.

pearl image both the noise and the *craquelure* are removed while the geometry of shadows is preserved. By taking a look at the particular showed in Figure 3.6, we see that around the nostril the variation of grey level in the shadowed area is over-segmented by the \mathbf{s} function (shadow-like trait in the \mathbf{s} map) but it is correctly outlined by the \mathbf{z} function. In fact, boundaries of shadows are characterized by a transient zone of luminance variation which is a ramp and not as sharp as a jump. The **aerial** image is smoothed out and the contrast between the ground and human-made objects is more evident in the segmented image. A similar behavior is observed for **airport**, where the smoothing removes the noise but is able to keep the geometry of cars in the parking area. By looking at the particular in Figure 3.7 we see that almost every trait in \mathbf{s} seems to be *doubled* in \mathbf{z} . Again, this happens because in luminance images the transition of intensity between two areas with different values is usually not purely a jump. The doubled trait is due to the fact that the \mathbf{z} function is able to detect both sides of the transition ramp.

In the **barracks** dataset, the capability of the model of detecting second-order edge

Table 3.2: Accuracy in the approximations given by the GS, BCDA and BCDA_c with respect to the *ideal* solution for the considered datasets.

dataset	method	$d_1(\mathbf{s}, \mathbf{s}^*)$	$d_1(\mathbf{z}, \mathbf{z}^*)$	$d_G(\mathbf{u}, \mathbf{u}^*)$
pearl	GS	$3.40e - 03$	$6.45e - 03$	$4.02e - 04$
	BCDA	$4.00e - 03$	$7.59e - 03$	$4.93e - 04$
	BCDA _c	$1.38e - 02$	$1.84e - 02$	$4.78e - 03$
aerial	GS	$9.76e - 03$	$1.27e - 02$	$1.13e - 03$
	BCDA	$9.44e - 03$	$1.22e - 02$	$1.14e - 03$
	BCDA _c	$3.33e - 02$	$3.18e - 02$	$8.35e - 03$
airport	GS	$5.41e - 03$	$7.46e - 03$	$6.59e - 04$
	BCDA	$5.74e - 03$	$7.80e - 03$	$7.43e - 04$
	BCDA _c	$1.86e - 02$	$1.83e - 02$	$6.88e - 03$
barracks	GS	$2.47e - 03$	$1.65e - 02$	$3.57e - 05$
	BCDA	$2.47e - 03$	$1.62e - 02$	$3.63e - 05$
	BCDA _c	$6.48e - 03$	$2.62e - 02$	$1.80e - 04$

boundaries is clear. In the surface models the geometry is real and the structure of many man-made objects is linear (buildings for instance). By looking at the particulars in Figure 3.8 we see that the noise is removed and the edges that define the roof planes are preserved and correctly detected by the \mathbf{z} function. Notice also the substantial difference between the functions \mathbf{s} and \mathbf{z} . This difference is not so evident in the images as variations of gradient of luminance are usually as not as sharp as variations of gradient of height in surface models of urban areas. This means that discontinuities that are purely of second-order are difficult to find in images, but not in DSMs. Lastly, we remark here that in a discrete setting a jump is also a crease. In view of the second-order differential operators used in (3.8) a jump is traced along 1-pixel wide curve in \mathbf{s} , whereas it is traced along a 2-pixel wide curve by \mathbf{z} (cfr. Section 3.3.3).

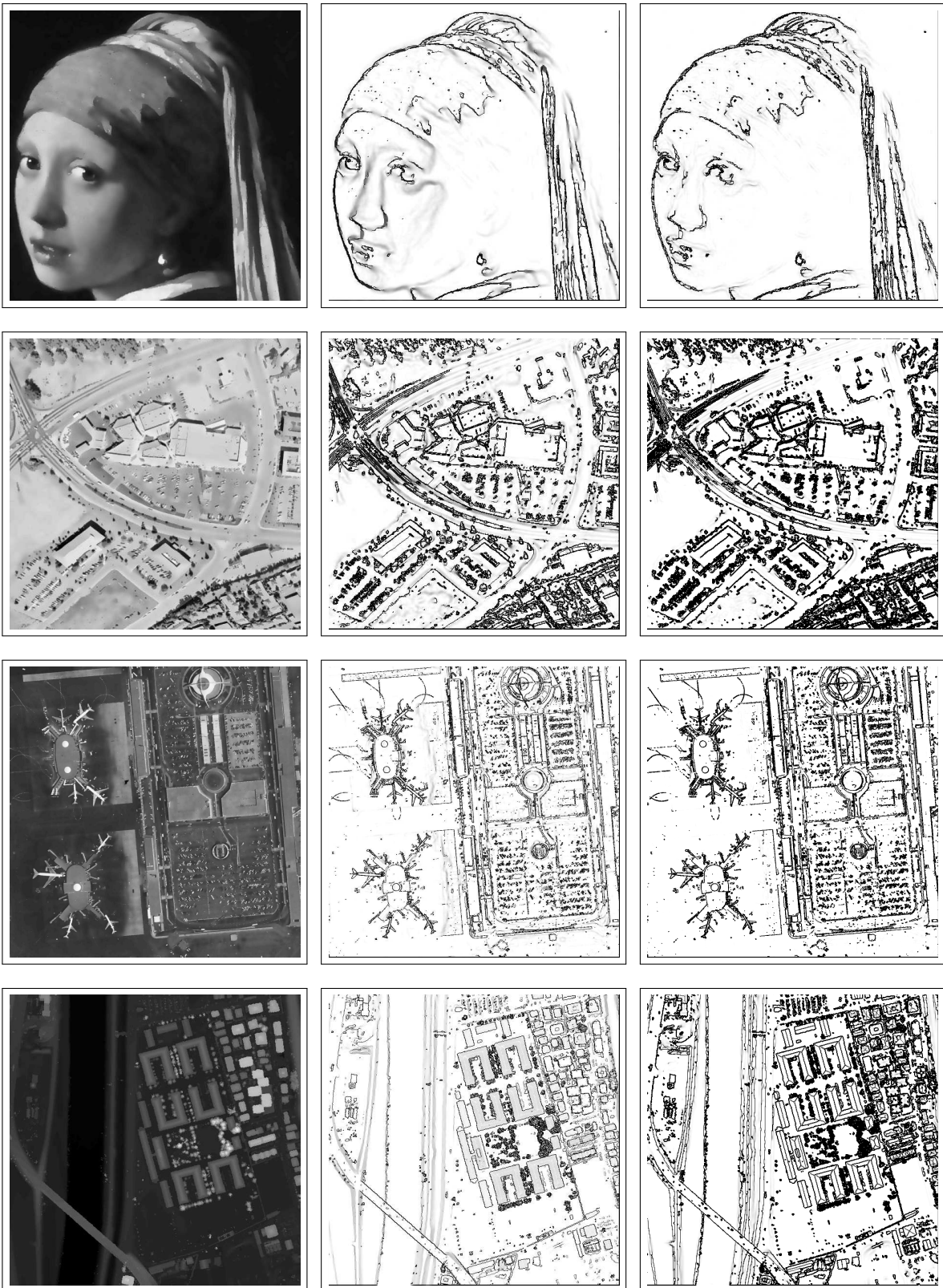
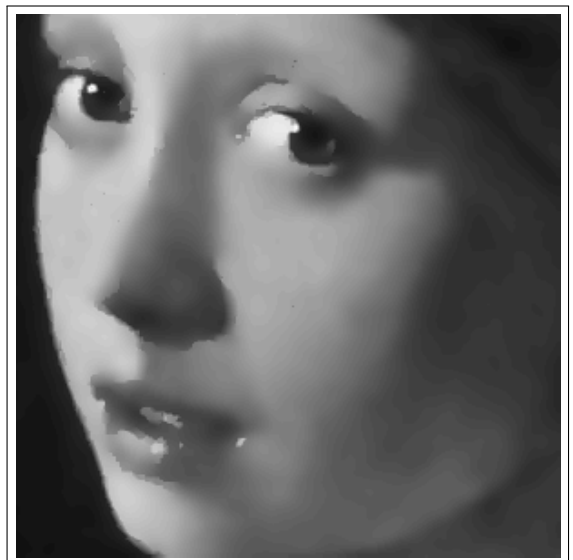


Figure 3.5: Results of the BCDA method for the considered datasets. First column is the smooth approximation \mathbf{u} , second column is the edge-map \mathbf{s} and third column is the edge/crease-map \mathbf{z} .



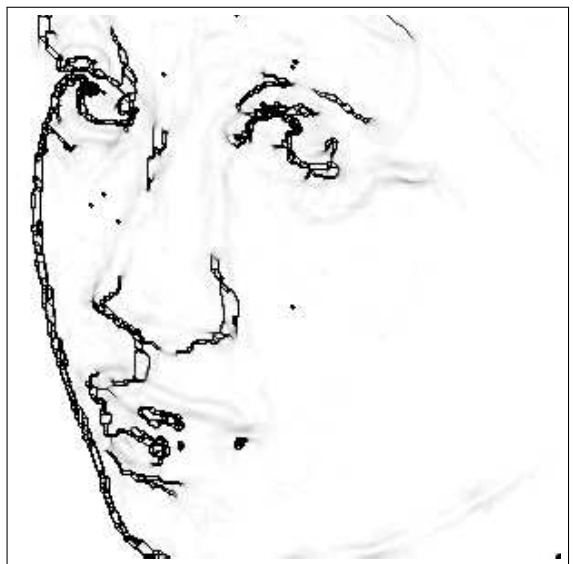
(a) g



(b) u



(c) s



(d) z

Figure 3.6: Particulars of the segmentation for the dataset pearl.

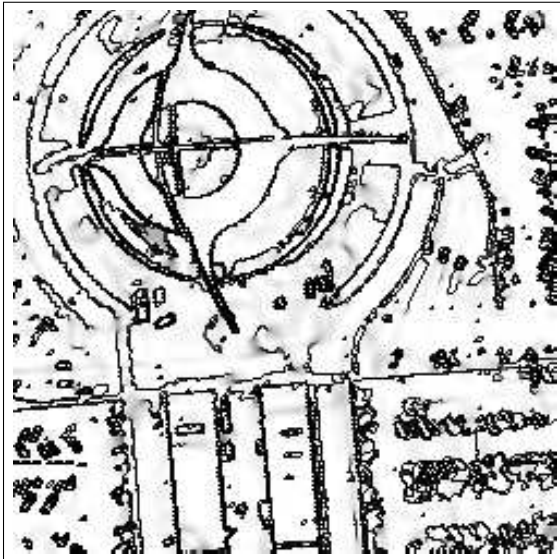
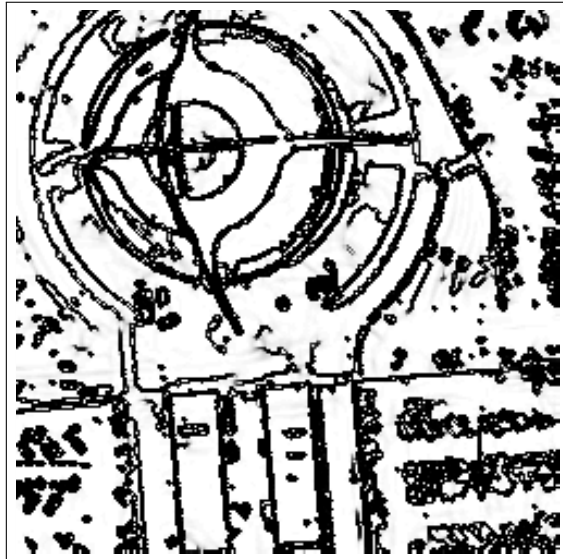
(a) g (b) u (c) s (d) z

Figure 3.7: Particulars of the segmentation for the dataset airport.

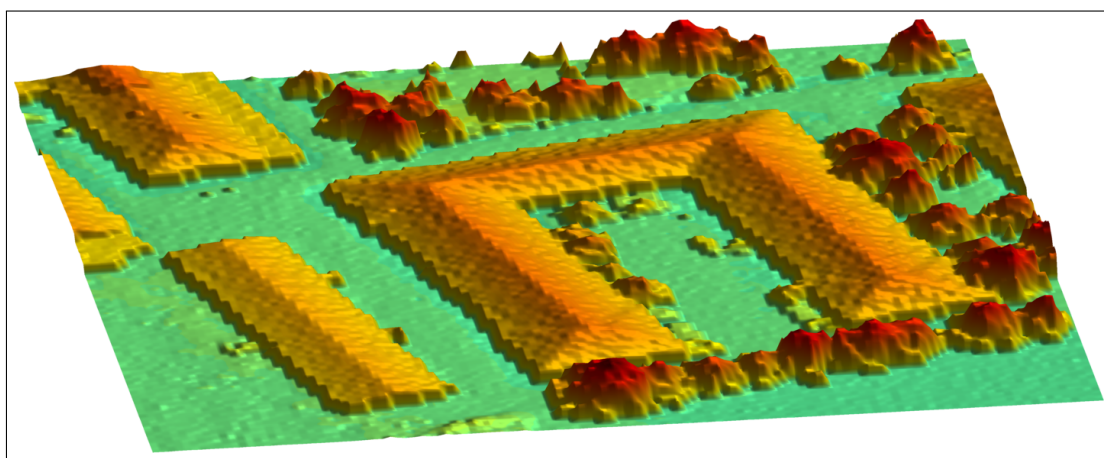
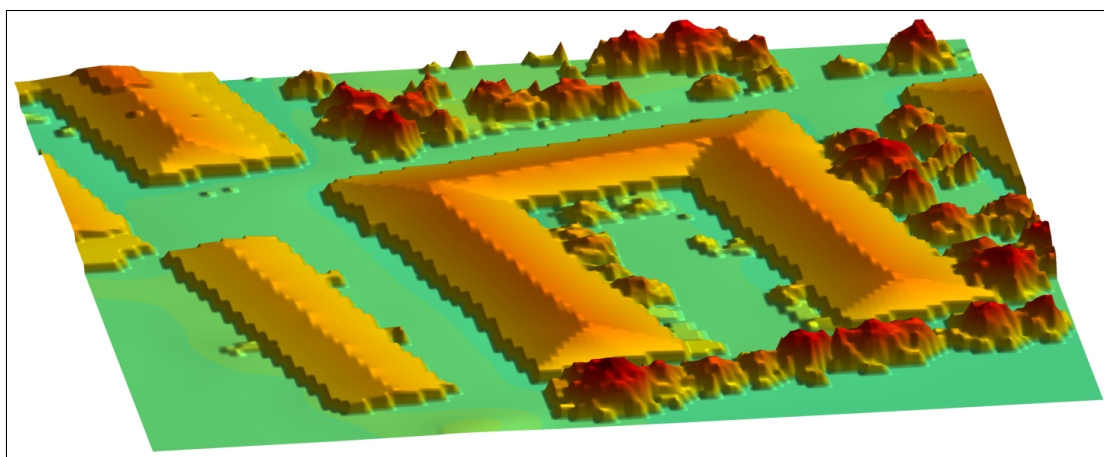
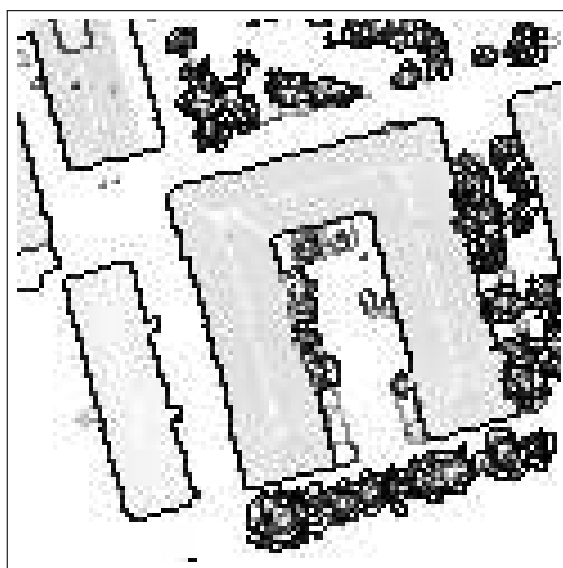
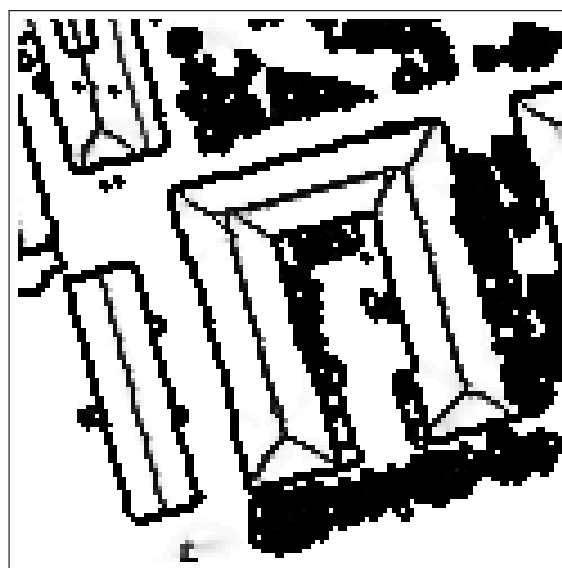
(a) g (b) u (c) s (d) z

Figure 3.8: Particulars of the segmentation for the dataset barracks. Above, there are the 3D renderings of the surface model g and its smooth approximation u . In the bottom, there are the edge map s and the edge+crease map z .

3.3.2 Noise reduction and effect of Block-Preconditioner for the BCDA

The Step 3 in Algorithm 1 attempts to minimize the objective function $F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u})$ with respect to the variable block \mathbf{u} . The performance of this step depends on both the size and the noise level of the image. The inverse of the parameter μ of the objective function represents the scale at which variations in \mathbf{g} are considered as noise. At each outer iteration k , variations that are in order to be smoothed out are signaled by the points where $\mathbf{s}^k, \mathbf{z}^k$ are close to 1. Smoothing takes place when the system involving \mathbf{A}_u is solved and returns \mathbf{u}^k . The computational burden related to this step is determined by the scale of the noise in the image: the greater the variations are related to the noise, the more effort is required to smooth out such variations. Assuming that interesting features (edge-boundaries) and noise are at different scales in the image, the parameter μ must be chosen at an intermediate level in such a way that the former ones are preserved by the model whereas the latter ones are smoothed out. However, from (3.11) one can see that small values of μ reduce the definite positiveness of matrix \mathbf{A}_u . Thus, μ should be optimal, i.e., as greater as possible but not such that noise is preserved.

The first test we present concerns with the relationship between the choice of μ , the noise reduction and the performance of the minimization performed by the proposed BCDA. The basic element of the datasets considered in the following is a synthetic 100×100 grey-scale image presenting geometrical features of first and second order consisting in a truncated pyramid. Images of greater dimensions are obtained by assembling several basic elements of the same type. Test images are corrupted by artificial additive Gaussian noise with 0 mean and varying standard deviation σ (top-left image in Figure 3.9 represents one pyramidal element, in the same row other elements with added noise). Along the experiments, functional parameters except μ are fixed to $\alpha = 2, \beta = 1, \delta = 30, \epsilon = 0.01, t = 1$. After running the algorithm several times, we observed that that for values $\mu < 0.05$ the smoothing also affected the interesting features of the image (pyramid edges), whereas for $\mu > 0.15$ no smoothing at all was observed. Therefore, the results given in the following relate to $0.05 \leq \mu \leq 0.15$. Performance of the BCDA with respect to noise and μ is evaluated in two experiments.

- In the first experiment the minimization is performed on one pyramidal element for different levels of noise $\sigma = 0, 0.5, 1, 2$ ($\sigma = 0$ means that no noise is added) and for $\mu = 0.15$. In Figure 3.9, the input image \mathbf{g} , the difference between \mathbf{g} and its smooth approximation \mathbf{u} , the edge-detection function \mathbf{s} and the edge/crease-detection function \mathbf{z} are showed. Notice from the plots of $\mathbf{g} - \mathbf{u}$ how the noise is uniformly detected and removed by the model in all cases. The detection of first and second order features of the image is sufficiently accurate, and only for $\sigma = 2$

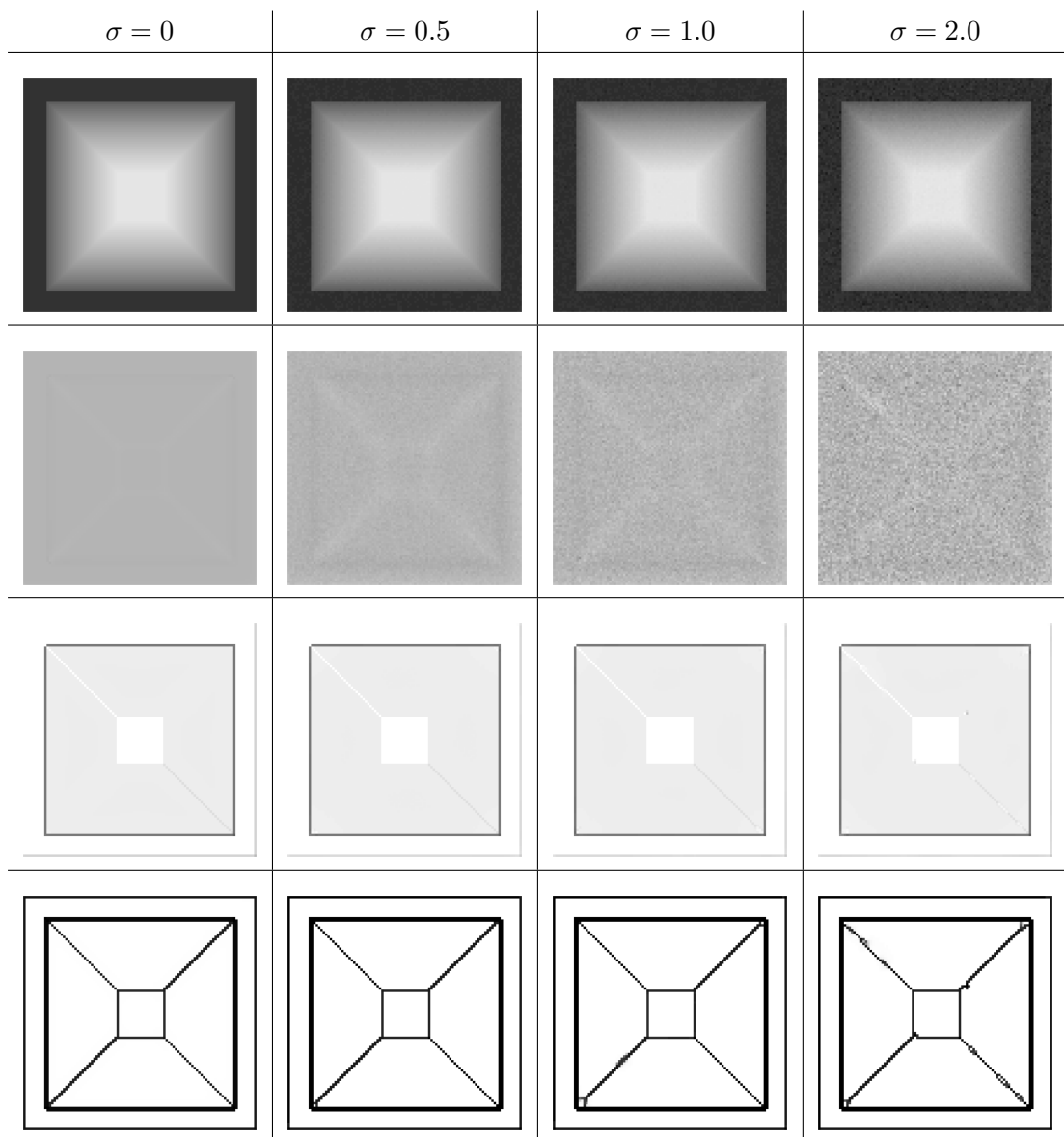


Figure 3.9: Smoothing on synthetic images with different noise levels, image size is 100×100 . Row 1: input noisy images \mathbf{g} . Row 2: difference $\mathbf{g} - \mathbf{u}$. Row 3: edge-detection functions \mathbf{s} . Row 4: edge/crease-detection functions \mathbf{z} .

the scale of the noise slightly affects the detection of pyramid edges. Let us denote \mathbf{g}_σ and \mathbf{u}_σ the input image and its smooth approximation for different values of σ . Quantitatively, the capability of the model of removing the noise is given in terms of the distance between the smooth approximations \mathbf{u}_σ and the original noise-free image \mathbf{g}_0 . We obtained $d_G(\mathbf{g}_0, \mathbf{u}_0) = 2.90e - 03$, $d_G(\mathbf{g}_0, \mathbf{u}_{0.5}) = 2.81e - 03$, $d_G(\mathbf{g}_0, \mathbf{u}_1) = 2.88e - 03$, $d_G(\mathbf{g}_0, \mathbf{u}_2) = 3.13e - 03$, where we used $G = \|\mathbf{g}_0\|_\infty$.

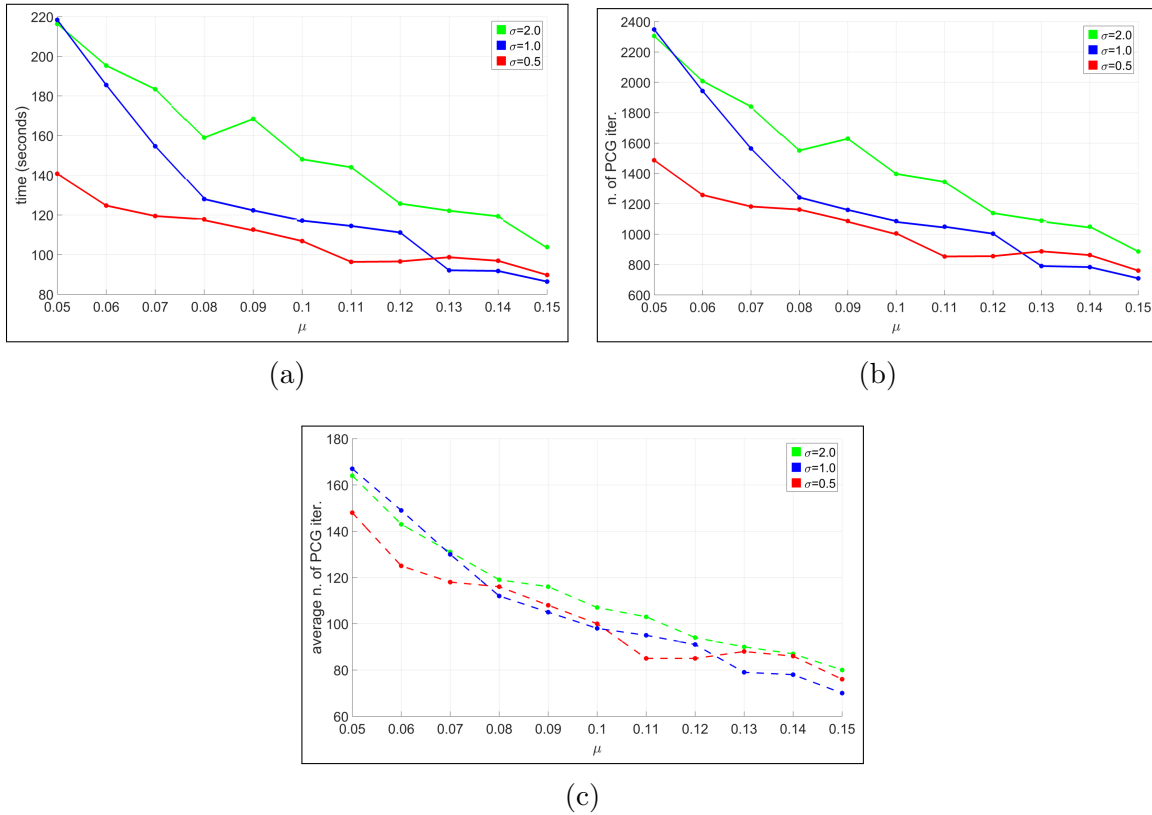


Figure 3.10: Performance details of BCDA by varying μ and for different variances of the noise. Test images are 1000×1000 pixels compositions of the pyramidal elements showed in Figure 3.9. For each value of μ we plot (a) the execution time, (b) the total number, and (c) the average number of iterations of the PCG solver related to \mathbf{u} .

- In the second experiment, the effect of varying μ on the performance of the minimization method is tested with respect to different levels of noise. Test images have size of 1000×1000 pixels. Times of computation and average number of iterations of the PCG related to \mathbf{u} are given in Figure 3.10. As we can see, for a fixed value of μ the time required for smoothing the data increases with σ . Moreover, since the average number of PCG iterations for \mathbf{u} does not significantly change with σ , we conclude that more outer iterations are needed to smooth out the noise. Notice also that, as expected, increasing values of μ resulted in less computational time due to the fact that the definite positiveness of matrix \mathbf{A}_u increases.

As we have seen in the previous experiments, the noise removal task can be very expensive. Moreover, the execution time also depends on the size of the input image. The second test we present aims at showing that the general performance of BCDA can be enhanced if a block-diagonal preconditioner, instead of a point-diagonal one, is used. In

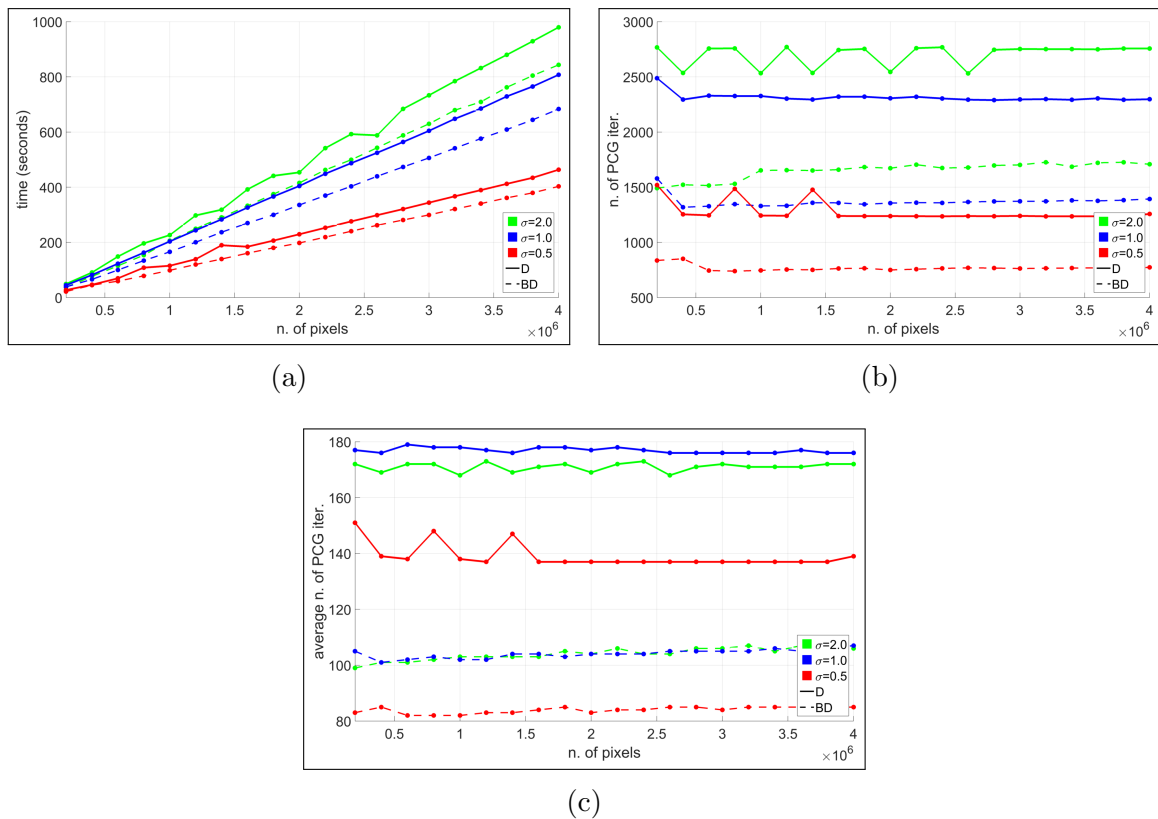


Figure 3.11: Performance of BCDA with Diagonal (D) and Block-Diagonal (BD) preconditioners for the PCG solver related to \mathbf{u} , versus the size of \mathbf{g} . We plot (a) the execution time, (b) the total number, and (c) the average number of iterations of the PCG solver related to \mathbf{u} .

the experiments, the algorithm is run on synthetic images composed by several pyramidal elements corrupted with noise with different variances. Analyzed images have a number of pixels ranging from $2 \cdot 10^5$ to $4 \cdot 10^6$. From Figure 3.11 we see that the execution time linearly increases with the size of the image. Moreover, the use of a block-preconditioner reduces the time of approximately 14% regardless the noise variance. It is also confirmed from the graphs of total and average number of PGC iterations that, for increasing sizes of the image the number of iterations does not significantly change. The computational burden is instead in the time for completing each iteration.

3.3.3 Parameter ϵ , grid resolution and sensitivity of the discontinuity functions

The explicit construction of a recovery sequence in the Γ -convergence proof of [14] allows for having an essential prediction of geometrical properties of the discontinuity functions s_ϵ, z_ϵ approximating the discontinuity sets $S_u, S_{\nabla u}$, for $\epsilon > 0$ (see Figure 3.2 and related discussion). The geometrical behavior of s_ϵ, z_ϵ is theoretically determined by the convergence parameter ϵ and geometrical features of the input image g . However, no analytical expression of the functions in proximity of the discontinuities is given. These facts give rise to some numerical problems that must be taken into account when the functional \mathcal{F}_ϵ is minimized and the numerical approximations of s_ϵ, z_ϵ (given by \mathbf{s}, \mathbf{z} , respectively) are computed. On the one hand, the discrete sampling of the domain Ω_0 with steps t_x, t_y must resolve the tubular neighborhoods of the discontinuity sets, which shrink as $\epsilon \rightarrow 0$. On the other hand, increasing values of $|\nabla g|$ and $|\text{Hg}|$ will force sharper transitions of the discontinuity functions from 1 to 0, whereas decreasing values will result in softer transitions, therefore limiting the capability of detecting edge boundaries.

In order to have a clear understanding of the geometrical behavior of functions \mathbf{s}, \mathbf{z} , in the following the sensitivity of the model with respect to the parameter ϵ and the variations of $|\nabla g|$ and $|\text{Hg}|$ is tested. In the tests the two synthetic images showed in Figure 3.12 are used. The first image contains a uniform jump (discontinuity) of grey value with variable height h . The second image contains a crease (gradient discontinuity) between a flat area and a uniform slope with variable angle θ .

Let us discuss the results obtained on the test image with a jump (Figure 3.12a), firstly by varying parameter ϵ , then by varying h . In the first experiment all the functional and discretization parameters, except ϵ , are fixed to $\delta = 30$, $\alpha = 2$, $\beta = 1$, $\mu = 1$, $t_x = t_y = 1$, the step height is $h = 90$. A sufficiently wide range of behaviors of the discontinuity functions can be depicted by minimizing the functional F_ϵ for values of ϵ ranging from 10^{-5} to 5. In Figures 3.13a and 3.13b the plots of slices of the minimizing functions \mathbf{s} and \mathbf{z} in correspondence of the jump, are given. As we can see, the width of the tubular neighborhoods of the discontinuity sets increases with ϵ . For $\epsilon = 10^{-5}$ the grid is too coarse for detecting the transition phase of both \mathbf{s} and \mathbf{z} , which are uniformly 1. The optimal choice of the Γ -convergence parameter is $\epsilon = 10^{-2}$, as it corresponds to a detection of the jump which is 1 grid-point wide in the case of \mathbf{s} , and two grid-point wide in case of \mathbf{z} (detection is optimal in view of the differential discrete operators described in Section 3.2.1). For greater values of ϵ the detection of the jump is increasingly over-estimated. In particular, notice the difference in the rate at which \mathbf{s} and \mathbf{z} become flat in the neighbors of the jump.

In the second experiment, the same parameters as before are used and ϵ is fixed at the

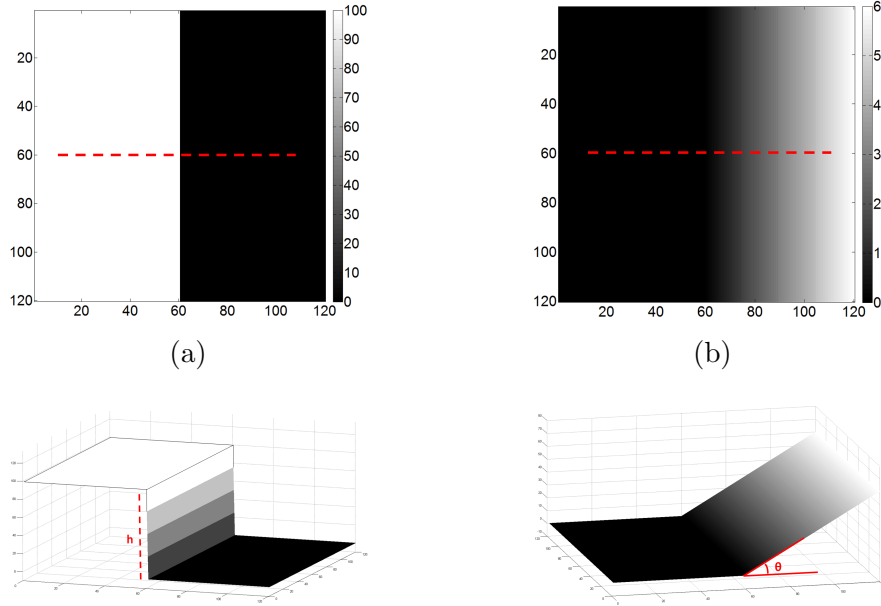


Figure 3.12: Test images. (a) Image with jump of variable height h . (b) Image with a crease of variable slope θ . The slices of functions \mathbf{s}, \mathbf{z} plotted in Figures 3.13 and 3.14 are located in correspondence of the red dashed lines.

optimal value $\epsilon = 0.01$. Slices of discontinuity functions are plotted in Figures 3.13c and 3.13d for values of h ranging from 1 to 90. In the case of $h = 90$ the gradient ∇g is very high, therefore functions \mathbf{s} and \mathbf{z} are forced to inhibit the costly gradient contribution to the energy by taking the 0 value in correspondence of the jump. By decreasing h , the discontinuity functions become less sensitive to the jump and their values approach 1 gradually. In the limit case, $h = 1$, the step is not detected at all. Since here we are studying the behavior of minimizers as ∇g changes, we have to remark that a similar behavior is observed if h is fixed and the discretization parameters t_x and t_y are changed; this corresponds to a re-scaling of the image that only changes derivatives.

Now, we present the results obtained on the test image with a gradient discontinuity (Figure 3.12b), again by varying parameter ϵ , and then by varying the inclination angle θ . Functional parameters are the same as in the previous tests, except that $\delta = 300$ (crease detection is enforced).

In the first experiment the range of values of ϵ is the same as in the previous one, and $\tan(\theta) = 1$ is fixed. In Figures 3.14a and 3.14b the plots of slices of the minimizing functions \mathbf{s} and \mathbf{z} in correspondence of the crease, are given. For $\epsilon > 10^{-2}$ the ramp is over-segmented by \mathbf{s} and \mathbf{z} . In particular, the larger is ϵ , the softer is the transition of \mathbf{z} from 1 to 0. The value $\epsilon = 10^{-2}$ is again optimal as \mathbf{s} is uniformly 1 (no over-

segmentation) and the crease is correctly detected with sharp variation of \mathbf{z} from 1 to 0. Again, for smaller values of ϵ the tubular neighborhood of the discontinuity of the gradient is too thin for being resolved by the grid.

In the second experiment we tested the sensitivity of the discontinuity functions with respect to the slope of the ramp. Parameters are the same as in the previous experiment. Results are plotted in Figures 3.14c and 3.14d for values of $\tan(\theta)$ ranging from 0.2 to 5. The steepest is the ramp, the better is the detection of the gradient discontinuity given by \mathbf{z} . Notice that for $\tan(\theta) = 5$, the function \mathbf{s} tends to over-segment the ramp. This fact can be explained by noticing that for such θ , the difference of grey value from one pixel to an adjacent one in the ramp is in the order of a detectable jump. Since $t_x = 1$, the jump in this case is 5. Compare the value of function \mathbf{s} with the value of \mathbf{s} in Figure 3.13c corresponding to a jump of height 6: they are both close to 0.8. As a last remark we point out that, as in the previous example, a change in the step widths t_x, t_y has the only obvious effect of re-scaling the derivatives in the image, thus the behavior of \mathbf{s} and \mathbf{z} is similar to the one presented here with variable θ .

These tests described the relationship between the geometry of the discontinuity functions returned by the minimization of F_ϵ , the geometry of the image g and the Γ -convergence parameter ϵ . By increasing the parameter ϵ the width of the tubular neighborhood of the discontinuity sets increases, until reaching a full over-segmentation of the image. Conversely, for small values of ϵ , the tubular neighborhood is narrower and narrower, until it is not resolved by the grid. A correct value of the Γ -convergence parameter should be chosen accordingly. The parameters that affect gradient and Hessian are the height of the jumps, crease angles, and also the grid steps t_x, t_y . By varying such quantities the width of the tubular neighborhoods does not change. However, functions \mathbf{s} and \mathbf{z} are sensitive to their variations and they approach 0 and 1 at different rates. We conclude that the width of the tubular neighborhood of the discontinuity sets depends only on ϵ , whereas other geometrical properties of the image affect the rate at which these functions approach 0 and 1.

3.4 Conclusions

In this chapter we proposed an efficient block-coordinate descent method for the numerical minimization of a variational approximation of the Blake-Zisserman functional given by Ambrosio, Faina and March. The Blake-Zisserman variational model for segmentation is a second-order model based on free discontinuities and free gradient discontinuities that is able to both detect first and second order edge-boundaries in images and produce a piecewise linear approximation of the input image. Therefore, the model presents several

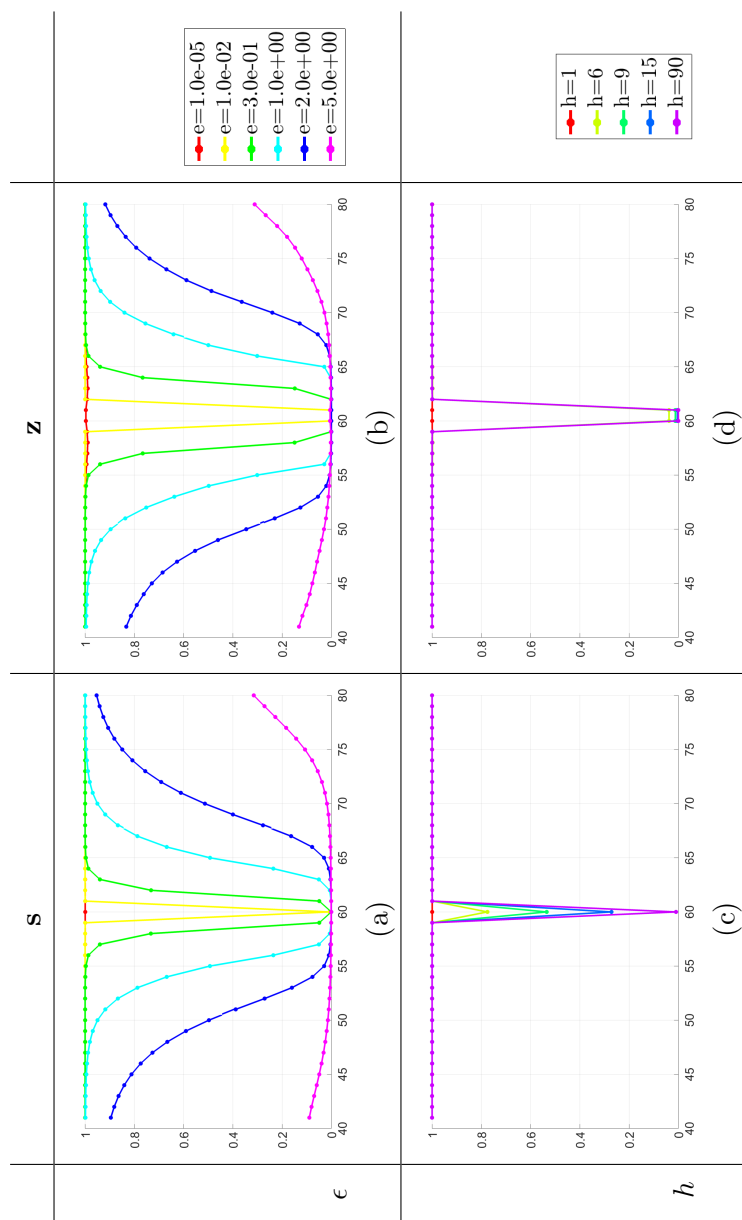


Figure 3.13: Slices of functions s and z obtained by minimizing F_ϵ in the case of test image with a jump (Figure 3.12a) for different values of ϵ (a,b) and h (c,d).

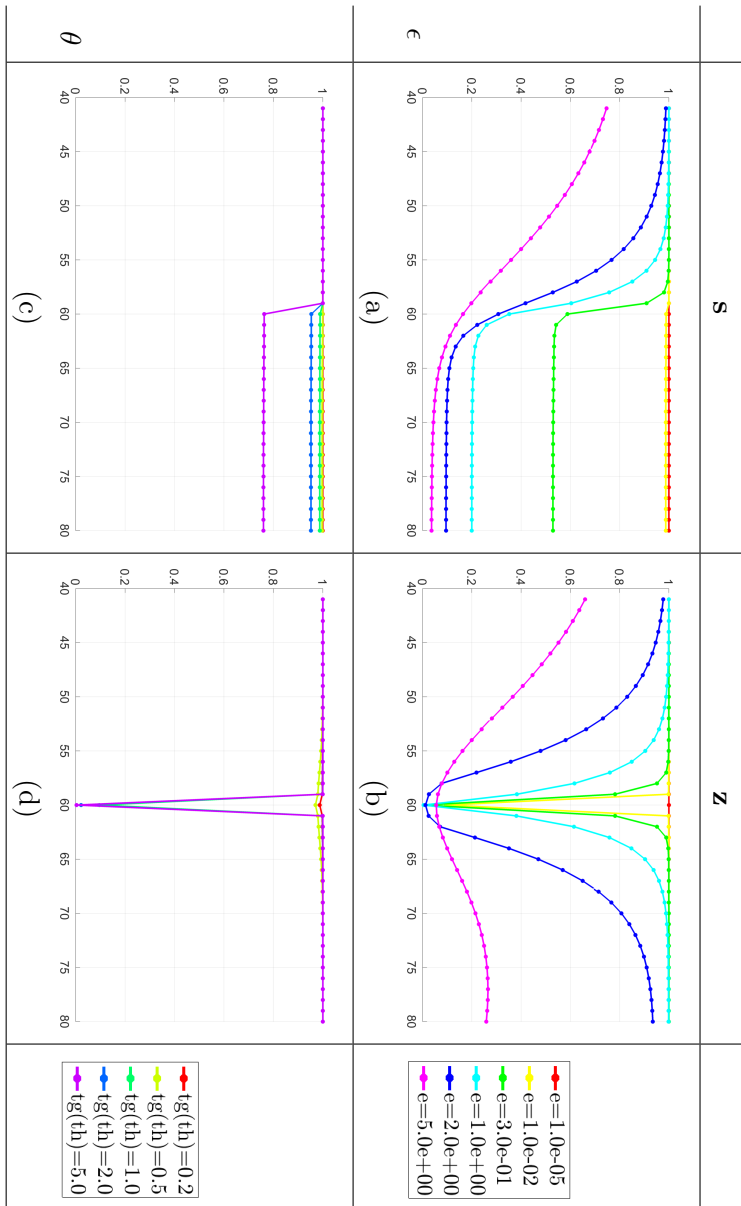


Figure 3.14: Slices of functions s and z obtained by minimizing F_ϵ in the case of test image with a gradient discontinuity (Figure 3.12b) for different values of ϵ (a,b) and $\tan(\theta)$ (c,d).

advantages if compared with other methods for image approximation. The proposed algorithm exploits the structure of the minimization problem, allowing the segmentation of large images, with a satisfying performance in terms of accuracy and computational time. In particular, it outperforms a standard Gauss-Seidel of 75%. Moreover, the use of a block-diagonal preconditioner always increases performance of about 14%. Numerical experiments presented in the chapter involve very different types of datasets. We consider also the segmentation of Digital Surface Models (DSMs), showing that the application of the Blake-Zisserman functional allows to reconstruct, and locate the boundary, of planar objects. In the experimental part we also focused the attention on the geometrical parameter of the functional related to noise reduction and the behaviour of the discontinuity functions with respect to the Γ -convergence parameter ϵ and geometrical properties of the input image. Preliminary tests showed that the proposed method can be combined with a domain decomposition technique, enabling the segmentation of huge images by a tiling strategy. Further studies will concern the deepening of this attractive feature in the framework of the multiprocessing computing with the aim to design a code for modern parallel architectures.

Acknowledgments

This research is partially supported by the project *Learning meets time: A new computational approach for learning in dynamic systems*, contract RBFR12M3A. The project is funded by the Italian Ministry of Education, University and Research *Futuro in Ricerca 2012* program.

Chapter 4

Variational approximation of vector-valued images and curves

Variational models are known to work well for addressing image restoration/regularization problems. However, most of the methods proposed in literature are defined for scalar inputs and are used on multiband images (such as RGB or multispectral imagery) by the composition of a simple bandwise processing. This involves suboptimal results and may introduce artifacts. Only in a few cases variational models are extended to the case of vector-valued inputs. However, the known implementations are restricted to 1st-order models, while 2nd-order models are never considered. Thus, typical problems of 1st-order models such as the staircasing effect cannot be overtaken. In this chapter we use a 2nd-order functional model to function approximation with free discontinuities given by Blake-Zisserman and we propose an efficient minimization algorithm in the case of vector-valued inputs. In the BZ model, the Hessian of the solution is penalized outside a set of finite length, therefore the solution is forced to be piecewise linear. Moreover, the model allows the formation of free discontinuities and free gradient discontinuities. The proposed algorithm is applied to difficult color image restoration/regularization problems and to piecewise linear approximation of curves in space.

4.1 Introduction

Typical approaches to image restoration/regularization assume the image g recorded by an optical sensor to be a noisy variation of a regular signal u . Mathematical methods to image approximation aim at recovering such u by either solving an associated Partial Differential Equation (PDE) or by minimizing a specific variational energy, both depending on g . In particular, edge-driven methods recognize the portions of the image contoured

by sharp variations of intensity (discontinuities) and associate them to different objects constituting the image subject. Therefore, meaningful approximations are obtained from edge-based methods when they are able to: (1) discriminate between intensity variations due to noise and those due to the presence of object edges, and, (2) return regular approximations where smoothing only reduces noise contaminations without affecting relevant edges. PDE methods are mainly based on diffusion equations. In the Total Variation (TV) [46] model diffusion is fully isotropic, therefore smoothing affects all image features indiscriminately. This drawback is avoided in anisotropic diffusion (AD) [8] models, where diffusion is inhibited according to local features of the image. In general, PDE methods do not allow the solutions to have free discontinuities and the physical meaning of the equations parameters are not fully understood [10, 11]. As a consequence, the progressive modification of PDEs to obtain more meaningful solutions moved towards variational representations. Indeed, in most of the cases PDEs can be seen as flows generated by variational energies.

To address these issues, Mumford and Shah [3] proposed a flexible variational model to image approximation based on *free discontinuities*, see Section 2 for details. From a numerical viewpoint, the explicit computation of a solution is a difficult problem. Among many strategies that have been proposed in literature to solve this problem (see for instance [22–24]), we recall the Ambrosio-Tortorelli (AT) elliptic approximation via Γ -convergence [17], which is numerically tractable [20]. Being a 1st-order model, the MS has some drawbacks. In particular, the staircasing effect [26, 47] is of major relevance, as it often limits the applicability of this model in practical situations. Briefly, this phenomenon can be explained as follows. The minimization of the gradient energy forces the solution to have a piecewise constant behaviour. Therefore, steep gradients are approximated by step-wise functions with many fictitious discontinuities, as the solution is not allowed to have 1st-order variations of high magnitude. This problem can be solved by replacing the gradient term of the energy by a 2nd-order operator. Indeed, this is the solution introduced by Blake and Zisserman [5], who proposed to penalize the Hessian (instead of the gradient) and the *size* of K_0, K_1 , the discontinuity and the gradient discontinuity sets of u , respectively. The BZ approximation can be found by minimizing

$$\begin{aligned} \mathcal{BZ}(K_0, K_1, u) = & \int_{\Omega_0 \setminus (K_0 \cup K_1)} \left(|Hu|^2 + \mu|u - g|^2 \right) dx \\ & + \alpha \mathcal{H}^1(K_0) + \beta \mathcal{H}^1(K_1) \end{aligned} \quad (4.1)$$

where μ, α, β are positive parameters. Hessian penalization allows the solution to have 1st-order variations outside $K_0 \cup K_1$, yielding to a piecewise linear approximation of the input image. A recent survey presents a summary and future perspectives about the study

of the Blake-Zisserman (BZ) variational model for segmentation, including theoretical results for existence and regularity of solutions [33]. To address numerical minimization, elliptic approximations of the functional exploiting the AT technique (used in the MS case) were given by Bellettini and Coscia [40] and Ambrosio, Faina and March (AFM) [14]. The first numerical implementations are given in dimension one [19] to piecewise linear approximation of signals and in dimension two [14] to segmentation of stereo images. Recently, the problem of numerically minimizing the AFM approximation of the BZ functional on large images has been addressed in [4], where the objective functional is written in a compact matricial form and optimization is performed by means of a special version of the block-coordinate descent algorithm (BCDA) [43] that exploits the partial convexity of the functional.

The PDE approaches to image approximation mentioned before have been successfully generalized to the case of vector-valued inputs. In [48], a vector-valued version of image restoration based on TV norm has been proposed for color images. A general framework for AD to vector-valued image restoration/enhancement has been proposed in [49], which is applicable to both color images and to other vector-valued image representations (e.g., stacks of image features like texture, motion, etc.). Curvature-preserving tensor-driven PDEs have been also designed to enhance regularity of edge boundaries [50]. Regarding variational methods, large attention has been devoted to the study of the MS model in the vector-valued case. By generalizing the AT approach, fundamental theoretical results have been proven for the MS problem in the vector-valued case [51,52], allowing also for explicit computations [47,53]. Other approaches have been also developed for the vector-valued MS by exploiting convex representation [54], by extending the active contour algorithm [55] or by combinatorial optimization [56]. In [57] the local AT approximation of the MS is extended to non-local formulation accounting for texture information. Numerical experiments in the cited works confirmed for the improved capability of approximation of variational models w.r.t. to PDE approaches such as TV and AD.

On the counterpart, the current limitation of the literature is the absence of methods for addressing the approximation of vector-valued inputs based on 2nd-order models. In particular, the minimization of the BZ functional for vector-valued functions is not considered at all. In 1st-order models such as MS, the gradient penalization forces the solution to be piecewise constant. If this has useful implications to segmentation purposes, it makes image regularization/restoration unfeasible as a locally flat approximation is generally too coarse. In this chapter, we propose a numerical approach to solve the image approximation problem based on the 2nd-order Blake-Zisserman functional for vector-valued functions. From a theoretical viewpoint, we prove that the discrete version of the objective functional

involving tensor differential operators retains partial convexity with respect to the new variable blocks associated to co-domain dimensions. To demonstrate its effectiveness, the proposed method is applied to difficult color image denoising/restoration problems and to the recovery of polygonal boundaries from discrete noisy sampling.

The plan for this chapter is as follows. In Section 4.2, we firstly introduce the elliptical AFM approximation of the BZ functional and then we propose two numerical approaches to address minimization in the case of vector-valued images and curves. Numerical experiments are presented in the next two sections. In Section 4.3, the proposed algorithm to piecewise linear approximation of vector-valued images is applied to denoising/restoration of color images and compared with the well-known MS model. In Section 4.4, the proposed method is applied to the recovery of polygonal shapes from discrete noisy sampling. In Section 4.5 we draw the conclusions.

4.2 The Blake-Zisserman model for the approximation of vector-valued images and curves

The numerical handling of the 2nd-order variational model to segmentation proposed by Blake and Zisserman is unfeasible in its original strong formulation as in (4.1) [5]. As in the MS case, the strong formulation does not allow to prove existence of solutions: because of the set unknowns K_0, K_1 , the functional lacks in lower semicontinuity. Therefore, the functional has been rewritten in the weaker space of Generalized Special Functions of Bounded Variation [27]

$$\begin{aligned} \mathcal{F}(u) = & \int_{\Omega_0} \left(\mu |u - g|^2 + |Hu|^2 \right) dx \\ & + (\alpha - \beta) \mathcal{H}^1(S_u) + \beta \mathcal{H}^1(S_{\nabla u} \cup S_u) \end{aligned} \quad (4.2)$$

where it loses its explicit dependency on the discontinuity sets, which can be regarded as geometrical properties of the only variable function u being $S_u, S_{\nabla u}$ the discontinuity and gradient discontinuity set of u , respectively. Also in the weak form the numerical minimization is still a hard issue. For this reason, elliptic approximations of the functional have been proven via Γ -convergence by exploiting the seminal idea of Ambrosio-Tortorelli for the approximation of the MS functional. This is done in dimension one by Bellettini and Coscia [40], and in dimension two (with partial results in any finite dimension) by Ambrosio, Faina and March [14].

Let $\Omega_0 \subset \mathbb{R}^k$ be an open set, $k = 1, 2$ and $g \in L^\infty(\Omega_0)$ the function to be approximated (either a signal or an image). The authors of [14] and [40] have introduced two

auxiliary functions $s, z : \Omega_0 \rightarrow [0, 1]$ aimed at approximating the indicator functions of the discontinuity sets K_0, K_1 and proposed a Γ -convergence approximation of the weak functional via the family of uniformly elliptic functionals

$$\begin{aligned} \mathcal{F}_\epsilon(s, z, u) &= \delta \int_{\Omega_0} z^2 |Hu|^2 dx + \xi_\epsilon \int_{\Omega_0} (s^2 + o_\epsilon) |\nabla u|^2 dx \\ &+ (\alpha - \beta) \mathcal{AT}_\epsilon(s) + \beta \mathcal{AT}_\epsilon(z) + \mu \int_{\Omega_0} |u - g|^2 dx, \end{aligned} \quad (4.3)$$

where s, z, u are in proper Sobolev spaces and \mathcal{AT} is the Ambrosio-Tortorelli component

$$\mathcal{AT}_\epsilon(v) = \int_{\Omega_0} \epsilon |\nabla v|^2 + \frac{1}{4\epsilon} (v - 1)^2 dx. \quad (4.4)$$

Here ϵ is the convergence continuous parameter, ξ_ϵ, o_ϵ are infinitesimals and the convergence is intended for $\epsilon \rightarrow 0$. For each $\epsilon > 0$ the functional \mathcal{F}_ϵ admits a minimizing triplet $(s_\epsilon, z_\epsilon, u_\epsilon)$. Γ -convergence properties ensure that the sequence of minimizers $\{(s_\epsilon, z_\epsilon, u_\epsilon)\}_{\epsilon \rightarrow 0}$ strongly converges to a minimizer of the weak functional (4.2). Fixed $\epsilon > 0$, the geometrical behaviour of a minimizing triplet $(s_\epsilon, z_\epsilon, u_\epsilon)$ is as follows. Due to the presence of the distance term $|u - g|^2$ the function u_ϵ is forced to be close to the input and smoothing constraints are given by the integral terms containing $|\nabla u|^2$ and $|Hu|^2$. For $0 < \epsilon \ll 1$ we have $1/4\epsilon \gg 1$, thus s_ϵ and z_ϵ must be 1 almost everywhere. Transitions from 1 to 0 are only energetically convenient to suppress high values of $|\nabla u|^2$ and $|Hu|^2$. Functional parameters $\delta, \mu, \alpha, \beta$ regulate the penalization of each term individually.

In the following, the AFM approximation of the BZ functional is considered in the general case of vector valued inputs, both for $k = 2$ (the case of vector-valued images) and for $k = 1$ (the case of curves in general space). The functionals are then discretized, written in matricial compact formulation and an efficient numerical algorithm to minimization is proposed.

4.2.1 Approximation of vector-valued images

Let $\Omega_0 \subset \mathbb{R}^2$ be a rectangular domain and $g : \Omega_0 \rightarrow \mathbb{R}^B$, with $B \geq 1$, a vector-valued image (such as RGB or multi-spectral image). Our aim is to derive a discretization and minimization approach to find $u : \Omega_0 \rightarrow \mathbb{R}^B$ and $s, z : \Omega_0 \rightarrow [0, 1]$ that minimize the functional (4.3). Of course, differential operators appearing in the functional must be intended as for vector valued functions. Thus, for a (sufficiently) differentiable function $v : x \mapsto (v_1(x), \dots, v_B(x))$, the symbols ∇v and Hv refer to the vector-valued gradient (Jacobian matrix) and Hessian tensor of the function v , respectively. More specifically,

$$\begin{aligned} [\nabla v]_{bk} &= \partial_k v_b \\ [Hv]_{bkh} &= \partial_{kh} v_b \end{aligned} \quad (4.5)$$

where $k, h = 1, 2$ represent derivative order and $b = 1, \dots, B$ is the coordinate component of the variable v . The squared Euclidean norm of ∇v and Hv is the sum of each squared tensor element

$$\begin{aligned} |\nabla v|^2 &= \sum_{b=1}^B \sum_{k=1}^2 (\partial_k v_b)^2, \\ |Hv|^2 &= \sum_{b=1}^B \sum_{k,h=1}^2 (\partial_{kh} v_b)^2. \end{aligned} \quad (4.6)$$

Discretization

We use the same notation as in Section 3.2.1. We can obtain a convenient discrete version of the objective functional by a 2-D composite rectangular rule. Indeed, by virtue of expressions (4.5), the decomposition of the vector-valued variables $\mathbf{g} = (\mathbf{g}_1, \dots, \mathbf{g}_B)$ and $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_B)$ can be conveniently split over image bands, so that we can write the discrete functional generalizing (4.3) to the vector-valued case as

$$\begin{aligned} F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u}) &= \\ &\sum_{b=1}^B \left\{ \delta \mathbf{u}_b^T [\mathbf{D}_{xx}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xx} + \mathbf{D}_{yy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{yy} + 2\mathbf{D}_{xy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xy}] \mathbf{u}_b \right. \\ &\quad \left. + \xi_\epsilon \mathbf{u}_b^T [\mathbf{D}_x^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_x + \mathbf{D}_y^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_y] \mathbf{u}_b + \mu (\mathbf{u}_b - \mathbf{g})^T (\mathbf{u}_b - \mathbf{g}) \right\} \\ &\quad + (\alpha - \beta) \left[\epsilon \mathbf{s}^T (\mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y) \mathbf{s} + \frac{1}{4\epsilon} (\mathbf{s} - \mathbf{e})^T (\mathbf{s} - \mathbf{e}) \right] \\ &\quad + \beta \left[\epsilon \mathbf{z}^T (\mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y) \mathbf{z} + \frac{1}{4\epsilon} (\mathbf{z} - \mathbf{e})^T (\mathbf{z} - \mathbf{e}) \right]. \end{aligned} \quad (4.7)$$

The functional presents an evident partially quadratic structure as it can be written in the following way

$$\begin{aligned} F_\epsilon(\mathbf{s}, \mathbf{z}, \mathbf{u}) &= \\ &= \frac{1}{2} (\mathbf{s}^T \ \mathbf{z}^T) \begin{pmatrix} \mathbf{A}_s & 0 \\ 0 & \mathbf{A}_z \end{pmatrix} \begin{pmatrix} \mathbf{s} \\ \mathbf{z} \end{pmatrix} - (\mathbf{s}^T \ \mathbf{z}^T) \begin{pmatrix} \mathbf{b}_s \\ \mathbf{b}_z \end{pmatrix} \\ &= \frac{1}{2} (\mathbf{u}_1^T, \dots, \mathbf{u}_B^T) \begin{pmatrix} \mathbf{A}_u & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{A}_u \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_B \end{pmatrix} \\ &\quad - (\mathbf{u}_1^T, \dots, \mathbf{u}_B^T) \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_B \end{pmatrix} \end{aligned} \quad (4.8)$$

where $\mathbf{A}_s = \mathbf{A}_s(\mathbf{u})$, $\mathbf{A}_z = \mathbf{A}_z(\mathbf{u})$, $\mathbf{A}_u = \mathbf{A}_u(\mathbf{s}, \mathbf{z})$ and $\mathbf{b}_s, \mathbf{b}_z, \mathbf{b}_b$ for $b = 1, \dots, B$, are given by

$$\begin{aligned}
\mathbf{A}_s &= 2\xi_\epsilon \sum_{b=1}^B \mathbf{R}_{|\nabla \mathbf{u}_b|^2} + 2\epsilon(\alpha - \beta)(\mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y) + \frac{\alpha - \beta}{2\epsilon} \mathbf{I} \\
\mathbf{b}_s &= \frac{\alpha - \beta}{2\epsilon} \mathbf{e} \\
\mathbf{A}_z &= 2\delta \sum_{b=1}^B \mathbf{R}_{|\mathbf{H}\mathbf{u}_b|^2} + 2\epsilon\beta(\mathbf{D}_x^T \mathbf{D}_x + \mathbf{D}_y^T \mathbf{D}_y) + \frac{\beta}{2\epsilon} \mathbf{I} \\
\mathbf{b}_z &= \frac{\beta}{2\epsilon} \mathbf{e} \\
\mathbf{A}_u &= 2\delta(\mathbf{D}_{xx}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xx} + \mathbf{D}_{yy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{yy} + 2\mathbf{D}_{xy}^T \mathbf{R}_{\mathbf{z}^2} \mathbf{D}_{xy}) + \\
&\quad 2\xi_\epsilon(\mathbf{D}_x^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_x + \mathbf{D}_y^T \mathbf{R}_{\mathbf{s}^2} \mathbf{D}_y) + 2\mu \mathbf{I} \\
\mathbf{b}_b &= 2\mu \mathbf{g}_b
\end{aligned} \tag{4.9}$$

It is worth mentioning here two facts. First, it is well recognized that, when dealing with vector-valued images the composition of separate processing on image components (i.e., bands) introduces artifacts [58]. Due to the tensor nature of the differential operators used in our model, we can avoid this phenomenon (see also [59, 60]). In fact, we have that matrices $\mathbf{A}_s, \mathbf{A}_z$ incorporate information from all image bands as Hessian and gradient norms from all bands are summed up. Thus, functions \mathbf{s}, \mathbf{z} are able to detect discontinuity and gradient discontinuity points gathering information from all bands of the input image. Second, the discrete functional retains quadratic structure with respect to each variable block \mathbf{u}_b , for $b = 1, \dots, B$. Moreover, each partially quadratic slice of the functional depends on the same matrix \mathbf{A}_u and only the constant terms \mathbf{b}_b vary among them. This fact has some relevance when numerical minimization is performed, as eigenvalue analysis to determine convergence parameters can be performed only once for each outer iteration.

Minimization strategy

The partially quadratic structure expressed in (4.8) allows us to address the functional minimization by following a Gauss-Seidel (GS) approach

$$\begin{cases} \mathbf{s}^{k+1} = \arg \min_{\mathbf{s}} F_\epsilon(\mathbf{s}, \mathbf{z}^k, \mathbf{u}^k) \\ \mathbf{z}^{k+1} = \arg \min_{\mathbf{z}} F_\epsilon(\mathbf{s}^{k+1}, \mathbf{z}, \mathbf{u}^k) \\ \mathbf{u}_b^{k+1} = \arg \min_{\mathbf{u}_b} F_\epsilon(\mathbf{s}^{k+1}, \mathbf{z}^{k+1}, \dots, \mathbf{u}_b, \dots) \end{cases} \tag{4.10}$$

where $b = 1, \dots, B$. Indeed, partial descent can be implemented along each variable block with respect to which the functional is quadratic. In order to enhance computing performance, an inexact approach based on Block-Coordinate Descent Algorithm (BCDA)

Algorithm 2 BCDA

Input: $\mathbf{s}^0, \mathbf{z}^0, \mathbf{u}^0, \gamma_s = \gamma_z = 1, \gamma_u = 1.5$;**Step 1:** $k = 0$;**Step 2:** Inexact minimization with respect to \mathbf{s} and \mathbf{z} :

- compute the search directions \mathbf{d}_s^k and \mathbf{d}_z^k ;
- compute $\alpha_s^k = \gamma_s \frac{-(\mathbf{A}_s^k \mathbf{s}^k - \mathbf{b}_s)^T \mathbf{d}_s^k}{\mathbf{d}_s^{kT} \mathbf{A}_s^k \mathbf{d}_s^k}$;
- update $\mathbf{s}^{k+1} = \mathbf{s}^k + \alpha_s^k \mathbf{d}_s^k$;
- compute $\alpha_z^k = \gamma_z \frac{-(\mathbf{A}_z^k \mathbf{z}^k - \mathbf{b}_z)^T \mathbf{d}_z^k}{\mathbf{d}_z^{kT} \mathbf{A}_z^k \mathbf{d}_z^k}$;
- update $\mathbf{z}^{k+1} = \mathbf{z}^k + \alpha_z^k \mathbf{d}_z^k$.

Step 3: Inexact minimization with respect to $\mathbf{u}_1, \dots, \mathbf{u}_B$. For each $b = 1, \dots, B$:

- compute the search direction $\mathbf{d}_{u_b}^k$;
- compute $\alpha_{u_b}^k = \gamma_u \frac{-(\mathbf{A}_u^k \mathbf{u}_b^k - \mathbf{b}_b)^T \mathbf{d}_{u_b}^k}{\mathbf{d}_{u_b}^{kT} \mathbf{A}_u^k \mathbf{d}_{u_b}^k}$;
- update $\mathbf{u}_b^{k+1} = \mathbf{u}_b^k + \alpha_{u_b}^k \mathbf{d}_{u_b}^k$.

Step 4: Set $k = k + 1$ and go to Step 2, until convergence;

[43] can be used with a modification accounting for the separable B problems involving the \mathbf{u} variable. The proposed modified scheme is outlined in Algorithm 2.

In order to find suitable gradient related search directions \mathbf{d}_s^k , \mathbf{d}_z^k and $\mathbf{d}_{u_b}^k$, a few iterations of a PCG solver can be applied to the linear systems $\mathbf{A}_s^k \mathbf{d}_s = \mathbf{b}_s - \mathbf{A}_s^k \mathbf{s}^k$, $\mathbf{A}_z^k \mathbf{d}_z = \mathbf{b}_z - \mathbf{A}_z^k \mathbf{z}^k$ and $\mathbf{A}_u^k \mathbf{d}_{u_b} = \mathbf{b}_b - \mathbf{A}_u^k \mathbf{u}_b^k$, where $b = 1, \dots, B$. The inexact solution of these systems can be stopped according to tolerance values that guarantee the convergence of the overall algorithm to a stationary point of the objective energy (4.7). The calculation of such tolerance values can be easily done by following the approach proposed in [4] (see also Section 3.2.2), which is based on bound estimates of eigenvalues of matrices $\mathbf{A}_s, \mathbf{A}_z, \mathbf{A}_u$. As previously mentioned, all B problems related to the \mathbf{u} variable depend on the same matrix \mathbf{A}_u , thus one single tolerance value can be used to solve all B partial minimization steps based on PCG in the \mathbf{u} variable. Moreover, being each one of these B sub-problems independent from the others, the computational burden of Step 3 in Algorithm 2 can be split over multiple cores (if available) in a parallel way. Being the global energy non-convex, the initialization step is crucial as it has strong impact on the significance of the final results. A valid strategy for initialization is presented in

Section 3.2.2 (see also [4, 14]).

4.2.2 Approximation of vector-valued curves

The capability of the BZ model to recover piecewise linear approximation of data can be exploited also to approximate curves in N -space. As possible interesting applications we mention here: (1) curve (or signal) rectification, and, (2) recovering of polygonal shapes from noisy sampling. Theoretical results and a first implementation about the Blake-Zisserman model in dimension one have only been given for scalar functions (i.e., signals) in [19, 40]. In this section we propose a more general framework for vector-valued curves and an efficient numerical algorithm that exploits the results in Section 4.2.1. Therefore, we consider here vector fields where domain dimension is one, that is functions of the type $g : \Pi \rightarrow \mathbb{R}^N$, with $\Pi \subset \mathbb{R}$ a closed connected interval on the real line and $N > 0$ integer. Let us utilize the more usual notation for derivatives in dimension one. Given a (sufficiently) derivable function $v : \Pi \rightarrow \mathbb{R}^N$, $v : t \mapsto (v_1(t), \dots, v_N(t))$, let us denote first and second order derivatives by

$$\begin{aligned} [v'(t)]_n &= d_t v_n(t) \\ [v''(t)]_n &= d_{tt} v_n(t) \end{aligned} \tag{4.11}$$

with $n = 1, \dots, N$. The functional model that we consider in the following is a reduced version of (4.3), that does not include the gradient term and the s variable. This reduced model proved to be very useful in the specific task of recovering polygonal closed curves as it is not affected by the slight staircasing effect induced by the gradient term (cfr. with the discussion in Section 4.3.2 and analysis of numerical experiments in Section 4.4). However, for the sake of generality we remark here that all the arguments presented in the following can be easily generalized by taking into account the full version of the functional. Given a curve $g : \Pi \rightarrow \mathbb{R}^N$ in N -space, we attempt to find a piecewise linear approximation of g by looking for $u : \Pi \rightarrow \mathbb{R}^N$ and $z : \Pi \rightarrow [0, 1]$ that minimize

$$\begin{aligned} \mathcal{F}_\epsilon(z, u) &= \int_{\Pi} z^2 |u''|^2 dt + \lambda \int_{\Pi} |u - g|^2 dt \\ &+ \eta \int_{\Pi} \left\{ \epsilon |z'|^2 + \frac{1}{4\epsilon} (z - 1)^2 \right\} dt \end{aligned} \tag{4.12}$$

where λ, η are positive parameters regualting the penalization of the corresponding terms.

Discretization and minimization strategy

The discretization of the functional (4.12) follows the same principles as in Section 4.2.1, with the slight simplification that no Kronecker product is needed when defining the

matricial differential operators.¹ In a discrete setting, the function domain is a set of points $t_1, \dots, t_P \in \Pi$ with $t_p < t_{p+1}$ for all $p = 1, \dots, P - 1$. The discrete variables representing the curves (or signals in case of $N = 1$) are denoted by $\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_N)$ with $\mathbf{v} = \mathbf{g}, \mathbf{u}$, and their values at each coordinate n over a point t_p are notated as $[\mathbf{v}_n]_p$.

Let \mathbf{W} be the diagonal matrix with diagonal entries $[\mathbf{W}]_{p,p} = t_{p+1} - t_p$, for $p = 1, \dots, P - 1$, and $[\mathbf{W}]_{P,P} = -t_P$. By considering the difference schemes of size $P \times P$ as in (3.8), the discrete operators implementing first and second order derivatives can be defined as $\mathbf{D}_t = \mathbf{W}^{-1} \mathbf{A}_P^1$ and $\mathbf{D}_{tt} = \mathbf{W}^{-2} \mathbf{A}_P^2$. Simple modifications to account for different types of boundary conditions are possible. Some useful examples are:

Null-Dirichlet: no modification.

Null-Neumann: $[\mathbf{D}_t]_{P,P} = 0$, $[\mathbf{D}_{tt}]_{1,1} = [\mathbf{D}_{tt}]_{P,P} = -1$.

Periodic: $[\mathbf{D}_t]_{P,1} = 1$, $[\mathbf{D}_{tt}]_{1,P} = [\mathbf{D}_{tt}]_{P,1} = 1$.

We are now able to write the discrete version of the BZ functional for vector-valued curves as

$$\begin{aligned}
 F_\epsilon(\mathbf{z}, \mathbf{u}) = & \\
 & \sum_{n=1}^N \left\{ \mathbf{u}_n^T \mathbf{D}_{tt}^T \mathbf{R}_{z^2} \mathbf{D}_{tt} \mathbf{u}_n + \lambda (\mathbf{u}_n - \mathbf{g}_n)^T (\mathbf{u}_n - \mathbf{g}_n) \right\} + \\
 & \eta \left\{ \epsilon [\mathbf{z}^T \mathbf{D}_t^T \mathbf{D}_t \mathbf{z}] + \frac{1}{4\epsilon} (\mathbf{z} - \mathbf{e})^T (\mathbf{z} - \mathbf{e}) \right\}.
 \end{aligned} \tag{4.13}$$

Similarly to the case of vector-valued images, this functional is quadratic with respect to the variables \mathbf{z} and \mathbf{u}_n , in fact it can be written as

$$\begin{aligned}
 F_\epsilon(\mathbf{z}, \mathbf{u}) & \\
 & = \frac{1}{2} \mathbf{z}^T \mathbf{A}_z \mathbf{z} - \mathbf{z}^T \mathbf{b}_z \\
 & = \frac{1}{2} (\mathbf{u}_1^T, \dots, \mathbf{u}_N^T) \begin{pmatrix} \mathbf{A}_u & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{A}_u \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_N \end{pmatrix} \\
 & \quad - (\mathbf{u}_1^T, \dots, \mathbf{u}_N^T) \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_N \end{pmatrix}
 \end{aligned} \tag{4.14}$$

¹Kronecker product is used in Section 4.2.1 to exploit one dimensional difference schemes to work as partial derivatives on (column-wise or row-wise) vectorized images.

where $\mathbf{A}_z = \mathbf{A}_z(\mathbf{u})$, $\mathbf{A}_u = \mathbf{A}_u(\mathbf{z})$, and $\mathbf{b}_z, \mathbf{b}_n$ for $n = 1, \dots, N$ are given by

$$\begin{aligned} \mathbf{A}_z &= 2 \sum_{n=1}^N \mathbf{R}_{|\mathbf{u}_n''|^2} + 2\epsilon\eta \mathbf{D}_t^T \mathbf{D}_t + \frac{\eta}{2\epsilon} \mathbf{I} \\ \mathbf{b}_z &= \frac{\eta}{2\epsilon} \mathbf{e} \\ \mathbf{A}_u &= 2\mathbf{D}_{tt}^T \mathbf{R}_{z^2} \mathbf{D}_{tt} + 2\lambda \mathbf{I} \\ \mathbf{b}_n &= 2\lambda \mathbf{g}_n \end{aligned} \tag{4.15}$$

and $|\mathbf{u}_n''|^2 := (\mathbf{D}_{tt} \mathbf{u}_n)^2$. To minimize this functional the same approach proposed in the previous section can be used, with obvious adaptations. For the sake of completeness, we remind here that the general minimization approach is based on a sequential partial minimization of the type

$$\begin{cases} \mathbf{z}^{k+1} = \arg \min_{\mathbf{z}} F_{\epsilon}(\mathbf{z}, \mathbf{u}^k) \\ \mathbf{u}_n^{k+1} = \arg \min_{\mathbf{u}_n} F_{\epsilon}(\mathbf{z}^{k+1}, \dots, \mathbf{u}_n, \dots) \end{cases} \tag{4.16}$$

for $n = 1, \dots, N$, in either an exact or inexact fashion. For technicalities we refer the reader to the previous section.

4.3 Experimental results: piecewise linear approximation of vector-valued images

We propose in this section a comparative analysis between the MS and the proposed BZ approaches to vector-valued image approximation. In particular, test cases are focused on difficult color image restoration/regularization tasks highlighting limitations of the MS model such as the staircasing effect and the crack-tip problem and demonstrating how the BZ model overcomes these issues. Computations are performed using MATLAB® R2015b, hardware is Intel® Core™ i5-4750 CPU @3.20 GHz, 16.00 GB Ram.

4.3.1 Functional models considered for comparison

Let us recall, for the sake of clarity, the functional model given by Ambrosio-Faina-March as in (4.3) (to which we will refer in the following as **AFM-BZap**):

$$\begin{aligned} \mathcal{F}_{\epsilon}(s, z, u) &= \delta \int_{\Omega_0} z^2 |Hu|^2 dx + \xi_{\epsilon} \int_{\Omega_0} (s^2 + o_{\epsilon}) |\nabla u|^2 dx \\ &+ (\alpha - \beta) \mathcal{AT}_{\epsilon}(s) + \beta \mathcal{AT}_{\epsilon}(z) + \mu \int_{\Omega_0} |u - g|^2 dx. \end{aligned} \tag{4.17}$$

From a numerical point of view, the presence of the gradient term in the **AFM-BZap** functional model introduces a slight staircasing effect in the solution. Although its influence vanishes asymptotically for $\epsilon \rightarrow 0$ due to the presence of the infinitesimal ξ_ϵ , in numerical applications ϵ cannot be 0. To account for this, a *reduced* version of the **AFM-BZap** functional is considered in the numerical experiments that does not present the staircasing effect. More specifically, the reduced functional (to which we will refer in the following as **AFM-BZ**) is

$$\mathcal{F}_\epsilon(z, u) = \int_{\Omega_0} z^2 |\text{Hu}|^2 dx + \beta \mathcal{AT}_\epsilon(z) + \mu \int_{\Omega_0} |u - g|^2 dx. \quad (4.18)$$

In this reduced version, the functional does not depend anymore on the gradient of u and the function s . Therefore, the solution is allowed to have 1st-order variations without gradient penalization. We recall that also for the reduced functional a full Γ -convergence result to the weak BZ functional holds true [14]. It is worth noting that, the reduced functional **AFM-BZ** can be directly compared with the well-known Ambrosio-Tortorelli approximation of the Mumford-Shah functional (to which we will refer in the following as **AT-MS**)

$$\mathcal{F}_\epsilon(s, u) = \int_{\Omega_0} s^2 |\nabla u|^2 dx + \alpha \mathcal{AT}_\epsilon(s) + \mu \int_{\Omega_0} |u - g|^2 dx \quad (4.19)$$

as it corresponds to the same functional where the Hessian term is replaced by the gradient. Notice that both the functional models **AFM-BZ** and **AT-MS** can be obtained from **AFM-BZap** by setting functional parameters to

$$\begin{aligned} \mathbf{AFM-BZ} : \quad & \xi_\epsilon = 0, \alpha = \beta, \delta = 1 \\ \mathbf{AT-MS} : \quad & \xi_\epsilon = 1, \alpha_\epsilon = 0, \beta = 0, \delta = 0. \end{aligned}$$

An important consequence of this, is that numerical methods to solve the minimization problems associated to the two functional models **AFM-BZ** and **AT-MS** can be derived with straightforward modifications from the method proposed in Section 4.2.1.

4.3.2 Restoration of color images

The study dataset is a color image representing a portion of the oil painting *Girl with a pearl earring*, by Johannes van der Meer, see Figure 4.1a. The image has size 600 × 600 pixels and depth 8-bits. As we can see, the painting is severely affected by the *craquelure*.² In the numerical experiments that follow, we compare the results obtained

²Craquelure is the fine pattern of dense "cracking" formed on the surface of the oil as part of the process of ageing.

by using the three functional models to image approximation presented in Section 4.3.1 where parameters are specifically selected to remove the craquelure effect.

In order to simplify the comparison among the three considered functional models (**AT-MS**, **AFM-BZap** and **AFM-BZ**), we will set some parameters to common values. First, the discretization parameters common to all the three models are the grid sizes t_x, t_y and the Γ -convergence parameter ϵ . They can be set to standard values $t_x = t_y = 1$ and $\epsilon = 0.01$ (see [4] for details). The remaining ones are the functional parameters: (**AT-MS**) μ, α , (**AFM-BZap**) $\mu, \delta, \alpha, \beta$, and (**AFM-BZ**) μ, β . To keep the maximum similarity among the three functional models, we set the remaining parameters as:

$$\mathbf{AT-MS} : \quad \alpha = 1$$

$$\mathbf{AFM-BZap} : \quad \delta = 1, \alpha = 2, \beta = 1$$

$$\mathbf{AFM-BZ} : \quad \beta = 1$$

The *smoothing* parameter μ is set with respect to the chosen application. By tuning the parameter μ , we forced the smoothing of the image until the approximating image was not showing any craquelure feature. By decreasing the value of this parameter the smoothing effect is increased. We started with $\mu = 1$ and we decreased it by negative powers of 10, i.e., $\mu = 1, 0.1, 0.01, 0.001, \dots$. The first value at which no craquelure was observable in all the three approximation results was $\mu = 0.01$, so results are showed according to this parameter value.

Comparative analysis of the approximations

The major challenge in the given image is to both remove the craquelure and still preserve the smooth color variations between shadowed and lightened regions of the girl's face.

In the **AT-MS** case the staircasing effect is very evident. Many coarse patches of constant color are clearly visible and the gradients of color are abruptly approximated by the edges of such patches, see Figure 4.1b. The map of these edges is given in the plot of the function \mathbf{s} , in Figure 4.2a. In the solution of **AFM-BZap** the over-segmentation is much less evident, as the gradient component is weighted by the infinitesimal ξ_ϵ . However, we can see mainly two problems in the computed solution \mathbf{u} , as displayed in Figure 4.1c. First, sharp edges are still visible in some portions of the image. See for example both sides of the nose ridge and the lips boundaries. These unnatural gradients of color are mapped by the edge function \mathbf{s} of **AFM-BZ**, as shown in Figures 4.2b and 4.2c. Second, many complex transitions between shadow and light (that are more evident on the girl's left cheek) are lost. As we can see from the plot of the edge function \mathbf{s} , this bad effect happens in correspondence of the gray shaded regions of the \mathbf{s} map. Here, the functional has penalized the gradient all over these wide regions instead of only along 1-dimensional

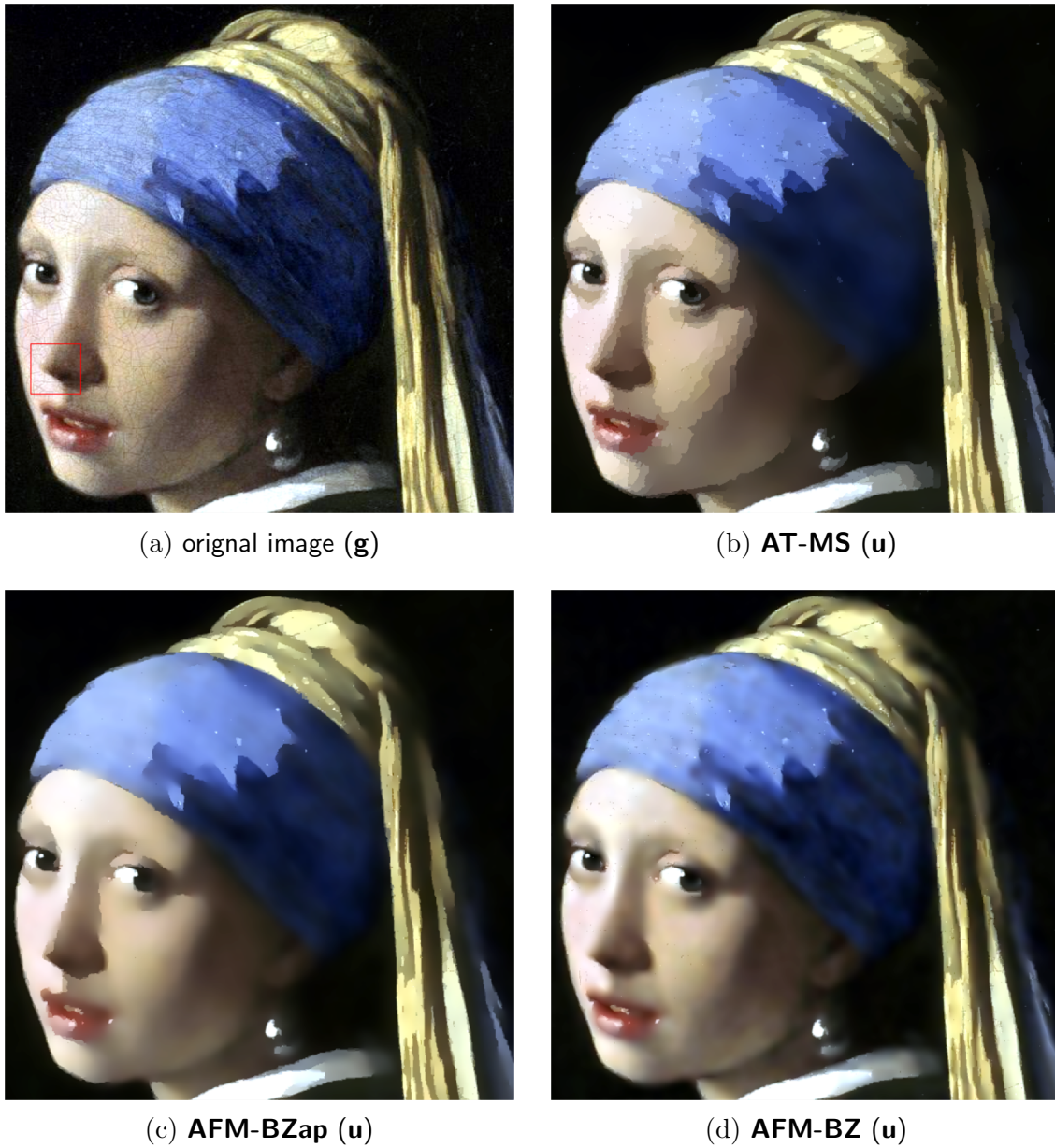


Figure 4.1: Craquelure removal via MS and BZ approaches. (a) The input color image represents a portion of the oil painting *Girl with a pearl earring*. Its approximations are obtained by the three different models: (b) **AT-MS**, (c) **AFM-BZap**, (d) **AFM-BZ**.

edges. As a consequence, the image is badly approximated by a too smooth function and the complex shadow geometry is destroyed. All these problems do not show up in the solution of the **AFM-BZ** model, see Figure 4.1d. Here, only 2nd-order information is penalized and the solution is allowed to have first order variations. As a result, the image is better approximated and no unwanted artifacts such as unnatural edges and too coarse

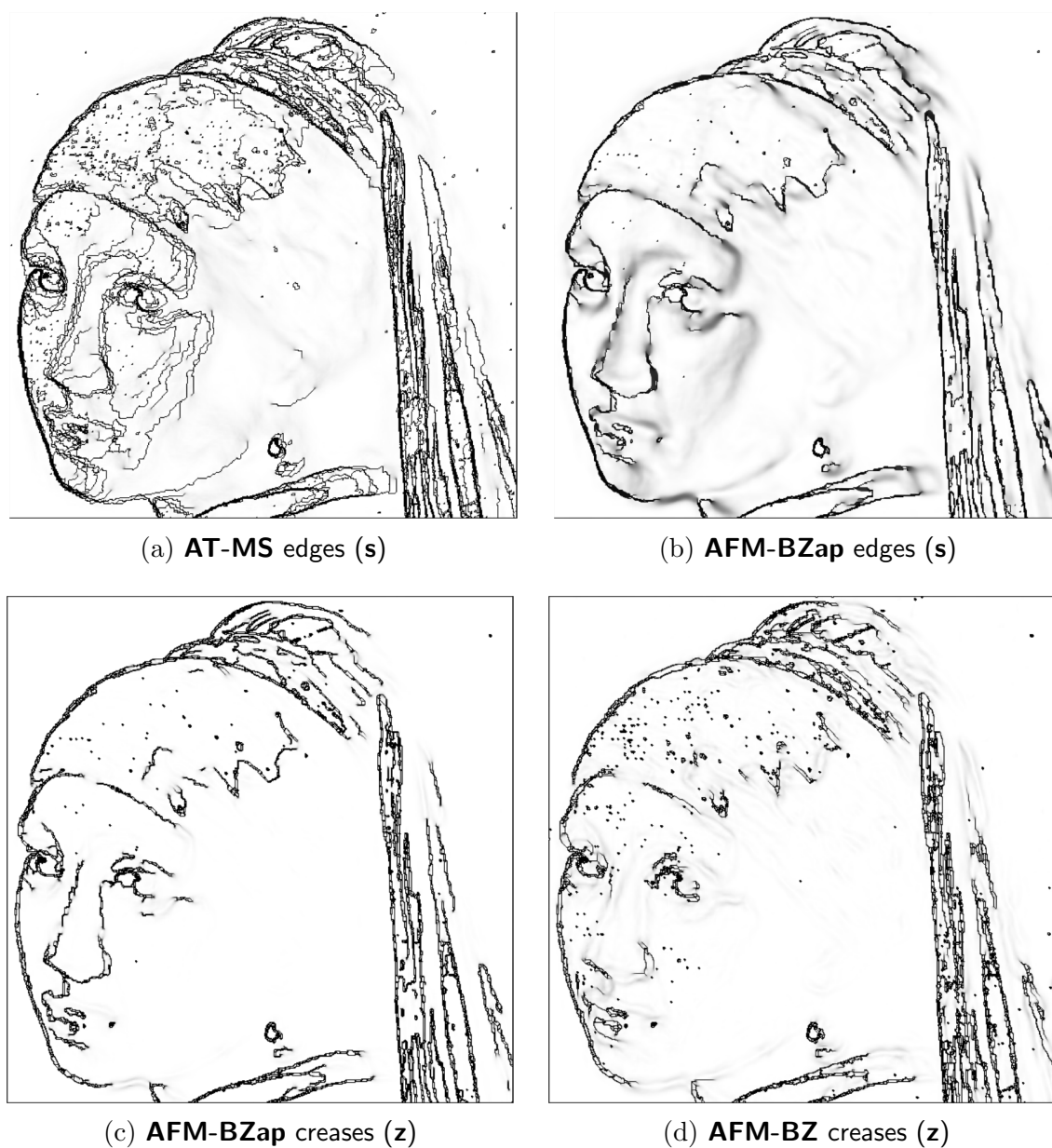


Figure 4.2: Discontinuity functions computed for the three functional models. (a) **AT-MS**, (b,c) **AFM-BZap**, (d) **AFM-BZ**. White corresponds to 1 and black to 0, gray values are in between.

shadowed areas are present. As an example, we can see from the map of detected gradient discontinuities \mathbf{z} in Figure 4.2d that the girl's nose is not contoured by any sharp edge, neither the left cheek is over-segmented.

In order to better illustrate the geometrical behaviour of these solutions and to emphasize the piecewise linear approximation, we will show the plots of a particular of the

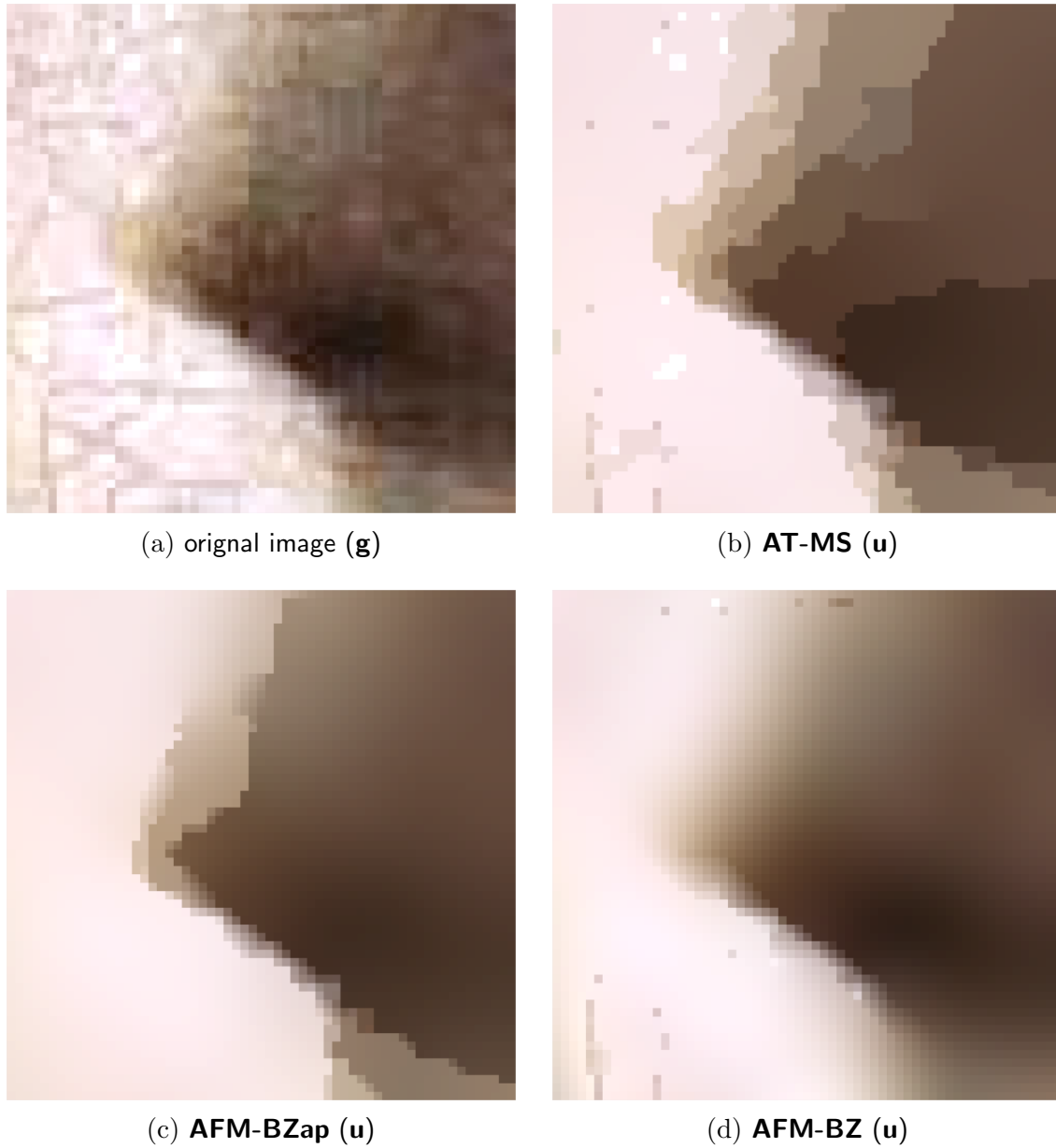


Figure 4.3: Particular of the approximations by zooming the area in the red square in Figure 4.1. (a) Original image, and the results for: (b) **AT-MS**, (c) **AFM-BZap**, (d) **AFM-BZ**. Notice the over-segmentation effect in (b,c).

images as embedded surfaces in the RGB space. This image portion corresponds to the red square of size 60×60 pixels in Figure 4.1a. Magnifications of this part in the original image and in the three computed solutions are showed in Figure 4.3. The scatter plots of the embedded surfaces are illustrated in Figure 4.4. The high level of noise of the original image results in a scatter plot where points are almost uniformly distributed

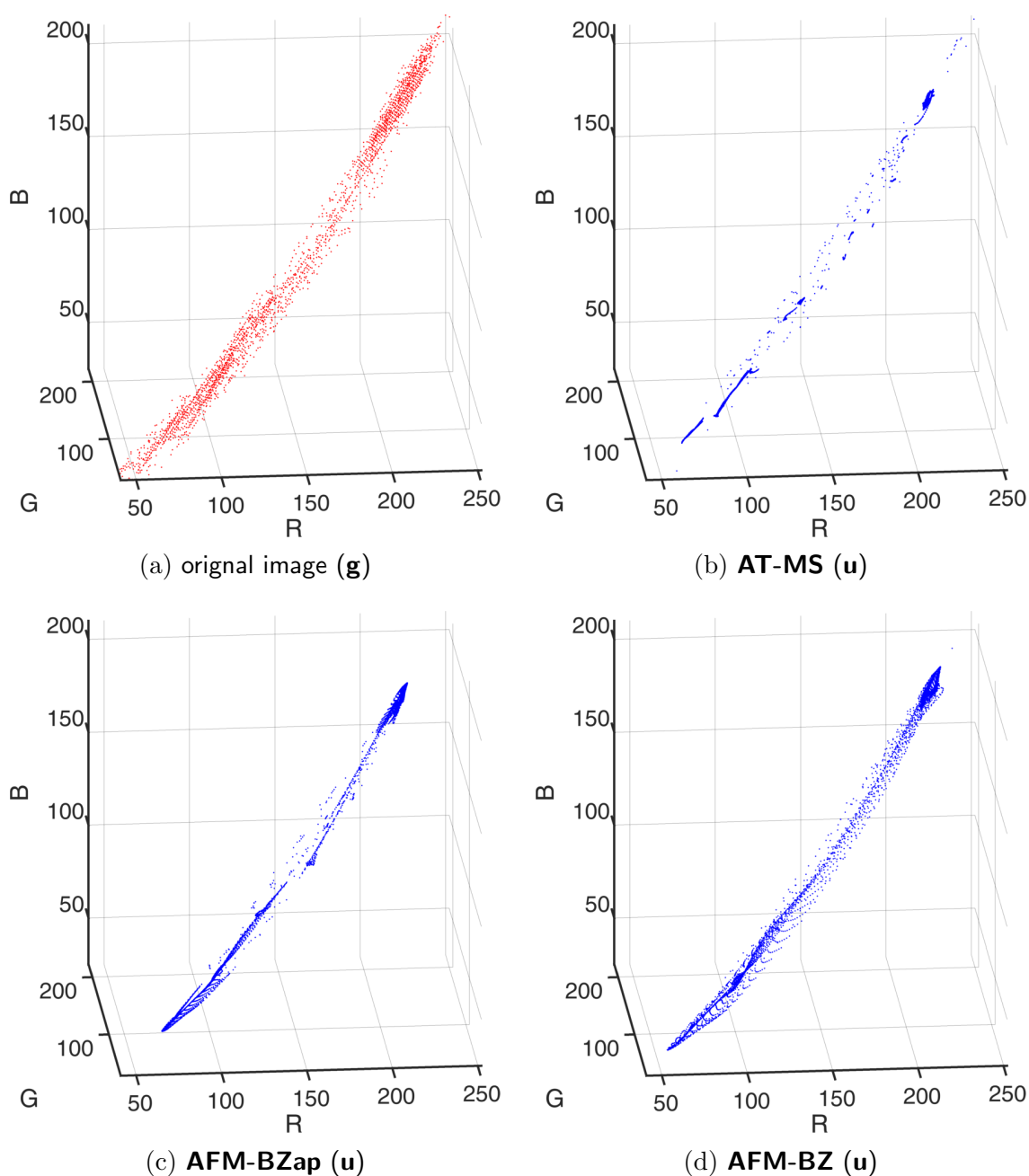


Figure 4.4: Pixel scatterplots of the image portions represented in Figure 4.3. (a) Original image, (b) **AT-MS**, (c) **AFM-BZap**, (d) **AFM-BZ**.

all around the RGB space’s main diagonal, with two regions where points are slightly denser (corresponding to the dark brown and the light pink regions of the image). In the **AT-MS** case, as a result of the severe over-segmentation effect, pixels are clustered in different almost isolated portions of the RGB space. In the **AFM-BZap** case these

Table 4.1: Outer(k)/inner(iter) iterations and execution time observed in the minimization of the three functional models.

model	k	s	z	iter			time (secs)
				u_1	u_2	u_3	
AT-MS	21	22	–	526	300	385	38.45
AFM-BZap	16	17	17	1280	742	889	98.75
AFM-BZ	21	–	22	1630	877	1110	121.24

clusters are slightly enlarged, however they are still distinguishable and sharply separated from background pixels. Much more regular is the scatter plot in the **AFM-BZ** case. Here, the dense clusters are visible but they are more displaced in space and surrounding pixels are uniformly and regularly distributed.

Numerical minimization performance

Let us analyze the details about the iterations in the minimization of the three functional models and compare them. In Table 4.1 are recorded the number of outer (k) and inner (totiter) iterations and the execution time. Inner iterations relate to the PCG solvers (triggered with diagonal preconditioner) applied to find (for each outer iteration) the gradient related search directions w.r.t. the variable blocks. As discussed in Section 4.2.1, the minimization w.r.t. the u variable can be separated into three convex sub-problems. Therefore, the search of a global gradient related descent direction can be split along the three sub-directions u_1, u_2, u_3 (the target image is color image, the three directions correspond to the R,G,B bands), independently. As we can see, in the case of the s and z variable blocks PCGs stopped in one iteration, whereas the u_i variable blocks required more iterations. This can be explained in terms of positive definiteness of the matrix \mathbf{A}_u . It follows from the numerical expression of \mathbf{A}_u in (4.9) that the positive definiteness of this matrix decreases with the parameter μ . As a consequence, for small values of μ the convexity of the quadratic form associated to \mathbf{A}_u reduces in magnitude, thus, the descent requires more iterations.

Another important aspect of the minimization is the competition among the functional terms induced by the parameters choice. The dynamics of this competition can be better understood by looking at the plots in Figure 4.5. The main fact that can be observed is that all the terms are decreasing except for the distance term. In particular, the AT components have higher decreasing rates in the first iterations, meaning that the contrast of the corresponding solutions is heavily decreasing.³ The behaviour of the distance term

³By premature stopping of the algorithms we could notice many discontinuities that are not present in the

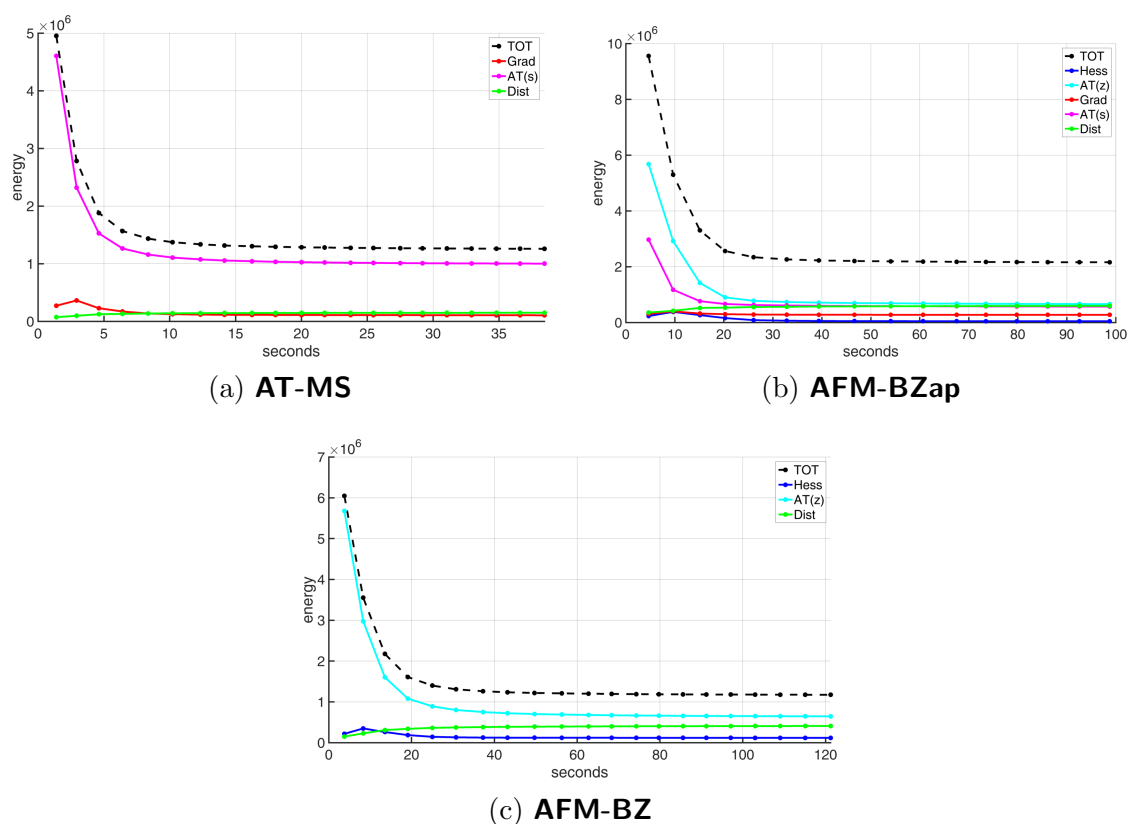


Figure 4.5: Energy-versus-time at each outer iteration for the three functional minimization cases: (a) **AT-MS**, (b) **AFM-BZap**, (c) **AFM-BZ**. The plots illustrate the descent of each additive term in the functional models. The black dashed line is the total energy. Blue is the Hessian component and Cyan is the AT component associated to the Hessian (present only in **AFM-BZap** and **AFM-BZ**). Red is the gradient component and Magenta is the AT component associated to the gradient (present only in **AT-MS** and **AFM-BZap**). Green is the distance term.

is as expected: due to the strong noise removal level induced by the parameters, the solution becomes more and more distant (in the Euclidean sense) with respect to the input image.

4.3.3 Reduction of white Gaussian additive noise

Recovering an image degraded by additive noise is a well-known inverse problem in image processing. Variational methods have been justified in this framework as the MS model can be properly derived by following a Bayesian rationale as an additive noise reduction model [61]. However, as we have seen also in Section 4.3.2, the MS approximation can

final results.

irremediably deteriorate some important features of the image because of the staircasing effect. In this experimental section, we aim at showing that the proposed **AFM-BZ** model actually outperforms the **AT-MS** also in terms of noise reduction.

In order to do this, we consider here a synthetic color image corrupted by different levels of noise. The image is reconstructed by using the **AT-MS** and the **AFM-BZ** models and in both cases the variance of the removed noise is estimated from the difference image. The synthetic 8-bits (per channel) color image is 300×300 pixels and contains two challenges: (1) a crack-tip (with circular gradient) in the red band, and (2) two very smooth creases in the green and blue bands with vertical and horizontal directions, respectively. Functional parameters are the same as in the previous section. The only difference is that the smoothing parameter has been set to a smaller value $\mu = 0.001$, as the additive noise added in the experiments resulted to be more difficult to suppress. Given that the image is at 8-bits, we added additive 3-dimensional 0-mean Gaussian noise with covariance matrix given by $\Sigma = \sigma I_3$, where I_3 is the identity matrix of size 3×3 , and in three different trials we set $\sigma = 50, 100, 200$.

The results of noise variance estimation are reported in Table 4.2. We can easily see that in all the three cases the **AT-MS** model returned very bad approximations of the noise variance, while the reconstruction given by **AFM-BZ** allowed for very precise estimates. This happened because the color geometry of the test image is highly non-constant, thus, the MS fails in approximating both the crack-tip and the orthogonal smooth creases. This fact can be clearly seen by looking at the images of the final approximations obtained via the two functional models, showed in Figure 4.6 (the images show the results obtained for $\sigma = 200$, being this the most critical case). Notice in particular the behaviour of the MS approximation at the end of the crack-tip (Figure 4.6e), showing the well-known phenomenon of the triple-points. In principle this phenomenon is due to the penalization of the discontinuity set, that induces the discontinuity edges to displace in optimal configurations with minimum length. This happens when they meet at $2/3\pi$ wide angles. Mumford and Shah conjectured in their seminal work [3] that the discontinuity set of a MS minimizer is the union of C^1 arcs that can only end at interior points (pure crack-tips) or meet with equal angles. It has been proven in [62] by using the calibration technique that a function with a triple-point discontinuity is a local minimizer of the homogeneous MS functional. It is worth noting that triple-points do not show up in the BZ approximation, where the solution properly follows the complex crack-tip geometry with surrounding circular gradient (Figure 4.6f).

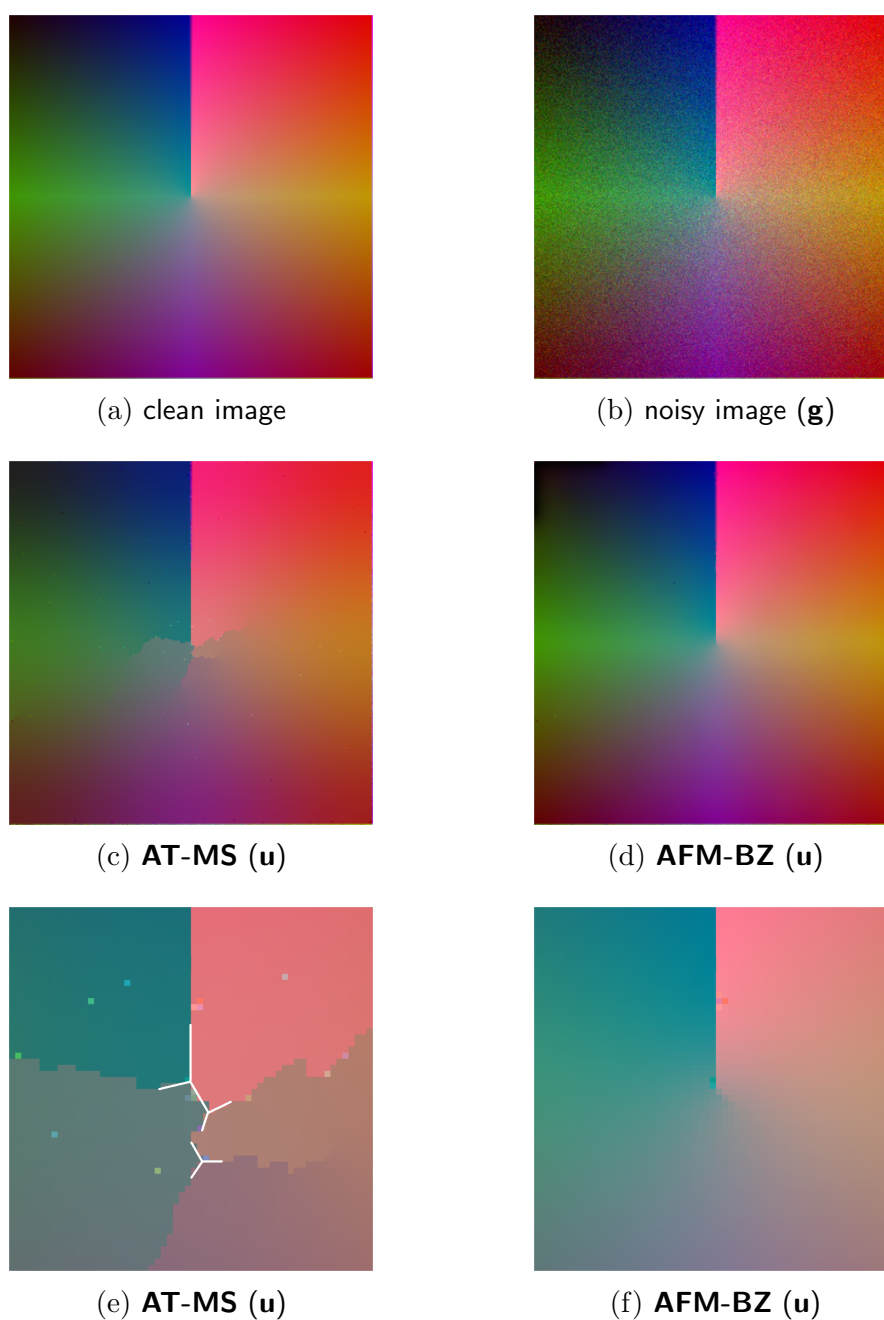


Figure 4.6: Estimation of Gaussian additive noise in color image containing challenging geometries. (a) Synthetic noise-free generated image, (b) noisy image. Reconstructions of the noisy image are obtained by the (c) **AT-MS** model, (d) **AFM-BZ** model. Particulars zoomed at the crack-tip end for the (e) **AT-MS** solution (white traits emphasize the main directions of the discontinuity edges) and the (f) **AFM-BZ** solution.

Table 4.2: Estimation of additive Gaussian noise variance in color images using the MS and the BZred models.

σ	model	estimated σ from			k	time (secs)
		u_1	u_2	u_3		
50	AT-MS	74.25	216.33	217.09	18	19.82
	AFM-BZ	53.69	49.84	50.30	17	49.63
100	AT-MS	121.98	265.68	266.17	10	8.97
	AFM-BZ	103.69	99.59	99.96	16	48.05
200	AT-MS	215.97	360.38	362.43	13	10.87
	AFM-BZ	200.83	197.48	199.80	21	49.31

4.4 Experimental results: polygonal approximation of planar closed curves

In this section, we show how the feature of the BZ model that allows the formation of free gradient discontinuities is fundamental in the task of recovering the shape of closed curves from discrete noisy sampling. Like other models such as cubic smoothing splines (CSSPs), the BZ model is able to provide a smooth approximation of the curve. However, as additional feature and unlikely other methods can do, the BZ model allows also to retrieve polygonal shapes (i.e., curves with gradient discontinuities).

As a specific application, in the following we recover the polygonal shape of building footprints from discrete noisy approximations obtained in the processing of low resolution Digital Surface Models (DSMs). DSMs are 2-dimensional scalar-valued rasters and they are obtained by interpolating raw LiDAR (Light Detection and Ranging) unstructured point clouds into regular grids [45]. The value of the DSM at each grid point corresponds to the height of the object hit by the laser pulse in the location of the grid point. It is common in the remote sensing literature to extract building edges in DSMs in order to recover a discrete approximation of the building footprints. However, if the DSM resolution is low (e.g., 1m), the discrete points forming the detected edges can be far from a polygonal shape (Figure 4.7b) and post-processing is required to recover the building footprint. An example of edge detection in urban DSM is given in [4, 35], where the BZ model for gray-scale images is applied to the DSM and discrete approximations of the building footprints are mapped by the edge-detection function \mathbf{s} . We can see in Figure 4.7 a 3-D rendering of the DSM of an old barrack and the mapping function of its detected edges. The DSM is at spatial resolution of 1m and the shape is not oriented parallel to the x, y -axis, therefore the discrete representation of the boundary is broken into many

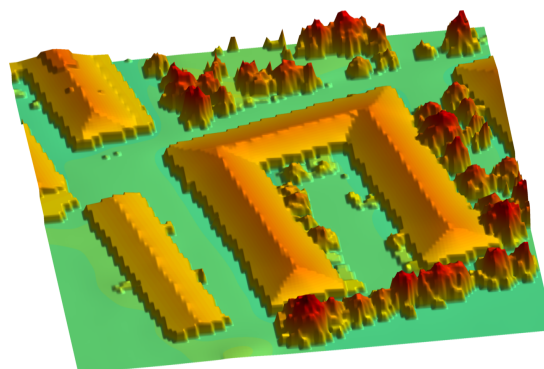
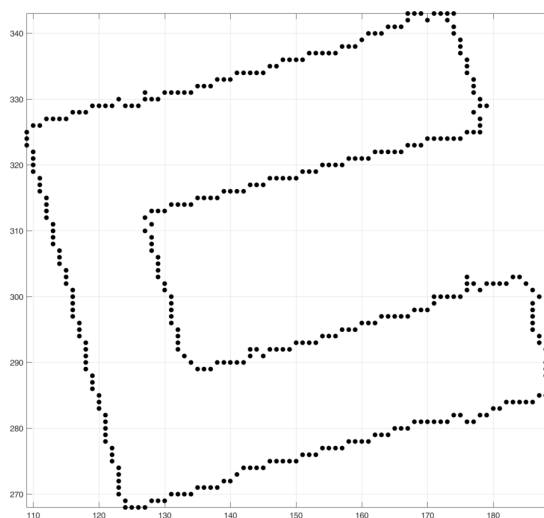
(a) DSM (\mathbf{g})(b) edge map (\mathbf{s})

Figure 4.7: Extraction of building edges from DSM. (a) 3d rendering of a DSM representing an old barrack. (b) Edge map of the main (U-shaped) building obtained by segmenting the DSM using the BZ model for gray-scale images [4]. The points correspond to the pixels where the function \mathbf{s} (the edge detection function) is 0.

segments oriented parallel to the x, y -axis.

To recover the polygonal shape approximating these points we exploit the framework proposed in Section 4.2.2. Input data is the set of P two-dimensional points $\{(x_p, y_p)\}_{p=1}^P$ representing the discrete noisy sampling of the unknown polygonal shape (the black points in Figure 4.7b). We set $N = 2$ and we construct the discrete variable $\mathbf{g} \in \mathbb{R}^{P \times 2}$ representing the discrete planar curve by simply assigning $[\mathbf{g}_1]_p = x_p$ and $[\mathbf{g}_2]_p = y_p$, for all $p = 1, \dots, P$. To recover a closed curve, we minimize the functional (4.13) with periodic boundary conditions (cfr. Section 4.2.2). The BZ model (for brevity **BZ**), is tested against the typical approach to curve approximation of Cubic Smoothing Splines (for

brevity **CSSP**). We recall that for **CSSP**, the solution that approximates the points \mathbf{g} is the piecewise cubic function \mathbf{v} that minimizes the functional expression

$$G(\mathbf{v}) = q \sum_{p=1}^P |\mathbf{v}(t_p) - \mathbf{g}_p|^2 + (1 - q) \int_{t_1}^{t_P} \mathbf{v}''(t) dt \quad (4.20)$$

where t is a parameteric variable and $q \in [0, 1]$ is a parameter that penalizes data fitting (q near 1) or data smoothness (q near 0). For brevity we omit details on **CSSP** implementation, we only recall that the solution can be found in a closed form. For details we refer the reader to [63, 64].

The **BZ** model depends on the two parameters λ and ν , penalizing the distance of the solution to the original data \mathbf{g} and the size of the gradient discontinuity set, respectively. The **CSSP** model only depends on the parameter q and the solution is not allowed to have gradient discontinuities. In order to better understand the behaviour of the proposed models for a large variety of parameter selections, we defined a grid of values for λ , ν and q . The whole range of possible behaviours of the resulting approximating curves (from over- to under- fitting) has been obtained for $\lambda = 10^{-4}, 10^{-5}, 10^{-6}$ and $\nu = 10^{-k}$, with $k = 1, \dots, 5$, and for several values of q between 0 and 1.

The results of curve approximation is illustrated for both **CSSP** and **BZ** models in Figures 4.8 and 4.9, respectively. Note from the results of **CSSP** that, by variation of the parameter q from 1 to 0, the behaviour of the solution is from complete over-fitting to very smooth (and poor) approximation of the points. Notice that the smoothing effect of the splines does not allow to well represent the right angles of the main corners of the shape. Instead, by varying the parameters of the **BZ** we still obtain different behaviours from over- to under-fitting, but for some parameter choices we have polygonal solutions. In fact, by decreasing the contrast parameter ν we allow the size of the gradient discontinuity set to be larger, thus allowing the formation of free gradient discontinuities. Polygon corners correspond to the points where the gradient of the solution is discontinuous. On the counter part, if we fix the value of ν the solutions from the top row to the bottom row show increasing under-fitting. Indeed, by decreasing the parameter λ the solution is allowed to be *distant* from the original curve, thus the minimization penalizes the discontinuity set and produces very smooth curves. Among all the solutions, we can say that the best polygonal approximation is obtained for $\lambda = 10^{-5}$ and $\nu = 10^{-4}$ as the curve segments are straight segments forming right angles. It follows from the globality of the geometrical parameters in (4.12), that polygonal shapes at the same scale and corrupted by the same level of noise can be recovered by using identical parameter selections.

The computational burden to obtain all the approximations can be considered as negligible, as in all the cases algorithms converged in less than one second. Hardware and

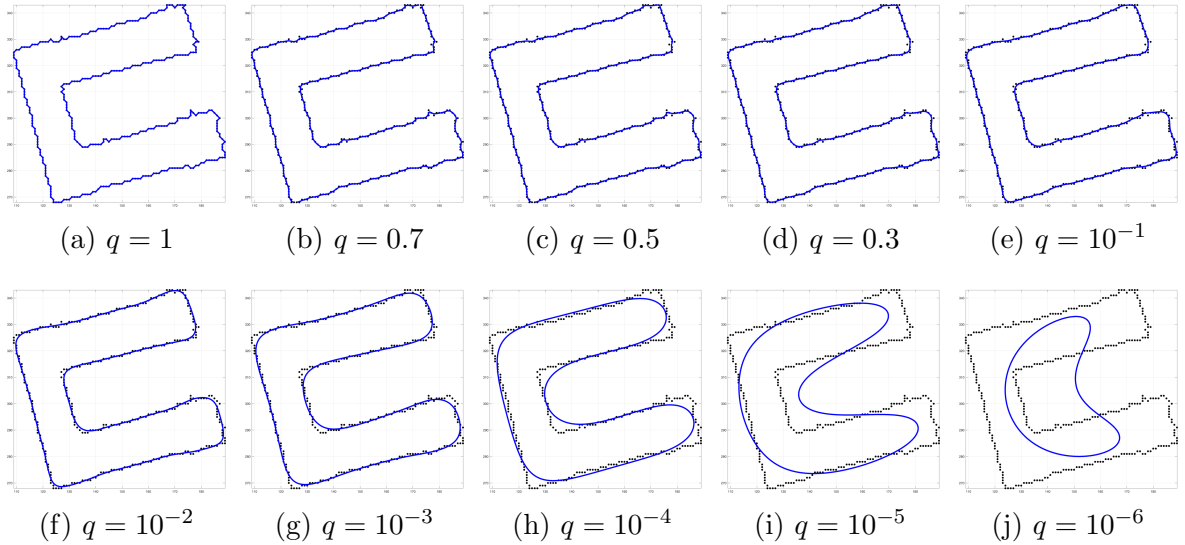


Figure 4.8: Curve approximation results obtained for different parameter choices of the **CSSP** model. The range of values used in the experiments allowed us to explore the behaviour of the solution from over- to under- fitting. No gradient discontinuity is allowed by the model.

software used is the same as in Section 4.3.

4.5 Conclusions

In the framework of variational methods to image approximation, the 1st-order model by Mumford-Shah is very popular. However, some intrinsic problems due to its 1st-order nature (such as the staircasing effect and the triple point cracks) limit its applicability to solve complex problems such as image denoising and restoration. To solve for this, 2nd-order methods can be used, but no attempts to implement vector-valued versions of 2nd-order models can be found in literature. This is critical when color or multispectral images need to be analyzed. In this chapter we considered a 2nd-order variational model to the approximation of vector-fields in dimension two (e.g., multiband images) and one (e.g., curves in space) and we proposed efficient numerical implementations of the associated minimization problems. Specifically, we focused on the numerical minimization of the Ambrosio-Faina-March (AFM) elliptic approximation of the Blake-Zisserman (BZ) functional as it is particularly prone to numerical implementation. In the proposed numerical formulation the objective functional is written in a compact matricial form and the minimization is decomposed into quadratic sparse convex sub-problems. We proved that the minimization sub-problem associated to the vector-valued variable (\mathbf{u}) can be split into B further quadratic sub-problems (where B is the vector size) all depending on the same

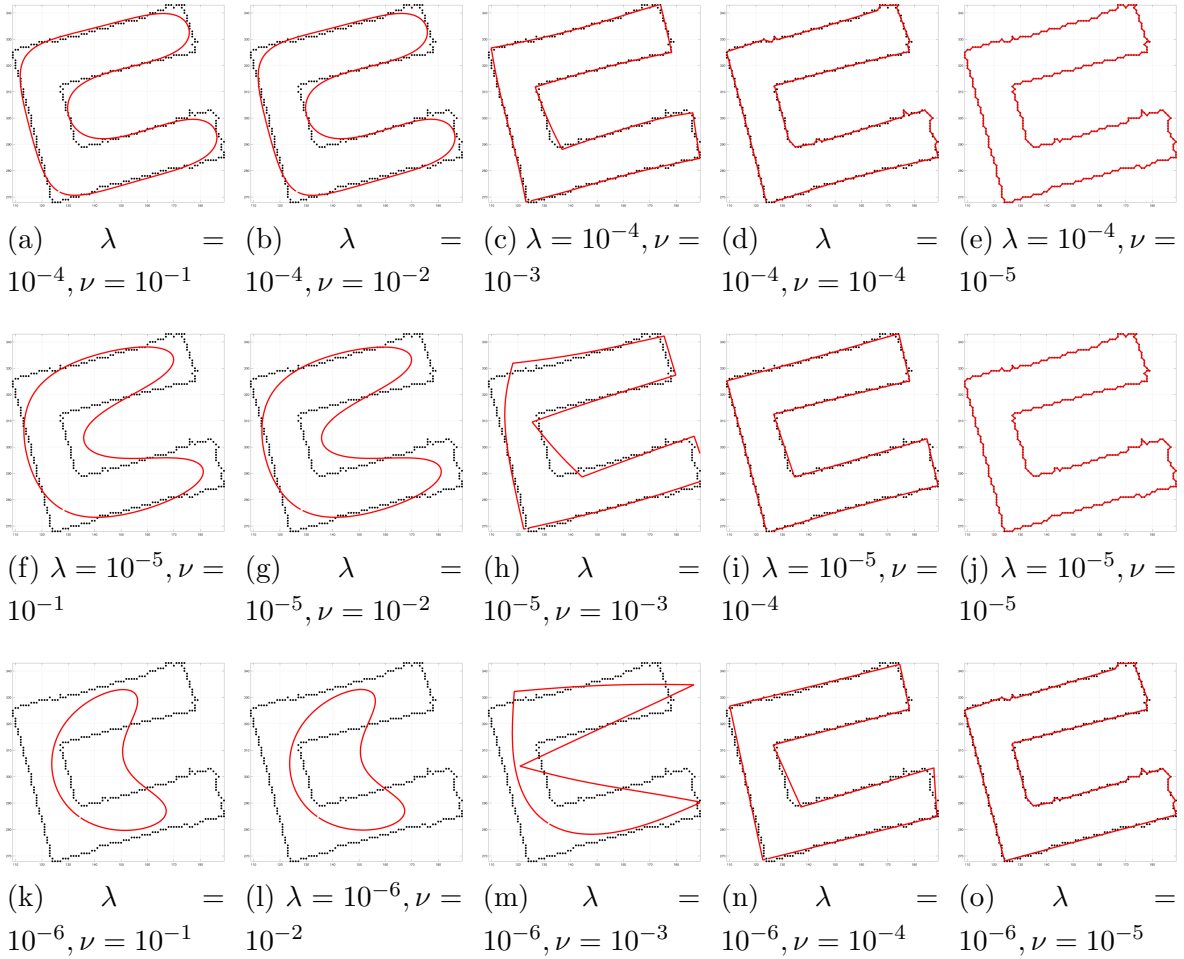


Figure 4.9: Curve approximation results obtained for different parameter choices of the **BZ** model. The range of values used in the experiments allowed us to explore the behaviour of the solution from under- to over- fitting passing also through polygonal solutions. The best polygonal approximation is (i), i.e., for parameters $\lambda = 10^{-5}$ and $\nu = 10^{-4}$.

matrix. Different types of experimental studies have been done to assess the effectiveness of the proposed numerical formulation. In the first experimental part, we proposed a comparative analysis of the BZ model against the MS on difficult image restoration/denoising problems. The results show that the capability of the BZ model to approximate the input image in a piecewise linear manner produces more natural (in terms of visual interpretation) and precise (in terms of noise reduction/estimation) reconstructions of corrupted color images. In the second experimental part, we applied the BZ model to the approximation of closed curves from discrete noisy sampling. Differently from other typical curve approximation models such as Cubic Smoothing Splines, the BZ model allows the formation of free gradient discontinuities. Thus, polygonal approximation can be obtained for

suitable choices of the parameters.

Part II

Novel statistical models for change detection in multispectral images

Chapter 5

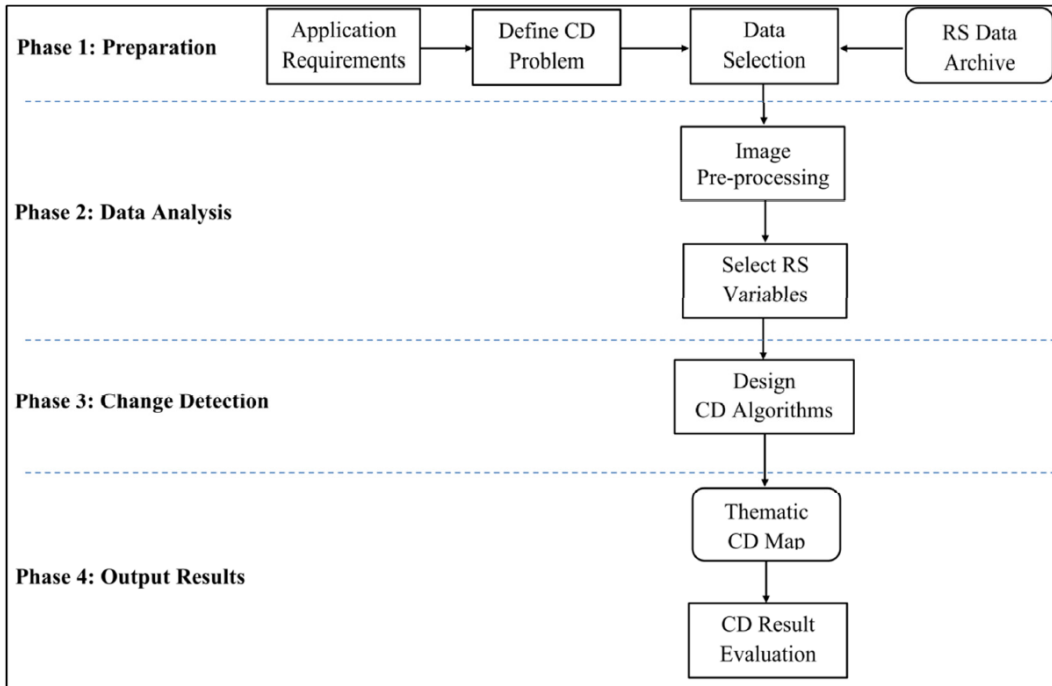
Background

This chapter introduces to the content of Part II. First, a review of the change detection problem and the techniques proposed in the literature to solve it are proposed. The discussion mainly concerns with change detection methods that apply to multispectral images at moderate/high spatial resolution, as these are the main focus of the thesis. However, the change detection problem on very high resolution multispectral images and main related issues is also presented. The last section of this chapter presents a detailed explanation of the novel contributions presented in the following chapters of the thesis.

5.1 Overview of the change detection problem

The modern society organization strongly relies on the cyclic monitoring-planning paradigm for an efficient control and management of natural resources and human infrastructure development. The remote sensing community assists this fundamental process by steadily providing up-to-date technologies to collect data over the globe and the know-how to extract the useful information from it. A comprehensive and systematic understanding of the global change is possible thanks to one of the leading remote sensing applications: the change detection [65], which substantially embodies the main principles of this paradigm. Indeed, CD is gaining more and more attention due to its strong impact in our daily life and many efforts and resources have been, and are going to be, made at the purpose of providing accurate, up-to-date and reliable operational products such as land-cover maps, land-use change detection maps and geophysical variables variations [66].

Technically, CD is the process that identifies changes occurred between two (or more) dates in a specific geographical location by analysis of RS images acquired over the investigated area [67]. It is worth noting that change detection is a comprehensive procedure that requires a set of analysis and processing phases [68], see Figure 5.1. Given the intrinsic non-objective nature of the *change* meaning [69,70], the CD procedure often starts



(a)

Figure 5.1: General high-level scheme for the design and the implementation of a change detection procedure.

with a problem understanding phase that includes the change definition. This first stage always includes a selection step where suitable RS data types are chosen with the aim of emphasizing the change characteristics by increasing their discriminability. The design of the CD algorithm is generally driven by the application. Typically, the practitioner aims at finding, and exploiting, a strong correlation between the variation of image properties (e.g., pixel radiance value, texture, and shape) and the land-cover material changes on the ground. However, other changes may be also detected caused by some factors like variation in atmospheric conditions, sensor conditions, illumination differences and seasonal effects. The detection of these specific kinds of changes, which are often considered sources of noise, strictly depends on the real applications requirements and goals. The final outcome of the CD algorithm is usually a map, superimposable to the image raster, in which changed pixels are labeled. A successful CD procedure should present high CD accuracy. The detection accuracy is calculated by intersection of the output change map with available ground truth information and/or by qualitative assessment by comparison with alternative, auxiliary data available.

5.2 Change detection methods for multispectral images

In the literature many approaches have been developed to detect changes in multitemporal images. Extensive overviews are available both from the image processing [71] and the remote sensing [65, 72] viewpoints.

A common way to group the majority of the CD methods proposed in literature is based upon the availability of reference information (labeled data) for the considered scene. According to this principle there are supervised and unsupervised approaches to change detection. The former group includes methods that are based on the use of supervised classifiers, see for instance [73–75]. Supervised methods require ground reference information for the training phase. Such information may be required for all available acquisitions, or only for one of them (e.g., when semisupervised and domain adaptation methods are considered [76, 77]) and is typically obtained from highly expensive in-situ campaigns or by dedicated photo-interpretation procedures. In light of the unprecedented huge amount of data now available in archives and the given rate at which data are nowadays collected, the cost of obtaining ground truth information cannot be afforded. By means of automatic techniques one can gradually reduce the need for conventional field investigations with drastical reduction of the operational cost. As a consequence, the most attractive CD techniques are the ones that aim at automatically detecting the changes. This is the case of the unsupervised methods [78–84], to which we only focus on in the remainder of this thesis.

5.2.1 Mid/high spatial resolution images

Generally speaking, unsupervised methods to CD are mainly able to detect the presence/absence of changes [82, 84, 85]. In some cases they can distinguish among different changes [86], but given the unavailability of reference data they usually cannot associate semantic information (e.g., a specific land cover transition) to them. Many unsupervised methods are based on pixelwise comparison [67, 80, 87]. The comparison phase, which usually consists in applying a mathematical operation to each pixel value, returns a change index. The most common change index is the spectral difference of a bitemporal image pair. Inference about the presence or absence of changes based on the change index is typically done via thresholding or clustering approaches.

Threshold-based methods differ among each other based on how the threshold value is assigned. In some cases the threshold is empirical and can be obtained by application of well-known image processing algorithms such as the Kittler-Illingworth algorithm [88], the Otsu algorithm [89] or maximum entropy algorithm [90]. In some other cases, a more change-detection oriented formulation of the problem is given and some theoretical

models have been proposed for the definition of the threshold. A prominent approach in this direction is based on the maximum-a-posteriori (MAP) assignment, which follows from a statistical modeling of the unchange and change classes [78, 80, 82]. What is typically observed after image differencing is that unchanged pixels have pixel intensity which is approximately null. On the contrary, changed pixels intensity is likely non-null. In [82], two Gaussian distributions are associated to the unchange and the change classes, respectively. The threshold value that separates between unchanged and changed pixels is then defined in a MAP Bayesian framework to produce the minimum overall error of decision. In the case one wants to assign different weights to commission and omission errors, a minimum cost approach also applies [80]. The statistical study of the binary detection of changes can be further refined by analyzing the magnitude of the spectral difference. In [78], a theoretical model based on the Rayleigh-Rice mixture has been proposed to represent the typical bi-modal shape of the magnitude of the difference image. An adaptive multiple-threshold approach is used in [85], where the CD problem is seen as an hypothesis test. Here, the unchange/change state of each pixel is decided adaptively on the basis of a significance test.

Differently to threshold-based approaches, clustering algorithms try to separate between unchanged and changed pixels according to some homogeneity principle that characterize these two classes. We can find in literature methodological approaches based on: k-means and kernel k-means clustering [91], fuzzy c-means and Gustafson-Kessel clustering [92], non linear support vector clustering [93], fuzzy clustering [94], and many others.

The detection of changes can be carried on in the original spectral domain (this is the case of all the methods mentioned above) or in a transformed domain. In the latter case, the transformation is meant to enhance the change representation and in many cases it is driven by the data. Well-known transformation techniques successfully applied to change detection are canonical correlation [95], Principal Component Analysis (PCA) [96,97] and Independent Component Analysis (ICA) [98]. In [99] PCA is applied to the difference image and changes are recognized as the first few principal components. In [100], an orthogonal representation of the changes based on the Gram-Schmidt procedure was used. In [101], the tasselled cap transformation was applied to detect the vegetation change from Landsat images. It is worth mentioning also the Multivariate Alteration Detection (MAD) technique [102], which is based on the canonical correlation analysis [95]. An improved version named Iterative Reweighted MAD (IR-MAD) was proposed in [103] to provide more reliable output components thus emphasizing changes. However, the main disadvantage of the above mentioned transformation-based CD approaches is that they require a strong interaction with the end-users to select the most informative components emphasizing the specific changes of interest, which is usually time consuming. On the

other hand, the transformation-based methods do not provide a clear number of changes. The number of detected changes highly depends on the selected number of components and the change information represented in those components. Some changes might be still mixed and unidentified in a given component. Therefore, the transformation-based approaches are good at extracting features for enhancing the detection of specific kinds of changes in CD, but in general they are not suitable for detecting all the possible change classes. Among pixel-based approaches to change detection we also find methods based on fuzzy set theory [99, 104] and similarity measures [105]. All the aforementioned methods demonstrated their effectiveness on moderate and high geometrical resolution images (from tens to hundreds meter spatial resolution) and in several applications.

5.2.2 Very high spatial resolution images

A major drawback of pixel-based methods to CD is that they only consider the change/unchange information within a single pixel, even if its neighbors contain significant information. Implicitly, these methods consider the pixels to be spatially independent. Although this is an acceptable assumption for moderate/high spatial resolution images, it is not for images that exhibit very high spatial resolution (VHR). Indeed, when the geometrical resolution of images increases up to less than one meter, spatial correlation among pixels becomes non-negligible and the straightforward application of the above mentioned methods may easily fail to give acceptable results. To account for geometrical additional information, many change detection methods have been adapted to a spatial-contextual framework. Spatial-context information can be modeled by applying: fixed-shape neighborhood systems for texture information extraction [106–108], Markov Random Fields [82, 109] and morphological filters [83, 110]. More advanced methods perform a context-sensitive analysis by considering adaptive neighborhoods modeled by multitemporal parcels [81, 111, 112] and object properties [113–117]. They better capture the spatial correlation information present in the scene and become particularly promising for VHR images showing complex objects (e.g., buildings and other man-made structures). In order to effectively model objects in the images at different scales, some of the concepts employed in the previously mentioned papers such as morphological filters, multitemporal parcels, and even Markov Random Fields can be adapted and used in multiscale/multilevel analysis [83, 109, 111, 118] together with specific multiscale/multilevel representation tools such as Wavelet transform [119–121]. When dealing with unsupervised change detection, pre-processing of multitemporal images becomes highly important because most of the approaches suffer of differences in image radiometry and/or geometry. In the literature, studies exist on the effects of residual misregistration on the change detection results [122–124], and techniques devoted to mitigate these effects (and thus change detection errors associated to

them) have been proposed [125–127]. Other papers are devoted to the analysis of effects of pansharpening algorithms [128] and radiometric differences [129, 130] on the change detection performance.

5.3 Challenges and novel contributions

Many studies of the change detection problem based on the analysis of the spectral vector difference can be grouped as specializations derived from a general framework well-known as Change Vector Analysis (CVA) [131]. CVA was introduced in [132] and it gives a theoretical foundation for the representation of change vectors both in cartesian [133] or polar coordinates [78]. In the remote sensing application, depending on which representation system is used, specific properties of the change vectors are put in relationship with the change class on the ground and methodologies are developed to identify and extract (mostly in an automatic way) difference vectors belonging to the same class. Very typical is the application of CVA to binary change detection, which has been applied both in the cartesian [82] and the polar [78] coordinate systems. In particular, in the latter case the magnitude representation of the change vectors allows to implement a simple, intuitive but yet effective concept: change vectors having small magnitude most-likely represent unchanged pixels, whereas those ones having high magnitude values can be considered as changed pixels. From a statistical point of view, one can assume the unchange and the change classes to follow specific distributions, estimate distribution parameters from data and then perform decision according to a Bayes decision rule. In [82] the bi-modal distribution of the magnitude is empirically approximated by a mixture of Gaussians. In [78], by assuming that the unchange and the change classes are Gaussian distributed in the difference image, a theoretical model based on the Rayleigh-Rice mixture has been proposed to better fit the real distribution of the magnitude. It has been conjectured that this approach could have improved the binary detection of changes in multispectral images. However, no actual implementation was possible due to the complexity related to the derivation of a parameter estimation algorithm involving the Rayleigh and Rician distributions.

Following this direction, in Chapter 6 we propose a theoretical derivation of a parameter estimation method based on the EM algorithm which is specifically tailored for the Rayleigh-Rice mixture. The numerical method is computationally convenient as it does not require any complex optimization routine because it is only based on an iterative updating of the parameters. Along with this, also the equation for the calculation of the Bayesian optimal threshold for minimum overall error is derived. In the experimental part, the binary detection method based on the Rayleigh-Rice model has been applied to

both synthetic and real multispectral data. The result shows that the proposed method outperforms the classical change detection approaches based on Gaussian mixtures.

In Chapter 7, a novel compound multiclass model to the representation of the difference image is proposed with the intent of improving both: the fitting capabilities of the statistical representation of the unchanged and changed pixels and the change detection performance. In fact, a careful analysis of the change detection results obtained by modeling the unchange and change classes as single classes shows that in some cases the estimated distributions do not fit very well the histogram of the magnitude and the computed threshold values are not always close to the optimal ones, especially when images acquired by last generation sensors are considered. This suggests that the real nature of the unchange/change classes is intrinsically more complex. The proposed multiclass model is aimed at better capturing and modeling the real statistical behaviour of the unchange/change classes. In particular, these two are seen as macro classes including different subclasses that represent on their turn different statistical behaviours for both unchanged and changed pixels. It is worth noting that, the proposed multiclass model is not imposed a-priori on the difference image, instead, it is derived from very general assumptions on the bi-temporal images. To demonstrate its effectiveness, the proposed model is applied in the context of binary detection based on the magnitude information. Numerical experiments on a large variety of datasets (including mostly images acquired by last generation sensors) show how the capability of modeling the unchange and the change classes as multiple can significantly improve the fitting and the detection performance.

We introduce in Chapter 8 a class-wise context-based model to multispectral image simplification that is based on 1st-order variational approximation. The model provides a new interpretation of the solution of the Mumford-Shah problem as the expected image (in a maximum-likelihood sense) under the given statistical formulation. Having at disposal the processing capability of computing piecewise smooth approximations of multispectral images developed in the previous chapters, the model is then applied to change detection in multispectral images. Here, the approximation phase greatly helps in reducing the statistical variability of classes by at the same time preserving their spatial distribution. As a result, the decision phase of the change detection becomes more easy and the detection performance sensibly increases. To enforce the remote sensing application, the method is coupled with a tiling approach that allows the processing of large images by exploiting a parallel implementation.

Chapter 6

The Rayleigh-Rice mixture model for binary change detection

The problem of estimating the parameters of a Rayleigh-Rice mixture density is often encountered in image analysis (e.g., remote sensing and medical image processing). In this chapter¹ we address this general problem in the framework of change detection (CD) in multitemporal and multispectral images. One widely used approach to change detection in multispectral images is based on Change Vector Analysis (CVA). Here, the distribution of the magnitude of the difference image can be theoretically modeled by a Rayleigh-Rice mixture density. However, given the complexity of this model, in applications a Gaussian-mixture approximation is often considered, which may affect the change detection results. Hereafter we present a novel technique for parameter estimation of the Rayleigh-Rice density that is based on a specific definition of the Expectation-Maximization (EM) algorithm. The proposed technique, which is characterized by good theoretical properties, iteratively updates the parameters and does not depend on specific optimization routines. Several numerical experiments on synthetic data demonstrate the effectiveness of the method which is general and can be applied to any image processing problem involving the Rayleigh-Rice mixture density. In the change detection context, the Rayleigh-Rice model (which is theoretically derived) outperforms other empirical models. Experiments on real multitemporal and multispectral remote sensing images confirm the validity of the model by returning significantly higher change detection accuracies than those obtained by using state-of-the-art approaches.

¹parts of this chapter appear in [1]

6.1 Introduction

In this chapter, the attention is focused on the CVA technique (and its derivations) [131] and the use of automatic statistical modeling and thresholding within the CVA framework. CVA demonstrated to be a valuable and flexible tool for the detection of changes in several contexts (e.g., Remote Sensing [133], Medical Diagnosis and Treatment [134]). This technique is based on the representation in polar coordinates of the difference image (which is obtained by subtracting two images representing the same scene at different times). In the polar feature space, pixels having high magnitude values are likely to be changed and their separation into different kinds of change can be performed by means of their direction values [78, 135]. As mentioned, the magnitude of the multispectral difference image carries information about presence/absence of changes. Thus, the magnitude variable can be employed to separate changed from unchanged samples, eliminate the latter and perform further analysis only on the former ones [135]. Usually, the information in the magnitude variable is extracted by means of thresholding procedures [97, 112]. In [82, 136], the statistical distribution of the magnitude as a mixture model representing the classes of unchanged and changed pixels is approximated by a Gaussian mixture, then decision is made using a Bayesian rule. However, recent studies [78] showed that the precise model of this distribution can be theoretically derived, thus opening the way to a theoretical well-founded method instead of an empirical one. The model relies mainly on two hypotheses: (1) natural classes are Gaussian distributed within each band of the multispectral images (a reasonable assumption for images obtained from passive sensors) and, (2) pixels are spatially independent (this assumption is usually done for remote sensing images at medium resolution). Under these hypotheses, the magnitude of the difference image can be theoretically described by a Rayleigh-Rice mixture density. For this reason we address the problem of defining an EM-type algorithm to the estimation of shape and mixture parameters of the Rayleigh-Rice density in order to accurately solve the binary CD problem on multispectral multitemporal images.

Problems involving the Rician distribution often arise in engineering applications, in particular in Remote Sensing [78] and Magnetic Resonance Imaging (MRI) [136–140]. Since when the Rician distribution was introduced for the modeling of the magnitude of Gaussian densities [141], many efforts have been made for developing algorithms for the estimation of its parameters. Important results have been achieved by using the method of moments (MOM) [139, 142], and Maximum-Likelihood (ML) approaches [142, 143]. Unfortunately, the MOM is shown to be inefficient at low signal-to-noise ratios (SNRs) [138, 140], while ML equations do not have in general a unique solution [144]. Because of the latter property, the solution of the ML problem becomes an optimization problem.

Some papers propose adaptive techniques for selecting the initial starting values [138,140], while in other cases slightly different Bayesian estimators are proposed in order to stabilize the problem [145].

In spite of these results, the problem of parameter estimation in the case of mixture densities including the Rician distribution is still poorly investigated. When the non-centrality of the signal is high, the Rician density is often approximated with a Gaussian one [136]. Other papers directly address the approximation of the parameters of a mixture model involving one (optional) Rayleigh and J Rician distributions all of them having a common scale parameter σ . In [146], the authors proposed a fitting procedure followed by an approximated Expectation-Maximization (EM) estimation of σ . In [147] an additional parameter is included in the model as missing information, leading to a substantial simplification in the maximization step. In practice, in [146] and [147] the parameter σ describes a common characteristic of the noise, which is supposed to appear with different non-centralities and same scale parameter. Therefore the estimation of σ from J Rician components makes the model robust. Other methods are based on local noise estimation [148] and wavelet-based noise estimation [149]. In many application problems, forcing the components of the Rayleigh-Rice mixture to have the same scale parameters is a strong assumption. In particular, in CD on remote sensing real images such similarity is rarely observed, and neither it has a theoretical justification. Therefore, an empirical use of the algorithms in [146], [147] to fit the distribution of the magnitude of the difference image, and therefore to solve the binary CD problem, is expected to present limitations. In the following we develop an EM-type novel method for the estimation of all shape and mixture parameters of a Rayleigh-Rice mixture density. Asymptotic properties of the considered statistical model enable us to define an iterative method based on subsequent updates of the parameters. By providing a variational interpretation of the EM algorithm as a problem related to a fixed point equation we both: (1) derive explicit formulas for implementing the updates, and (2) establish a lower bound on the speed of convergence of the iterations for reaching a maximum of the expectation. The algorithm is robust and it can be initialized using standard techniques of preliminary thresholding.

On the one hand, if compared to already existing methods, in our model all the statistical parameters of the mixture density are free (we let the two mixture components to have different scale parameters). As a result, a large variety of practical problems can be addressed. On the other hand, we emphasize that the algorithm does not need any optimization routine (differently from [146], [147]), thus many issues related to the choice of optimal maximization strategies and their impact on the solution are avoided.

The chapter is structured as follows. Section 6.2 introduces the Bayesian framework for binary change detection based on the Rayleigh-Rice mixture density as a model describing

the magnitude of the difference image. In Section 6.3 we first present an overview of the EM algorithm and then we provide explicit formulas for the approximation of mixture and shape parameters of the Rayleigh-Rice mixture density. Convergence analysis is also provided. The experimental results are given in Section 6.4 where the proposed method is applied to synthetic and real change detection problems on multispectral images. Section 6.5 draws the conclusions. An Appendix including mathematical notions and technical results is also provided.

6.2 A problem of binary change detection

In this section, following [78], we recall how under reasonable assumptions the statistical distribution of the magnitude of a bi-temporal difference image acquired by passive sensors is described by a Rayleigh-Rice mixture density. The magnitude of the unchanged pixels follows a Rayleigh distribution, while the magnitude of the changed pixels follows a Rician distribution. Then, a framework for the classification of the pixels according to their class is given in terms of Bayesian decision theory.

6.2.1 The Rayleigh-Rice mixture model for the magnitude of the difference image

Let us consider two multispectral images $\mathbf{y}^1, \mathbf{y}^2$ acquired by passive remote sensing sensors at different times t_1, t_2 , respectively, and representing the same geographical area. Let us assume that the two images are co-registered and radiometrically corrected, and that there has been (only) one relevant change in the scene between the two dates. Therefore, the pixels can be divided into two classes only: ω_n (unchanged pixels) and ω_c (changed pixels). The aim is to discriminate between changed and unchanged pixels in an unsupervised way. The detection of the changes occurred between t_1 and t_2 is based on the study of the so-called difference image

$$\mathbf{d} := \mathbf{y}^2 - \mathbf{y}^1. \quad (6.1)$$

When images acquired by passive sensors are considered, the statistical distribution of natural classes within each spectral band can be reasonably modeled by Gaussian densities [150]. From now on, our analysis will be restricted to considering two bands among all the available ones.² The following theoretical analysis holds for any multi-band image where classes can be modeled as Gaussian densities, both in the case where they are independent

²This is just to simplify the notation and the exposition and it is not a technical restriction. The generalization of the proposed theory to the case where an arbitrary number of bands n is considered can be found in Appendix 6.A.5.

or jointly-distributed as Gaussian. In these cases we have that also the classes ω_n, ω_c in the difference image are Gaussian distributed, thus let us assume they are modeled by $\mathcal{N}(\mu_{b,n}, \sigma_{b,n})$ and $\mathcal{N}(\mu_{b,c}, \sigma_{b,c})$, for each band $b = 1, 2$.

In polar coordinates, the magnitude of pixels can be exploited for discriminating between unchanged and changed pixels, so we are interested in describing how this model can be represented when the difference image is transformed. Since we are considering two-band images, we have that every pixel $\mathbf{d}(i, j)$ is a two-dimensional vector. Hence, it is uniquely determined by its magnitude $\rho(i, j)$ and direction $\theta(i, j)$ with respect to a fixed reference direction. Given the pixel in spatial position (i, j) , if no change has occurred on the ground between the two dates t_1 and t_2 , then the magnitude $\rho(i, j)$ is expected to be close to zero. Conversely, whenever a change has occurred, the magnitude is expected to be significantly different from zero.

We are now interested in modeling the theoretical distribution of the magnitude. To this aim, let us denote the random variable that describes this feature by ρ . Because of the above-mentioned assumption of normality of classes within bands, the distribution of ρ is theoretically given by the distribution of the magnitude of a two-dimensional point whose coordinates are Gaussian distributed random variables. The obtained model is quite complex, but with some additional reasonable assumption it can be greatly simplified [141]. A first crucial assumption is spatial independence of pixels. From the application viewpoint, this assumption is reasonable and widely supported in literature if optical images at medium spatial resolution (e.g., 30 mt) are considered [82], [67]. Then, since the two images are co-registered and radiometrically corrected, we can assume that in those areas where pixels are not changed the distributions are not significantly different, therefore

$$\begin{aligned}\mu_{1,n} &= \mu_{2,n} = 0 \\ \sigma_{1,n} &= \sigma_{2,n} =: \sigma_n\end{aligned}\tag{6.2}$$

and we have that the magnitude of unchanged pixels is modeled by a Rayleigh distribution

$$p(\rho|\omega_n) = \frac{\rho}{b_n^2} \exp\left(-\frac{\rho^2}{2b_n^2}\right) \quad \rho \geq 0,\tag{6.3}$$

where the parameter $b_n = \sigma_n$. In the case of changed pixels, we still assume that the distributions have the same variance, but they can have different non-zero means. Hence,

$$\begin{aligned}\mu_{1,c} &\neq \mu_{2,c} \quad \mu_{1,c}, \mu_{2,c} \neq 0 \\ \sigma_{1,c} &= \sigma_{2,c} =: \sigma_c.\end{aligned}\tag{6.4}$$

and the magnitude of changed pixels follows the more general Rician distribution, [141]

$$p(\rho|\omega_c) = \frac{\rho}{\sigma_c^2} \exp\left(-\frac{\rho^2 + \nu_c^2}{2\sigma_c^2}\right) I_0\left(\frac{\rho\nu_c}{\sigma_c^2}\right) \quad \rho \geq 0.\tag{6.5}$$

Here $\nu_c = \sqrt{\mu_{1,c}^2 + \mu_{2,c}^2}$ is the so called non-centrality parameter, and $I_0(\cdot)$ is the 0-th order modified Bessel function of first kind [151]. In conclusion, the theoretical mixture density that models the distribution of the magnitude of pixels is given by

$$p(\rho) = p(\omega_n) p(\rho|\omega_n) + p(\omega_c) p(\rho|\omega_c) \quad (6.6)$$

where $p(\omega_h)$, $h = n, c$, are the prior probabilities of classes. According to the given assumptions, the magnitude image $\boldsymbol{\rho} := \{\rho(i, j) : i, j\}$ can be considered as a set of i.i.d. samples drawn from the theoretical distribution (6.6). For images where the i.i.d. assumption is not reasonable, further modelization of the spatial-contextual dependence of pixels is required. For example this can be done using Markov Random Fields (MRFs) [82], [112].

6.2.2 A framework for automatic binary change detection

The theoretical formulation of the distribution of the pixel magnitude as a mixture model allows for a formal characterization of a threshold that separates pixels into two classes according to their magnitude, by means of the Bayesian decision theory. The two probability models involved in the mixture model, the Rayleigh and the Rician distributions, have got a non-empty intersection. Accordingly, it is not possible to exactly decide which class a pixel with a given magnitude belongs to. This means that, for every possible classification, there is always a classification error with probability e_n, e_c (as shown in Table 6.1). It is well known that the overall error $e_n + e_c$ can be minimized by selecting the separating threshold as the solution $\rho = T$ (solving for ρ) of the equation

$$\frac{p(\omega_c)}{p(\omega_n)} = \frac{p(\rho|\omega_n)}{p(\rho|\omega_c)} \quad \rho \geq 0 \quad (6.7)$$

that corresponds to the intersection of the two curves $p(\omega_n) p(\rho|\omega_n)$ and $p(\omega_c) p(\rho|\omega_c)$ which lies between the two modes (it is worth noting that this equation generally has more than one solution). This equation can be equivalently written as

$$\left(\frac{1}{2b_n^2} - \frac{1}{2\sigma_c^2} \right) \rho^2 + \log I_0 \left(\frac{\rho\nu_c}{\sigma_c} \right) = \frac{\nu_c}{2b_n^2} + \log \frac{\sigma_c^2 p(\omega_n)}{b_n^2 p(\omega_c)}. \quad (6.8)$$

The classification that corresponds to the minimum overall error is given by

$$\begin{aligned} W_n &= \{(i, j) : \boldsymbol{\rho}(i, j) \leq T\} \\ W_c &= \{(i, j) : \boldsymbol{\rho}(i, j) > T\}, \end{aligned} \quad (6.9)$$

where W_n, W_c are the sets of predicted unchanged/changed pixels, respectively.

Table 6.1: Classification error in a two-class decision problem. ω_n unchanged pixels, ω_c changed pixels

		Predicted classes	
		ω_n	ω_c
Actual classes	ω_n	–	e_n
	ω_c	e_c	–

In Section 6.3 we provide an EM-type algorithm for finding an accurate estimation of the parameters b_n, ν_c, σ_c and the prior probabilities $p(\omega_n), p(\omega_c)$, in such a way that the change detection can be performed by thresholding the magnitude of the difference image at T , which is obtained by solving (6.8).

6.3 The EM algorithm for parameter estimation of the Rayleigh-Rice mixture density

Here we consider the problem of estimating the parameters describing the probability density function of a continuous non-negative random variable ρ . In the considered model, the population represented by ρ is a mixture of two components following a Rayleigh and a Rician distribution, respectively. Given a set of i.i.d. samples drawn from this distribution, the estimation problem is solved by following the principles of the EM algorithm (see [152, 153], for a comprehensive overview of the theory).

Firstly, an overview of the EM theory is given. Then, by exploiting the asymptotic properties of the considered model, an explicit iterative method for solving the estimation problem is presented. Being iterative and explicit, the proposed method does not require any optimization strategy. Detailed discussion on the convergence of the method is given.

6.3.1 EM approach to parameter estimation

Let us consider the family of density functions depending on the set of parameters $\Psi = (\alpha, \Theta)$, where $\Theta = (\theta_1, \theta_2)$ with $\theta_1 = b$ and $\theta_2 = (\nu, \sigma)$, given by

$$p(\rho|\Psi) = \alpha_1 p(\rho|\omega_1, \theta_1) + \alpha_2 p(\rho|\omega_2, \theta_2), \tag{6.10}$$

where $\alpha_1 = \alpha$, $\alpha_2 = 1 - \alpha$ with $0 < \alpha < 1$, and

$$p(\rho|\omega_1, \theta_1) = \frac{\rho}{b^2} \exp\left(-\frac{\rho^2}{2b^2}\right) \quad (6.11)$$

$$p(\rho|\omega_2, \theta_2) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2 + \nu^2}{2\sigma^2}\right) I_0\left(\frac{\rho\nu}{\sigma^2}\right). \quad (6.12)$$

Let \mathbf{x} be a set of N i.i.d. samples drawn from the distribution (6.10) determined by the set of parameters $\bar{\Psi}$, i.e., $p(\rho|\bar{\Psi})$. The aim of the EM algorithm is to estimate the real values $\bar{\Psi}$ using the samples \mathbf{x} . In particular, here we consider the case where each sample $x \in \mathbf{x}$ is unlabeled, thus the sampling is incomplete. We recall that a complete sample would be of the form $y = (x, \omega)$ where $\omega \in \{\omega_1, \omega_2\}$ is the *label* of the sample x (i.e., the index representing the population from which the observed value x comes from). The EM algorithm gives, under certain hypotheses, an estimation of $\bar{\Psi}$ as a local maximizer $\hat{\Psi}$ of the so-called log-likelihood of the samples \mathbf{x}

$$L(\Psi) = \sum_{x \in \mathbf{x}} \log p(x|\Psi). \quad (6.13)$$

The key point of the EM algorithm is [152]: given Ψ' , then $L(\Psi) \geq L(\Psi')$ if Ψ maximizes the *conditional log-expectation*

$$Q(\Psi|\Psi') = \sum_{x \in \mathbf{x}} \sum_{h=1,2} p(\omega_h|x, \Psi') \log(\alpha_h p(x|\omega_h, \theta_h)), \quad (6.14)$$

where for $h = 1, 2$ and for each $x \in \mathbf{x}$, the weight

$$p(\omega_h|x, \Psi') := \frac{\alpha'_h p(x|\omega_h, \theta'_h)}{p(x|\Psi')} \quad (6.15)$$

is the posterior probability that x originated in the h -th component of the population, given Ψ' (see Figure 6.1). In its more general fashion, the EM algorithm is implemented as follows: an approximation Ψ^0 is firstly chosen, then for $k = 0, 1, \dots$ the conditional log-expectation $Q(\Psi|\Psi^k)$ is evaluated (E-step) and the next iterate is found (M-step) as $\Psi^{k+1} := \arg \max_{\Psi} Q(\Psi|\Psi^k)$. A recursive formula for updating the mixing proportions can be formally derived and the M-step can be also formulated as³

$$\begin{aligned} \alpha^{k+1} &:= N^{-1} \sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi^k) \\ \Theta^{k+1} &:= \arg \max_{\Theta} Q((\alpha^k, \Theta)|\Psi^k). \end{aligned} \quad (6.16)$$

³We present also this formulation of the EM algorithm since it helps us in making a clear distinction between our approach and the one in [146].

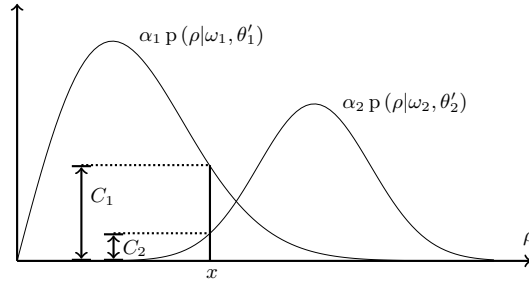


Figure 6.1: Geometrical interpretation of the posterior probability that x originated in the h -th component given Ψ' , as $p(\omega_h|x, \Psi') = C_h/(C_1 + C_2)$ for $h = 1, 2$.

Exploiting information theory, the EM algorithm to the parameter estimation of (6.10) can be further simplified. In the considered statistical model all parameters b, ν, σ are mutually independent (see Appendix 6.A.3). Thus, following [153] the EM algorithm can be split into separated maximization steps

$$\begin{aligned}
 \alpha^{k+1} &:= N^{-1} \sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi^k) \\
 b^{k+1} &:= \arg \max_b Q((\alpha^k, b, \nu^k, \sigma^k)|\Psi^k) \\
 \nu^{k+1} &:= \arg \max_\nu Q((\alpha^k, b^k, \nu, \sigma^k)|\Psi^k) \\
 \sigma^{k+1} &:= \arg \max_\sigma Q((\alpha^k, b^k, \nu^k, \sigma)|\Psi^k).
 \end{aligned} \tag{6.17}$$

Note that the difference between (6.16) and (6.17) is substantial. Without the parameter independence assumption, the iterative search of maximizers cannot be separated and the implementation must rely on ad-hoc optimization techniques. For example this is the case of [146] with $J = 2$, where $\theta_1 = (\nu_1, \sigma)$ and $\theta_2 = (\nu_2, \sigma)$, the parameters of two Rician distributions, are obviously not independent. In the sequel, we take advantage of (6.17) and we show that each partial maximization step can be performed by updating the corresponding variable according to an iterative rule.

6.3.2 Iterative equations for the EM algorithm

In this section the problem of numerically implementing (6.17) is addressed. By defining $\ell(\Psi) := Q(\Psi|\Psi)$, the iterative procedure (6.17) can be instantiated by an iterative method attempting to solve

$$\nabla \ell(\Psi) = 0. \tag{6.18}$$

By writing gradient equations and performing math (see Appendix 6.A.2 for details) we get

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha} &= \sum_{x \in \mathbf{x}} \frac{1}{\alpha} p(\omega_1|x, \Psi) - \frac{1}{1-\alpha} p(\omega_2|x, \Psi) \\
\frac{\partial \ell}{\partial b} &= \sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi) \left[\frac{x^2}{b^3} - \frac{2}{b} \right] \\
\frac{\partial \ell}{\partial \nu} &= \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left[\frac{x}{\sigma^2} \frac{I_1\left(\frac{x\nu}{\sigma^2}\right)}{I_0\left(\frac{x\nu}{\sigma^2}\right)} - \frac{\nu}{\sigma^2} \right] \\
\frac{\partial \ell}{\partial \sigma} &= \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left[\frac{x^2 + \nu^2}{\sigma^3} - \frac{2}{\sigma} - \frac{2x\nu}{\sigma^3} \frac{I_1\left(\frac{x\nu}{\sigma^2}\right)}{I_0\left(\frac{x\nu}{\sigma^2}\right)} \right]
\end{aligned} \tag{6.19}$$

where $I_1(\cdot)$ is the 1-st order modified Bessel function of first kind [151]. As we can see, gradient equations are highly non-linear. Formally, we derive a set of iterative equations for approximating the solution of (6.18) according to the method of subsequent approximations (see Appendix 6.A.4 for details). After some analytical manipulations we get the following iterative rules

$$\begin{aligned}
\alpha^{k+1} &= N^{-1} \sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi^k) \\
(b^2)^{k+1} &= \frac{\sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi^k) x^2}{2 \sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi^k)} \\
\nu^{k+1} &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi^k) \frac{I_1\left(\frac{x\nu^k}{(\sigma^k)^2}\right)}{I_0\left(\frac{x\nu^k}{(\sigma^k)^2}\right)} x}{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi^k)} \\
(\sigma^2)^{k+1} &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi^k) \left[x^2 + (\nu^k)^2 - 2x\nu^k \frac{I_1\left(\frac{x\nu^k}{(\sigma^k)^2}\right)}{I_0\left(\frac{x\nu^k}{(\sigma^k)^2}\right)} \right]}{2 \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi^k)}.
\end{aligned} \tag{6.20}$$

The above formulas fully determine our algorithm. In general, the method of subsequent approximations converges at least with linear speed provided the spectral radius of the Jacobian matrix of the iterative function, computed at the exact solution, is strictly less than one. Of course, convergence properties of the algorithm strongly depend on the choice of the first iterate.

6.3.3 Initialization of the algorithm

In order to increase the probability to converge to an optimal stationary point of the objective energy $\ell(\Psi)$, an adequate initial approximation Ψ^0 must be found. A standard approach to this aim is based on a first raw classification of the data followed by maximum likelihood (ML) estimates of the parameters. Let us assume that the samples drawn from the mixture are divided into the two approximate classes W_1 and W_2 , in such a way that $\mathbf{x} = W_1 \cup W_2$. ML estimates of shape parameters (in the following denoted by the ML superscript) are derived as the solutions of the so-called log-likelihood equations. Samples in W_1 are used to approximate b , whereas samples in W_2 are used to approximate ν, σ . For an explicit computation of their values we refer the reader to the existing literature [138, 140, 142, 145]. Once the ML estimates $b^{ML}, \nu^{ML}, \sigma^{ML}$ are computed, we use their values as the initial set of parameters Θ^0 for triggering the EM algorithm:

$$\Theta^0 = (b^{ML}, (\nu^{ML}, \sigma^{ML})). \quad (6.21)$$

As initial value for the mixing proportion we use the ratio ($\#$ denotes the number of samples)

$$\alpha^0 = \frac{\#W_1}{\#X}. \quad (6.22)$$

According to the method that is used to populate the approximate classes W_1, W_2 , we have different initializations of the EM algorithm. The more these classes are good representatives of the true classes ω_1, ω_2 , the more the ML parameter estimates of the two distributions $p(\rho|\omega_h, \theta_h)$, $h = 1, 2$ will be close to the real values. A simple yet effective way to populate such approximate classes is given by thresholding the values of the samples \mathbf{x} (in [82] a similar approach is used in the case of a mixture of Gaussian densities). Let us define, for any fixed value $T \geq 0$, the approximate classes W_1, W_2 as follows

$$\begin{aligned} W_1 &:= \{x \in X : x \leq T\}, \\ W_2 &:= \{x \in X : x > T\}. \end{aligned} \quad (6.23)$$

In terms of Bayes decision theory, the choice of the threshold T can be interpreted as an attempt of approximating a solution of (6.7). In this context, the problem of defining the approximate classes W_1, W_2 is turned into the problem of choosing the separating threshold T . At this stage, proper knowledge on the specific dataset should be used in order to simplify the task of computing T . On the one hand, such T should be able to give at least a coarse discrimination of the data, e.g., by properly exploiting the bi-modal behavior of the histogram. On the other hand, the computational complexity of this step, being itself a preliminary step, should be kept low. A choice often encountered in

applications that meets the two above mentioned important requirements [82], is that of using

$$T = T_{mid} := \frac{\max X - \min X}{2}. \quad (6.24)$$

Whenever the two modes of the mixture described by (6.10) are well separated by T_{mid} , we expect to have sufficiently accurate preliminary ML estimates of the mixture parameters for triggering the EM algorithm. Of course, other strategies can be used.

6.3.4 Convergence analysis

In this section we analyze convergence properties of the parameter estimation algorithm defined by (6.20). For a better understanding of the quantitative analysis of the results, we specify that computations are performed using MATLAB[®] on a standard workstation. Hardware is Intel(R) Core(TM) i5-4750 CPU @3.20 GHz, 8.00 GB Ram.

According to [153], we aim at showing that the performance of the proposed EM-type iterative algorithm strongly depends on the *separability* of the two mixture components. The proposed iterative algorithm is run several times on synthetic samples \mathbf{x} generated by the inverse transform sampling method applied to mixture densities of the type $p(\rho|\Psi)$, where $\Psi = (\alpha, b, \nu, \sigma)$ are fixed sets of parameters. The size of each sample is $N = 10^4$. In order to parameterize the separability of the mixture components, in the tests all parameters are fixed except ν . Significance of the resulting estimates is ensured by stopping the algorithm at the iteration k such that the maximum relative error in approximating all the parameters is

$$\max_{i=1,\dots,4} \left| \frac{\Psi_i - \Psi_i^k}{\Psi_i} \right| < 0.05.$$

For each test, we recorded the number of iterations k , the relative variation of the objective energy $|\ell(\Psi^k) - \ell(\Psi^{k-1})|/|\ell(\Psi^{k-1})|$, the spectral radius of the Jacobian of the iterative function calculated at Ψ and the time of computation. The components of the Jacobian matrix are calculated by implementing (6.34).

The results relate to two sets of parameters Ψ . In the first case the fixed values are $\alpha = 0.4$, $b = 1$, $\sigma = 1$, in the second case the Rician scale parameter is increased to $\sigma = 2$. In both cases the Rician non-centrality parameter ranges from $\nu = 2.0$ to $\nu = 10.0$. Results of computations (the mean of ten runs) are shown in Tables 6.2 and 6.3. Histograms of the corresponding samples are given in Figures 6.2 and 6.3. Results are in agreement with the expected performance of the EM algorithm. Let us discuss more in detail the outcome of the experiments.

- An analysis of Figures 6.2 and 6.3 enables us to discriminate the separability of the corresponding mixtures. In the case of $\sigma = 1$, being also $b = 1$, the separation

Table 6.2: Iteration details of the proposed EM algorithm on Rayleigh-Rice mixtures with parameters $\alpha = 0.4$, $b = 1$, $\sigma = 1$ and for different values of ν .

ν	k	$\frac{ \ell(\Psi^k) - \ell(\Psi^{k-1}) }{ \ell(\Psi^{k-1}) }$	$rad(J\varphi(\Psi))$	$time(secs)$
2.0	3694	$1.07 \cdot 10^{-5}$	0.7728	35.93
2.2	2969	$2.26 \cdot 10^{-6}$	0.7419	30.13
2.4	1217	$1.14 \cdot 10^{-5}$	0.7034	13.65
2.6	800	$1.33 \cdot 10^{-5}$	0.6723	8.44
2.8	427	$9.14 \cdot 10^{-6}$	0.6521	4.84
3.0	266	$1.10 \cdot 10^{-5}$	0.6282	3.12
3.5	75	$2.03 \cdot 10^{-5}$	0.5963	1.07
4.0	22	$4.90 \cdot 10^{-4}$	0.5701	0.69
5.0	8	$1.29 \cdot 10^{-3}$	0.5426	0.55
10.0	1	$7.36 \cdot 10^{-5}$	0.5133	0.47

Table 6.3: Iterations details of the proposed EM algorithm on Rayleigh-Rice mixtures with parameters $\alpha = 0.4$, $b = 1$, $\sigma = 2$ and for different values of ν .

ν	k	$\frac{ \ell(\Psi^k) - \ell(\Psi^{k-1}) }{ \ell(\Psi^{k-1}) }$	$rad(J\varphi(\Psi))$	$time(secs)$
2.0	352	$4.37 \cdot 10^{-6}$	0.8250	3.28
2.2	408	$5.90 \cdot 10^{-6}$	0.8424	2.99
2.4	304	$1.03 \cdot 10^{-5}$	0.8766	2.86
2.6	229	$2.34 \cdot 10^{-5}$	0.8827	2.21
2.8	186	$4.23 \cdot 10^{-5}$	0.8796	2.29
3.0	130	$9.09 \cdot 10^{-5}$	0.8625	1.72
3.5	91	$1.57 \cdot 10^{-4}$	0.8232	1.20
4.0	74	$1.06 \cdot 10^{-4}$	0.7781	1.13
5.0	28	$5.49 \cdot 10^{-4}$	0.6874	0.74
10.0	4	$5.41 \cdot 10^{-4}$	0.5447	0.54

becomes more evident only for $\nu \geq 2.8$. In the case of $\sigma = 2$, the difference in the scale parameters allows for a better discrimination for any $\nu \geq 2.0$.

- As expected, the number of iterations for reaching sufficient approximations of the target parameters sensibly decreases as ν increases. In particular, in Table 6.2 the order of k moves from 10^3 ($\nu = 2.0$) to 10^0 ($\nu = 5.0, 10.0$), whereas in Table 6.3 this number is one order of magnitude less. This substantial difference is due to the

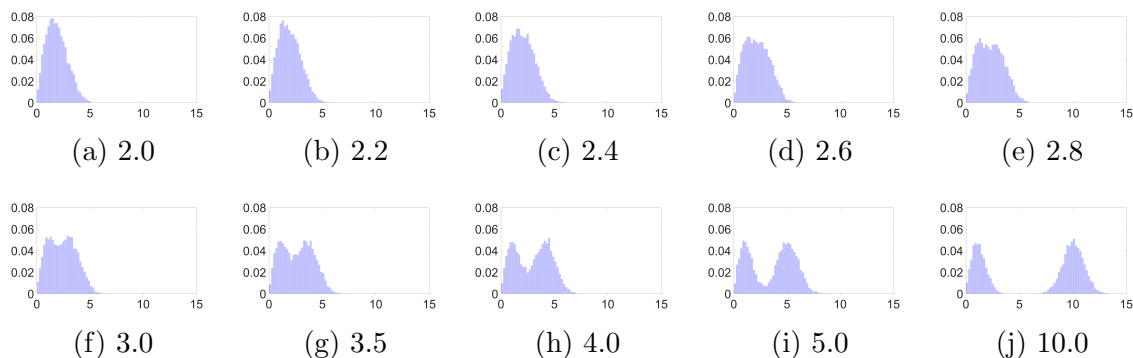


Figure 6.2: Histograms of samples \mathbf{x} generated from $p(\rho|\Psi)$ with $\alpha = 0.4$, $b = 1$, $\sigma = 1$ and for different values of ν .

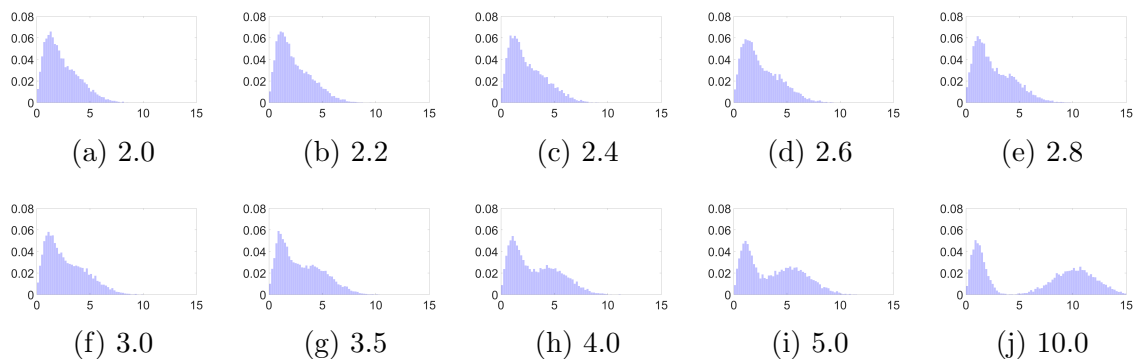


Figure 6.3: Histograms of samples \mathbf{x} generated from $p(\rho|\Psi)$ with $\alpha = 0.4$, $b = 1$, $\sigma = 2$ and for different values of ν .

fact that, in the second case, the separation between the two mixture components is possible because of the difference in the scale parameters.

- In all tests the spectral radius of the Jacobian of the iterative function is less than one, thus the algorithm enjoys at least linear convergence. Notice that for increasing ν , the spectral radius (which provides a quantitative estimation of the factor by which errors are reduced from one iteration to the next) decreases.
- Results from the tables suggest that a reasonable threshold value on the relative variation of the objective energy $\ell(\Psi)$ for stopping the algorithm is $tol = 10^{-6}$.

6.4 Experimental results on multispectral images

In this section, after presenting the datasets and the details of the experimental setup, we analyze the performance of the proposed method. Firstly, we give a quantitative measure-

ment of data fitting by means of two divergence measures between the data and different statistical models estimated from the data. Then, the change detection performance is analyzed. A comparison in terms of change detection errors between the proposed method and the state-of-the-art one based on Gaussian mixtures is given.

6.4.1 Datasets description

The datasets considered in the experiments are both synthetic and real multispectral images. Real datasets (consisting of couples of multitemporal multispectral images) are accompanied by reference maps⁴ of the changes, i.e., binary maps representing the classes ω_n, ω_c . Notice that minor changes (i.e., mapped changes that do not belong to the main change class of interest) are depicted in the reference maps in red (Figures 6.5c and 6.6c).

Dataset A (synthetic)

This first dataset is a synthetically generated two-band difference image with statistical properties as defined in Section 6.2.1, see Figure 6.4. Classes ω_n, ω_c are Gaussian distributed within each band with parameters: $\mu_{1,n} = \mu_{2,n} = 0$, $\sigma_n = 2.5$, $\mu_{1,c} = -50.0$, $\mu_{2,c} = -20.0$, $\sigma_c = 25$. The size of the image is 700×600 pixels. The proportions between the number of pixels in simulated classes and the total number of pixels (class priors) are

$$p(\omega_n) = \frac{336000}{420000} = 0.8, \quad p(\omega_c) = \frac{84000}{420000} = 0.2.$$

Simulated changed pixels are located in the bottom-right corner of the image.

Dataset B

The dataset is made up of two multispectral images acquired by the Thematic Mapper (TM) multispectral sensor of the Landsat 5 satellite, see Figure 6.5. The images are co-registered and radiometrically corrected. The scene represents an area including Lake Mulargia (Sardinia Island, Italy), at a resolution of 30 m. The image consists of 300×412 pixels (a total of 123600 pixels). The dates of acquisition are September 1995 (t_1) and July 1996 (t_2). Between the two dates one most relevant change, which is related to the extension of the lake surface, occurred in the study area. The scene presents 116120 unchanged and 7480 changed pixels. It follows that the prior probabilities of class are given by

$$p(\omega_n) = \frac{116120}{123600} = 0.94, \quad p(\omega_c) = \frac{7480}{123600} = 0.06.$$

⁴Reference maps are obtained via photo-interpretation and they are used only for validation/comparison purposes.

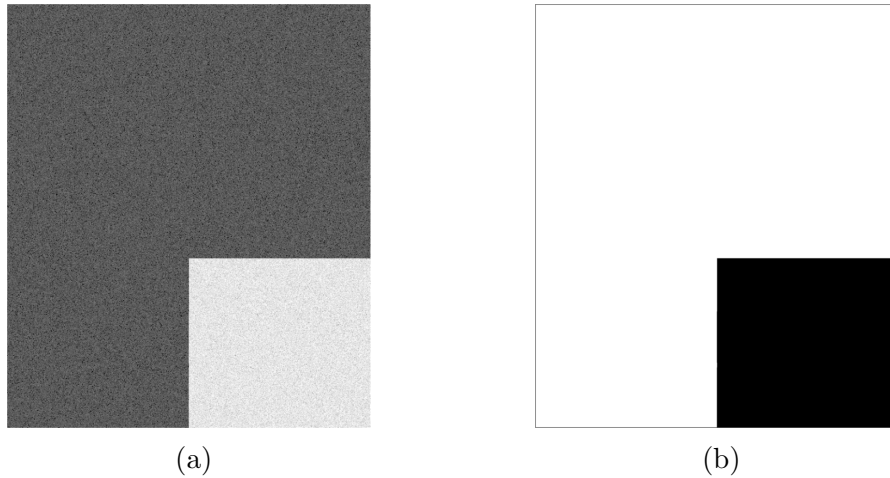


Figure 6.4: Dataset A. Synthetic two-band difference image. (a) Magnitude of the difference image, (b) map of simulated changed pixels (black).

The two most representative bands of the changes are 4 and 7, the near infrared (NIR) and the middle infrared (MIR) (see [78] for details on the band selection). The ML parameter estimations of the normal distributions of classes ω_n, ω_c within each band of the difference image are $\mu_{1,n} = 3.57$, $\sigma_{1,n} = 10.26$, $\mu_{1,c} = -55.37$, $\sigma_{1,c} = 8.90$, and $\mu_{2,n} = 2.63$, $\sigma_{2,n} = 8.73$, $\mu_{2,c} = -40.84$, $\sigma_{2,c} = 10.67$. The numbers show that the initial assumptions (6.2) and (6.4) are approximately satisfied. The means $\mu_{1,n}$ and $\mu_{2,n}$ are both close to 0, $\sigma_{1,n}$ and $\sigma_{2,n}$ are very close each other and their mean can be used as an approximation of the variance of the unchanged pixels, i.e., $\sigma_n = 9.49$. A similar argument holds for the variance of the changed pixels, which can be approximated by $\sigma_c = 9.79$, the mean value of $\sigma_{1,c}$ and $\sigma_{2,c}$.

Dataset C

The dataset consists of a couple of multispectral images acquired by the Operational Land Imager (OLI) multispectral sensor of the Landsat 8 satellite, see Figure 6.6. The investigated area includes Lake Omodeo and a portion of Tirso River (Sardinia Island, Italy). The image consists of 700×650 pixels (a total of 455000 pixels) at a resolution of 30 m. The dates of acquisition are 25th July 2013 (t_1) and 10th August 2013 (t_2). The change we are interested to estimate is a fire occurred between August 7th and 9th in the south of Ghilarza village. The post-event image is acquired just one day after the fire was extinguished. The area affected by the fire is mostly agricultural, with an extension of approximately 100 ha. The images are co-registered and radiometrically corrected. According to the reference map, the scene presents 420227 unchanged and 34773 changed

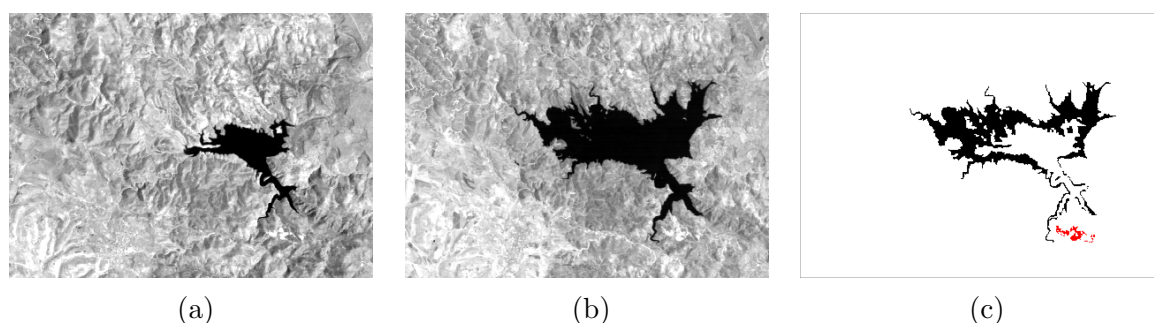


Figure 6.5: Dataset B: images of Lake Mulargia (Italy) acquired by the Thematic Mapper sensor of the Landsat 5 satellite: (a) channel 4 of the image acquired in September 1995; (b) channel 4 of the image acquired in July 1996; (c) change reference map indicating the enlargement of the lake (black) and an open quarry (red).

pixels, 1636 of them are related to small clouds and variations of the lake surface. The class prior probabilities are given by

$$p(\omega_n) = \frac{420227}{455000} = 0.92, \quad p(\omega_c) = \frac{34773}{455000} = 0.08.$$

The two bands selected as most representative of the changes are bands 5 and 6, the near infrared (NIR) and the first short wavelength infrared (SWIR1). The choice is made by considering the band-pair that produced the smallest amount of overall errors in the change detection.

6.4.2 Experimental setup

Given two multispectral two-band images $\mathbf{y}^1, \mathbf{y}^2$, by following the framework given in Section 6.2 the magnitude ρ of the difference image $\mathbf{d} = \mathbf{y}^2 - \mathbf{y}^1$ is modeled by the mixture density (6.6). The parameters of this density are estimated via EM algorithm. Then, a threshold for binary decision is calculated by following a Bayes rule and change detection is performed accordingly.

In the experiments presented in Section 6.4.3, the fitting of the proposed Rayleigh-Rice mixture model is tested against two empirical models for parameter estimation already present in literature: the first one is based on a Gaussian mixture, the second one is based on a mixture of Rician distributions with common scale parameter. It is worth noting that, the former represents the state-of-the-art in binary change detection based on the statistical modeling of the magnitude information, whereas the latter has been proposed by Maitra and Faden for the estimation of noise variance in MR images [146]. For clarity of notation, let us summarize the parameter notation used in the following:

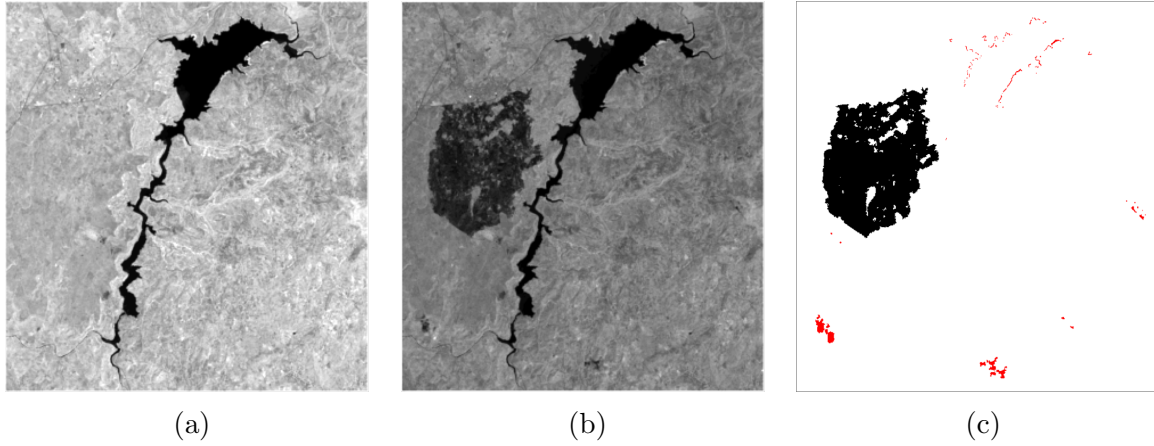


Figure 6.6: Dataset C: images of Lake Omodeo and surrounding area (Italy) acquired by the Operational Land Imager sensor of the Landsat 8 satellite: (a) channel 5 of the image acquired in July 2013; (b) channel 5 of the image acquired in August 2013; (c) change reference map indicating the burned area extension (black) and other minor changes related to clouds and water (red).

Rayleigh-Rice mixture (RR):

$$p(\rho|\omega_n) = \text{Rayl}(b), \quad p(\rho|\omega_c) = \text{Rice}(\nu, \sigma).$$

Gaussian mixture (GG):

$$p(\rho|\omega_n) = \mathcal{N}(\mu_1, \sigma_1), \quad p(\rho|\omega_c) = \mathcal{N}(\mu_2, \sigma_2).$$

Rician mixture with common scale parameter (MF):

$$p(\rho|\omega_n) = \text{Rice}(\nu_1, \beta), \quad p(\rho|\omega_c) = \text{Rice}(\nu_2, \beta).$$

Iterative formulas for implementing parameter estimation in the GG model can be found in [86]. Implementation for parameter estimation in the MF case is presented in [146]. Initialization follows the principles outlined in Section 6.3.3. Data fitting is measured in terms of the χ^2_P (χ^2 -Pearson) divergence and the Kolmogorov-Smirnov (KS) distance between the data and the estimated densities.

In Section 6.4.4 the binary change detection problem is solved using the framework described in Section 6.2. The final parameter estimates of RR and GG are used to compute the magnitude thresholds T^{RR} and T^{GG} that correspond to the Bayes Decision Rule (BDR) of minimum overall error of classification, then change detection is performed by thresholding the difference image. By using a standard trial-and-error selection based on the reference map it has been possible to compute for all datasets the

optimal threshold value T^{MOE} of minimum overall error (MOE). The results of change detection are compared to such optimal values. In order to make the results on the different datasets comparable, the detection errors are also given in terms of percentage according to $e_m := \frac{\text{missed}}{\text{changed}} \cdot 100\%$, $e_f := \frac{\text{false}}{\text{unchanged}} \cdot 100\%$ and $e_o := \frac{\text{overall}}{\text{total}} \cdot 100\%$. The steps of the whole procedure are summarized:

1. Populate the approximate classes W_1, W_2 and initialize the EM algorithm (in the MF case the algorithm is initialized as described in [146]).
2. Apply EM algorithm and assess fitting properties of the estimated mixture densities.
3. Calculate the threshold values according to both the BDR of minimum overall error and the optimal choice of MOE obtained via the reference map.
4. Populate the change detection classes W_n, W_c by thresholding the magnitude image.
5. Assess change detection performance in terms of false and missed alarms by comparing the estimated classes W_n, W_c with the true classes ω_n, ω_c .

6.4.3 Data fitting results

Let us present and discuss the results of EM parameter estimation and data fitting in the considered datasets. Numerical values of the estimated parameters and fitting performance in terms of statistical divergences χ_p^2 and KS are showed in Table 6.4. For a qualitative understanding of the fitting of the estimated models, a plot of the histograms of the magnitude images with superimposed estimated densities is given in Figure 6.7.

In general, the real prior probability of the unchange class $p(\omega_n)$ is always well approximated by α as no significant differences are observed among the trials. As expected, the fitting of RR is very precise in the case of Dataset A, as this dataset represents the ideal case of a difference image having the properties described in Section 6.2. Nonetheless, the RR model results in the best fitting also in the real remote sensing datasets B and C, confirming that this model is much more suitable for representing the real distribution of the magnitude. The GG model is flexible enough to follow the bimodal behavior of the histogram, but it is never as precise as the RR. The limitations in using the MF model are evident. From the EM estimates of the RR model we can see that the scale parameters of the mixture components are significantly different in all datasets. This strongly affects the fitting of the MF model, which assumes the same scale parameter for both the mixture components. Data fitting measurements obtained with this model show that this strong assumption leads to very poor approximations of the real distributions (this can be

Table 6.4: Parameter estimation via EM algorithm and data fitting evaluation for the three considered mixture densities.

mix	estimated parameters					χ_P^2	KS
Dataset A							
RR	α	b	ν	σ		0.0002	0.0004
	0.80	2.50	53.89	25.04			
GG	α	μ_1	σ_1	μ_2	σ_2	0.0184	0.0561
	0.80	3.11	1.61	59.34	23.95		
MF	α	ν_1	ν_2	β		0.9045	0.5142
	0.83	≈ 0	66.32	7.20			
Dataset B							
RR	α	b	ν	σ		0.0136	0.0362
	0.92	9.36	60.24	17.37			
GG	α	μ_1	σ_1	μ_2	σ_2	0.0420	0.0836
	0.90	11.10	5.97	52.73	22.57		
MF	α	ν_1	ν_2	β		0.0239	0.0590
	0.93	0.06	66.91	9.78			
Dataset C							
RR	α	b	ν	σ		0.0215	0.0400
	0.90	2.67	23.84	14.90			
GG	α	μ_1	σ_1	μ_2	σ_2	0.0500	0.0778
	0.89	3.25	1.70	26.35	14.13		
MF	α	ν_1	ν_2	β		0.1414	0.1880
	0.93	≈ 0	34.82	3.41			

seen very clearly from Figure 6.7). Thus, the model is inadequate to address the change detection problem. Notice also that the non-centrality parameter ν_1 converged approximately to 0 in all cases, confirming that the first component of the mixture (related to the distribution of unchanged pixels) is Rayleigh in real data.

Let us now check the consistency of the RR parameter estimates with the theoretical properties described in Section 6.2.1 that express the relationships between b_n, ν_c, σ_c (approximated by b, ν, σ) and the parameters of the normal distribution of classes within bands: σ_n, σ_c and $\mu_{b,c}$, for $b = 1, 2$. The values related to Dataset A matches perfectly as

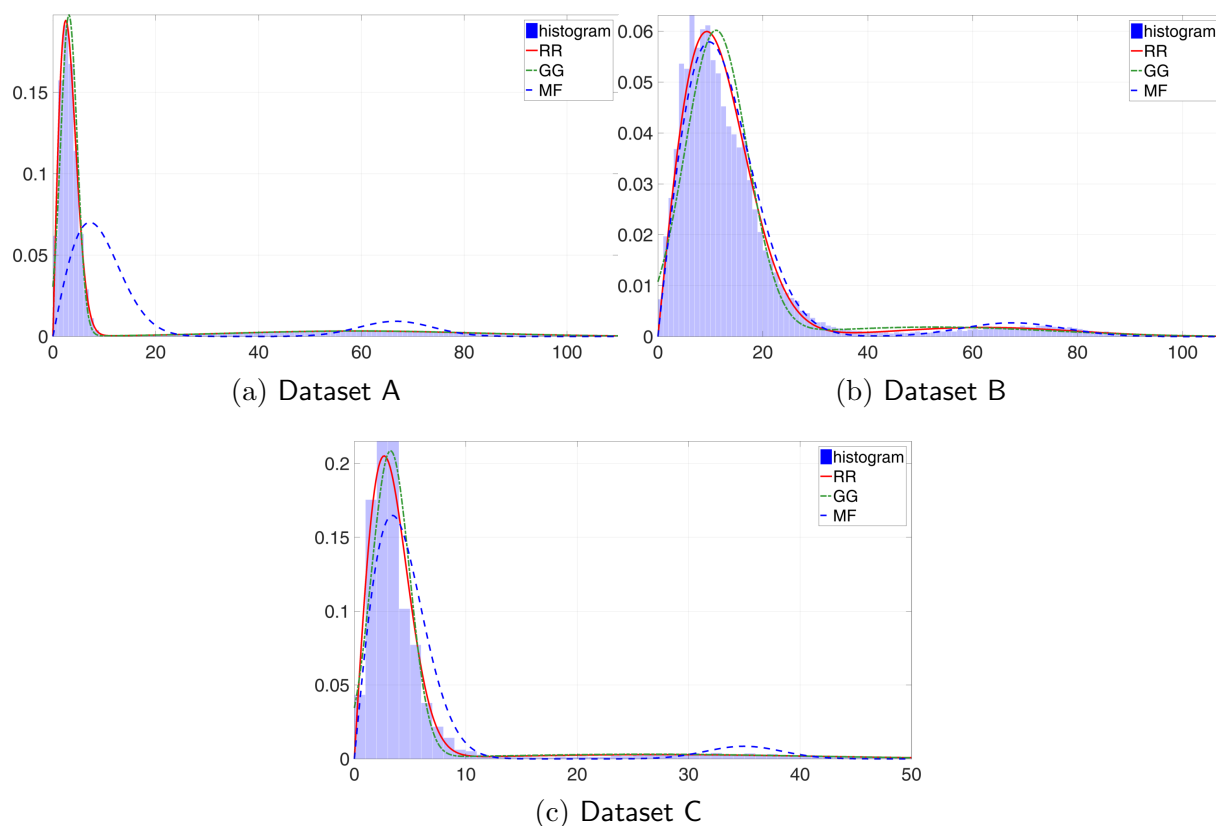


Figure 6.7: Histograms of the magnitude of the difference image and plot of the estimated densities.

the hypotheses are fulfilled by definition: we have $\sigma_n = b = 2.5$, $\sigma_c = 25$, $\sigma = 25.04$, and $\nu = 53.89$, $\sqrt{\mu_{1,c}^2 + \mu_{2,c}^2} = 53.85$. In case of Dataset B, the Rayleigh parameter $b = 9.33$ matches with the variance $\sigma_n = 9.44$. The Rice parameters are slightly unmatched, compare $\sigma = 17.37$ with the variance $\sigma_c = 9.79$ and $\nu = 60.24$ with the non-centrality measurement $\sqrt{\mu_{1,c}^2 + \mu_{2,c}^2} = 68.75$. A similar behavior is observed for Dataset C. The observed differences in real datasets are due to the Gaussian approximation of the distribution of classes within bands.

6.4.4 Change detection results

In this section the change detection results on the considered datasets are presented. The outcome of the experiments is detailed in Table 6.5, where the CD performance is evaluated in terms of false and missed alarms and overall errors. It follows from the analysis in Section 6.4.3 that the mixture component representing ω_n is better approximated by a Rayleigh density. The high non-symmetry of the Rayleigh component forces its Gaussian

approximation to be slightly overestimated on the right side, thus, its right descending slope is steeper than the Rayleigh version. It follows that the decision threshold T^{GG} is smaller than T^{RR} in general. A first consequence of this is that the overall error of RR is always smaller than the error of GG. Moreover, the change detection by using the GG mixture presents less missed alarms, but many more false alarms with respect to the RR case. This is confirmed by looking at the results related to Dataset A (synthetic). It is not surprising that in this case the BDR threshold related to the RR model is almost identical to the optimal choice of minimum overall error: $T^{RR} = 10.12$ and $T^{MOE} = 10.40$. As expected, the threshold returned by the GG model, $T^{GG} = 8.82$, is smaller than T^{RR} and corresponds to an increasing overall error that moves from 0.19% (optimal case) to 0.29%. Notice that in this case the total number of wrongly detected pixels increases of approximately 1/3 (from 791 to 1211).

In the case of real datasets error percentages slightly increase, accounting for the approximation in assuming the Gaussian distribution of the classes within bands. Note that, though in some cases the BDR thresholds are not very close to the optimal values, the RR model always returned a better approximation of the optimal threshold with respect to the standard GG model. In Dataset B, the optimal threshold in terms of overall error corresponds to 770 overall errors, and the RR and the GG models returned 1820 and 3982 overall errors, respectively. Therefore, the overall error is more than halved. A similar result is obtained in the case of Dataset C, where again the RR model performed much better than GG. The number of overall errors is 4621 in the optimal case, whereas it is 8761 in case of RR and 13583 in the case of GG. Again we observe that the overall error is more than halved by using the proposed RR model. Figure 6.8 shows the change detection maps obtained by thresholding the magnitude of the difference image in the three datasets. Notice that in general the smaller number of missed alarms given by T^{RR} results in a much less noisy change map compared to the one given by T^{GG} . In the optimal case the change maps are very clean. Since in these cases the number of missed alarms is the highest one, some details of the changes are lost (e.g., the thin boundaries of the lake in Dataset B and a portion of the fire in Dataset C). A possible improvement of the change detection maps could be obtained by using the distributions of the classes obtained by the proposed technique within a context-sensitive approach (e.g., based on MRFs [82]).

6.5 Conclusions

In this chapter we addressed the problem often encountered in image analysis of the estimation of the parameters of a Rayleigh-Rice mixture density. The problem has been

Table 6.5: Performance of change detection based on the thresholding of the magnitude image for different values of the threshold.

Threshold	missed (e_m)	false (e_f)	overall (e_o)
Dataset A			
$T^{RR} = 10.12$	699 (0.83%)	99 (0.03%)	798 (0.19%)
$T^{GG} = 8.82$	530 (0.63%)	681 (0.20%)	1211 (0.29%)
$T^{MOE} = 10.40$	735 (0.88%)	56 (0.02%)	791 (0.19%)
Dataset B			
$T^{RR} = 33.95$	36 (0.48%)	1784 (1.54%)	1820 (1.47%)
$T^{GG} = 28.02$	9 (0.12%)	3973 (3.42%)	3982 (3.22%)
$T^{MOE} = 47.11$	356 (4.76%)	414 (0.36%)	770 (0.62%)
Dataset C			
$T^{RR} = 9.92$	956 (2.75%)	7805 (1.86%)	8761 (1.93%)
$T^{GG} = 8.63$	727 (2.09%)	12856 (3.06%)	13583 (2.99%)
$T^{MOE} = 15.23$	1964 (5.65%)	2657 (0.63%)	4621 (1.02%)

studied in the framework of Change Vector Analysis (CVA) for binary change detection in multitemporal and multispectral images. Here, under proper hypotheses, the distribution of the magnitude of the difference image can be theoretically modeled by a Rayleigh-Rice mixture density. The Rayleigh density describes the distribution of unchanged pixels, whereas the Rice density describes the distribution of the changed pixels. Parameter estimates are used to solve the binary change detection problem in a Bayesian context.

The chapter presents a general implementation of an EM-type algorithm for the estimation of mixture and shape parameters of a Rayleigh-Rice mixture density. The proposed method enjoys good theoretical properties. First, statistical independence of parameters allowed us to define the algorithm in an iterative way, which results fast, easy to implement and not depending on specific optimization routines. Detailed analysis of accuracy and convergence properties of the algorithm is given. Second, in the considered model all the statistical parameters are free. We remark here that, because of its flexibility, the proposed method can be used for addressing a large variety of practical problems involving the Rayleigh-Rice density.

In the experimental part of the chapter, the effectiveness of the Rayleigh-Rice model in solving the binary CD problem is demonstrated. Tests have been conducted on both synthetic and real datasets consisting of bi-temporal pairs of multispectral remote sensing

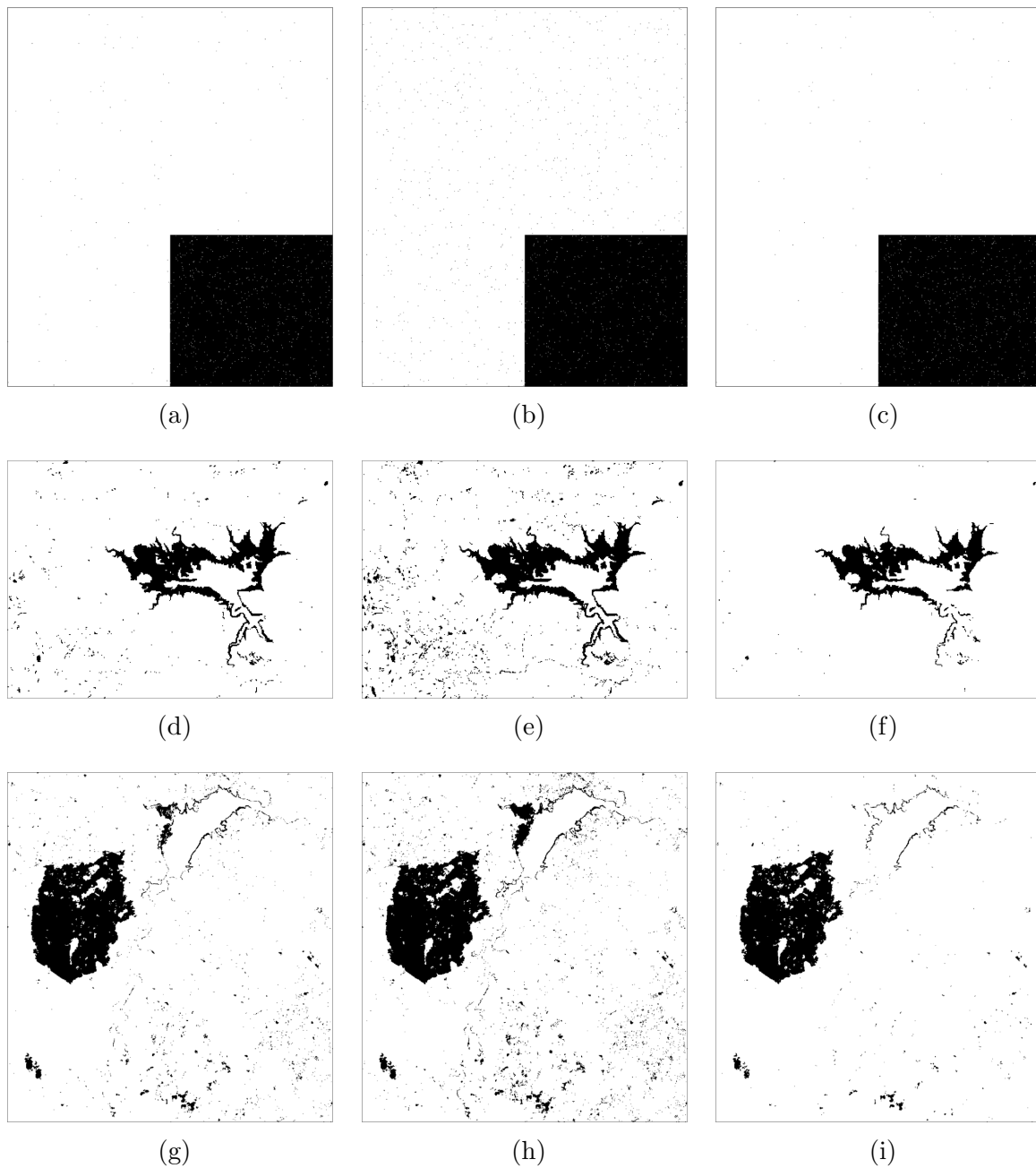


Figure 6.8: Change detection maps obtained by thresholding the magnitude image using T^{RR} (a,d,g), T^{GG} (b,e,h) and T^{MOE} (c,f,i). In black are the estimated changed pixels W_c , in white the estimated unchanged pixels W_n .

images. Among other statistical models proposed in literature, the Rayleigh-Rice proved to be the one that better fits the distribution of the magnitude difference image in CVA. The change detection is significantly improved when compared to state-of-the-art method

based on Gaussian modeling. In particular, the overall error of detection is always approximately halved. Moreover, even though the theoretical model is given for one single change class, the algorithm is robust to the presence of minor additional changes.

6.6 Acknowledgments

This work was supported by the Italian Ministry of Education, University, and Research (MIUR) under Grant 2012L48PE5 (PRIN 2012).

6.A Appendices

6.A.1 Basic notions about modified Bessel functions

We recall here some facts about modified Bessel functions which are crucial for the development of the method that is presented in this chapter. Let $I_m(x)$ be the m -th order modified Bessel function of first kind [151]. Derivatives of $I_m(x)$ satisfy (among many other) the following recurrence rules:

$$\frac{d}{dx} I_0(x) = I_1(x), \quad (6.25)$$

$$\frac{d}{dx} (x^m I_m(x)) = x^m I_{m-1}(x). \quad (6.26)$$

In order to simplify the presentation of the results of next sections, we define two functions that appear frequently in the computations:

$$J_1 := \frac{I_1}{I_0} \text{ and } J_2 := J_1^2 - 1. \quad (6.27)$$

6.A.2 Derivation of the conditional log-expectation

Let us derive the energy $\ell(\Psi)$ with respect to its parameters to obtain (6.39). Computations for $\partial/\partial b$ are the same to that of $\partial/\partial\sigma$ if σ is replaced by b and $\nu = 0$. It is a trivial fact that by deriving $\ell(\Psi)$ with respect to the Rice parameters the terms related to the Rayleigh distribution vanish. Therefore, without loss of generality we can consider

$$\ell(\Psi) = \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \log((1 - \alpha) p(x|\omega_2, \theta_2)). \quad (6.28)$$

Because of summation over x and the log, we can write

$$\ell(\Psi) = \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) (\log A + \log B), \quad (6.29)$$

where $A := \frac{x}{\sigma^2} \exp\left(-\frac{x^2 + \nu^2}{2\sigma^2}\right)$ and $B := I_0\left(\frac{x\nu}{\sigma^2}\right)$. Derivations involving A and B can be computed by using (6.25) and by applying the chain rule. We get

$$\begin{aligned} \frac{\partial \log A}{\partial \nu} &= -\frac{\nu}{\sigma^2}, & \frac{\partial \log A}{\partial \sigma} &= \frac{x^2 + \nu^2}{\sigma^3} - \frac{2}{\sigma}, \\ \frac{\partial \log B}{\partial \nu} &= \frac{x}{\sigma^2} J_1\left(\frac{x\nu}{\sigma^2}\right), & \frac{\partial \log B}{\partial \sigma} &= -\frac{2x\nu}{\sigma^3} J_1\left(\frac{x\nu}{\sigma^2}\right). \end{aligned}$$

Final derivatives are obtained by putting together the results.

6.A.3 Asymptotic analysis of the maximum-likelihood estimation

The Rayleigh-Rice mixture density $p(\rho|\Psi)$ is twice differentiable with respect to its parameters and enjoys some regularity properties that allow us to write its information matrix (see Lemma 5.3 of [154] and its generalization to the multi-parameter case, p.125) as

$$I(\Psi)_{i,j} = -E \left[\frac{\partial^2}{\partial \psi_i \partial \psi_j} \log p(\rho|\Psi) \Big| \Psi \right] \quad i, j = 1, \dots, 4 \quad (6.30)$$

where we recall that $\Psi = (\alpha, b, \nu, \sigma)$. Since the information matrix of the Rician distribution is diagonal [137], also $I(\Psi)$ is diagonal. Thus, the parameters of the mixture density are mutually orthogonal.

Orthogonality brings some nice statistical properties [155]: (1) asymptotic estimates of the parameters are independent; (2) the asymptotic standard error for estimating one parameter does not depend on the knowledge of the others; (3) the maximum likelihood estimate of ψ_i given ψ_j varies only slowly with ψ_j . All these properties are fundamental for the formalization of an algorithm that iteratively updates the ML estimates of the mixture parameters.

6.A.4 Iterative method for the solution of the non linear gradient system

Iterative formulas for approximating the solution of the gradient system (6.18) are given according to the method of subsequent approximations applied to a fixed point equation [156]. After some analytical manipulations of equations (6.39) we can see that a point Ψ^* satisfying $\nabla \mathcal{L}(\Psi^*) = 0$ also satisfies $\Psi^* = \varphi(\Psi^*)$, where $\varphi(\Psi)$, the so-called iterative

function, is made up of four components:

$$\begin{aligned}
 \varphi_1(\Psi) &= \frac{\sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi)}{\sum_{x \in \mathbf{x}} 1} \\
 \varphi_2(\Psi) &= \sqrt{\frac{\sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi)x^2}{2 \sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi)}} \\
 \varphi_3(\Psi) &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi)J_1\left(\frac{x\nu}{\sigma^2}\right)x}{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi)} \\
 \varphi_4(\Psi) &= \sqrt{\frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Theta) \left[x^2 + \nu^2 - 2x\nu J_1\left(\frac{x\nu}{\sigma^2}\right)\right]}{2 \sum_{x \in \mathbf{x}} p(\omega_2|x, \Theta)}}.
 \end{aligned} \tag{6.31}$$

A small remark is needed here. Since $0 < I_1(y)/I_0(y) < 1$ for every $y > 0$, the argument of the square root in (6.31) is always non-negative, hence φ_4 is well defined.

Given an initial point Ψ^0 , the iterative method for solving the fixed point equation $\Psi = \varphi(\Psi)$ consists in applying the iterative rule

$$\Psi^{k+1} = \varphi(\Psi^k). \tag{6.32}$$

The method of subsequent approximations is shown to be convergent with at least linear speed whenever the spectral radius of the Jacobian matrix $J\varphi$, which in our case reduces to

$$J\varphi = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial \varphi_3}{\partial \nu} & \frac{\partial \varphi_3}{\partial \sigma} \\ 0 & 0 & \frac{\partial \varphi_4}{\partial \nu} & \frac{\partial \varphi_4}{\partial \sigma} \end{pmatrix}, \tag{6.33}$$

computed at the exact solution Ψ^* , is less than one. The computation of these non-zero

derivatives can be performed by applying (6.26). We have

$$\begin{aligned}
\frac{\partial \varphi_3}{\partial \nu}(\Psi) &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left\{ -\frac{x^2}{\sigma^2} J_2\left(\frac{x\nu}{\sigma^2}\right) - \frac{x}{\nu} J_1\left(\frac{x\nu}{\sigma^2}\right) \right\}}{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi)} \\
\frac{\partial \varphi_3}{\partial \sigma}(\Psi) &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left\{ \frac{2x^2\nu}{\sigma^3} J_2\left(\frac{x\nu}{\sigma^2}\right) + \frac{2x}{\sigma} J_1\left(\frac{x\nu}{\sigma^2}\right) \right\}}{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi)} \\
\frac{\partial \varphi_4}{\partial \nu}(\Psi) &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left\{ \frac{x^2\nu}{\sigma^2} J_2\left(\frac{x\nu}{\sigma^2}\right) + 1 \right\}}{\varphi_4(\Psi) \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi)} \\
\frac{\partial \varphi_4}{\partial \sigma}(\Psi) &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left\{ \frac{x^2\nu^2}{\sigma^3} J_2\left(\frac{x\nu}{\sigma^2}\right) + \frac{x\nu}{\sigma} J_1\left(\frac{x\nu}{\sigma^2}\right) \right\}}{\varphi_4(\Psi) \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi)}
\end{aligned} \tag{6.34}$$

The spectral radius $rad(J\varphi(\Psi))$ is defined as the maximum absolute value of the eigenvalues of $J\varphi(\Psi)$. The spectral radius is a measure of the rate at which the error in approximating the fixed point reduces between two consecutive iterations.

The fixed point iteration can be accelerated to quadratic convergence by using the more general Newton's method (or one of its many derivations)

$$\Psi^{k+1} = \Psi^k - H\ell(\Psi^k)^{-1}\ell(\Psi^k), \tag{6.35}$$

where $H(\cdot)$ is the Hessian operator. However, a low number of iterations is at the cost of one matrix inversion for each iteration, which may not be desirable in applications. In particular, for CD in multispectral images the simple iteration (6.32) performed excellently without the need of accelerators.

6.A.5 The EM algorithm for a mixture of generalized Rayleigh and Rician distributions

Let $y = (y_1, \dots, y_n)$ a random normal vector $y \sim \mathcal{N}(\mu, \Sigma)$ with scalar covariance matrix $\Sigma = \sigma^2 I_n$. The distribution of the non-negative random variable $x = |y|$ is derived in [157] and presents two analytical forms depending on whether the mean μ is null or not. If $\mu = 0$, the distribution of x is generalized Rayleigh

$$p(x) = \frac{2x^{n-1}}{(2\sigma^2)^m \Gamma(m)} \exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{6.36}$$

whereas it is generalized Rician if $\mu \neq 0$

$$p(x) = \frac{x^m}{\sigma^2 \nu^{m-1}} \exp\left(-\frac{x^2 + \nu^2}{2\sigma^2}\right) I_{m-1}\left(\frac{x\nu}{\sigma^2}\right) \tag{6.37}$$

where $m = n/2$, $\nu = |\mu|$ and $I_k(\cdot)$ is the k -th order modified Bessel function of first kind. One can also obtain (6.37) from the analytical expressions given in [158] Page 12, and using the result in [151] Page 80 formula (10) to simplify them. Some studies are proposed in literature to the estimation of the parameters of the generalized Rayleigh and Rician distributions, see for instance [159]. However, the case of a mixture of them is never considered. Therefore, we propose hereafter an extension of the method derived in Section 6.3 to the case of the generalized forms of the distributions.

Let us consider a mixture distribution where classes ω_1, ω_2 are generalized Rayleigh and Rician distributed, respectively. Then

$$p(x) = \alpha p(x|\omega_1) + (1 - \alpha) p(x|\omega_2) \quad (6.38)$$

where $0 < \alpha < 1$ is the mixing parameter. Given a set of N samples \mathbf{x} drawn from the distribution (6.38), one can estimate distribution parameters $\Psi = (\alpha, b, \nu, \sigma)$ (where b is the scale parameter of the generalized Rayleigh) by solving the non linear gradient equations associated the log-likelihood of the samples

$$\ell(\Psi) := Q(\Psi|\Psi).$$

Using useful properties of Bessel functions derivatives [151], after some math gradient equations will result as

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= \sum_{x \in \mathbf{x}} \frac{1}{\alpha} p(\omega_1|x, \Psi) - \frac{1}{1 - \alpha} p(\omega_2|x, \Psi) \\ \frac{\partial \ell}{\partial b} &= \sum_{x \in \mathbf{x}} p(\omega_1|x, \Psi) \left[\frac{x^2}{b^3} - \frac{n}{b} \right] \\ \frac{\partial \ell}{\partial \nu} &= \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left[\frac{x}{\sigma^2} \frac{I_m\left(\frac{x\nu}{\sigma^2}\right)}{I_{m-1}\left(\frac{x\nu}{\sigma^2}\right)} - \frac{\nu}{\sigma^2} \right] \\ \frac{\partial \ell}{\partial \sigma} &= \sum_{x \in \mathbf{x}} p(\omega_2|x, \Psi) \left[\frac{x^2 + \nu^2}{\sigma^3} - \frac{n}{\sigma} - \frac{2x\nu}{\sigma^3} \frac{I_m\left(\frac{x\nu}{\sigma^2}\right)}{I_{m-1}\left(\frac{x\nu}{\sigma^2}\right)} \right] \end{aligned} \quad (6.39)$$

where $I_k(\cdot)$ is the k -th order modified Bessel function of first kind. Formally, we derive a set of iterative equations for approximating the solution of $\nabla \ell(\Psi) = 0$ according to the method of subsequent approximations. After some analytical manipulations we get the

following iterative rules

$$\begin{aligned}
\alpha^{k+1} &= N^{-1} \sum_{x \in \mathbf{x}} p(\omega_1 | x, \Psi^k) \\
(b^2)^{k+1} &= \frac{\sum_{x \in \mathbf{x}} p(\omega_1 | x, \Psi^k) x^2}{n \sum_{x \in \mathbf{x}} p(\omega_1 | x, \Psi^k)} \\
\nu^{k+1} &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2 | x, \Psi^k) \frac{I_m\left(\frac{x\nu^k}{(\sigma^k)^2}\right)}{I_{m-1}\left(\frac{x\nu^k}{(\sigma^k)^2}\right)} x}{\sum_{x \in \mathbf{x}} p(\omega_2 | x, \Psi^k)} \\
(\sigma^2)^{k+1} &= \frac{\sum_{x \in \mathbf{x}} p(\omega_2 | x, \Psi^k) \left[x^2 + (\nu^k)^2 - 2x\nu^k \frac{I_m\left(\frac{x\nu^k}{(\sigma^k)^2}\right)}{I_{m-1}\left(\frac{x\nu^k}{(\sigma^k)^2}\right)} \right]}{n \sum_{x \in \mathbf{x}} p(\omega_2 | x, \Psi^k)}.
\end{aligned} \tag{6.40}$$

The above formulas fully determine our algorithm. In general, the method of subsequent approximations converges at least with linear speed provided the spectral radius of the Jacobian matrix of the iterative function, computed at the exact solution, is strictly less than one. Of course, convergence properties of the algorithm strongly depend on the choice of the first iterate.

Chapter 7

A compound multi-class mixture model for change detection

The advent of new generation multispectral sensors with high radiometric resolution has given rise to new challenges in the development of automatic unsupervised change detection (CD) techniques. In particular, typical approaches to CD are not able to well model and properly exploit the effects of the increased radiometric resolution that characterizes new data, which results in a larger number of natural classes visible in the distribution of each single spectral band. This chapter¹ addresses this issue by defining a general model for change detection in multitemporal multispectral images. Specifically, a multiclass model of the difference image based on a compound representation of the multitemporal data is derived, which provides both a general description of the classes that populate the difference image and an analytical representation of their statistical distributions. The model is exploited in the context of binary CD based on the magnitude of the difference image derived by the change vector analysis (CVA) technique. In this framework, under some simplifying assumptions, a multiclass distribution of the magnitude is derived and a numerical method based on the Expectation-Maximization (EM) algorithm and Bayes decision is defined. The effectiveness of the proposed method is demonstrated on a large variety of datasets from different multispectral sensors, including those having high radiometric resolution. Numerical tests confirm that: 1. the fitting of the magnitude distribution significantly improves if compared to already existing models, 2. the change detection overall error is smaller than that obtained by traditional models and is comparable to optimal performance.

¹parts of this chapter appear in [160]

7.1 Introduction

Multispectral (MS) images provide information about the observed scene both in the spatial and the spectral domains. Given two coregistered multitemporal images acquired over the same geographical area, classical CD techniques assign to each pixel a label of type *unchanged* or *changed* [161]. In the context of multiple CD, changed pixels are further divided into categories representing different kinds of changes [86, 103]. Being CD essentially a classification problem, it can be addressed by both supervised and unsupervised approaches. In MS images natural classes can be described in statistical terms. Indeed, radiance values show evident statistical accordance given their originating natural classes (not necessarily semantic classes). Post-classification methods first classify the two images separately and then perform CD by comparing the two classification maps [162]. The main issue is that accuracy is not very high, as this simple approach does not take into account the temporal dependence of the image pair. Therefore, compound-classification methods are introduced, that incorporate temporal dependency in the statistical formulation of the classification problem [74, 163]. Further improvements can be obtained by applying active-learning techniques for the training phase [77], or domain-transfer techniques if reference information is available only for one date [164].

The above-mentioned methods do not exploit a direct comparison of the single time images. In particular, the spectral difference at the pixel level is not explicitly considered. Nevertheless, the difference image carries useful information for the solution of the CD problem, especially in the context of unsupervised methods. In change vector analysis (CVA), changes are associated to spectral variations of pixels after image differencing [131, 133]. The orientation of the spectral difference vectors can be used to separate between different kinds of changes [86]. However, much attention is devoted to CVA for the specific purpose of unsupervised binary change detection, i.e., discriminating between unchanged and changed pixels in a bitemporal image pair. Former change detection approaches assume the difference image as populated by two general classes representing the unchanged and changed pixels. Decision can be performed by using different techniques, including: 1) a-posteriori inference based on the Expectation-Maximization (EM) algorithm [82], 2) significance test [85], or, 3) local gradual descent [165]. In the context of CVA, the two-class model is used for the representation of the magnitude of the difference image [1, 78, 112]. Driven by the general intuition that in the difference image pixels having small magnitude are likely to be unchanged and pixels with high magnitude are changed, the magnitude information shows to be very informative. Indeed, many studies about binary change detection involve such feature, and the classical bi-modal behaviour of the histogram of the magnitude of the difference image is usually interpreted as follows:

the left mode (low magnitude) represents the population of unchanged pixels, whereas the right mode (high magnitude) represents the changed pixels. In [112], the distribution of the magnitude is empirically approximated by a mixture of two Gaussians, and inference is justified by means of Bayes decision. The binary decision is performed by thresholding the magnitude at the value that produces the minimum overall error of classification based on the Gaussian mixture (GG) model. Recently, it has been shown [78] that by assuming the Gaussian distribution of natural classes in the original images, the magnitude distribution can be approximated as Rayleigh for the unchange class and Rice for the change class. Such model is deeply investigated in [1], where a numerical method based on the EM algorithm for estimating the parameters of a Rayleigh-Rice (RR) mixture is developed. The parameter estimates are then used to compute a threshold by following a Bayes decision rule. Experiments on both real and synthetic datasets show that the RR model outperforms the classical GG. However, it is also observed that estimated densities do not fit accurately the magnitude distribution and the computed thresholds for binary decision are still different from the optimal ones. In this perspective, it seems that the two-class unchange/change model has some limitations due to the non-negligible effects caused by some subtle components that actually populate the mixture. Indeed, new generation sensors like those mounted on Landsat 8 and Sentinel 2 satellites, are characterized by a sensibly increased radiometric resolution compared to the old ones. As a consequence, multispectral data provided by modern satellite missions present greater statistical variability and the typical simplifying two-class model [1, 78, 82, 112, 165] for the difference image may no longer be valid to well describe the CD problem in last generation data. At the best of our knowledge, there is no study that extends the statistical interpretation of CVA to the multi-class case. In the sequel we address this specific problem.

In the first part of this chapter, we present a statistical study of the difference image and we propose a general definition of the change detection problem in the framework of CVA. In the proposed model, both the unchange and the change classes are multiple, and their statistical distributions are explicitly derived starting from a well-known joint distribution model. On the one hand, the proposed model has strong connections with the compound-classification approach to CD [74, 163], and allows for a formal extension of this method for the study of the difference image. On the other hand, being a generalization of the statistical description of CVA, our model can be used to generalize the already existing methods based on CVA to the multiclass case. In the second part of this work, we propose an extension of the binary change detection based on the magnitude of the difference image. Under some simplifying assumptions, we derive the theoretical distribution of the magnitude of the difference image in the general multiclass case. Then, we develop a framework for binary decision based on Bayes theory. Lastly, we

present a numerical method for parameter estimation and magnitude thresholding that is a generalization of the one presented in [1]. It is worth noting that, we do not make any a-priori assumption directly on the difference image. Instead, our model is derived from the proposed theoretical study of the difference image, which has more chances to fit with real data and does not constraint unchanged pixels to have the same statistical behaviour. Numerical experiments involve both synthetic and real multispectral images from Landsat 5, Landsat 7, Landsat 8 and Sentinel 2 satellites. Results show that, the proposed method better fits the real data and improves the detection performance of the method based on the RR mixture. Moreover, in all experiments it is observed that the computed threshold is very close to the optimal one.

The chapter is structured as follows. In Section 7.2 the statistical model describing the difference image is proposed and analyzed in detail. In Section 7.3, firstly we derive the statistical distribution of the magnitude difference image in the multi-class case, and then, we propose a numerical method based on the EM algorithm and Bayes decision to perform binary change detection. Numerical results are presented in Section 7.4. Finally, Section 7.5 draws the conclusions of the work.

7.2 Statistical study of the difference image

In this section, we provide a theoretical derivation of the statistical distribution of the difference image starting from the hypothesis of Gaussian distribution of natural classes in the single time images. Based on this model, we give a formal definition of the multiple unchange and change classes. On the one hand, the proposed model allows for formally describing some approaches already present in literature. On the other hand, it also provides a framework for extending the statistical interpretation of change detection to a larger set of cases.

7.2.1 A general mixture model for the difference image

Let \mathbf{y}^t , with $t = 1, 2$, be two multispectral images acquired over the same area at two different times. Each image has B bands, therefore each pixel value at location (i, j) is a B -dimensional vector $\mathbf{y}^t(i, j) \in \mathbb{R}^B$, where $t = 1, 2$. In a general statistical interpretation, the multispectral image formation is modeled as joint realization of certain B -dimensional random variables. To each pixel location (i, j) is assigned a random variable (y^t, Φ^t) where $y^t \in \mathbb{R}^B$ is associated to the observed spectral values of the pixel at time t , and $\Phi^t \in \{\phi_1, \dots, \phi_C\}$ is associated to its class label at time t . Hereafter, we will assume that these random variables are i.i.d., thus, the statistical distribution of the images can be fully described by defining one random variable (y^t, Φ^t) for each single time t . Two remarks

about the notation used hereafter. First, probability density functions are denoted by $p(\cdot)$ for convenience. Second, given $t = 1, 2$ and $h = 1, \dots, C$, the probabilistic event $\Phi^t = \phi_h$ is written in a more compact way as ϕ_h^t .

A typical and commonly used [74, 77, 163, 164] joint probabilistic model for y^1, y^2 assumes that the couple (y^1, y^2) depends on (Φ^1, Φ^2) . By marginalizing w.r.t. the class variables, this leads to the following distribution

$$p(y^1, y^2) = \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) p(y^1, y^2 | \phi_h^1, \phi_k^2). \quad (7.1)$$

It is reasonable to assume that the realization of the variable y^t only depends on its associated class label Φ^t , for each $t = 1, 2$. Therefore, we slightly simplify our joint model by assuming the following conditional independencies² [163]:

$$\begin{aligned} y^1 \perp y^2 &| (\Phi^1, \Phi^2) \\ y^1 \perp \Phi^2 &| \Phi^1 \\ y^2 \perp \Phi^1 &| \Phi^2 \end{aligned} \quad (7.2)$$

With these assumptions we can split the joint conditional distribution appearing in (7.1) to obtain

$$p(y^1, y^2) = \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) p(y^1 | \phi_h^1) p(y^2 | \phi_k^2). \quad (7.3)$$

In the most general case, the two images might not present the same set of observable classes: we assume that $\{\phi_1, \dots, \phi_C\}$ is the joint set of class labels which are observable in both images, where $C \geq 1$ is the total number of classes. The fact that some classes might not be observed in one of the single time images is formalized by setting to zero the corresponding class prior probabilities. For instance, if class ϕ_h is not observable at time $t = 1$, then $p(\phi_h^1, \phi_k^2) = 0$ for all $k = 1, \dots, C$. A similar argument works for a class that is not observable at time $t = 2$. Therefore, we let $0 \leq p(\phi_h^1, \phi_k^2) \leq 1$ and

$$\sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) = 1 \quad (7.4)$$

in order for (7.3) to be consistent as a probability model. Now, let us assume that natural classes are distributed as multidimensional Gaussians in the two images. This is a common and reasonable assumption in the literature when multispectral images are

²Following a standard notation, for $A \perp B | C$ we say that A and B are conditionally independent given C . In words, the knowledge of C makes A and B independent.

considered [150]. A natural class that is observed at both times might be described by different statistical parameters in the two images due to seasonal effects, different radiometric conditions, co-registration errors, etc. Therefore, we model them as

$$y^t | \phi_h^t \sim \mathcal{N}(y^t; \mu_h^t, \Sigma_h^t) \quad (7.5)$$

where μ_h^t, Σ_h^t are the mean vectors and the covariance matrices of the class ϕ_h observed at time $t = 1, 2$.

Under the hypothesis of a joint model as in (7.3)–(7.5), the distribution of the difference $d := y^1 - y^2$ can be written as a mixture of (at most) C^2 Gaussian components (technical details are given in Appendix 7.A.1). The density function of the difference turns out to be

$$p(d) = \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) \mathcal{N}(d; \mu_h^1 - \mu_k^2, \Sigma_h^1 + \Sigma_k^2) \quad (7.6)$$

where the mixture terms are the class prior probabilities $p(\phi_h^1, \phi_k^2)$.

7.2.2 Physical interpretation of the difference mixture model

The distribution (7.6) formalizes the intuitive concept that the difference image is populated by a set of classes, each one representing a possible class-by-class matching among the classes that populate the single time images. As mentioned above, we allow the same class ϕ_h to have different parameters μ_h^t and Σ_h^t in correspondence of the two acquisitions at times $t = 1, 2$. Every pixel that can be considered as drawn from the distribution $\mathcal{N}(\mu_h^1, \Sigma_h^1)$ in \mathbf{y}^1 and from the distribution $\mathcal{N}(\mu_h^2, \Sigma_h^2)$ in \mathbf{y}^2 is an unchanged pixel, as it belongs to the same natural class ϕ_h . It follows that each class ϕ_h has its own *unchange* behaviour if and only if the class is observed in both images. Conversely, a pixel that can be considered as drawn from the distribution $\mathcal{N}(\mu_h^1, \Sigma_h^1)$ in \mathbf{y}^1 and from the distribution $\mathcal{N}(\mu_k^2, \Sigma_k^2)$ in \mathbf{y}^2 , with $h \neq k$, is a changed pixel that changed its class from ϕ_h to ϕ_k . As a consequence, each class may have at most $C - 1$ different *change* behaviours.

In order to better understand which classes are changed and which ones are not changed, we can reason about priors. Each mixing term $p(\phi_h^1, \phi_k^2)$ is exactly the joint probability that class ϕ_h is observed at time $t = 1$ and class ϕ_k is observed at time $t = 2$. These probabilities can be arranged in a useful way into a square matrix

$$\mathbf{Q} = \begin{pmatrix} p(\phi_1^1, \phi_1^2) & p(\phi_1^1, \phi_2^2) & \dots & p(\phi_1^1, \phi_C^2) \\ p(\phi_2^1, \phi_1^2) & p(\phi_2^1, \phi_2^2) & \dots & p(\phi_2^1, \phi_C^2) \\ \vdots & \vdots & \ddots & \vdots \\ p(\phi_C^1, \phi_1^2) & p(\phi_C^1, \phi_2^2) & \dots & p(\phi_C^1, \phi_C^2) \end{pmatrix}. \quad (7.7)$$

If all the entries of the matrix \mathbf{Q} are non-null, the distribution function $p(d)$ has exactly C^2 mixture class components. If class ϕ_h is not observable at time $t = 1$, then the h -th row of the matrix is null. Similarly, if class ϕ_k is not observable at time $t = 2$, then the k -th column of matrix \mathbf{Q} is null. Other simplifications of matrix \mathbf{Q} can be done (i.e., setting to zero some of its entries) if prior knowledge about the studied scene is available. Just as an example, if we know that in the considered image pair it is impossible to find a change from class ϕ_h to class ϕ_k , then its corresponding prior probability can be set to zero, i.e., $p(\phi_h^1, \phi_k^2) = 0$. All the non-null entries of the matrix \mathbf{Q} define the prior probabilities of the mixture components that populate the difference image. Each of these mixture components describe the statistical distribution of the pixels that belong to class ϕ_h at time $t = 1$ and to class ϕ_k at time $t = 2$. It follows that, if $h = k$, then these pixels are not changed, whereas if $h \neq k$, then the pixels have changed their class from ϕ_h to ϕ_k . In both cases, their distributions are given explicitly. This interpretation allows us to give a formal definition of the unchange and change classes. in the difference image, as multiple classes.

Definition 1 (Unchange class). *Each pixel that belongs to class ϕ_h both at times $t = 1$ and $t = 2$ is said to belong to the unchange class ω_h , where $h \in \{1, \dots, C\}$. The distribution of the unchange class ω_h in the difference image is Gaussian with parameters*

$$d|\omega_h \sim \mathcal{N}(d; \mu_h^1 - \mu_h^2, \Sigma_h^1 + \Sigma_h^2).$$

The set of all unchange classes is denoted by Ω_n , and it contains all classes that are associated to non-null diagonal entries of the matrix \mathbf{Q} :

$$\Omega_n := \{\omega_h : p(\phi_h^1, \phi_h^2) \neq 0\}.$$

Definition 2 (Change class). *Each pixel that belongs to class ϕ_h at time $t = 1$ and to class ϕ_k at time $t = 2$ is said to belong to the change class ω_{hk} , where $h, k \in \{1, \dots, C\}$ and $h \neq k$. The distribution of the change class ω_{hk} in the difference image is Gaussian with parameters*

$$d|\omega_{hk} \sim \mathcal{N}(d; \mu_h^1 - \mu_k^2, \Sigma_h^1 + \Sigma_k^2).$$

The set of all change classes is denoted by Ω_c , and it contains all classes that are associated to non-null off-diagonal entries of the matrix \mathbf{Q} :

$$\Omega_c := \{\omega_{hk} : p(\phi_h^1, \phi_k^2) \neq 0, h \neq k\}.$$

As a simple consequence of the two definitions above, we have that the distribution of the difference image can be splitted into two parts as

$$p(d) = p(d|\Omega_n) + p(d|\Omega_c), \quad (7.8)$$

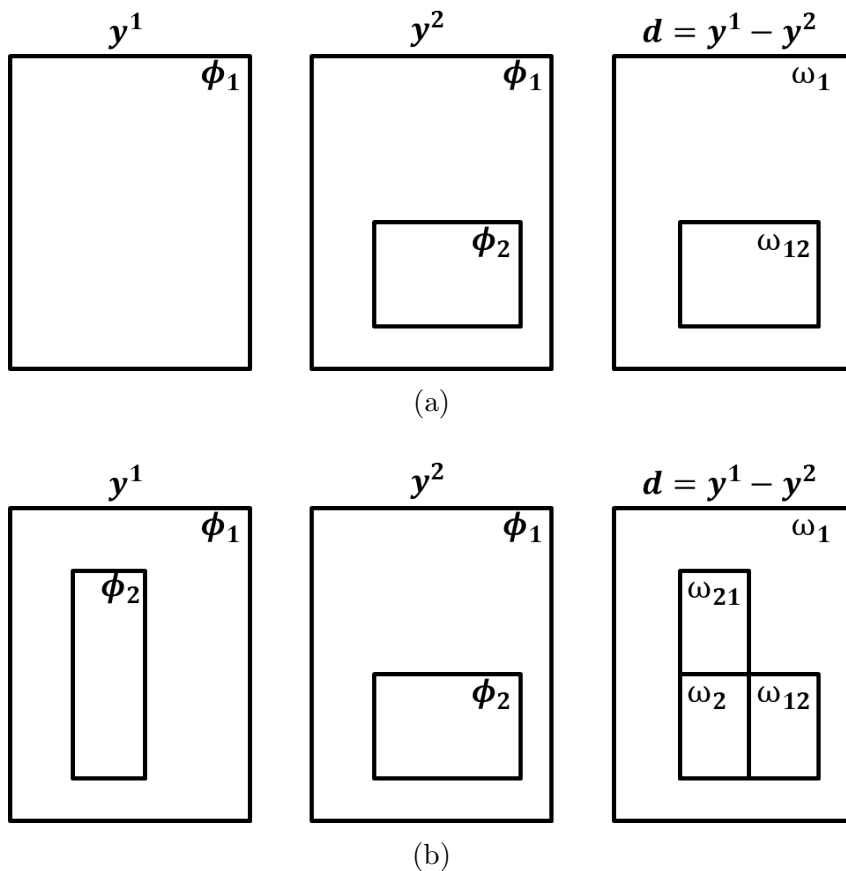


Figure 7.1: Example of statistical dependency of the difference image with respect to the input pair when the number of natural classes is $C = 2$. (a) The natural class ϕ_2 is only observable in image y^2 , thus the resulting difference image has only one unchange behaviour ω_1 and one change behaviour ω_{12} (from class ϕ_1 to ϕ_2). (b) Both natural classes ϕ_1, ϕ_2 are observable in the two images. Therefore, they have their own unchange statistical behaviours ω_1, ω_2 , and mutual change behaviours ω_{12} (from class ϕ_1 to ϕ_2), and ω_{21} (from class ϕ_2 to ϕ_1).

where the first term describes the statistical distribution of the multiple unchange class:

$$p(d|\Omega_n) := \sum_{h=1}^C p(\phi_h^1, \phi_h^2) \mathcal{N}(d; \mu_h^1 - \mu_h^2, \Sigma_h^1 + \Sigma_h^2), \quad (7.9)$$

and the second term describes the statistical distribution of the multiple change class:

$$p(d|\Omega_c) := \sum_{\substack{h,k=1 \\ h \neq k}}^C p(\phi_h^1, \phi_k^2) \mathcal{N}(d; \mu_h^1 - \mu_k^2, \Sigma_h^1 + \Sigma_k^2). \quad (7.10)$$

Some important remarks are needed here. It is common practice in literature to model the change class as a mixture of different components [166]. However, the underlying

assumption is always that the principal changes (the more statistically evident) are a small subset of all the ones that can be formulated theoretically. In particular this affects also the modelization of the unchange class, which is commonly reduced to be a single class. The increased radiometric resolution of last generation multispectral sensors typically results in more natural classes than the ones typically observed in older images. Thus, the mentioned approaches may not be appropriate anymore (a simple example about this distinction when $C = 2$, i.e., when images can be represented by two natural classes, is depicted in Figure 7.1). At the best of our knowledge, there is no attempt in literature to address the change detection problem by explicitly keeping into consideration the potential statistical variability of last generation multispectral data. In this work, we propose a general model that takes into account all the complexities arising by considering that both the unchange and change classes are multiple classes (it is worth noting that the proposed model also includes as particular cases some of the typical approaches to change detection already given in literature). With respect to the existent literature, the statistical model in (7.6) (and the discussion in this section), provide us additional quantitative information about:

1. the physical meaning of the classes that populate the difference image, and,
2. an explicit analytical description of their statistical distribution.

7.3 Derivation of a method for binary decision based on the difference mixture model

As an application of the framework presented in Section 7.2, here we present the derivation of a fully automatic numerical method for the calculation of the optimal decision threshold of the magnitude (binary change detection) that takes into account the complex multi-class nature of its distribution. Firstly, with some simplifying assumptions, the theoretical distribution of the magnitude of the difference image is derived from the density (7.6). Then, a framework for binary change detection is developed by following a multi-class maximum a-posteriori (MAP) approach. Lastly, a numerical method based on the EM algorithm [153] for parameter estimation of the derived mixture density is presented. Parameter estimates are finally used to compute the optimal value for thresholding the magnitude image and perform the binary change detection.

7.3.1 The magnitude distribution in the multi-class case

The framework presented in Section 7.2 is general and flexible enough to fit many possible real scenarios of change detection. On the counterpart, the model has several free

parameters to be estimated from data, thus the development of unsupervised methods exploiting this model (e.g., multivariate Gaussian fitting) could present critical issues (e.g., initialization criteria, convergence of parameter estimation methods, decision rules, limited robustness, etc.) and may result too numerically expensive to apply in reasonable time. Nevertheless, when a problem such as binary CD is considered, the number of parameters can be drastically reduced. The theoretical setting given in the previous section allows us to derive precise relationships between the parameters of the model in the original feature space and those modeling the reduced space. Here, we aim at defining an unsupervised method for binary decision based on a multiclass model for the magnitude distribution. In order to do this, some simplifying assumptions on the proposed general model proposed in Section 7.2 are made. Firstly we assume similar radiometric conditions and no seasonal differences in the two images $\mathbf{y}^1, \mathbf{y}^2$. Therefore we expect the parameters of the same class to not vary between the two acquisitions, in such a way that

$$\begin{aligned}\mu_h^1 &= \mu_h^2 =: \mu_h \\ \Sigma_h^1 &= \Sigma_h^2 =: \Sigma_h\end{aligned}\tag{7.11}$$

for all $h = 1, \dots, C$. Secondly, since our purpose is to exploit the magnitude information in the decision process, some regularity assumptions on the structure of the covariance matrices Σ_h are needed in order to derive the theoretical distribution of the magnitude as a Rice-like mixture type [141]. In practice, we assume that covariances are scalar matrices of the type

$$\Sigma_h = \sigma_h^2 \mathbf{I}_B\tag{7.12}$$

where \mathbf{I}_B is the identity matrix of size $B \times B$ and σ_h are positive scalars representing the standard deviation of each class. It is common practice in change detection to select a few bands (the most representatives of the changes of interest) that correspond to distant frequencies in the spectrum. Being them highly uncorrelated, we see that the assumption of diagonality of the matrices Σ_h is not very strict. For simplicity, in the sequel only two bands are considered (i.e., $B = 2$) in order to get the well-known Rician distributions associated to the magnitude variable. However, all the arguments presented can be generalized to any finite dimension by keeping into account that Rician distributions have more complex analytical expressions when $B > 2$. This technical extension is considered in Appendix 6.A.5. Lastly, we recall again (following from the discussion in Section 7.2), that spatial independence of pixels is assumed here. Therefore, each pixel value of the difference image is considered to be drawn from the distribution (7.6) independently from the others.

By taking into account (7.11), the difference image distribution $p(d)$ can be written

in the simpler form

$$p(d) = \sum_{h,k=1}^C q_{hk} \mathcal{N}(d; \mu_h - \mu_k, \Sigma_h + \Sigma_k). \quad (7.13)$$

where the mixture terms q_{hk} are the entries of matrix Q given in (7.7), i.e., $q_{hk} := p(\phi_h^1, \phi_k^2)$. Notice that each class component of the multiple unchange class is 0-mean and its covariance matrix is proportional (by factor 2) to the covariance matrix of the associated (unchanged) natural class

$$p(d|\Omega_n) = \sum_{h=1}^C q_{hh} \mathcal{N}(d; 0, 2\Sigma_h). \quad (7.14)$$

Regarding the class components of the multiple change class, they are mutually symmetric with respect to the origin

$$p(d|\Omega_c) = \sum_{\substack{h,k=1 \\ h \neq k}}^C q_{hk} \mathcal{N}(d; \mu_h - \mu_k, \Sigma_h + \Sigma_k). \quad (7.15)$$

Indeed, we have that $\mu_h - \mu_k = -(\mu_k - \mu_h)$ and, obviously, $\Sigma_h + \Sigma_k = \Sigma_k + \Sigma_h$.

These considerations enable us to describe the distribution $p(d)$ when coordinates are changed to magnitude. Indeed, given the previous assumptions, the theoretical derivation of the distribution of the random variable $\rho := |d| = |y^1 - y^2|$ is possible (the operator $|\cdot|$ denotes the Euclidean norm). It follows that the magnitude image $\boldsymbol{\rho} := \{|\mathbf{d}(i, j)| : (i, j)\}$ can be considered as a sample drawn from this distribution. Let us recall the general form of the Rician distribution that is given by

$$\mathcal{R}(\rho; \nu, \delta) = \frac{\rho}{\delta^2} \exp\left(-\frac{\rho^2 + \nu^2}{2\delta^2}\right) I_0\left(\frac{\rho\nu}{\delta^2}\right) \quad \rho \geq 0 \quad (7.16)$$

where $I_0(\cdot)$ is the 0-th order modified Bessel function of first kind [151]. By defining the magnitude variable as $\rho := |d|$, we have that the distribution of ρ follows

$$p(\rho) = \sum_{h,k=1}^C q_{hk} \mathcal{R}(\rho; \nu_{hk}, \delta_{hk}) \quad (7.17)$$

where $\nu_{hk} = |\mu_h - \mu_k|$ and $\delta_{hk} = \sqrt{\sigma_h^2 + \sigma_k^2}$. Let us point out some important remarks here. Since $\nu_{hh} = 0$ for every $h = 1, \dots, C$, we easily have that any component related to the unchange class is actually Rayleigh. Moreover, the scale parameters of these Rayleigh distributions only depend on the variances of the associated unchanged classes. More

specifically, $\delta_{hh} = \sigma_h \sqrt{2}$ for each h . The unchange class in the magnitude space is therefore described by

$$p(\rho|\Omega_n) = \sum_{h=1}^C q_{hh} \mathcal{R}(\rho; 0, \sigma_h \sqrt{2}). \quad (7.18)$$

All the other cross terms in (7.17) are mutually equal as $\nu_{hk} = \nu_{kh}$ and $\delta_{hk} = \delta_{kh}$ for each couple (h, k) such that $h \neq k$. Moreover, each non-centrality term is in general $\nu_{hk} \neq 0$, thus all components related to the change class are Rician. It follows that all the components of the multiple change class group together two-by-two and their mixture parameters are summed up

$$p(\rho|\Omega_c) = \sum_{\substack{h,k=1 \\ h < k}}^C (q_{hk} + q_{kh}) \mathcal{R}(\rho; \nu_{hk}, \delta_{hk}). \quad (7.19)$$

This means that, in the magnitude space the change from class ϕ_h to ϕ_k (class ω_{hk}) and the opposite change (class ω_{kh}) are statistically equivalent. Therefore, the number (and the meaning) of change classes in the magnitude space is different from that in the difference space. In light of this, we change the notation for the unchange and change classes in the magnitude space by adding a superscript ρ , namely Ω_n^ρ and Ω_c^ρ .

Once the structure of the matrix \mathbf{Q} is known, i.e., when the indices of its non-null entries (and not necessarily their values) are known, the number of unchange and change class members in (7.17) can be easily obtained. Let us denote $C' := \#(\Omega_n^\rho)$ the number of members of the unchange class and $C'' := \#(\Omega_c^\rho)$ the number of members of the change class. On the one side, only classes that are observable in both images can have their unchange behaviour, therefore we have in general

$$0 \leq C' \leq C. \quad (7.20)$$

On the other side, each natural class can have at most $C - 1$ different change behaviours. Thus, in view of the above mentioned statistical equivalence of change classes ω_{hk} and ω_{kh} , we have that

$$0 \leq C'' \leq \frac{C(C-1)}{2}. \quad (7.21)$$

For ease of notation, in the following the mixture indices are re-ordered and mixture and shape parameters are re-named accordingly. We now write the distribution of the magnitude as

$$p(\rho) = \sum_{h=1}^H \alpha_h \mathcal{R}(\rho; \nu_h, \delta_h) \quad (7.22)$$

where $H = C' + C''$ and, of course, $\sum_{h=1}^H \alpha_h = 1$. In the same manner we denote class labels as $\omega_h \in \Omega_n^\rho$ for $h = 1, \dots, C'$, and $\omega_h \in \Omega_c^\rho$ for $h = C' + 1, \dots, H$.

7.3.2 A Bayesian multi-class approach to binary decision

The theoretical model (7.22) describes the distribution of the magnitude of the difference image in the multi-class case and enables us to define a Bayesian framework for binary decision based on the maximum a-posteriori principle (MAP) and the thresholding of the magnitude variable. By following the MAP approach, for any given value of the target variable, one selects as the most likely originating class the one that maximizes the posterior probability of observing that value among all classes. It is well-known that this procedure minimizes the overall probability of committing classification errors. Therefore, given the mixture model (7.22) we define the MAP classification function

$$W[\rho] := \arg \max_{\omega_h \in \Omega^\rho} \alpha_h p(\rho|\omega_h) \quad (7.23)$$

where $\Omega^\rho := \Omega_n^\rho \cup \Omega_c^\rho = \{\omega_1, \dots, \omega_H\}$ is the set of all classes. It is worth noting that, being each competing class either $\omega_h \in \Omega_n^\rho$ or $\omega_h \in \Omega_c^\rho$, this approach has an intrinsic binary interpretation. Indeed, each value $W[\rho]$ identifies (and associates to ρ) one member of the two multiple unchange or change classes. Thus, the binary classification based on the thresholding of the magnitude of the difference image can be defined by

$$\begin{aligned} W_n &= \{(i, j) : W[\rho(i, j)] \in \Omega_n^\rho\} \\ W_c &= \{(i, j) : W[\rho(i, j)] \in \Omega_c^\rho\}, \end{aligned} \quad (7.24)$$

where W_n, W_c are the sets of predicted unchanged/changed pixels, respectively.

The Bayes decision method presented here generalizes the two-class approach for binary change detection presented in [1], as this last one can be derived from (7.23) and (7.24) when two natural classes are considered (i.e., $C = 2$), with the additional assumption that one class is observable only in one of the two input images (i.e., $C' = C'' = 1$). An example of this situation is illustrated in Figure 7.1a, whereas in Figure 7.1b a more general case involving all possible class unchange and change behaviours is illustrated.

7.3.3 Basic assumptions for the binary change detection problem

The binary change detection problem is usually addressed in literature by assuming the unchange class to be a single class [1, 78, 82, 112, 165]. Sometimes, such assumption is extended also to the change class [1, 78], while in other cases, the change class is modeled as multiple [86]. By a careful analysis of matrix \mathbf{Q} and its relationship with the mixture components in (7.22), we see that:

- A model that assumes both the unchange and change classes to be single, is represented (in the simplest case) by a matrix of size 2×2 . More specifically, the possible cases are:

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} q_{11} & 0 \\ q_{21} & 0 \end{pmatrix}, \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & 0 \end{pmatrix},$$

where we recall that the changes ω_{12} and ω_{21} are statistically equivalent in the magnitude space, thus the last case relates to a single change.

- The assumption of having more than one change in the magnitude space puts a constraint on the number of natural classes in the model, which must be $C \geq 3$. The description of all possible cases becomes more and more complex when C increases.

As we can see, the modelization of the problem can be arbitrarily complex if we do not have any prior knowledge about the classes involved. Some information about the structure of the matrix \mathbf{Q} can be recovered in a supervised or semi-supervised way, by separately analyzing the single time images $\mathbf{y}^1, \mathbf{y}^2$ and their difference \mathbf{d} . However, this kind of approach may result too articulated for the specific purpose of binary change detection, whereas it seems more appropriate to be developed in the context of multiple change detection. On the counter part, a model can also be fixed a priori, and a change detection method can be derived accordingly. In formulating the a priori assumptions for the development of our method, we mainly consider that: 1) simple models proved to be effective for addressing the specific binary change detection problem; 2) typically, the solution of the EM algorithm is sub-optimal (local max), thus the introduction of several additional parameters will decrease the significance of the solution itself. In view of these points, let us consider the case where only one relevant change can be identified in the image pair, so that we can focus the attention on the multiple unchange class. Therefore, the matrix \mathbf{Q} can be assumed of the form

$$\mathbf{Q} = \begin{pmatrix} q_{11} & & & & \\ & \ddots & & & \\ & & & q_{hk} & \\ & & & \ddots & \\ q_{kh} & & & & \ddots \\ & & & & & q_{CC} \end{pmatrix}, \quad (7.25)$$

where $h, k \in \{1, \dots, C\}$ are fixed. We stress the fact that for the derivation of the statistical model in (7.22), some regularity assumptions on the natural classes are made, see (7.11) and (7.12). In addition to allowing the derivation of a precise model for the magnitude distribution, such assumptions have the remarkably positive effect of collapsing

many subtle natural classes into a few dominant ones. Indeed, according to (7.14), *all* natural classes that are described by the same covariance matrix can be grouped into one single unchange class in the difference space as their distributions coincide. More specifically, let us assume that for a fixed $p \in \{1, \dots, C\}$ there is a set of classes $\mathcal{B}_p \subset \{1, \dots, C\}$ such that $\Sigma_h = \Sigma_p$, for all $h \in \mathcal{B}_p$. Then,

$$\sum_{h \in \mathcal{B}_p} q_{hh} \mathcal{N}(d; 0, 2\Sigma_h) = \mathcal{N}(d; 0, 2\Sigma_p) \sum_{h \in \mathcal{B}_p} q_{hh}. \quad (7.26)$$

This simple mathematical fact proves that the complexity of the unchange class (in terms of number of distinguishable class members) does not depend on the real number of natural classes, but only on those that have different covariance matrices. Notice in particular that, under the considered assumptions, the mean vector of each class (i.e., the feature vector that mainly characterizes the spectral signature of each class) does not play any role in the definition of the unchange class members.

Accordingly, we devise a distribution model that is able to represent the unchange multiple class in terms of dominant classes, while it does not increase too much the overall complexity of the associated estimation problem. Without putting a constraint on the *actual* number C of natural classes that populate the images, we just assume that two main groups \mathcal{B}_p and \mathcal{B}_q , with $p, q \in \{1, \dots, C\}$, can be formed out of all classes, which are homogeneous in terms of covariance matrices. Thus, we can write

$$p(d|\Omega_n) = \mathcal{N}(d; 0, 2\Sigma_p) \sum_{h \in \mathcal{B}_p} q_{hh} + \mathcal{N}(d; 0, 2\Sigma_q) \sum_{h \in \mathcal{B}_q} q_{hh}. \quad (7.27)$$

In principle, this simplification could be extended to any number of groups. Nonetheless, we will show in the experimental part of this chapter how the simple choice of assuming two dominant classes proved to be sufficiently general to provide a change detection performance which is nearly optimal.

7.3.4 A numerical method for binary change detection based on the EM algorithm

The magnitude of the difference image $\boldsymbol{\rho}$ is considered to be a set of samples independently drawn from distribution (7.22). In order to apply the MAP approach for decision, parameters of this distribution must be numerically estimated. In the theoretical model presented in Section 7.3.1, the parameters are constrained by their dependency on the natural class parametrization. In order to partially recover some flexibility that we lost in imposing conditions such as in (7.11) and (7.12), we implement a version of the EM algorithm without parameter constraints. In addition to greater flexibility, this also gives

the advantage of implementing an unconstrained optimization. Given this, from the a priori model (7.25) coupled with (7.27) and applied to the distribution (7.22), we can write the probability distribution of the magnitude of the difference image as

$$p(\rho) = \alpha_1 \mathcal{R}(\rho; 0, \delta_1) + \alpha_2 \mathcal{R}(\rho; 0, \delta_2) + \alpha_3 \mathcal{R}(\rho; \nu, \delta) \quad (7.28)$$

where $\alpha_3 = 1 - \alpha_1 - \alpha_2$. For notation simplicity we omitted the subscripts of the parameters of the Rician component (the third term in the summation). Let $\Psi = (\alpha_1, \alpha_2, \delta_1, \delta_2, \nu, \delta)$ be the vector of shape and mixture parameters involved in (7.28), then the dependency of this probability distribution with respect to the parameters can be expressed by writing $p(\rho|\Psi)$. For any $r \in \boldsymbol{\rho}$ we define

$$p(\omega_h|r, \Psi') := \frac{\alpha'_h p(r|\omega_h, \Psi')}{p(r|\Psi')} \quad (7.29)$$

to be the posterior probability that the sample r originated in the h -th component of the population, given Ψ' .

Iterative formulas

By the same arguments developed in [1], it is possible to define an iterative version of the EM algorithm for the estimation of the parameters of (7.28) as follows

$$\begin{aligned} \alpha_i^{k+1} &= N^{-1} \sum_{r \in \boldsymbol{\rho}} p(\omega_i|r, \Psi^k) \\ (\delta_i^2)^{k+1} &= \frac{\sum_{r \in \boldsymbol{\rho}} p(\omega_i|r, \Psi^k) r^2}{2 \sum_{r \in \boldsymbol{\rho}} p(\omega_i|r, \Psi^k)} \\ \nu^{k+1} &= \frac{\sum_{r \in \boldsymbol{\rho}} p(\omega_3|r, \Psi^k) \frac{I_1\left(\frac{r\nu^k}{(\delta^k)^2}\right)}{I_0\left(\frac{r\nu^k}{(\delta^k)^2}\right)} r}{\sum_{r \in \boldsymbol{\rho}} p(\omega_3|r, \Psi^k)} \\ (\delta^2)^{k+1} &= \frac{\sum_{r \in \boldsymbol{\rho}} p(\omega_3|r, \Psi^k) \left[r^2 + (\nu^k)^2 - 2r\nu^k \frac{I_1\left(\frac{r\nu^k}{(\delta^k)^2}\right)}{I_0\left(\frac{r\nu^k}{(\delta^k)^2}\right)} \right]}{2 \sum_{r \in \boldsymbol{\rho}} p(\omega_3|r, \Psi^k)}. \end{aligned} \quad (7.30)$$

for $i = 1, 2$, where $I_1(\cdot)$ is the 1-th order modified Bessel function of first kind [151].

Initialization

Iterative versions of the EM algorithm are known to be typically slow [153]. Moreover, the proposed algorithm iteratively updates the parameters of three different mixture components, for a total of two mixture and four shape parameters. This makes the initialization

of the algorithm a very crucial point, as a bad strategy may easily result in a stagnation of the parameters updates. In particular, after many numerical experiments (both on synthetic and real data) we observed that the typical approach of median thresholding [1, 82] cannot be used for the preliminary Maximum Likelihood (ML) estimate of the parameters.

For a correct initialization, we found that the typical local minimum that separates the two main modes of the histogram of the magnitude must be precisely identified in order to have a first separation of the samples into two approximate classes: W_n^0 (unchanged samples) and W_c^0 (changed samples). Under the hypothesis that this local minimum is the stationary point of the underlying pdf with less curvature, it can be found as the point of lower variation of the corresponding cdf. A more stable numerical computation can be done by calculating the maximum point of the derivative of the quantile function. More specifically, let $Q : [0, 1] \rightarrow [0, 1]$ be the quantile function of the magnitude histogram and let $t^0 := \arg \max_{t \in [0, 1]} Q'(t)$. Then, the two approximate classes are defined as

$$\begin{aligned} W_n^0 &= \{(i, j) : \rho(i, j) \leq t^0\} \\ W_c^0 &= \{(i, j) : \rho(i, j) > t^0\}, \end{aligned} \tag{7.31}$$

Since in the target mixture the components that represent unchanged pixels are two, the samples in W_n^0 are further divided into two sets $W_{n,1}^0, W_{n,2}^0$ by median thresholding. Finally, shape and mixture parameters of the mixture are estimated via ML approach (refer to [142]) using the samples in the approximate classes. The shape parameters δ_i are estimated using the samples in $W_{n,i}^0$, with $i = 1, 2$, and ν, δ are estimated using the samples in W_c^0 . The mixture parameters α_i are approximated by their corresponding proportions $\#(W_{n,i}^0)/\#(\mathbb{R})$, for $i = 1, 2$.

Convergence

The iterative algorithm defined by (7.30) is gradient-based, thus it can be stopped when slow variations of the objective energy (the ML function of the given set of magnitude samples) are observed. More specifically, the algorithm is stopped at the first iteration k such that the relative variation of the energy is less than 10^{-6} [1].

7.4 Experimental results

In this section we present an extensive experimental analysis of the proposed technique for binary change detection based on the EM algorithm. Firstly, we describe the experimental setup, which is designed for comparing the proposed technique against the one recently proposed in [1], which represents a reliable reference in statistical methods

for binary change detection. Secondly, a brief description of the datasets analyzed in the tests (including image pairs from different multispectral sensors) is given. Then, the performance of the EM algorithm for parameter estimation is analyzed in detail, together with the fitting capability of the models. Lastly, the change detection is performed and the results are compared in order to assess the effectiveness of the proposed method with respect to the state-of-the-art techniques.

7.4.1 Experimental setup

The proposed algorithm is designed with the specific intent of improving the change detection performance of the method proposed in [1] (which is denoted in the following by \mathbf{rR}) in particular in the case where images from last generation multispectral sensors are considered. In the sequel, parameter notation related to this method is the same as in the cited reference. The proposed algorithm is denoted by \mathbf{rrR} and its parameters are named according to Section 7.3.4.

Let us present the outline of the experimental setup. Firstly, the image pair is differenced and the magnitude image is calculated. Then, the magnitude values are used to estimate in an unsupervised way the parameters of both the \mathbf{rR} and \mathbf{rrR} statistical models via EM algorithm. Iteration details are given and the fitting performance of the estimated densities are analyzed. Lastly, the estimated parameters are used to perform Bayes decision according to the MAP approach. In particular, the binary change detection is performed by thresholding the magnitude image using the threshold value T resulting in the of minimum estimated overall error calculated for both the \mathbf{rR} and \mathbf{rrR} cases. The change detection performance is evaluated in terms of missed/false alarms and overall errors.

7.4.2 Datasets description

The datasets considered hereafter consist of bitemporal image pairs from different multispectral sensors. The images are co-registered and radiometrically corrected. Images are followed by reference maps of the changes obtained by photo-interpretation. The datasets are illustrated in Figure 7.2 and briefly described in the following.

Dataset A

Two multispectral images acquired by the Thematic Mapper (TM) multispectral sensor of the Landsat 5 satellite. The scene represents an area including Lake Mulargia (Sardinia Island, Italy), at a resolution of 30 m. The images consist of 300×412 pixels. The dates of acquisition are September 1995 (t_1) and July 1996 (t_2). Between the two dates one

most relevant change, which is related to the extension of the lake surface, occurred in the study area. The two bands that better represents the change of interest are TM4 and TM7 (channels 4 and 7), the near infrared (NIR) and the middle infrared (MIR) [78], respectively.

Dataset B

Two multispectral images acquired by the Enhanced Thematic Mapper Plus (ETM+) multispectral sensor of the Landsat 7 satellite. The scene represents the same subject of Dataset A (Lake Mulargia, Sardinia). Spatial resolution is 30 m. The images consist of 310×449 pixels. The dates of acquisition are May 2001 (t_1) and May 2003 (t_2). As for dataset A, bands ETM+4 (NIR) and ETM+7 (MIR) are selected as most representative of the change (lake enlargement).

Dataset C

Two multispectral images acquired by the Operational Land Imager (OLI) multispectral sensor of the Landsat 8 satellite. The investigated area includes Lake Omodeo and a portion of Tirso River (Sardinia Island, Italy). The images consist of 700×650 pixels at a resolution of 30 m. The dates of acquisition are 25th July 2013 (t_1) and 10th August 2013 (t_2). The main change is a fire occurred between August 7th and 9th in the south of Ghilarza village. The two bands selected as most representative of the changes are bands 5 and 6, the near infrared (NIR) and the first short wavelength infrared (SWIR1) [1], respectively.

Dataset D

Two multispectral images acquired by the Multispectral Instrument (MSI) of the Sentinel 2a satellite. The images consist of 300×400 pixels at a resolution of 10 m. The dates of acquisition are 31st December 2015 (t_1) and 20th January 2016 (t_2). The investigated area includes the surrounding of Tesis village (Pordenone, Italy). The main changes are all related to variations in soil conditions due to human activities on crops. The two bands selected as most representative of the changes are bands 3 and 4, the red and the NIR channels, respectively.

7.4.3 EM algorithm and fitting performance

In the run of the iterative algorithm for parameter estimation of the \mathbf{rR} mixture, the classical median thresholding technique for the initialization worked well. However, it

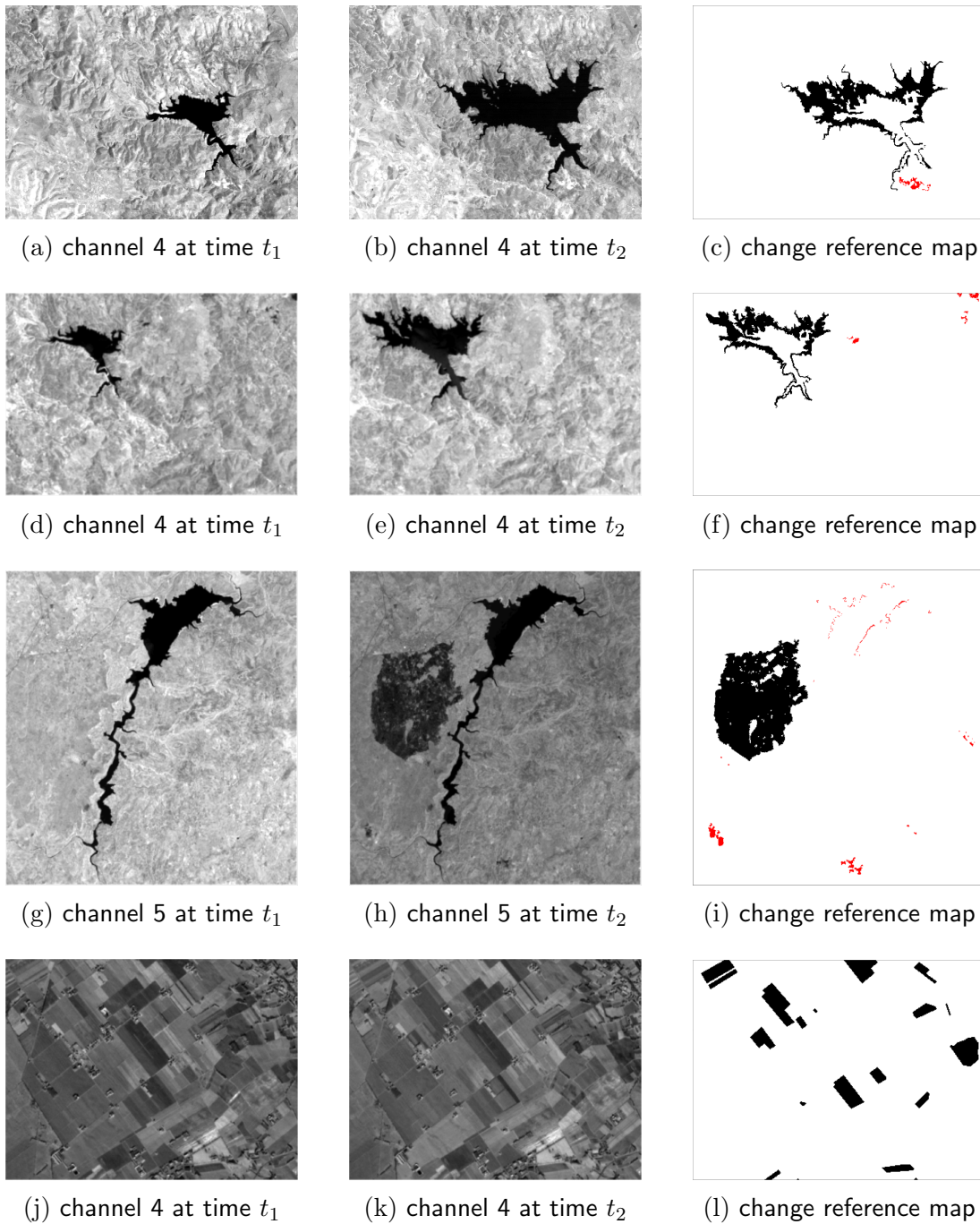


Figure 7.2: Illustration of the datasets analyzed in the experiments. The pictures show pre and post images in one band and a reference map of the changes (red pixels are minor changes with respect to the main changes that are represented by black pixels) for (a,b,c) dataset A, (d,e,f) dataset B, (g,h,i) dataset C, and (j,k,l) dataset D.

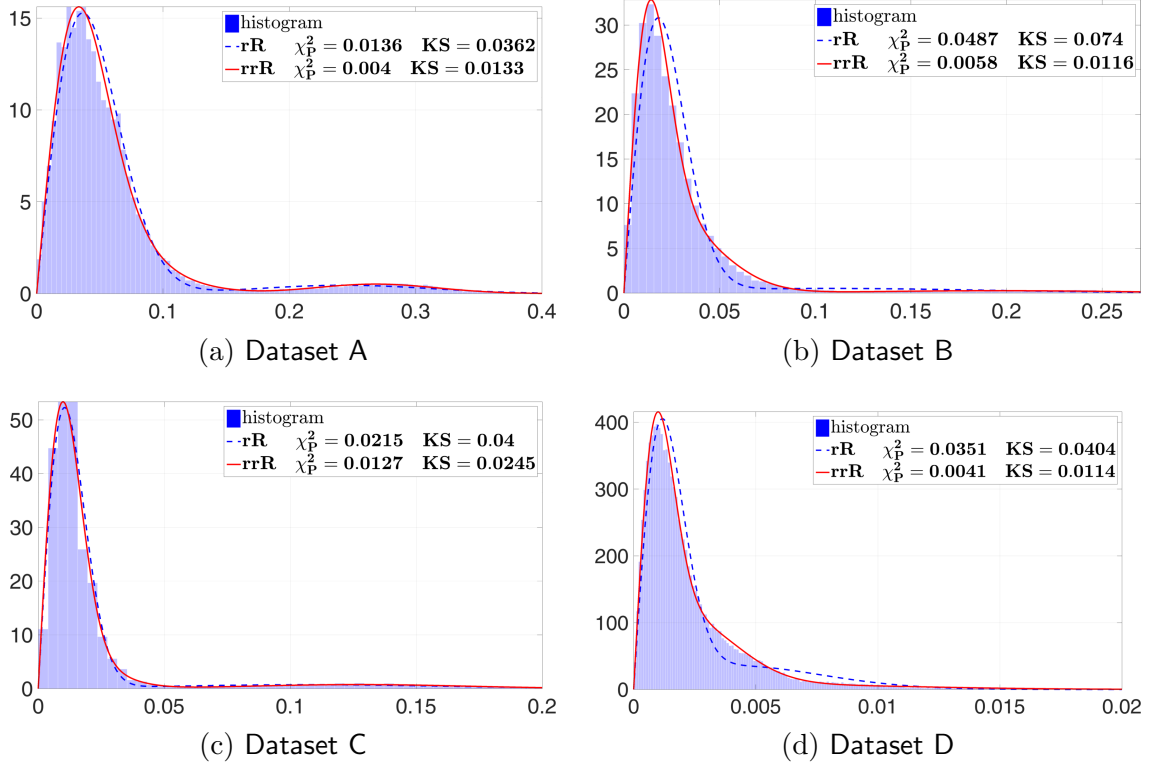


Figure 7.3: Fitting capability of the considered methods. Estimated densities are superimposed to the histograms of the magnitude samples. In the legends also the distance metrics between the estimated densities and the histograms are given.

resulted in never-ending iterations (the relative variation of the objective energy has never reached values below tolerance) in the **rrR** case, for all the three datasets. Therefore, the more appropriate initialization strategy based on the quantile function (described in Section 7.3.4) has been used. The initial approximate threshold values obtained in the calculations are $t^0 = 0.1477, 0.1091, 0.0559, 0.0100$ for Dataset A, B, C and D, respectively. These values are placed in the central portion of the histograms (cfr. with histograms in Figure 7.3) between the two principal modes. Notice that this does not hold true in the case of the median thresholding, indeed the corresponding values obtained in the computations are $t^{mid} = 0.2105, 0.1833, 0.2511, 0.0216$. As we can see, such values are significantly shifted on the right in the corresponding histograms. This proved to be a limitation for the initial estimation of the two Rayleigh components in the **rrR** mixture as the unchange class is sensibly over-estimated.

Iteration details and the initial and final values of the estimated parameters are shown in Table 7.1. As expected, the number of iterations and the time of computation increases in the case of **rrR**, as there are two additional parameters to be estimated in the model.

Table 7.1: EM algorithm iteration details and parameter estimations for both the considered mixtures.

mix	k	estimated parameters						secs
Dataset A								
rR	0	α	-	b	-	ν	σ	4.0
	60	0.93	-	0.04	-	0.25	0.06	
rrR	0	α_1	α_2	δ_1	δ_2	ν	δ	73.2
	1258	0.79	0.14	0.03	0.07	0.25	0.06	
Dataset B								
rR	0	α	-	b	-	ν	σ	3.5
	66	0.96	-	0.02	-	0.20	0.06	
rrR	0	α_1	α_2	δ_1	δ_2	ν	δ	13.9
	270	0.90	0.06	0.02	0.05	0.20	0.06	
Dataset C								
rR	0	α	-	b	-	ν	σ	10.6
	57	0.92	-	0.01	-	0.12	0.04	
rrR	0	α_1	α_2	δ_1	δ_2	ν	δ	65.4
	400	0.90	-	0.01	-	0.09	0.06	
Dataset D								
rR	0	α	-	b	-	ν	σ	2.15
	51	0.98	-	0.002	-	0.013	0.004	
rrR	0	α_1	α_2	δ_1	δ_2	ν	δ	6.01
	400	0.75	-	0.001	-	0.001	0.004	

Notice that, between the first and the last iteration, the parameters that are more significantly changed are the prior probabilities of classes, i.e., α in the **rR** case and α_1, α_2 in the **rrR** case.

Let us now study the fitting properties of the two methods. A qualitative analysis of the fitting performance can be done by looking at the plots of the estimated densities \mathbf{rR} and \mathbf{rrR} , which are superimposed to the histogram of the magnitude in Figure 7.3. The figure also provides quantitative evaluation of the fitting in terms of χ^2 -Pearson divergence and Kolmogorov-Smirnov (KS) distance between data and estimated densities. From the results we can see that the proposed \mathbf{rrR} model fits the data in a satisfying way, and much better than \mathbf{rR} .

7.4.4 Change detection results

The whole theoretical framework presented has been developed with the aim of better representing the typical distributions that characterize multispectral data acquired by new generation sensors. Following the expected (and confirmed) increased fitting performance of the resulting model, it was also expected an improvement in the change detection performance due to the increased capability of the model of representing the specific unchange and change multiple classes. Indeed, this has been confirmed in the binary change detection phase of our tests. Table 7.2 reports the computed threshold value T and the change detection performance in terms of missed/false alarms and overall errors for each dataset and each considered mixture model. The results can be compared with optimal performance based on a usual trial-and-error procedure applied to the reference maps of the changes.

As we can see, the proposed \mathbf{rrR} mixture model for binary change detection proved to be sufficiently flexible to well model the unchange and change classes in the datasets as it is able to return the same thresholds (with precision at the second decimal digit) that are obtained in the optimal case. A qualitative understanding of the performance is possible by looking at the change detection maps illustrated in Figure 7.4. With respect to the \mathbf{rR} case, the proposed model presents much less false alarms at the expense of very few additional missed alarms. Notice that, the change maps obtained in the optimal case are not illustrated in the picture as their differences w.r.t. to the maps obtained with the \mathbf{rrR} model cannot be seen just visually.

7.5 Discussion and conclusion

The new generation of multispectral sensors mounted on satellite missions such as Landsat 8 and Sentinel 2 offers a unique opportunity of studying and monitoring the Earth surface by providing images at very short revisit time. Moreover, the new technologies allow to represent the spectral signature of the observed objects with higher radiometric resolutions. This augments the statistical variability of the multispectral data and typical

Table 7.2: Comparison of the change detection performance of the method in [1] and the proposed method with respect to optimal performance.

mix	T	MA (%)	FA (%)	OE (%)
Dataset A				
rR	0.1332	139 (1.81)	1672 (1.44)	1811 (1.47)
rrR	0.1741	402 (5.22)	498 (0.43)	900 (0.73)
opt	0.1825	491 (6.38)	392 (0.34)	883 (0.71)
Dataset B				
rR	0.0624	451 (7.57)	5421 (4.07)	5872 (4.22)
rrR	0.1028	975 (16.36)	801 (0.60)	1776 (1.28)
opt	0.1119	1115 (18.71)	567 (0.43)	1682 (1.21)
Dataset C				
rR	0.0390	1104 (3.03)	6309 (1.51)	7413 (1.63)
rrR	0.0548	2029 (5.57)	1789 (0.43)	3818 (0.84)
opt	0.0570	2140 (5.88)	1617 (0.39)	3757 (0.83)
Dataset D				
rR	0.0033	841 (11.79)	18963 (16.80)	19804 (16.50)
rrR	0.0071	2563 (35.94)	1074 (0.95)	3637 (3.03)
opt	0.0072	2626 (36.82)	1002 (0.89)	3628 (3.02)

approaches to change detection presented in literature that worked well on data acquired by older satellite missions are no more able to address the new challenges arising in this important field of remote sensing.

In order to fill this gap, in the first part of this chapter we presented a theoretical study of the change detection problem in a full multiclass framework. In particular, a statistical model of the difference image is derived starting from a few basic assumptions that are usually made in multispectral image analysis. The aim of the proposed study is that of defining a model with sufficient degrees of freedom for well representing the intrinsic multiclass nature of multispectral data also in the CD problem. It is worth noting that, in the proposed model no a-priori assumptions are made directly on the difference image. Instead, it explicitly describes the dependency between the distributions of natural classes in the bitemporal image pair and the statistical model of the difference image. Being a generalization of the statistical description of CVA, our model can be used to generalize the already existent methods based on CVA to the multiclass case.

Following this direction, in the second part of the chapter our complete model of the difference image has been exploited in order to describe the statistical distribution of the

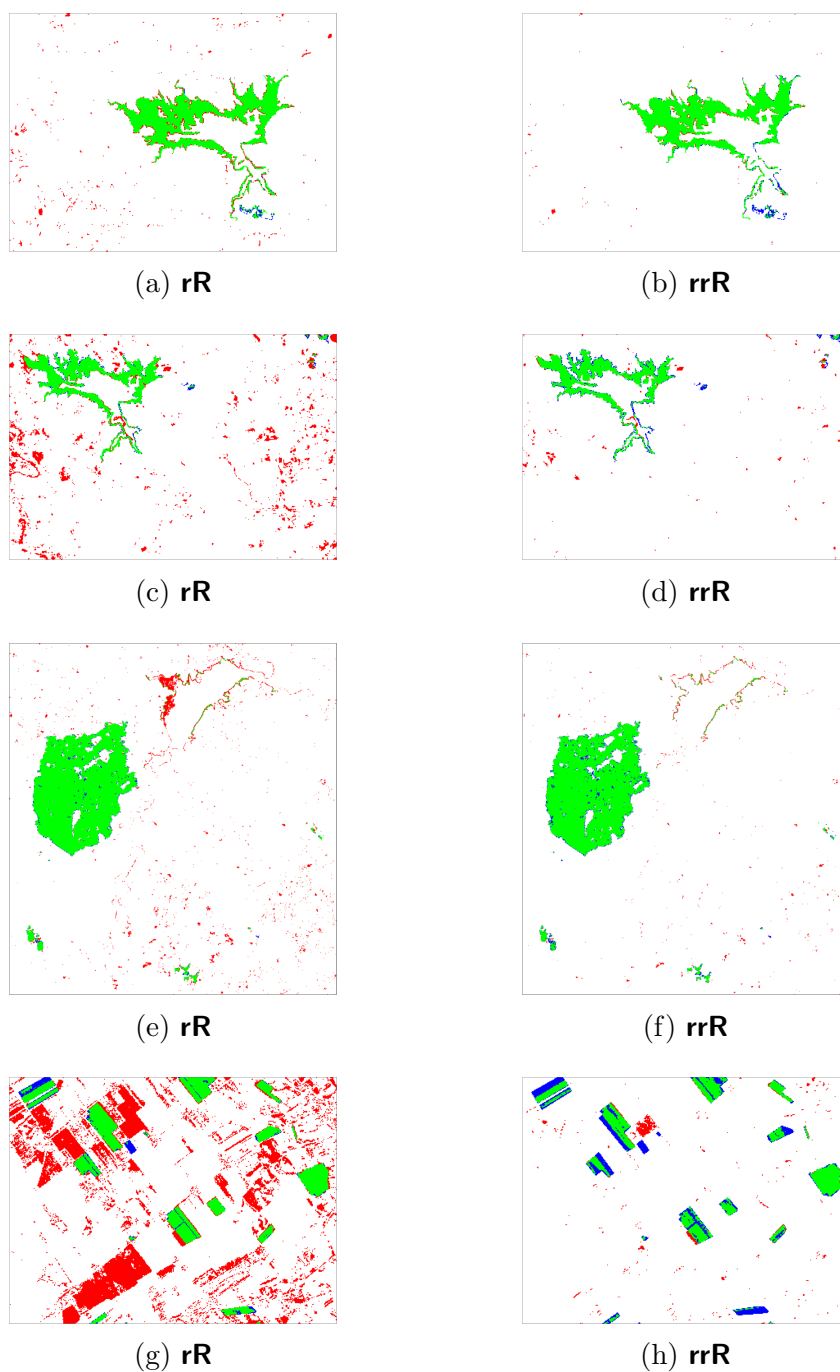


Figure 7.4: Change detection map obtained on (a,b) dataset A, (c,d) dataset B, (e,f) dataset C, and (g,h) dataset D. Blue pixels are missed alarms, red ones are false alarms and green ones are correctly detected changes.

magnitude of the difference image in a more general fashion. In the experimental part of this work, the parameters of this statistical model are estimated from data by using

an iterative version of the EM algorithm specifically devised for dealing with problems where the unchange class is multiple. Then, inference is done by using a standard MAP approach. Numerical results on a large variety of datasets demonstrated the effectiveness of the proposed method with a remarkable improvement with respect to state-of-the-art approaches. On the one hand, the fitting capability of the proposed model is increased, as expected. On the other hand, in all considered datasets the change detection performance is nearly optimal, as confirmed by a comparison with the performance that can be obtained by exploiting the reference change maps.

7.A Appendices

7.A.1 Statistical distribution of the difference of Gaussian mixtures

An explicit derivation of the statistical distribution of the difference of two random variables $y^1, y^2 \in \mathbb{R}^B$ jointly distributed as

$$p(y^1, y^2) = \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) p(y^1 | \phi_h^1) p(y^2 | \phi_k^2) \quad (7.32)$$

where $p(y^1 | \phi_h^1) = \mathcal{N}(y^1; \mu_h^1, \Sigma_h^1)$ and $p(y^2 | \phi_k^2) = \mathcal{N}(y^2; \mu_k^2, \Sigma_k^2)$ is possible by exploiting the characteristic function of linear combinations of random variables. For a random variable y admitting density function $p(y)$, the characteristic function (CF) of y is the inverse Fourier transform of its density function:

$$\varphi_y(t) := E \left[e^{it^T y} \right] = \int_{\mathbb{R}^B} e^{it^T y} p(y) dy. \quad (7.33)$$

The CF completely defines the probability distribution of the variable y . For a Gaussian distributed random variable y , its characteristic function can be described analytically in a closed form. Let us assume $y \sim \mathcal{N}(\mu, \Sigma)$, then we have that $\varphi_y(t) = \exp \{ it^T \mu - \frac{1}{2} t^T \Sigma t \}$, where i is the imaginary constant: $i^2 = -1$.

Let us now consider the above mentioned random variables y^1, y^2 , and let $v := a_1 y^1 + a_2 y^2$ be a linear combination of them with coefficients $a_1, a_2 \in \mathbb{R}$. By plugging (7.32) into (7.33) with $y = (y^1, y^2)$ and using the properties of the exponentials, the CF of v can be

calculated as

$$\begin{aligned}
\varphi_v(t) &= \varphi_{a_1 y^1 + a_2 y^2}(t) \\
&= E [\exp \{a_1 t^T y^1\} \exp \{a_2 t^T y^2\}] \\
&= \int \int \exp \{a_1 t^T y^1\} \exp \{a_2 t^T y^2\} p(y) dy \\
&= \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) \int \exp \{a_1 t^T y^1\} p(y^1 | \phi_h^1) dy^1 \times \\
&\quad \int \exp \{a_2 t^T y^2\} p(y^2 | \phi_k^2) dy^2 \\
&= \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) \exp \left\{ i a_1 t^T \mu_h^1 - \frac{1}{2} a_1^2 t^T \Sigma_h^1 t \right\} \times \\
&\quad \exp \left\{ i a_2 t^T \mu_k^2 - \frac{1}{2} a_2^2 t^T \Sigma_k^2 t \right\} \\
&= \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) \exp \left\{ i t^T (a_1 \mu_h^1 + a_2 \mu_k^2) + \right. \\
&\quad \left. - \frac{1}{2} t^T (a_1^2 \Sigma_h^1 + a_2^2 \Sigma_k^2) t \right\}.
\end{aligned} \tag{7.34}$$

This analytical expression uniquely identifies the distribution of v as a mixture of Gaussians

$$p(v) = \sum_{h,k=1}^C p(\phi_h^1, \phi_k^2) \mathcal{N}(v; a_1 \mu_h^1 + a_2 \mu_k^2, a_1^2 \Sigma_h^1 + a_2^2 \Sigma_k^2) \tag{7.35}$$

having C^2 mixture components. The distribution of the difference $y^1 - y^2$ can be easily obtained from (7.35) with $a_1 = 1$ and $a_2 = -1$.

Chapter 8

A class-wise spatial-contextual approach based on a free discontinuity model for change detection in multispectral images

The increased radiometric resolution of last generation multispectral sensors results in large statistical variability of classes represented in the image. However, classes present high spatial homogeneity. To preserve classes identity while simplifying their representation, in this chapter we propose a class-wise spatial-contextual method based on a variational model with free discontinuities that reduces the statistical variability of classes by emphasizing their spatial countours. To prove its effectiveness, the proposed method is applied in the context of change detection in multispectral images. Here, it is able to augment the discrimination between the unchange and the change classes and to improve the detection performance.

8.1 Introduction

Statistical minimum effort approaches have been successfully applied to change detection on multispectral (MS) images [1]. The study of large scenes poses some criticalities in the utilization of statistical models for MS image analysis, especially for the purpose of change detection [160]: 1) members of the same semantic class may present different statistical parametrizations in a whole large scene due to the large swath, especially in images acquired by optical sensors of last generation. 2) images acquired in different seasons or under different climatic-atmospheric conditions present a potential high vari-

ability in the data that may lead to erroneous discrimination of very similar classes or to difficult discrimination of theoretically distinguishable classes. A complete statistical representation of a whole scene would inevitably cause overfitting problems due to the many parameters needed to describe all the classes involved. This typically leads to cumbersome computations when parameters must be estimated from data. For example, the Expectation-Maximization (EM) optimization algorithm can easily get stuck into shallow critical points and fail to converge. In order to mitigate for this variability by at the same time avoiding the overfitting problem, we propose an effective image approximation framework that enjoys both relevant spatial and statistical simplification features. The proposed framework is based on a well-know variational model for image approximation in computer vision, to which we give a completely new statistical interpretation.

8.2 A spatial-contextual framework for class-wise statistical reduction of multispectral images

In order to introduce our statistical reduction method, we first need to define a novel spatially related statistical model for MS images which localizes a general multiclass distribution over homogeneous spatial regions of the image. For convenience we use a *continuous* (instead of discrete) model to the representation of the image. Thus, let us assume a d -band multispectral image to be a function of the type $y(x) : \Omega_0 \rightarrow \mathbb{R}^d$, where $\Omega_0 \subset \mathbb{R}^2$ is a rectangular domain and $x \in \Omega_0$ is the spatial coordinate. We assume the image formation to be a spatially localized statistical process of the form

$$p(y, x | \Omega, \Theta) = \frac{1}{|\Omega_0|} \sum_{i=1}^{n_\Omega} \chi_{\Omega_i}(x) p(y | \theta_i) \quad (8.1)$$

where:

- (i) $|\Omega_0|$ is the area of Ω_0 .
- (i) n_Ω is the number of classes in the image.
- (i) $\Omega := \{\Omega_i : i = 1, \dots, n_\Omega\}$ is a *regular* partition of Ω_0 , i.e., Ω_i are open connected sets, $\Omega_0 = K \cup (\cup_{i=1}^{n_\Omega} \Omega_i)$ and $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$. The set K is the total boundary of the partition. The function $\chi_{\Omega_i}(x)$ is the indicator of Ω_i .
- (i) $\Theta := \{\theta_i : i = 1, \dots, n_\Omega\}$ is a parametric set such that each $\theta_i = (\mu_i, \Sigma_i)$ and $p(y, \theta_i) = \mathcal{N}(y; \mu_i, \Sigma_i)$ is d -variate Gaussian with mean $\mu_i \in \mathbb{R}^d$ and covariance matrix $\Sigma_i \in \mathbb{R}^{d \times d}$.

With sufficient regularity assumptions on the boundary of the regions Ω_i , this model is aimed at concentrating pixels having similar statistical properties (under the Gaussian

assumption) inside them. This fact can be formally stated in a maximum-likelihood (ML) framework. Let Y, X be the random variables associated to the values y, x , then a simple computation shows that the conditional expectation $E[Y|X]$ is given by

$$E[Y|X] = \sum_{i=1}^{n_\Omega} \mu_i \chi_{\Omega_i}(x). \quad (8.2)$$

Therefore, under the statistical distribution $p(y, x|\Omega, \Theta)$, i.e., given the knowledge of Ω, Θ , the expected image is a piecewise constant function which takes the means μ_i as representative values for the corresponding classes.

8.2.1 A free-discontinuity variational model for boundary and parameter estimation

Given an image y drawn from the distribution $p(y, x|\Omega, \Theta)$, we turn into the problem of estimating the more likely partition Ω and the parametric set Θ that originated the data. In order to do that, we firstly show that if Ω is known, the means μ_i can be estimated from data by solving the following variational problem

$$\min_{u \in PC(\Omega)} E_0(u) := \int_{\Omega_0} |u - y|^2 dx \quad (8.3)$$

where $PC(\Omega)$ is the space of piecewise constant functions over the partition Ω . One can prove (we omit the mathematical proof) that the (unique) solution of (8.3) is $u^* = \overline{E[Y|X]} = \sum_{i=1}^{n_\Omega} \bar{\mu}_i \chi_{\Omega_i}(x)$, where $\bar{\mu}_i$ is the integral mean of y over Ω_i . Therefore, the solution is a piecewise constant function whose values are the class mean ML predictions estimated from data at each region of the image.

To explicitly estimate the location of the class boundaries, we introduce the partition Ω as a variable in the variational problem. In order to avoid meaningless solutions, we penalize the total length of the partition boundary $\mathcal{H}^1(K)$. In this way we come up with a problem similar to (8.3) with objective functional $E_0(u, K) := \int_{\Omega_0} |u - y|^2 dx + \nu \mathcal{H}^1(K)$, where ν is a positive parameter. This is the piecewise constant Mumford-Shah functional [3]. In literature there are numerical methods from computer vision and image processing to the minimization of functionals of this kind. We notice that convex relaxation methods [54] and primal-dual approaches [167] (faster than convex relaxation methods) cannot explicitly detect image discontinuities (class boundaries). Among methods that are able to detect image discontinuities we have the multiphase level set method [168] and the Ambrosio-Tortorelli approximation [17]. However, the multiphase level set method requires cumbersome computations and cannot be extended to more than four phases.

8.2.2 Numerical minimization approach

An effective matricial approach to address the numerical minimization of Ambrosio-Tortorelli approximations of free-discontinuity variational problems on large images has been proposed in Part I of this thesis. In the following this approach is used for numerically computing image approximations with low statistical variability and for explicitly detecting class boundaries, as described in Section 8.2.1. To do so, we need to relax the piecewise constant assumption by embedding the problem in a weaker space of (differentiable) functions and by adding a gradient penalization term to the objective energy. Therefore, we aim at minimizing

$$E(u, K) := \int_{\Omega_0 \setminus K} |\nabla u|^2 + \lambda |u - y|^2 dx + \nu \mathcal{H}^1(K), \quad (8.4)$$

the piecewise smooth Mumford-Shah functional [3]. It is worth noting that here u is not necessarily piecewise constant, but the gradient penalization forces it to be very smooth (nearly constant) outside the 1-dimensional set K , which is meant to be the set of discontinuities of u .

The numerical minimization of (8.4) can be addressed by following the approach proposed in Section 4 (see also [4]), which relies on a compact matricial formulation of the functional and an efficient iterative solver. This approach deals with an extended formulation of the variational problem depending also on second-order derivatives due to Blake and Zisserman, which includes as a particular case the Mumford-Shah functional. Straightforward modifications are in order to adapt the solver to the Mumford-Shah case. An interesting property of the proposed method is that it can be implemented in a parallel way by means of a simple interconnection rule, thus allowing the segmentation of large images in competitive time [169].

8.2.3 A parallel tiling approach

The segmentation of large images is typically addressed in the following manner: the input image is splitted into several tiles, a specific segmentation algorithm is run separately on each tile and the partial results are merged together. However, the global dependency of data of typical segmentation approaches produces undesired incompatibility of segments in the tile sides and some specific post-processing is often needed to remove artifacts even when some overlap between tiles is applied [170]. In variational models the solution is theoretically formulated as a minimizer of a global energy, see Section 2.2.2. However, being these energies non-convex, numerical methods for their solution always return local minimizers. In particular, the solutions are very dependent on initialization and they are energetically close to the initial data. This inherently weakens the constraint of global

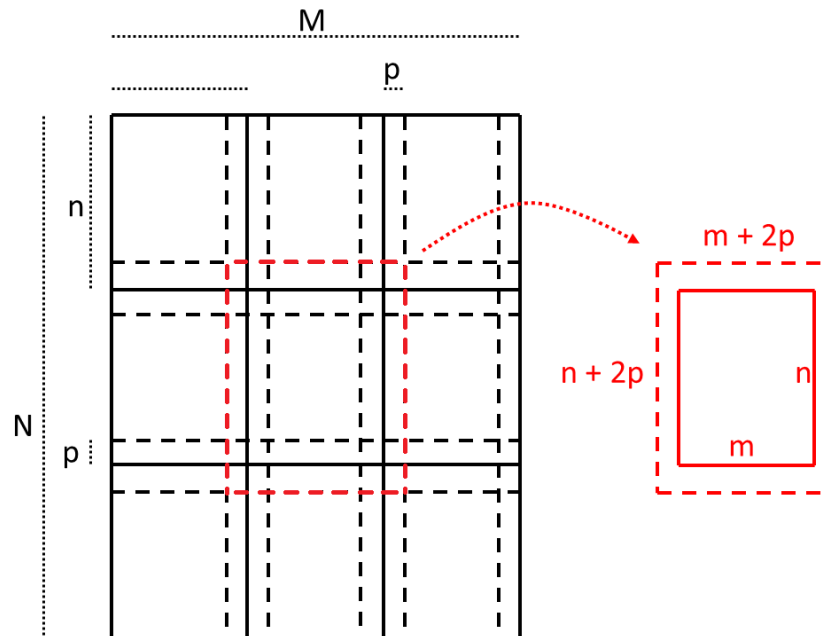


Figure 8.1: The proposed tiling scheme with related size parameters.

minimality of the solutions, and at the same time it causes local features of the initial datum to prevail. As a consequence, descent methods applied to variational problems such as MS and BZ are expected to have good locality properties in general. One particular property of this kind seems to be that local minimizers weakly depend on boundary conditions. So far, we only have experimental evidence of this, but theoretical studies are on going to formalize this concept. Therefore, the minimization can be (at least heuristically) addressed on separated tiles that are sufficiently large to let the effect of boundary conditions vanish far away from the boundary. We exploit this fact to show that a typical tiling procedure with partially overlapping tiles can be applied to a numerical method for the minimization of the BZ functional without any need of further processing on tiles junctions.

The functional is implemented with null boundary conditions, i.e., $v_{0,j} = v_{N+1,j} = v_{i,0} = v_{i,M+1} = 0$. Formally, the use of Dirichlet conditions corresponds to adding an artificial discontinuity all along the boundary of the image. During the functional minimization, the artificial discontinuity is detected and its influence vanishes elsewhere because of the localization of the functional (when $s, z = 1$, the discontinuity does not contribute to the energy but only the smoothing terms of the functional do). This justifies the use of a tiling approach for addressing the segmentation of large images, as long as we do not consider some pixels around the boundary. Indeed, if N and M are very large, the application of the method in [4] to the whole image might be too expensive. Firstly, the

domain Ω is splitted into partially overlapping tiles of size $(n+2p) \times (m+2p)$, with $n < N$ and $m < M$ and p is the number of overlapping pixels, see Figure 8.1. Straightforward modifications hold for boundary tiles. Then, the BCDA algorithm with Dirichlet (null) boundary conditions is run on each tile. Lastly, in the computed solutions, the stripes corresponding to overlapped pixels (that have width p) are cutted out of the tiles, and the resulting $n \times m$ pieces are merged. The use of Dirichlet boundary conditions makes the tiles independent of each other, as no boundary values must be stored for solving adjacent tiles. Therefore, very efficient implementation is possible on modern parallel architectures.

8.3 Binary change detection based on class-wise statistical reduction

Here we show how the proposed statistical reduction model based on the Mumford-Shah functional can be applied to the purpose of change detection on MS images. The nearly-piecewise constant approximation of MS images allows us to obtain a simplified version of the image with reduced statistical variability. As a consequence, the discriminative capability of statistical-based CD algorithms increases.

Let $y^1, y^2 : \Omega_0 \rightarrow \mathbb{R}^d$ two coregistered MS images acquired over the same area at different times. By similar arguments to those presented in [160], we can assume their difference $d = y^2 - y^1$ to follow a distribution of the type (8.1). In particular, if only one significant change occurred between the two observations, we have

$$p(d, x | \Omega, \Theta) = \frac{1}{|\Omega_0|} \sum_{i=1}^2 \chi_{\Omega_i}(x) p(d | \theta_i) \tag{8.5}$$

where: $\theta_1 = (\mu_1, \Sigma_1)$, with $\mu_1 = 0$, parametrizes the unchange class and $\theta_2 = (\mu_2, \Sigma_1)$ with $\mu_2 \neq 0$, parametrizes the change class. The binary dection of changes can be done by changing the coordinates to magnitude and performing the decion in a bayesian framework. The resulting statistical model describing the magnitude of the difference image is a Rayleigh-Rice mixture distribution [1]. However, the discrimination of the changes can be challenging if $\mu_2 \approx \mu_1$ (with possibly $\mu_1 \neq 0$) and Σ_1, Σ_2 have large norm. By reducing the statistical variability of the difference image using the approach proposed in Section 8.2 we can get a nearly piecewise constant image \bar{d} , where classes are highly distinguishable as covariance matrices have signicantly reduced norm. This is confirmed by numerical experiments that follows.

8.4 Experimental results

In this section we present the experimental results we obtained for both: validating the tiling approach proposed in Section 8.2.3, and demonstrating the effectiveness of the change detection method based on class-wise spatial-contextual approach proposed in Section 8.3.

8.4.1 Experimental validation of the tiling approach

The proposed algorithm has been successfully tested on many different datasets including medium and very-high resolution images and also Digital Surface Models (DSMs) [35]. Here we present a validation on one of them. The study dataset is a DSM¹ of Trento (Italy) with size 2020×2020 and spatial resolution of 1 mt (see also [4, 35]). We test the validity of the proposed tessellation scheme by considering as ground truth the solutions obtained by segmenting the whole dataset, as in [170]. Let g be the input DSM. We call u, s, z the solutions obtained on the whole dataset (ground truth), and u^t, s^t, z^t the solutions obtained by the tiling procedure described in Section 8.2.3. We set $n, m = 505$, therefore the tile grid is 4×4 , for a total of 16 tiles. The number of overlapping pixels is set to $p = 10$. The differences between the two solutions are analyzed both globally and locally, and an evaluation of the tiling approach can be given in terms of their closeness.

Global error analysis. The values of g range between $\min g = 183.7700$ and $\max g = 971.9160$ (meters), and the difference between u and the tiled solution u^t is only confined between $\min u - u^t = -0.5719$ and $\max u - u^t = 0.6179$. Moreover, more than 99,99% of the difference pixels ranges between -0.04 and 0.04 (thus, between -4 and 4 centimeters), and their distribution is given in Figure 8.2.² Similarly, we can also study the differences between discontinuity and gradient discontinuity mapping functions. Since the values of these functions range between 0 and 1, their difference is bounded in the interval $[-1, 1]$. From the histogram of $s - s^t$ and $z - z^t$ in Figure 8.2, we see that more than 99,99% of the difference pixels range between -0.025 and 0.025 (in case of $s - s^t$), and between -0.04 and 0.04 (in case of $z - z^t$). Thus, errors are not greater than the 5% of the pixel values, in all cases. Lastly, as global statistic, we report the mean of the difference images: $3.8414e - 05$ for $u - u^t$, $-1.1379e - 07$ for $s - s^t$, and $-1.7163e - 04$ for $z - z^t$.

Local error analysis. Dirichlet boundary conditions weakly affect the solutions of the minimization. In Figure 8.3, we plotted a portion of the difference $u - u^t$ and of the function z^t (gradient discontinuity map) in correspondence of the tile junctions. As we can see, the spatial distribution of peaks (high difference values) is not related with tile

¹Dataset details and download: <http://www.territorio.provincia.tn.it/portal/server.pt/community/lidar/847/lidar/23954>

²Since altimetric precision of input data is estimated at $\sigma = 30$ cm, the proposed tiling procedure does not alter the precision of data measurements.

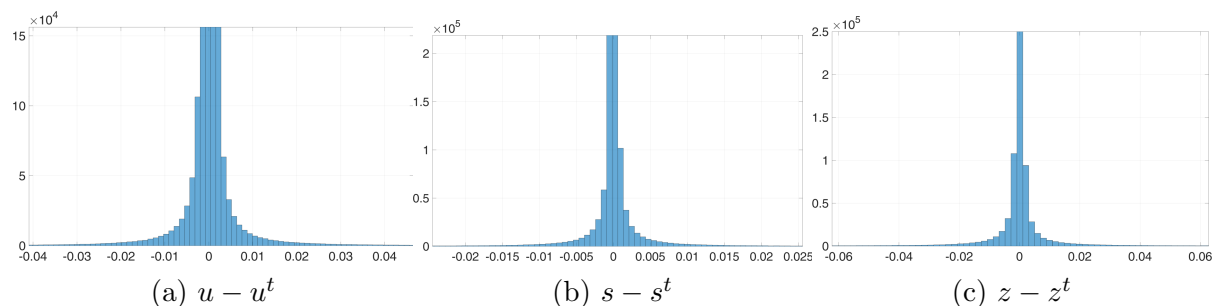


Figure 8.2: Hitograms of the differences between the solutions. x -axis (binned difference values) is restricted to a portion where more than 99,99% of pixels are counted. The central bars are cutted at the top.

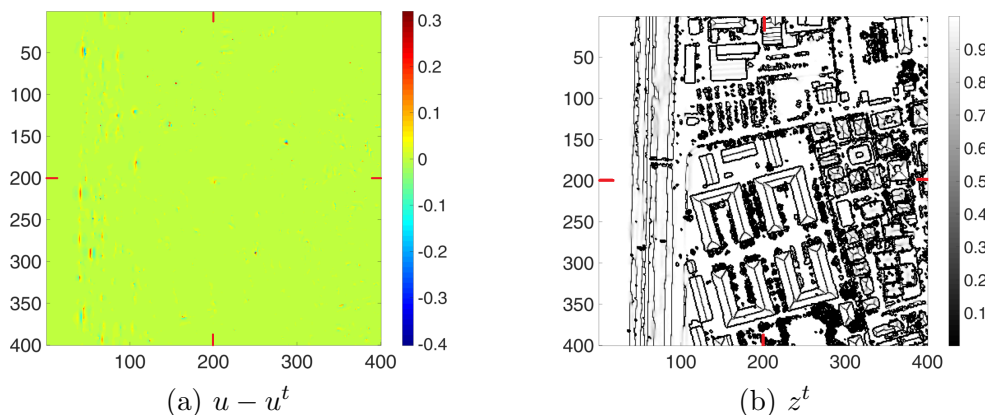


Figure 8.3: A portion of the difference $u - u^t$ and of the function z^t in correspondence of a cross tile junction (indicated by the red traits). The scene represents some old barracks and sourrounding area.

junctions, neither are discontinuity boundaries. For giving a quantitative measurement of this, let $D \subset \Lambda$ be the set of pixels (i, j) that surround all the tiles (i.e. for all j and fixed $i = 1, 505, 506, 1010, \dots, 2020$, and for all i with fixed $1, 505, 506, 1010, \dots, 2020$). For each variable $v \in \{u, s, z\}$ and for a given threshold value h , we define the quantity

$$P_v^h := \frac{\#\{(i, j) \in D : |u_{ij} - u_{ij}^t| > h\}}{\#\{(i, j) \in \Lambda : |u_{ij} - u_{ij}^t| > h\}}$$

that returns the ratio of the difference values greater that h that lie in D . We obtained $P_u^{0.01} = 0.0084$, $P_s^{0.005} = 0.0074$, and $P_z^{0.01} = 0.0069$. Thus, less than 1% of difference values greater than h are in proximity of tiles junctions. This shows the effectiveness of the proposed tiling strategy.

Time performance. We run computations with MATLAB[®] 2015b, on a standard workstation. Hardware is Intel(R) Core(TM) i5-4750 CPU @3.20 GHz, 8.00 GB Ram.

The BCDA algorithm converged in 261 seconds on the whole dataset, whereas it converged on average in 18 seconds on each tile. We see that the algorithm can benefit of ad-hoc parallel implementation, that would significantly enhance the performance in terms of execution time.

8.4.2 Change detection results

In this section we compare the results of binary change detection obtained in [1] with the results that can be obtained by prior application of class-wise statistical reduction to the difference image. The considered dataset is Dataset C from [1] and it consists of a couple of multispectral images acquired by the Operational Land Imager (OLI) multispectral sensor of the Landsat 8 satellite. The investigated area includes Lake Omodeo and a portion of Tirso River (Sardinia, Italy). The image size is 700×650 pixels (a total of 455000 pixels) at a spatial resolution of 30 m. The dates of acquisition are 25th July 2013 (pre) and 10th August 2013 (post). The change we are interested to estimate is a fire occurred between August 7th and 9th in the south of Ghilarza village. The post-event image is acquired just one day after the fire was extinguished. The area affected by the fire is mostly agricultural, with an extension of approximately 100 ha. The images are co-registered and radiometrically corrected. In Figure 8.4a,b the difference image and the reference map of the changes are shown.

The experiment is carried on as follows: the difference image d is calculated by pixel-wise subtraction, then the statistically reduced difference image \bar{d} is obtained by minimizing the functional (8.4) with d as input data (in place of y). Lastly, the binary detection is performed in the same manner as in [1]: the magnitude of \bar{d} is computed, the parameters of the Rayleigh-Rice mixture are estimated via EM algorithm and decision is performed according to a Bayes rule of minimum overall error. The minimization of the Mumford-Shah functional to approximate the difference image has been driven by a selection of the λ, μ parameters able to reduce the high variability of the two classes involved, i.e., the background (unchanged pixels) and the fire (changed pixels) and by enhancing their separability. In Figure 8.4c the difference image after class-wise spatial-contextual statistical reduction is illustrated. Notice how the smoothing effect that results from the minimization of the Mumford-Shah functional resembles a complex anisotropic application of the Gaussian filtering that affects only the spatial *interior* of the classes in the image, while it enhances their boundaries. Moreover, these boundaries are explicitly detected. We can see in Figure 8.4d that they accurately delineate the transition between the unchange and the change classes. As quantitative measurement of the reduction of class variability we report in Table 8.1 the variance of the pixel values (channels 5 and 6) for the change and the unchange classes in both the original and the reduced difference images.

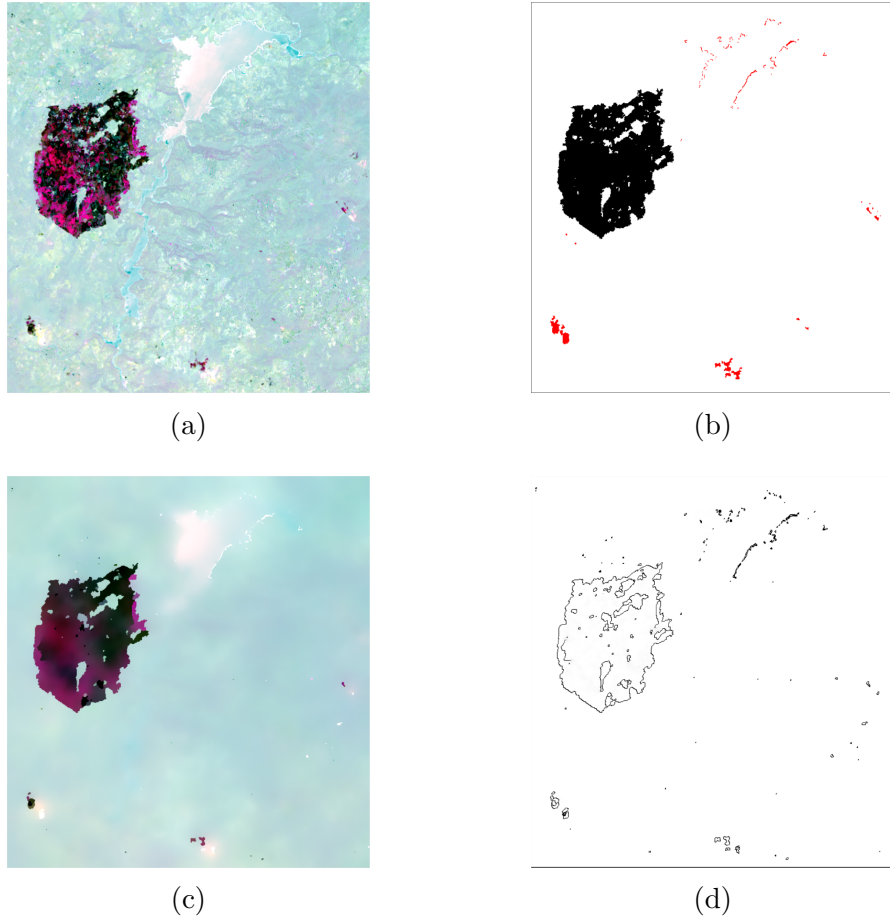


Figure 8.4: Landsat 8 dataset used in the experiments. (a) Original difference image (false color composition), see [1]. (b) Reference map of the changes. Black pixels belong to the fire, the red ones are minor changes unrelated to the fire. (c) Difference image after statistical reduction (false color composition), this image is used for computations in the experiments. (d) Map of the class boundaries extracted from the reduced image. The false color compositions are obtained by putting channels 3,5,6 of the Landsat image into R,G,B.

Table 8.1: Class variance calculated on both the original (d) and the reduced (\bar{d}) difference images using channels 5 and 6. The change class contains the fire, the unchange class all the rest.

channel	change		unchange	
	d	\bar{d}	d	\bar{d}
5	$8.30e^{-4}$	$3.88e^{-4}$	$1.15e^{-4}$	$5.02e^{-5}$
6	$1.60e^{-3}$	$5.47e^{-4}$	$1.10e^{-4}$	$4.98e^{-5}$

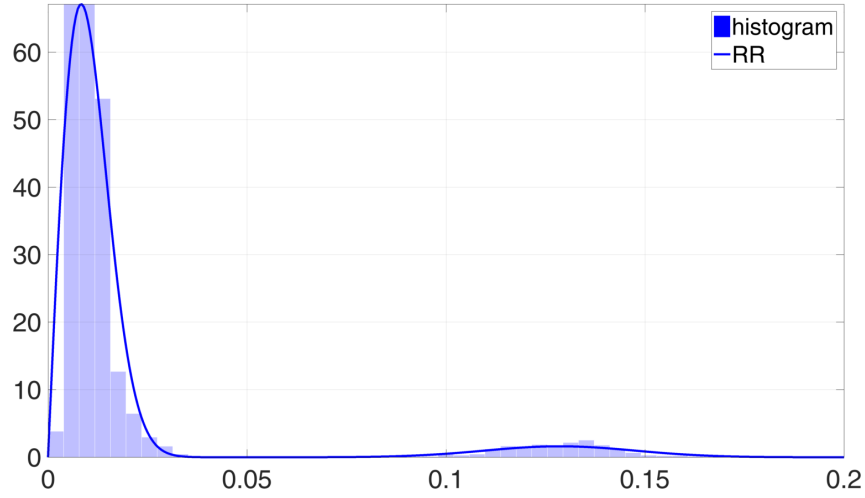


Figure 8.5: Histogram of the magnitude of the statistically reduced difference image and estimated density. Notice that the two classes are well separated.

Then, the magnitude of the reduced difference image \bar{d} is computed and the EM algorithm to parameter estimation of the Rayleigh-Rice mixture is run. Figure 8.6a shows the histogram of the magnitude of the reduced difference image with superimposed the estimated distribution. Notice that, the unchange (Rayleigh) and the change (Rice) classes are well separated by a large flat zone in the histogram. They are very distinguishable. Indeed, the EM algorithm converged in only one iteration. The computed threshold value for minimum error of classification associated to the Rayleigh-Rice (**rR**) model is 0.05. It is very close to the optimal one 0.07, which is calculated by trial and error procedure. As we can see from Table 8.2, both the calculated and the optimal threshold computed on the basis of the statistically reduced image \bar{d} return smaller overall percentage errors if compared to the same error measurements obtained in [1] (which are computed on the basis of the original difference image d). Without statistical reduction, the number of false alarms is much higher. This can be clearly seen also by looking at the change maps depicted in Figure 8.6, where we can notice that much less false alarms are present, if compared to [1] and the methods proposed in Chapters 6 and 7, where no reduction is applied.

8.5 Conclusion

In this chapter we proposed a novel approach to reduce the statistical variability of classes in multispectral images by means of a class-adaptive spatial-contextual approach. To prove its effectiveness, the class-wise statistical reduction approach is applied to change

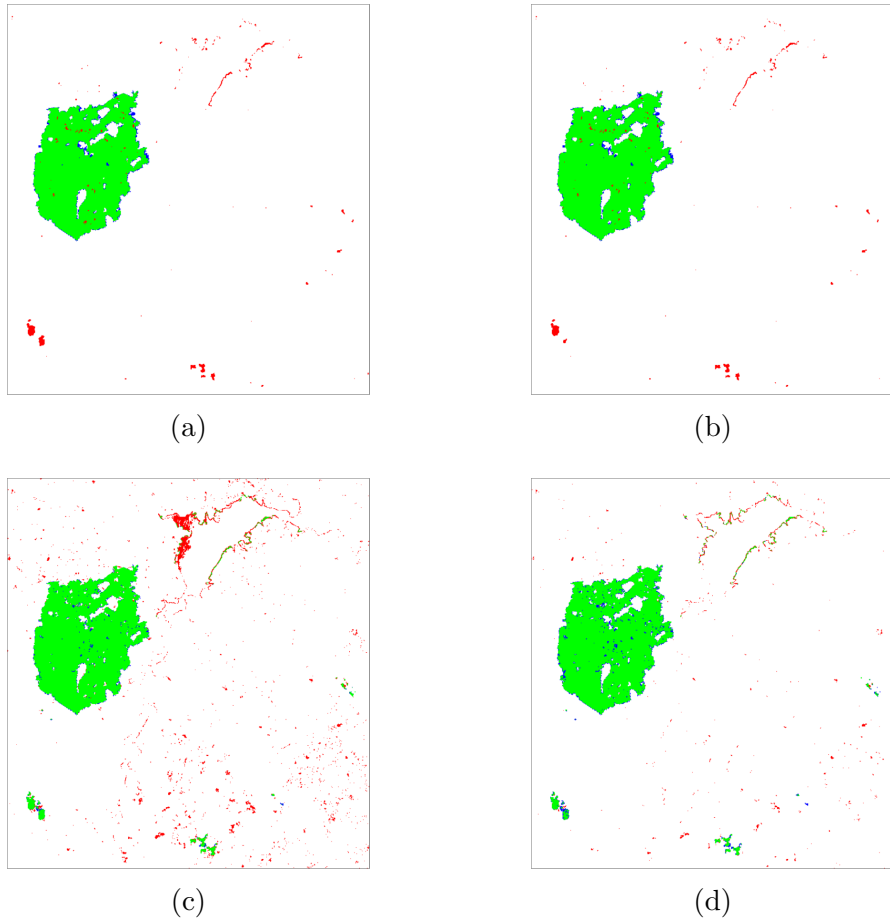


Figure 8.6: Results of the change detection. (a) CD map obtained with the Rayleigh-Rice binary detector (**rR**) applied to the reduced image, (b) CD map related to the optimal threshold (**opt**) applied to the reduced image. For comparison with the methods proposed in the two previous chapters: (c) CD map obtained with the **rR** binary detector (Chapter 6) applied to the image before reduction, and (d) CD map obtained with the **rrR** binary detector (Chapter 7) applied to the image before reduction. In all CD maps red, blue and green are false alarms, missed alarms and hits, respectively.

detection in multispectral images. After statistical reduction, the difference image presents two well distinguishable classes that are easily recognized by the binary detection algorithm. Thus, the change detection performance is significantly improved if reduction is applied.

Table 8.2: Change detection performance of the Rayleigh-Rice binary detector (**rR**) and optimal (**opt**) performance in the case of the original difference image d (see [1]) and the statistically reduced difference image \bar{d} . MA, FA and OE are missed alarms, false alarms and overall errors, respectively.

		MA (%)	FA (%)	OE (%)
d	rR	956 (2.75)	7805 (1.86)	8761 (1.93)
	opt	1964 (5.65)	2657 (0.63)	4621 (1.02)
\bar{d}	rR	1304 (3.75)	1600 (0.38)	2904 (0.64)
	opt	1331 (3.83)	1437 (0.34)	2768 (0.61)

Chapter 9

Conclusions and future developments

This chapter draws the conclusions of the research carried out in the PhD activity. It summarizes and discusses the results presented in this thesis and proposes some future research directions to follow for further development of the subjects proposed.

9.1 Conclusions

This thesis introduced novel mathematical models for the representation and analysis of remote sensing optical images, by focusing in particular on multispectral images. Multispectral images represent a precious source of information for global studies as they are collected with high frequency all over the globe by modern Earth observation satellites. This makes it possible to study the global change by analysis and comparison of multitemporal acquisitions. The ever increasing level of spatial/spectral detail available in last generation images requires novel and more flexible models for their representation. This is clearly seen in many applicative contexts. This thesis has proposed novel contributions to two important image processing domains: (1) mathematical methods to image approximation, and, (2) statistical methods to change detection in multispectral images.

In the first part of the thesis, the image model is presented in a mathematical way as a functional mechanism that maps the spectral response of the observed objects over a two-dimensional domain. Then, the concept of image approximation is introduced in a geometrical sense as the process that returns an ideal (simplified) representation of the image where the objects are clearly distinguishable. Being able to compute an approximation of an image is fundamental in many applicative areas. In this thesis we have investigated in great detail two variational models to image approximation due to Mumford-Shah and Blake-Zisserman. In the variational approach, the image is modeled as a composition of several pieces on which the image values are homogeneous in terms of derivatives and the boundary of image objects are identified as image discontinuities.

The solution of Mumford-Shah and Blake-Zisserman models involves the minimization of a variational functional depending on both function and set unknowns. Numerically, computing a solution is a difficult problem. In the thesis a numerical approach to the minimization of the two mentioned variational approaches to image approximation is proposed, with particular attention to the remote sensing application. Indeed, remote sensing highlights some well-known drawbacks of mathematical methods to image processing: the difficult scalability to large sized data and the extension of already existing models from their native scalar version to the vector-valued case. We addressed these important issues by proposing an efficient highly parallelizable numerical approach based on a compact matrixial formulation of the problem and a specifically tailored version of a block coordinate descent algorithm that guarantees convergence in competitive time. Several successful applications of variational methods have been proposed both in the remote sensing and the image processing domains.

In the second part of the thesis, we studied the change detection problem in multispectral images from a statistical viewpoint. In remote sensing, many change detection methods are based on the assumption that images can be modeled as realizations of a set of random variables following specific statistical distributions. Following the progresses in the development of modern multispectral sensors, more flexible models are needed to represent images as the statistical variability of last generation images is sensibly higher. Accordingly, it is observed that change detection methods that have worked well for previous generations of multispectral images do not provide the same quality of results if applied to more recent images. The theoretical models presented in the second part of this thesis are aimed at: (1) giving more degrees of freedom in the representation of multispectral images, (2) better exploiting the temporal correlation of the image pair for the physical interpretation of the changes, and, (3) extend the statistical model to a spatially related context. From these models, we derived some novel methods to change detection based on the study of the magnitude of the difference image. Firstly, an algorithm for binary change detection is derived based on a standard two-class (unchange/change) model. When coordinates are changed to magnitude unchanged and changed vectors are assumed to follow a Rayleigh and a Rician distribution, respectively. To perform change detection, parameter estimation of a mixture involving these distributions is crucial. Since no method was available in literature, in this thesis we proposed a theoretical derivation of a parameter estimation method based on the EM algorithm that is specifically tailored for the Rayleigh-Rice mixture. The experimental results presented in the thesis show that the binary detection of changes in multispectral images based on the Rayleigh-Rice mixture outperforms the classical change detection approaches based on Gaussian mixtures and other empirical thresholding approaches. Then, the two-class model is generalized

to a multiclass model. More specifically, we proposed a compound multiclass statistical model that, under certain assumptions, describes the relationship between the distribution of the classes populating the difference image and the distribution of natural classes in the image pair by taking into account the temporal correlation between the images. By taking advantage of the proposed compound multiclass approach the fitting and the detection performance are further improved. In particular, in several experimental tests on different image pairs a better modelization of the unchange class as a multiple class has returned an improved fitting of the magnitude histogram and a nearly optimal change detection accuracy. To account for the difficult application of statistical models for change detection on large images, we proposed an approach to image simplification aimed at reducing the statistical variability of classes while still preserving their spatial distribution. The proposed method relies on an extension of the multiclass image model to a spatially related context. This formulation allowed us to propose a new statistical interpretation of the piecewise smooth Mumford-Shah model, which becomes relevant in the solution of the change detection problem using statistical methods on large datasets. To enforce the remote sensing application, the method is coupled with a tiling approach that allows the processing of large images by exploiting a parallel implementation.

9.2 Future developments

The subjects proposed in this thesis lay the groundwork for several investigations. In the following, some planned research directions are proposed to deepen the proposed theoretical arguments and extend their applications.

Variational methods to image approximation demonstrated to be a valid tool in many different applications, e.g., statistical reduction of multispectral images, piecewise linear approximation of urban Digital Surface Models (DSMs), polygonal boundary approximation, scale space representation of VHR images, etc. In particular, the second-order model due to Blake-Zisserman has shown promising results if applied to DSMs for the detection of planar surfaces (e.g., building rooves) [35]. The automatic detection of geometric features, such as edges and creases, from objects represented by 3D point clouds is a very important issue in different application domains including urban monitoring and building reconstruction. A limitation of many methods in the literature is that they rely on rasterization or interpolation of the original grid, with consequent potential loss of detail. Currently, we are investigating the possibility of implementing a finite element algorithm (FEM) to variational approximation and edge detection on unstructured point clouds (e.g., LiDAR measurements, Tomographic SAR). Difficulties arise because the Hessian (2nd order operator) cannot be discretized by standard linearization techniques and

specific workarounds must be found. Preliminary results have been encouraging [171].

In Appendix 6.A.5 we derived a general version of the EM algorithm to parameter estimation of the mixture between generalized Rayleigh and Rician distributions to fit the distribution of the magnitude of a random variable described by a mixture of two B -variate Gaussians. In the change detection application, this allows us to exploit exactly B bands of the multispectral difference image to discriminate between unchanged and changed pixels. However, it is observed in some preliminary tests that, even by fixing all shape parameters, for increasing values of B the discriminative capability of the mixture decreases. In mathematical terms, we conjecture that the probability of misclassification error associated to the mixture tends to 1 as $B \rightarrow \infty$ (the curse of dimensionality due to the change of coordinates from cartesian (B -dimensional) to magnitude (1-dimensional)). In practical terms, a proof of the conjecture would imply the existence of an optimal selection of B (possibly depending on the data) such that the discriminative capability of the mixture is maximized.

The framework for class-wise spatial-contextual approximation of multispectral images based on the Mumford-Shah piecewise smooth approximation proposed in Chapter 8 allows us to extend the statistical approach to change detection on large images and gives us some advantages. On the one hand, the significant reduction of class variability (lower variances) augments the typical performance of parameter estimation algorithms as classes are more easily distinguishable. Therefore, many different classes have more chances to be discriminated and the statistical approach to change detection can be extended to the detection of different kinds of changes. On the other hand, the proposed approach represents a generalization of the statistical multiclass model for the difference image to a spatially related context. As we have seen in the thesis, the Mumford-Shah functional is in strict relationship with Gibbs distributions. Therefore, the proposed approach represents a valid alternative to change detection methods based on Markov Random Fields (MRFs) [82], or Spatially Variant Finite Mixtures (SVFM) [172]. A comparison of these approaches could clarify advantages and disadvantages of using these different approaches. Lastly, we are planning an integration of the proposed analysis of the difference image with the compound-classification approach to change detection [163], where the additional information about the physical meaning of the classes that originate after image differencing can be used to enforce the capability of the classifier to model temporal correlation in the data.

Bibliography

- [1] Massimo Zanetti, Francesca Bovolo, and Lorenzo Bruzzone. Rayleigh-Rice mixture parameter estimation via EM algorithm for change detection in multispectral images. *Image Processing, IEEE Transactions on*, 24(12):5004–5016, 2015.
- [2] M Drusch, U Del Bello, S Carlier, O Colin, V Fernandez, F Gascon, B Hoersch, C Isola, P Laberinti, P Martimort, et al. Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012.
- [3] David Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989.
- [4] Massimo Zanetti, Valeria Ruggiero, and Michele Miranda Jr. Numerical minimization of a second-order functional for image segmentation. *Communications in Nonlinear Science and Numerical Simulation*, 36:528–548, 2016.
- [5] Andrew Blake and Andrew Zisserman. *Visual reconstruction*. MIT Press Series in Artificial Intelligence. MIT Press, Cambridge, MA, 1987.
- [6] Andrew P Witkin. Scale-space filtering: A new approach to multi-scale description. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’84.*, volume 9, pages 150–153. IEEE, 1984.
- [7] Luis Alvarez, Frédéric Guichard, Pierre-Louis Lions, and Jean-Michel Morel. Axioms and fundamental equations of image processing. *Archive for rational mechanics and analysis*, 123(3):199–257, 1993.
- [8] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):629–639, 1990.
- [9] Leonid I Rudin and Stanley Osher. Total variation based image restoration with free local constraints. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 1, pages 31–35. IEEE, 1994.
- [10] Joachim Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.
- [11] Francine Catté, Pierre-Louis Lions, Jean-Michel Morel, and Tomeu Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM Journal on Numerical analysis*, 29(1):182–193, 1992.
- [12] Ennio De Giorgi. *Frontiers in Pure and Applied Mathematics, a collection of papers dedicated to J.L. Lions on the occasion of his 60th birthday*, chapter Free discontinuity problems in calculus of variations, pages 55–62. North-Holland P.C., 1991.
- [13] Bernd Kawohl. From Mumford–Shah to Perona–Malik in image processing. *Mathematical methods in the applied sciences*, 27(15):1803–1814, 2004.
- [14] Luigi Ambrosio, Loris Faina, and Riccardo March. Variational approximation of a second order free discontinuity problem in computer vision. *SIAM Journal on Mathematical Analysis*, 32(6):1171–1197, 2001.

- [15] E De Giorgi, M Carriero, and A Leaci. Existence theorem for a minimum problem with free discontinuity set. *Archive for Rational Mechanics and Analysis*, 108(4):195–218, 1989.
- [16] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. Oxford University Press, 2000.
- [17] Luigi Ambrosio and Vincenzo Maria Tortorelli. Approximation of functionals depending on jumps by elliptic functional via γ -convergence. *Communications on Pure and Applied Mathematics*, 43(8):999–1036, 1990.
- [18] L Modica and S Mortola. Un esempio di γ -convergenza. *Bollettino Unione Matematica Italiana*, 14(B):285–299, 1977.
- [19] G Bellettini and A Coscia. Discrete approximation of a free discontinuity problem. *Numerical Functional Analysis and Optimization*, 15(3-4):201–224, 1994.
- [20] Pasqua D’Ambra and Gaetano Tartaglione. Solution of Ambrosio–Tortorelli model for image segmentation by generalized relaxation method. *Communications in Nonlinear Science and Numerical Simulation*, 20(3):819–831, 2015.
- [21] A Chambolle and G Dal Maso. Discrete approximation of the Mumford-Shah functional in dimension two. *Mathematical Modelling and Numerical Analysis*, 33(4):651–672, 1999.
- [22] B Bourdin and A Chambolle. Implementation of an adaptive Finite-Element approximation of the Mumford-Shah functional. *Numerische Mathematik*, 85:609–646, 2000.
- [23] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1133–1140. IEEE, 2009.
- [24] Tony F Chan and Luminita A Vese. A level set algorithm for minimizing the Mumford-Shah functional in image processing. In *Variational and Level Set Methods in Computer Vision, 2001. Proceedings. IEEE Workshop on*, pages 161–168. IEEE, 2001.
- [25] Andrea Braides. *Approximation of free-discontinuity problems*. Number 1694. Springer-Verlag, 1998.
- [26] Michele Carriero, Antonio Farina, and Ivonne Sgura. Image segmentation in the framework of free discontinuity problems. In *Calculus of variations: topics from the mathematical heritage of E. De Giorgi*, volume 14 of *Quad. Mat.*, pages 86–133. Dept. Math., Seconda Univ. Napoli, Caserta, 2004.
- [27] Michele Carriero, Antonio Leaci, and Franco Tomarelli. A second order model in image segmentation: Blake & zisserman functional. In *Variational Methods for Discontinuous Structures*, volume 25 of *Progress in Nonlinear Differential Equations and Their Applications*, pages 57–72. Birkhäuser Basel, 1996.
- [28] Michele Carriero, Antonio Leaci, and Franco Tomarelli. Strong minimizers of blake & zisserman functional. In *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze*, volume 25, pages 257–285. Scuola Normale Superiore, Pisa, 1997.
- [29] Tommaso Boccellini and Franco Tomarelli. Generic uniqueness of minimizer for Blake & Zisserman functional. *Revista matemática computense*, 26(2):361–408, 2013.
- [30] M Carriero, A Leaci, and F Tomarelli. Uniform density estimates for Blake & Zisserman functional. *Discrete Contin. Dyn. Syst.*, 31(4):1129–1150, 2011.
- [31] Michele Carriero, Antonio Leaci, and Franco Tomarelli. Euler equations for Blake and Zisserman functional. *Calculus of Variations and Partial Differential Equations*, 32(1):81–110, 2008.
- [32] Michele Carriero, Antonio Leaci, and Franco Tomarelli. A candidate local minimizer of Blake and Zisserman functional. *Journal de mathématiques pures et appliquées*, 96(1):58–87, 2011.
- [33] Michele Carriero, Antonio Leaci, and Franco Tomarelli. A survey on the Blake–Zisserman functional. *Milan Journal of Mathematics*, 83(2):397–420, 2015.

- [34] Michele Carriero, Antonio Leaci, and Franco Tomarelli. Image inpainting via variational approximation of a Dirichlet problem with free discontinuity. *Advances in Calculus of Variations*, 7(3):267–295, 2014.
- [35] Massimo Zanetti and Alfonso Vitti. The Blake-Zisserman model for digital surface models segmentation. In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, volume II-5/W2, pages 355–360, 2013.
- [36] E De Giorgi and T Franzoni. Su un tipo di convergenza variazionale. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.*, 8(58):842–850, 1975.
- [37] Gianni Dal Maso. *An introduction to G -convergence*. Springer Science & Business Media, 1993.
- [38] Andrea Braides. *Γ -convergence for Beginners*, volume 22. Oxford University Press Oxford, 2002.
- [39] Gianpietro Del Piero, Giovanni Lancioni, and Riccardo March. A variational model for fracture mechanics: numerical experiments. *Journal of the Mechanics and Physics of Solids*, 55(12):2513–2537, 2007.
- [40] G. Bellettini and A. Coscia. Approximation of a functional depending on jumps and corners. *Boll. Un. Mat. Ital. B (7)*, 8(1):151–181, 1994.
- [41] P. C. Hansen, J. G. Nagy, and D. P. O’Leary. *Deblurring Images: Matrices, Spectra and Filtering*. SIAM, Philadelphia, 2006.
- [42] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to Total Variation for image analysis. *Radon Series Comp. Appl. Math.*, XX.
- [43] Luigi Grippo and Marco Sciandrone. Globally convergent block-coordinate techniques for unconstrained optimization. *Optimization Methods & Software*, 10(4):587–637, 1999.
- [44] R. A. Horn and C. R. Johnson. *Matrix Analysis (2nd ed.)*. Cambridge University Press, 2012.
- [45] Gianfranco Forlani, Carla Nardinocchi, Marco Scaioni, and Primo Zingaretti. Complete classification of raw LIDAR data and 3D reconstruction of buildings. *Pattern Analysis and Applications*, 8(4):357–374, 2006.
- [46] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
- [47] Alexander Brook, Ron Kimmel, and Nir A Sochen. Variational restoration and edge detection for color images. *Journal of Mathematical Imaging and Vision*, 18(3):247–268, 2003.
- [48] Peter Blomgren and Tony F Chan. Color TV: total variation methods for restoration of vector-valued images. *IEEE transactions on image processing*, 7(3):304–309, 1998.
- [49] Guillermo Sapiro and Dario L Ringach. Anisotropic diffusion of multivalued images with applications to color filtering. *IEEE transactions on image processing*, 5(11):1582–1586, 1996.
- [50] David Tschumperlé. Fast anisotropic smoothing of multi-valued images using curvature-preserving PDE’s. *International Journal of Computer Vision*, 68(1):65–82, 2006.
- [51] Matteo Focardi. On the variational approximation of free-discontinuity problems in the vectorial case. *Mathematical Models and Methods in Applied Sciences*, 11(04):663–684, 2001.
- [52] Matteo Focardi and Flaviana Iurlano. Ambrosio-Tortorelli approximation of cohesive fracture models in linearized elasticity. 2013.
- [53] Jonas Haehnle. Numerical approximations of the Mumford-Shah functional for unit vector fields. *Interfaces and Free Boundaries*, 13(3):297–326, 2011.
- [54] Evgeny Strekalovskiy, Antonin Chambolle, and Daniel Cremers. A convex representation for the vectorial Mumford-Shah functional. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1712–1719. IEEE, 2012.
- [55] Tony F Chan, B Yezriev Sandberg, and Luminita A Vese. Active contours without edges for vector-valued images. *Journal of Visual Communication and Image Representation*, 11(2):130–141, 2000.

- [56] Noha El-Zehiry, Prasanna Sahoo, and Adel Elmaghraby. Combinatorial optimization of the piecewise constant Mumford-Shah functional with application to scalar/vector valued and volumetric image segmentation. *Image and Vision Computing*, 29(6):365–381, 2011.
- [57] Miyoun Jung, Xavier Bresson, Tony F Chan, and Luminita A Vese. Nonlocal Mumford-Shah regularizers for color image restoration. *IEEE transactions on image processing*, 20(6):1583–1598, 2011.
- [58] Silvano Di Zenzo. A note on the gradient of a multi-image. *Computer vision, graphics, and image processing*, 33(1):116–125, 1986.
- [59] Raul Machuca and Keith Phillips. Applications of vector fields to image processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (3):316–329, 1983.
- [60] Ron Kimmel, Ravi Malladi, and Nir Sochen. Images as embedded maps and minimal surfaces: movies, color, texture, and volumetric medical images. *International Journal of Computer Vision*, 39(2):111–129, 2000.
- [61] David Mumford. Bayesian rationale for the variational formulation. In *Geometry-Driven Diffusion in Computer Vision*, pages 135–146. Springer, 1994.
- [62] Maria Giovanna Mora. Local calibrations for minimizers of the Mumford–Shah functional with a triple junction. *Communications in Contemporary Mathematics*, 4(02):297–326, 2002.
- [63] NY Graham. Smoothing with periodic cubic splines. *Bell System Technical Journal*, 62(1):101–110, 1983.
- [64] Yuedong Wang. *Smoothing splines: methods and applications*. CRC Press, 2011.
- [65] Pol Coppin, Inge Jonckheere, Kristiaan Nackaerts, Bart Muys, and Eric Lambin. Digital change detection methods in ecosystem monitoring: a review. *International journal of remote sensing*, 25(9):1565–1596, 2004.
- [66] Dengsheng Lu, P Mausel, E Brondizio, and Emilio Moran. Change detection techniques. *International journal of remote sensing*, 25(12):2365–2401, 2004.
- [67] Lorenzo Bruzzone and Francesca Bovolo. A novel framework for the design of change-detection systems for very-high-resolution remote sensing images. *Proceedings of the IEEE*, 101(3):609–630, 2013.
- [68] Dengsheng Lu, Guiying Li, and Emilio Moran. Current situation and needs of change detection techniques. *International Journal of Image and Data Fusion*, 5(1):13–38, 2014.
- [69] Ronald A Rensink. Change detection. *Annual review of psychology*, 53(1):245–277, 2002.
- [70] Sara Spotorno and Sylvane Faure. Change detection in complex scenes: hemispheric contribution and the role of perceptual and semantic factors. *Perception*, 40(1):5–22, 2011.
- [71] R. J. Radke, S. Andra, O. Al-Kofani, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, March 2005.
- [72] Y Ban and O Yousif. Change detection techniques: A review. In Y. Ban, editor, *Multitemporal Remote Sensing: Methods and Applications*. Springer, 2016.
- [73] Michele Volpi, Devis Tuia, Francesca Bovolo, Mikhail Kanevski, and Lorenzo Bruzzone. Supervised change detection in VHR images using contextual information and support vector machines. *International Journal of Applied Earth Observation and Geoinformation*, 20:77–85, 2013.
- [74] Lorenzo Bruzzone and Sebastiano B Serpico. An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):858–867, 1997.
- [75] Lorenzo Bruzzone, Roberto Cossu, and Gianni Vernazza. Detection of land-cover transitions by combining multirate classifiers. *Pattern Recognition Letters*, 25(13):1491–1500, 2004.

- [76] Lorenzo Bruzzone and Roberto Cossu. A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. *IEEE Transactions on Geoscience and Remote Sensing*, 40(9):1984–1996, 2002.
- [77] Begüm Demir, Francesca Bovolo, and Lorenzo Bruzzone. Detection of land-cover transitions in multitemporal remote sensing images with active-learning-based compound classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5):1930–1941, 2012.
- [78] F. Bovolo and L. Bruzzone. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1):218–236, January 2007.
- [79] Francesca Bovolo, Gustavo Camps-Valls, and Lorenzo Bruzzone. A support vector domain method for change detection in multitemporal images. *Pattern Recognition Letters*, 31(10):1148–1154, 2010.
- [80] L Bruzzone and D Fernandez Prieto. A minimum-cost thresholding technique for unsupervised change detection. *International Journal of Remote Sensing*, 21(18):3539–3544, 2000.
- [81] L Bruzzone and D Fernández Prieto. An adaptive parcel-based technique for unsupervised change detection. *International Journal of Remote Sensing*, 21(4):817–822, 2000.
- [82] L. Bruzzone and D. F. Prieto. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 38(3):1171–1182, May 2000.
- [83] Mauro Dalla Mura, Jon Atli Benediktsson, Francesca Bovolo, and Lorenzo Bruzzone. An unsupervised technique based on morphological filters for change detection in very high resolution images. *IEEE Geoscience and Remote Sensing Letters*, 5(3):433–437, 2008.
- [84] T Celik. Method for unsupervised change detection in satellite images. *Electron. Lett*, 46(9):624–626, 2010.
- [85] Til Aach and André Kaup. Bayesian algorithms for adaptive change detection in image sequences using Markov random fields. *Signal Processing: Image Communication*, 7(2):147–160, 1995.
- [86] F. Bovolo, S. Marchesi, and L. Bruzzone. A framework for automatic and unsupervised detection of multiple changes in multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2196–2212, May 2012.
- [87] Jin Chen, Xuehong Chen, Xihong Cui, and Jun Chen. Change vector analysis in posterior probability space: A new method for land cover change detection. *IEEE Geoscience and Remote Sensing Letters*, 8(2):317–321, 2011.
- [88] Josef Kittler and John Illingworth. Minimum error thresholding. *Pattern recognition*, 19(1):41–47, 1986.
- [89] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.
- [90] Prasanna K Sahoo, SAKC Soltani, and Andrew KC Wong. A survey of thresholding techniques. *Computer vision, graphics, and image processing*, 41(2):233–260, 1988.
- [91] Michele Volpi, Devis Tuia, Gustavo Camps-Valls, and Mikhail Kanevski. Unsupervised change detection with kernels. *IEEE Geoscience and Remote Sensing Letters*, 9(6):1026–1030, 2012.
- [92] Ashish Ghosh, Niladri Shekhar Mishra, and Susmita Ghosh. Fuzzy clustering algorithms for unsupervised change detection in remote sensing images. *Information Sciences*, 181(4):699–715, 2011.
- [93] Frank De Morsier, Devis Tuia, Volker Gass, Jean-Philippe Thiran, and Maurice Borgeaud. Unsupervised change detection via hierarchical support vector clustering. In *Pattern Recognition in Remote Sensing (PRRS), 2012 IAPR Workshop on*, pages 1–4. IEEE, 2012.

- [94] Maoguo Gong, Linzhi Su, Meng Jia, and Weisheng Chen. Fuzzy clustering with a modified MRF energy function for change detection in synthetic aperture radar images. *IEEE Transactions on Fuzzy Systems*, 22(1):98–109, 2014.
- [95] Allan Aasbjerg Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing*, 11(3):293–305, 2002.
- [96] T. Celik. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE Geoscience and Remote Sensing Letters*, 6(4):772–776, 2009.
- [97] M. Baisantry, D. S. Negi, and O. P. Manocha. Change vector analysis using enhanced PCA and inverse triangular function-based thresholding. *Defence Science Journal*, 62(4):236–242, 2012.
- [98] Silvia Marchesi and Lorenzo Bruzzone. ICA and kernel ICA for change detection in multispectral remote sensing images. In *IGARSS (2)*, pages 980–983, 2009.
- [99] Peng Gong. Change detection using principal component analysis and fuzzy set theory. *Canadian Journal of Remote Sensing*, 19(1):22–29, 1993.
- [100] John B Collins and Curtis E Woodcock. Change detection using the Gramm-Schmidt transformation applied to mapping forest mortality. *Remote Sensing of Environment*, 50(3):267–279, 1994.
- [101] John Rogan, Janet Franklin, and Dar A Roberts. A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery. *Remote Sensing of Environment*, 80(1):143–156, 2002.
- [102] A. A. Nielsen, K. Conradsen, and J. J. Simpson. Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64(1):1–19, 1998.
- [103] Allan Aasbjerg Nielsen. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image processing*, 16(2):463–478, 2007.
- [104] Amandine Robin, Lionel Moisan, and Sylvie Le Hégat-Masclé. An a-contrario approach for subpixel change detection in satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1977–1993, 2010.
- [105] Francesca Bovolo, Lorenzo Bruzzone, and Mattia Marconcini. A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7):2070–2082, 2008.
- [106] Paul C Smits and Alessandro Annoni. Updating land-cover maps by using texture information from very high-resolution space-borne imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3):1244–1254, 1999.
- [107] Glenn Healey and David Slater. Computing illumination-invariant descriptors of spatially filtered color image regions. *IEEE Transactions on Image Processing*, 6(7):1002–1013, 1997.
- [108] Liyuan Li and Maylor KH Leung. Integrating intensity and texture differences for robust change detection. *IEEE Transactions on image processing*, 11(2):105–112, 2002.
- [109] Gabriele Moser, Elena Angiati, and Sebastiano B Serpico. Multiscale unsupervised change detection on optical images by Markov random fields and wavelets. *IEEE Geoscience and Remote Sensing Letters*, 8(4):725–729, 2011.
- [110] Mauro Dalla Mura, Jón Atli Benediktsson, Björn Waske, and Lorenzo Bruzzone. Morphological attribute profiles for the analysis of very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(10):3747–3762, 2010.
- [111] Francesca Bovolo. A multilevel parcel-based approach to change detection in very high resolution multi-temporal images. *IEEE Geoscience and Remote Sensing Letters*, 6(1):33–37, 2009.

- [112] L. Bruzzone and D. F. Prieto. An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images. *IEEE Transactions on Image Processing*, 11(4):452–466, April 2002.
- [113] Ping Lu, André Stumpf, Norman Kerle, and Nicola Casagli. Object-oriented change detection for landslide rapid mapping. *IEEE Geoscience and remote sensing letters*, 8(4):701–705, 2011.
- [114] Chunlei Huo, Zhixin Zhou, Hanqing Lu, Chunhong Pan, and Keming Chen. Fast object-level change detection for VHR images. *IEEE Geoscience and Remote Sensing Letters*, 7(1):118–122, 2010.
- [115] Matthieu Molinier, Jorma Laaksonen, and Tuomas Hame. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):861–874, 2007.
- [116] Mark J Carlotto. A cluster-based approach for detecting man-made objects and changes in imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(2):374–387, 2005.
- [117] Geoffrey G Hazel. Object-level change detection in spectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3):553–561, 2001.
- [118] Francesca Bovolo and Lorenzo Bruzzone. An adaptive multiscale random field technique for unsupervised change detection in VHR multitemporal images. In *2009 IEEE International Geoscience and Remote Sensing Symposium*, volume 4, pages IV–777. IEEE, 2009.
- [119] Turgay Celik and Kai-Kuang Ma. Unsupervised change detection for satellite images using dual-tree complex wavelet transform. *IEEE Transactions on Geoscience and Remote Sensing*, 48(3):1199–1210, 2010.
- [120] Turgay Celik. Multiscale change detection in multitemporal satellite images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):820–824, 2009.
- [121] Turgay Celik and Kai-Kuang Ma. Multitemporal image change detection using undecimated discrete wavelet transform and active contours. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):706–716, 2011.
- [122] Xiaolong Dai and Siamak Khorram. The effects of image misregistration on the accuracy of remotely sensed change detection. *IEEE Transactions on Geoscience and Remote sensing*, 36(5):1566–1577, 1998.
- [123] Francesca Bovolo, Lorenzo Bruzzone, and Silvia Marchesi. Analysis and adaptive estimation of the registration noise distribution in multitemporal VHR images. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2658–2671, 2009.
- [124] David P Roy. The impact of misregistration upon composited wide field of view satellite data and implications for change detection. *IEEE Transactions on geoscience and remote sensing*, 38(4):2017–2032, 2000.
- [125] Silvia Marchesi, Francesca Bovolo, and Lorenzo Bruzzone. A context-sensitive technique robust to registration noise for change detection in VHR multispectral images. *IEEE Transactions on Image Processing*, 19(7):1877–1889, 2010.
- [126] Mingtao Ding, Zheng Tian, Zi Jin, Min Xu, and Chunxiang Cao. Registration using robust kernel principal component for object-based change detection. *IEEE Geoscience and Remote Sensing Letters*, 7(4):761–765, 2010.
- [127] Massimiliano Aldrighi and Fabio Dell’Acqua. Mode-based method for matching of pre-and postevent remotely sensed images. *IEEE Geoscience and Remote Sensing Letters*, 6(2):317–321, 2009.
- [128] Francesca Bovolo, Lorenzo Bruzzone, Luca Capobianco, Andrea Garzelli, Silvia Marchesi, and Filippo Nencini. Analysis of the effects of pansharpening in change detection on VHR images. *IEEE Geoscience and Remote Sensing Letters*, 7(1):53–57, 2010.

- [129] Xiajun Yang and CP Lo. Relative radiometric normalization performance for change detection from multi-date satellite images. *Photogrammetric Engineering and Remote Sensing*, 66(8):967–980, 2000.
- [130] Shilpa Inamdar, Francesca Bovolo, Lorenzo Bruzzone, and Subhasis Chaudhuri. Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4):1243–1252, 2008.
- [131] S. Singh and R. Talwar. Review on different change vector analysis algorithms based change detection techniques. In *Image Information Processing (ICIIP), 2013 IEEE Second International Conference on*, pages 136–141, 2013.
- [132] William A Malila. Change vector analysis: an approach for detecting forest changes with landsat. In *LARS Symposia*, page 385, 1980.
- [133] R. D. Johnson and E. S. Kasischke. Change vector analysis: A technique for the multispectral monitoring of land cover and condition. *International Journal of Remote Sensing*, 19(3):411–426, 1998.
- [134] R. Simoes and C. Slump. Change detection and classification in brain MR images using change vector analysis. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 7803–7807, 2011.
- [135] S. Liu, L. Bruzzone, F. Bovolo, and Peijun Du. Hierarchical unsupervised change detection in multitemporal hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1):244–260, January 2015.
- [136] A. Hennemuth, A. Seeger, O. Friman, S. Miller, B. Klumpp, and S. Oeltze H. O. Peitgen. A comprehensive approach to the analysis of contrast enhanced cardiac MR images. *IEEE Transactions on Medical Imaging*, 27(11):1592–1610, 2008.
- [137] Ole. T. Karlsen, R. Verhagen, and Wim M.M.J. Bovee. Parameter estimation from Rician-distributed data sets using a maximum likelihood estimator: Application to T1 and perfusion measurements. *Magnetic resonance in medicine*, 41(3):614–623, 1999.
- [138] T. Benedict and T. Soong. The joint estimation of signal and noise from the sum envelope. *IEEE Transactions on Information Theory*, 13(3):447–454, 1967.
- [139] J. Sijbers, A. J. Den Dekker, J. Van Audekerke, M. Verhoye, and D. Van Dyck. Estimation of the noise in magnitude MR images. *Magnetic Resonance Imaging*, 16(1):87–90, 1998.
- [140] J. J. Sijbers, A. J. den Dekker, P. Scheunders, and D. Van Dyck. Maximum-likelihood estimation of Rician distribution parameters. *IEEE Transactions on Medical Imaging*, 17(3):357–361, 1998.
- [141] S. O. Rice. Mathematical analysis of random noise. *Bell System Technical Journal*, 23(3):282–332, 1944.
- [142] K. K. Talukdar and W. D. Lawing. Estimation of the parameters of the Rice distribution. *The Journal of the Acoustical Society of America*, 89(3):1193–1197, 1991.
- [143] J. Sijbers and A. J. Den Dekker. Maximum likelihood estimation of signal amplitude and noise variance from MR data. *Magnetic Resonance in Medicine*, 51(3):586–594, 2004.
- [144] C. F. Carobbi and M. Cati. The absolute maximum of the likelihood function of the Rice distribution: Existence and uniqueness. *IEEE Transactions on Instrumentation and Measurement*, 57(4):682–689, 2008.
- [145] L. Lauwers, K. Barbé, W. Van Moer, and R. Pintelon. Estimating the parameters of a Rice distribution: A Bayesian approach. In *IEEE Instrumentation and Measurement Technology Conference (I2MTC'09)*, pages 114–117, May 2009.
- [146] R. Maitra and D. Faden. Noise estimation in magnitude MR datasets. *IEEE Transactions on Medical Imaging*, 28(10):1615–1622, 2009.
- [147] R. Maitra. On the Expectation-Maximization algorithm for Rice-Rayleigh mixtures with application to noise parameter estimation in magnitude MR datasets. *Sankhya B*, 75(2):293–318, 2013.

- [148] J. Rajan, D. Poot, J. Juntu, and J. Sijbers. Noise measurement from magnitude MRI using local estimates of variance and skewness. *Physics in medicine and biology*, 55(16):N441, 2010.
- [149] R. D. Nowak. Wavelet-based Rician noise removal for magnetic resonance imaging. *IEEE Transactions on Image Processing*, 8(10):1408–1419, 1999.
- [150] C. H. Chen. *Image Processing for Remote Sensing*. CRC Press, Boca Raton, FL, first edition, 2007.
- [151] G. N. Watson. *A treatise on the Theory of Bessel Functions*. Cambridge University Press, second edition, 1966.
- [152] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.
- [153] R. A. Redner and H. F. Walker. Mixture densities, Maximum Likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.
- [154] Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.
- [155] D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B*, 49(1):1–39, 1987. With a discussion.
- [156] R. L. Burden and J. D. Faires. *Numerical Analysis*. Brooks/Cole, Cengage Learning, Boston, MA, ninth edition, 2011.
- [157] Kenneth S Miller. *Multivariate distributions*. RE Krieger Pub. Co., 1975.
- [158] Marvin K Simon. *Probability distributions involving Gaussian random variables: A handbook for engineers and scientists*. Springer Science & Business Media, 2007.
- [159] Jerry W Vickers. A parameter estimation technique for the generalized Rayleigh-Rician distribution and Laha’s Bessel distribution. Technical report, DTIC Document, 1976.
- [160] Massimo Zanetti and Lorenzo Bruzzone. A generalized statistical model for binary change detection in multispectral images. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 3378–3381. IEEE, 2016.
- [161] Francesca Bovolo and Lorenzo Bruzzone. The time variable in data fusion: A change detection perspective. *Geoscience and Remote Sensing Magazine, IEEE*, 3(3):8–26, 2015.
- [162] Ashbindu Singh. Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003, 1989.
- [163] B. Demir, F. Bovolo, and L. Bruzzone. Classification of time series of multispectral images with limited training data. *IEEE Transactions on Image Processing*, 22(8):3219–3233, August 2013.
- [164] Begüm Demir, Francesca Bovolo, and Lorenzo Bruzzone. Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(1):300–312, 2013.
- [165] Zeki Yetgin. Unsupervised change detection of satellite images using local gradual descent. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(5):1919–1929, 2012.
- [166] Sicong Liu, Lorenzo Bruzzone, Francesca Bovolo, Massimo Zanetti, and Peijun Du. Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions on*, 53(8):4363–4378, 2015.
- [167] Evgeny Strelakovsky and Daniel Cremers. Real-time minimization of the piecewise smooth Mumford-Shah functional. In *European Conference on Computer Vision*, pages 127–141. Springer, 2014.

-
- [168] Luminita A Vese and Tony F Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International journal of computer vision*, 50(3):271–293, 2002.
- [169] Massimo Zanetti, Riccardo Zanella, and Lorenzo Bruzzone. A tiling procedure for second-order variational segmentation of large size remote sensing images. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 6902–6905. IEEE, 2016.
- [170] Pierre Lassalle, Jordi Inglada, Julien Michel, Manuel Grizonnet, and Julien Malik. Large scale region-merging segmentation using the local mutual best fitting concept. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pages 4887–4890. IEEE, 2014.
- [171] Massimo Zanetti and Lorenzo Bruzzone. Edge-crease detection and surface reconstruction from point clouds using a second-order variational model. In *SPIE Remote Sensing*, pages 924409–924409. International Society for Optics and Photonics, 2014.
- [172] S Sanjay-Gopal and Thomas J Hebert. Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm. *IEEE Transactions on Image Processing*, 7(7):1014–1028, 1998.