Hierarchical joint classification models for multi-resolution, multi-temporal and multi-sensor remote sensing images. Application to natural disasters

Ihsen Hedhli

UNIVERSITÉ NICE SOPHIA ANTIPOLIS
ECOLE DOCTORALE STIC
SCIENCES ET TECHNOLOGIES DE L'INFORMATION
ET DE LA COMMUNICATION

P H D T H E S I S

to obtain the title of

PhD of Science

of the University of Nice - Sophia Antipolis
Specialty : Signal and Image Processing

Defended by
Ihsen Hedhli

Hierarchical joint classification models for multi-resolution, multi-temporal and multi-sensor remote sensing images. Application to natural disasters

Thesis Advisors: Josiane Zerubia and Gabriele Moser
prepared at INRIA Sophia Antipolis, Ayin Team
defended on March 18, 2016

JURY

President :		-	
Reviewers :	Dr. Zoltan Kato	-	University of Szeged, Hungary
	Dr. Farid Melgani	-	University of Trento, Italy
Examinator :	Prof. Paolo Gamba	-	University of Pavia, Italy
Advisor :	Prof. Josiane Zerubia	-	INRIA (Ayin), France
Co-Advisor :	Dr. Gabriele Moser	-	University of Genoa (DITEN), Italy

UNIVERSITY OF GENOA

CICLO XXVIII

Hierarchical joint classification models for multi-resolution, multi-temporal and multi-sensor remote sensing images. Application to natural disasters

A Dissertation for the Degree of Doctor of Philosophy in

Monitoraggio dei sistemi e gestione dei rischi ambientali

March 2016

Author: Ihsen HEDHLI

Tutors: Dr. Gabriele MOSER and Prof. Josiane ZERUBIA

Chairperson: Dr. Roberto SACILE

Abstract:
The capabilities to monitor the Earth's surface, notably in urban and built-up areas, for example in the framework of the protection from environmental disasters such as floods or earthquakes, play important roles in multiple social, economic, and human viewpoints. In this framework, accurate and time-efficient classification methods are important tools required to support the rapid and reliable assessment of ground changes and damages induced by a disaster, in particular when an extensive area has been affected. Given the substantial amount and variety of data available currently from last generation very-high resolution (VHR) satellite missions such as Plèiades, COSMO-SkyMed, or RadarSat-2, the main methodological difficulty is to develop classifiers that are powerful and flexible enough to utilize the benefits of multi-band, multi-resolution, multi-date, and possibly multi-sensor input imagery. With the proposed approaches, multi-date/multi-sensor and multi-resolution fusion is based on explicit statistical modeling. The method combines a joint statistical model of multi-sensor and multi-temporal images through hierarchical Markov random field (MRF) modeling, leading to statistical supervised classification approaches. We have developed novel hierarchical Markov random field models, based on the marginal posterior modes (MPM) criterion, that support information extraction from multi-temporal and/or multi-sensor information and allow the joint supervised classification of multiple images taken over the same area at different times, from different sensors, and/or at different spatial resolutions. The developed methods have been experimentally validated with complex optical multi-spectral (Plèiades), X-band SAR (COSMO-Skymed), and C-band SAR (RadarSat-2) imagery taken from the Haiti site. The experimental results suggest that the methods are able to provide accurate classification maps from input heterogeneous imagery. Experimental results and comparisons with the state of the art suggests the effectiveness of the proposed approaches in fusing multiple information sources for classification purposes. Indeed, the proposed techniques were demonstrated to be advantageous in terms of classification accuracy on the test set, spatial regularity of the classification maps, minimization of spatial artifacts, and tradeoff with respect to computation time.

Keywords: Satellite images, image time series, multi-resolution, multi-sensor, quad-tree, classification, hierarchical Markov random fields, MPM.

Résumé

Les moyens mis en oeuvre pour surveiller la surface de la Terre, notamment les zones urbaines, en cas de catastrophes naturelles telles que les inondations ou les tremblements de terre, et pour évaluer l'impact de ces événements, jouent un rôle primordial du point de vue sociètal, économique et humain. Dans ce cadre, des méthodes de classification prècises et efficaces sont des outils particulièrement importants pour aider à l'évaluation rapide et fiable des changements au sol et des dommages provoqués. Étant données l'énorme quantité et la variété des données haute Résolution (HR) disponibles grâce aux missions satellitaires de dernière génération et de différents types telles que Pléiades, COSMO-SkyMed ou RadarSat-2 la principale difficulté est de trouver un classifieur qui puisse prendre en compte des données multi-bande, multi-résolution, multi-date et éventuellement multi-capteur tout en gardant un temps de calcul acceptable. Les approches de classification multi-date/multi-capteur et multi-résolution sont fondées sur une modélisation statistique explicite. En fait, le modèle développé consiste en un classifieur bayésien supervisé qui combine un modèle statistique conditionnel par classe intégrant des informations pixel par pixel à la même résolution et un champ de Markov hiérarchique fusionnant l'information spatio-temporelle et multi-résolutions, en se basant sur le critère des Modes Marginales a Posteriori (MPM en anglais), qui vise à affecter à chaque pixel l'étiquette optimale en maximisant récursivement la probabilité marginale a posteriori, étant donné l'ensemble des observations multi-temporelles ou multi-capteur. Les méthodes développées ont été validées expérimentalement avec des séries temporelles d'images optiques (Pléiades), Radar en bande X (COSMO-SkyMed), et Radar en bande C (RADARSAT-2) prisent sur Haiti. Lorsqu'elle est appliquée à des images à haute résolution, la méthode proposée donne une précision globale de la classification de valeur élevée avec un temps de calcul raisonnable grâce à la structure hiérarchique utilisée. En effet, les techniques proposées se sont montrées avantageuses en termes de précision de la classification sur l'ensemble des échantillons de validation, la régularité spatiale des cartes de classification, la minimisation des artefacts spatiales causé par la structure hiérarchique utilisée, et le compromis par rapport au temps de calcul.

Mots-clés: Images satellitaires, series temporelles, multi-résolution, multi-capteur, quad-arbre, classification, champs de Markov hiérarchiques, MPM.

Contents

List of Tables

List of Figures

Introduction

Earthquakes, tsunamis, floods, fires and volcanic eruptions are among the many natural disasters we are exposed to on Earth. As well as being dangerous to humans, these events play important roles in multiple social, economic, and human viewpoints. Faced with such phenomena, there is a crucial need for adequate and rapid damage assessment. However, when an extensive area is affected, the assessment of the damage becomes a complex and time consuming task, usually conduced by human experts. Nonetheless, the capabilities to monitor the Earth's surface using computer vision algorithms, notably in urban and built-up areas, for environmental disasters and to assess the ground effects and damage of such events play important roles. Given a monitored area, remote sensing allows elements at risk to be identified and their vulnerabilities to be quantified. After a disaster actually occurs, this prior extracted information can be combined with results of multi-temporal remote sensing image analysis, ancillary information, and possible physically-based modeling to estimate damage [Serpico et al., 2012].

In this framework, accurate and time-efficient classification methods for single-date and multi-temporal imagery are important tools required to support rapid and reliable extraction of information on a monitored region, especially when an extensive area is considered. Given the substantial amount and variety of data available currently from last-generation very-high resolution (VHR) satellite missions, the main difficulty is to develop a classifier that utilizes the benefits of multi-resolution, multi-date, multi-sensor, and multi-frequency input imagery. From an application-oriented viewpoint, the goal is to take advantage of this variety of input sources to maximize accuracy and effectiveness of the resulting thematic mapping products. From a methodological viewpoint, this goal claims for the development of novel data fusion techniques that are flexible enough to support the joint classification of images collected on the same area by different sensors at different times and associated with multiple spatial resolutions and wavelength ranges. In this manuscript, this fusion problem is addressed by developing four novel techniques, whose common trait is a mathematical formalization based on hierarchical Markov random field (MRF) models. On the one hand, the use of multi-resolution and multi-band imagery has been previously shown to optimize the classification results in terms of accuracy and computation time, and on the other hand, the integration of the temporal dimension into a classification scheme can both enhance the results in terms of reliability and capture the evolution in time of the monitored area. However, the joint problem of the fusion of several distinct data modalities (e.g., both multi-resolution and multi-temporal or multi-sensor) has been much more scarcely addressed in the remote sensing literature so far. The next paragraphs briefly summarize the key ideas of the developed methods and how their presentations are organized within the various chapters of the thesis.

Addressed problems and thesis organization

In this manuscript, a set of hierarchical methods to fuse multi-date, multi-resolution, multi-sensor, and multi-frequency remote sensing imagery for supervised classification is developed and experimentally validated with challenging multi-modal imagery from the topical test site of Haiti.

Chapter 1

This chapter presents an introduction to remote sensing images and the different sensors and their properties, with a special focus on optical and synthetic aperture radar (SAR) data: how do the sensors work, what are the possible sensor variants, what are their advantages, weaknesses, uses, and what properties of a given scene are observed with them, are among the items recalled in this chapter. The motivation to jointly use heterogeneous sensors is also discussed.

Chapter 2

This chapter summarizes various approaches to data fusion proposed in the literature with regard to multi-modal remote sensing imagery for classification purposes. First, a variety of multi-scale and multi-resolution fusion approaches are reviewed. Then, the complementarity between various sensors, especially when operating in the microwave and optical ranges, is stressed by presenting several models used in the literature. Finally, a thorough review of multi-temporal remote sensing image analysis methods is presented.

Chapter 3

This chapter provides all necessary background about MRFs, such as probabilistic image models and methodological tools for classification, and gives a Bayesian justification for the form of their energy functions. In addition, the graphical representation of MRFs is detailed in order to point out an important class of MRF models that rely on a probabilistic causality concept captured by the factorization of the prior distribution in term of causal transitions.

Chapter 4

The objective of this chapter is to describe multi-resolution classification techniques based on hierarchical MRFs using quad-tree structures. The use of quad-trees to model scale-to-scale interactions is justified by their causality properties over scale and by the possibility of employing a fast optimization method. However, these structures may induce blocky artifacts in the final classification map. This chapter also presents several techniques to mitigate such undesirable effect.

Chapter 5

During this chapter, a novel framework for joint multi-temporal and multi-resolution classification is presented. Multi-date and multi-resolution image classification is based on explicit statistical modeling through a hierarchical MRF. The proposed model allows both the input data collected at multiple resolutions and additional multi-scale features derived through wavelets to be fused. The described approach consists of a supervised Bayesian classifier that combines (i) a joint class-conditional statistical model for pixel-wise information and (ii) a hierarchical MRF for spatio-temporal and multi-resolution contextual information. Step (i) addresses the modeling of the statistics of the spectral channels acquired at each resolution and conditioned to each class. Step (ii) consists of integrating this statistical modeling in a hierarchical Markov random field for each date. A novel element of the proposed approach is the use of multiple quad-trees in cascade, each one being associated with each new available image at different dates, to characterize the temporal correlations associated with distinct images in the input time series. The transition probabilities between the scales and between different dates determine the hierarchical MRF because they formalize the causality of the statistical interactions involved. Inference and probability density modeling are accomplished by integrating the marginal posterior mode (MPM) criterion, the modified Metropolis dynamics (MMD) energy minimization technique, finite Gaussian mixture models, and the stochastic expectation-maximization (SEM) algorithm. Experimental results using multi-temporal and multi-resolution Pléiades images on Haiti by employing this new hierarchical model for time series classification are presented.

Chapter 6

In this chapter, two methods are proposed for the joint supervised classification of multi-sensor images including SAR and optical components and acquired at multiple spatial resolutions. The rationale of both approaches is to take benefit of the data fusion capabilities of hierarchical MRFs while the computation of the joint statistics of optical and SAR data, for which no closed form parametric models are available, is avoided. A novel approach, based on multiple quad-trees in cascade, to multi-sensor and multi-resolution fusion is described. In this framework, the first proposed method addresses the general problem of the joint classification of a SAR and an optical image acquired over the same area at high spatial resolutions. For each sensor, one of these two input images is associated with a separate quad-tree structure on the basis of its resolution. The proposed approach formalizes a supervised Bayesian classifier within this multiple quad-tree topology that combines a class-conditional statistical model for pixel-wise information and a hierarchical MRF for multi-sensor and multi-resolution contextual information. The second proposed method focuses on the specific case of the fusion of multi-frequency SAR data collected by the COSMO-SkyMed (X band) and RADARSAT-2 (C band) sensors together with optical Pléiades data. A multiple quad-tree structure is used again but optical and

SAR data are jointly included in each quad-tree to take benefit of both the best resolution available from each sensor and the synergy between the data offered by these opical space-borne missions for Earth observation.

Chapter 7

Throughout this chapter, the quad-tree approach is extended and a novel contextual multi-scale technique is proposed in order to classify multi-resolution remote sensing data, incorporate spatial contextual information, and mitigate possible blocky artifacts while keeping the causality of the hierarchical model. This is accomplished by proposing a novel integrated technique that combines the hierarchical MRF and Markov mesh approaches to take benefit of their causality properties in the scale and spatial domains, respectively. The focus is on the fusion of multi-resolution and spatial-contextual information for the supervised classification of single-date imagery. A novel formulation of the MPM criterion is developed and a preliminary experimental analysis is conducted.

Conclusion and perspectives

In this chapter, conclusions on the proposed methods are drawn, along with comments on their possible relevance in the current remote sensing framework, and future possible extensions are briefly outlined.

Chapter 1

Remote sensing and image analysis

Contents

1.1 Remote sensing system

Remote sensing is the discipline whose goal is to obtain information about an object, area or phenomenon through the analysis of data acquired by a device which is not in physical contact with the object being investigated [Lillesand et al., 2014]. In modern usage and in the application to Earth observation (EO) the term generally refers to the use of aerial or satellite sensor technologies, using various categories of sensors; we remotely collect data that may be analyzed to obtain information about the object under investigation. The remotely collected data can be of many forms that are potentially available from the energy filed arising from the Earth's surface including *spectral*, *spatial* and *temporal* fields [Landgrebe, 2005]. The variations of these fields implies variations in force distributions (for example measuring Gravity from a moving aircraft), acoustic wave distributions (for instance using torque and tire pressure sensors in automotive applications) or electromagnetic energy distributions (for example via radar sensors on-board artificial satellites). Though Gravity and acoustic fields are useful for this purpose within many applications, we will focus on the electromagnetic energy sensors that are currently being operated from airborne and space-borne platforms to gather information remotely from the Earth surface. Figure 1.1 schematically illustrates the generalized processes and elements involved in Earth observation system. As the figure shows, the system may be divided into three basic parts: the scene, the sensor, and the processing system.

The scene is the part of the system to be analyzed, it lies in front of the sensor and consists of the Earth's surface and its atmosphere. This portion of the system is characterized by not being under human control during all the observation process.

The sensor is the part of the system that collects the data to be analyzed. There are mainly two sensor categories : *Passive sensors* which measure reflected sunlight that was emitted from the sun and *active sensors* which have their own source of "illumination" and measure reflected (backscattered) energy. Furthermore, sensors may be carried on either an *airborne* (aircraft, drone) or *space-borne* (satellite) platforms depending on the use of the prospective images. The sensor system is usually under human control especially during the system design phase.

The processing system is the part of the system control whose goal is to extract information from control the data provided by the sensor. The capability of current remote sensors to generate data far exceeds the current capacity to handle these data. Processing data provided by sensors into an interpretable format can be an effort entailing considerable thought, instrumentation, time, experience, and reference data. Human intervention in data processing is and will mostly continue to be essential to the productive application of remote sensing information. This processing can be done partly on board and partly in ground stations.

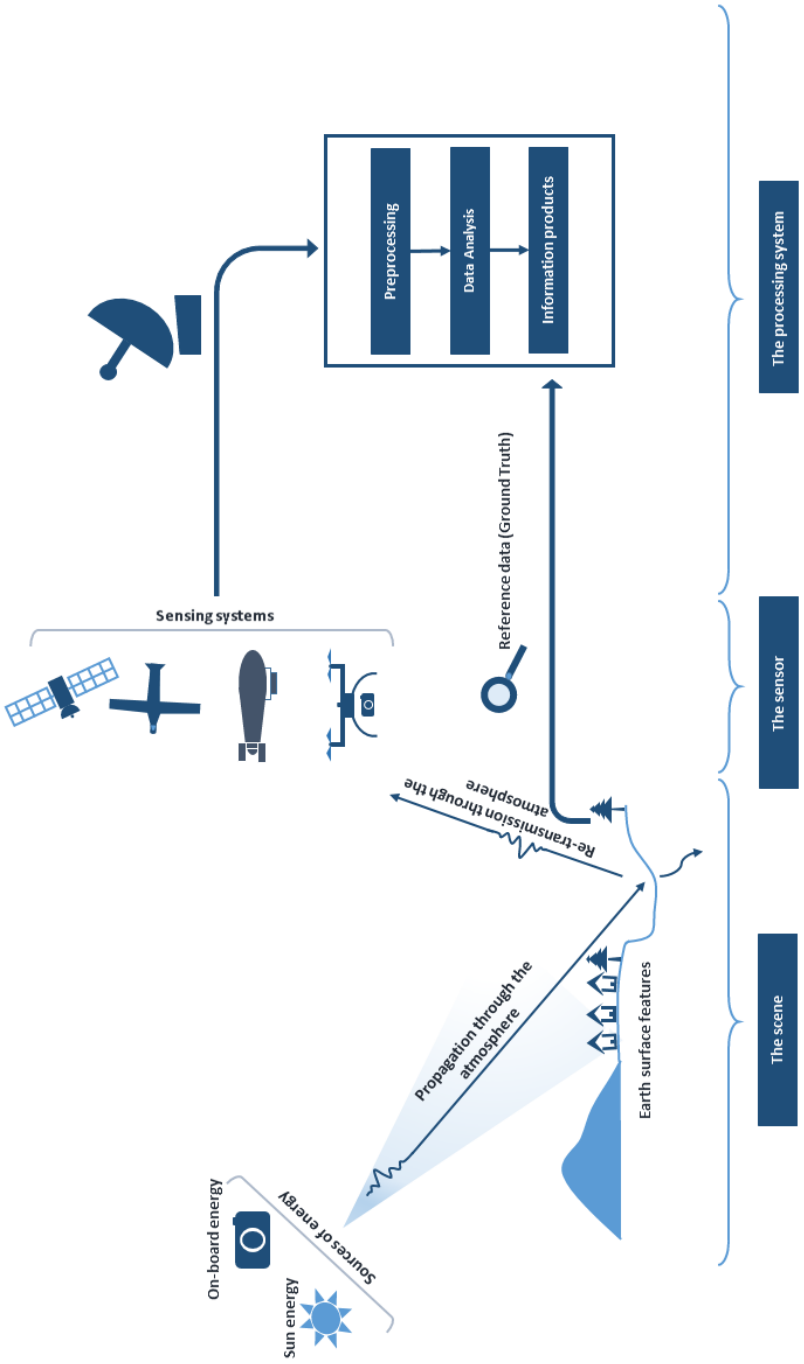


Figure 1.1: The generalized processes and elements involved in Earth observation system.

In the remainder of this chapter, the basic principles underlying the remote sensing process are discussed. The fundamentals of electromagnetic energy are considered by going through the interaction of the energy with the atmosphere and with Earth surface. These basics will permit to conceptualize the main technologies in remote sensing systems according to the electromagnetic wavelength range being used.

1.2 Energy sources and radiation principles

Visible light is only one of many forms of electromagnetic energy. Other familiar forms are invisible electromagnetic radiations, such as radio waves, ultraviolet rays and X-rays. All this energy follows the basic wave theory. As shown in Figure 1.2 this theory describes electromagnetic energy as a traveling wave through space in a rippling electric and magnetic field at the velocity of light, c. the distance from one wave peak to the next one is the wavelength λ and the number of peaks passing a fixed point in space per unit time is the wave frequency ν. The amplitude of an electromagnetic wave is the height of the wave crest above the undisturbed position.

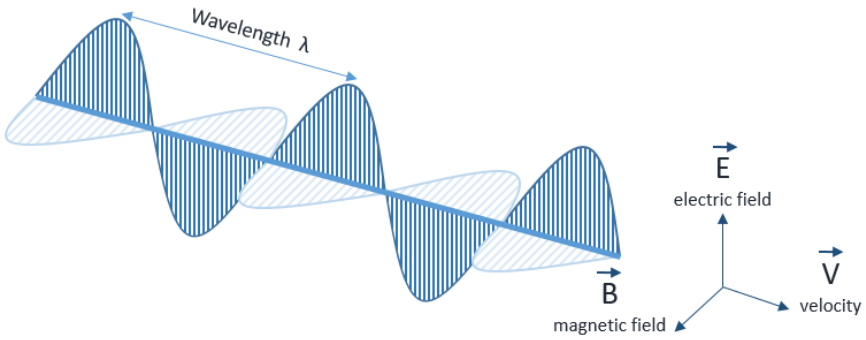


Figure 1.2: An electromagnetic wave.

In remote sensing it is most common to characterize electromagnetic waves by their wavelength location within the electromagnetic spectrum (Figure 1.3) which ranges from the shorter wavelengths (including Gamma and X-rays) to the longer wavelengths (including microwaves and broadcast radio waves). There are several regions of the electromagnetic spectrum which are useful for remote sensing because first, they can be related to Earth observation applications and second, the atmosphere is relatively transparent in those ranges (see below). The visible portion of the plot in Figure 1.3 is an extremely small one, since the spectral sensitivity of the human eye extends only from approximately 0.4 μm to 0.7 μm. There are a lot of radiations around us which are "invisible" to our eyes, but can be detected by other remote sensing instruments and used to our advantage.The portion of the spectrum

of more recent interest to remote sensing is the microwave region from about 1 mm to 1 m. This covers the longest wavelengths that are commonly used for remote sensing. The use of such a range is not arbitrary. In fact some wavelengths are inaccessible as a result of interaction of the wave with the atmosphere and Earth surface features.

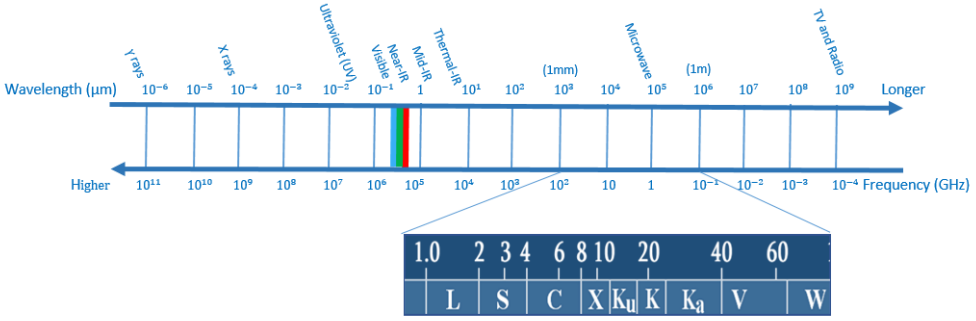


Figure 1.3: The electromagnetic spectum.

1.3 Interactions with the atmosphere

Irrespective of its source, the radiations used for remote sensing (i.e., passive VIS-NIR (radiation from the sun) and active radar (radiation from the sensor)) have to travel some distance through the atmosphere before they reach the Earth's surface. Because of their variety, particles and gases in the atmosphere can affect the incoming light and radiation. These effects are principally caused through the mechanisms of atmospheric *scattering* and *absorption* [Lillesand et al., 2014].

Scattering

Scattering is the (deterministically unpredictable) diffusion of the electromagnetic radiation caused by the interaction with particles in the atmosphere. According to the size of particles in interaction, there are three types of scattering.

1. *Rayleigh scattering* is common when radiation interacts with atmospheric molecules and other tiny particles that are very small in diameter compared to the wavelength of the interacting radiation. Rayleigh scatter is one of the primary causes of "haze" in imagery.
2. *Mie scattering* occurs when atmospheric particle diameters are just about the same size as the wavelength of the radiation being sensed. Water vapor and dust are common causes of Mie scattering which tends to influence longer wavelengths than those affected by Rayleigh scattering.

3. *Non-selective scattering* is the most bothersome phenomenon especially for optical sensors. This occurs when the diameters of particles causing scatter are much larger than the wavelength of the energy being sensed. Water droplets and large dust particles can cause this type of scattering. They commonly have a diameter in the range 5 to 100 μm and scatter all visible wavelength. Non-selective scattering gets its name from the fact that equal quantities of blue, green, and red lights are scattered, hence fog and clouds appear white in optical images.

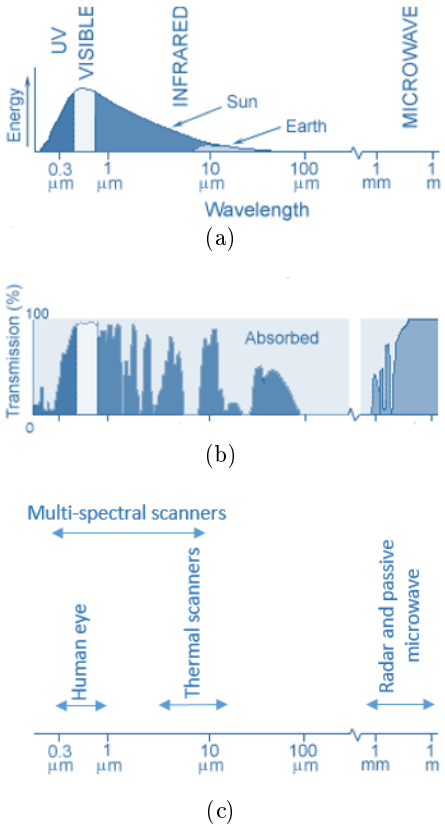


Figure 1.4: Spectral characteristics of (a) nominal black-body energy sources, (b) atmospheric effects, and (c) sensing systems.

Absorption

In contrast to scattering, this phenomenon causes molecules in the atmosphere to absorb energy at various wavelengths. The most efficient absorbers of solar radiation in this regard are water vapor, carbon dioxide, and ozone. Because these gases tend to absorb electromagnetic energy in specific wavelength bands, they strongly influence "where we look" spectrally with any given remote sens-

ing system. The wavelength ranges in which the atmosphere is particularly transmissive of energy are referred to as *atmospheric windows*. Figure 1.4 (a) shows the two most common energy/radiation sources (the sun and the Earth). In Figure 1.4 (b), spectral regions in which the atmosphere blocks energy are shaded. Accordingly, only non-blocked spectral regions could be used for remote sensing. The important point to note from Figure 1.4 is the interdependence between the primary sources of electromagnetic energy, the atmospheric windows through which source energy may be transmitted to and from surface features, and the spectral sensitivity of the sensors available to detect and record the energy.

1.4 Energy interactions with Earth surface features

Radiation that is not absorbed or scattered in the atmosphere can reach and interact with the Earth's surface, as shown in Figure 1.5. There are three fundamental energy interactions that can take place when energy is incident (I) upon the surface. Various fractions of the energy are reflected (R), absorbed (A), and/or transmitted (T). The total incident energy will interact with the surface in one or more of these three ways. The proportions of each will depend on the wavelength of the energy and the material and condition of the feature.

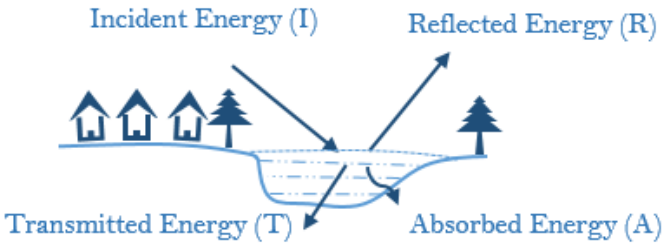


Figure 1.5: Basic interactions between electromagnetic energy and Earth surface feature.

By measuring the energy that is reflected (or emitted) by targets on the Earth's surface over a variety of different wavelengths, we can build up a spectral response for that object. By comparing the response patterns of different features we may be able to distinguish between them, where we might not be able to, if we only compared them at one wavelength. For example, water and vegetation may reflect somewhat similarly in the visible wavelengths but are almost always separable in the near-infrared (see Figure 1.6). Knowing where to "look" spectrally and understanding the factors which influence the spectral response of the features of interest are critical to correctly interpret the interaction of electromagnetic radiation with the surface.

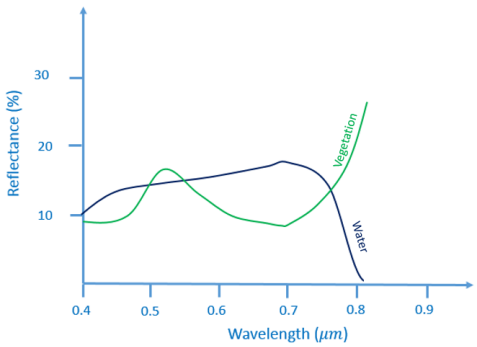


Figure 1.6: Spectral reflectance of vegetation and water.

1.5 Temporal domain

Heretofore, all these interaction (in the atmosphere or with the Earth surface features) are made in a fixed time. Despite our tendency to consider Earth as static, it is actually a dynamic and ever-changing planet. Spectral response can be quite variable, even for the same target type, and can also vary with time (e.g. "greenness" of leaves, see Figure 1.7) and location. Therefore the temporal domain play a primary role in remote sensing analysis

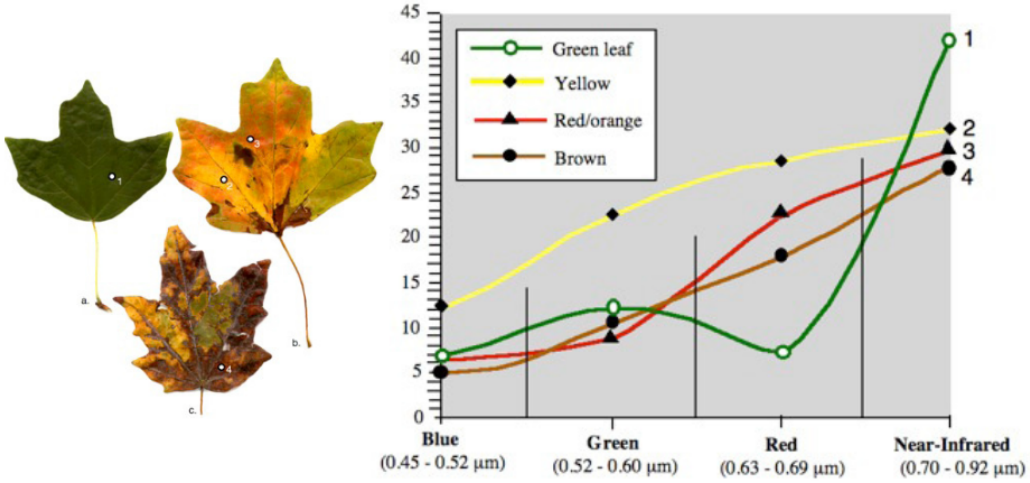


Figure 1.7: Spectral reflectance of oak leaves.

1.6 Satellites and sensors properties

1.6.1 Characteristics of remote sensing images

Before we go further in the remaining of this chapter, which focuses in more detail on sensors and their characteristics, we need to define and understand a few fundamental terms and concepts associated with remote sensing images.

Spatial Resolution, Pixel Size, and Scale

The detail discernible in an image is dependent on the *spatial resolution* of the sensor and refers to the size of the smallest possible spatial feature that can be detected. Indeed, a remote sensing image could be represented and displayed in a digital format by subdividing the image into small equal-sized and shaped areas, called picture elements or *pixels*, and representing the brightness of each area with a numeric value quantized on a finite set of levels. It is important to distinguish between *pixel size* and *spatial resolution* - they are not interchangeable. In fact, If a sensor has a spatial resolution of 1m and an image from that sensor is displayed at full resolution with ideally no blurring, each pixel represents an area of $1m \times 1m$ on the ground. In this case the pixel size and resolution are the same. However, it is possible to display an image with a pixel size different than the resolution. More generally, image acquisition introduces a (possibly small but nonzero) blurring due to the point spread function of the sensor. Images where only large features are visible are said to have *coarse* or *low resolution*. In *fine* or *high resolution* images, small objects can be detected.

Spectral Resolution

As shown in the beginning of the current section, different classes of features and details in an image taken by a passive sensor can often be distinguished on the basis of their responses over distinct wavelength ranges. Thus, one would require a sensor which could distinguish broad wavelength ranges. At this point, *spectral resolution* describes the ability of a sensor to define fine wavelength intervals. The finer the spectral resolution, the narrower the wavelength range for a particular channel or band.

Radiometric resolution

While the arrangement of pixels describes the spatial structure of an image, the *radiometric resolution* describe the precision with which the measurement taken on each pixel of an image is described. It is usually expressed in terms of the number of quantization levels or a the corresponding number of bits used to code these quantization levels with a binary format. Thus, the difference in the level of detail discernible depends on the number of bits used in representing the energy recorded.

Temporal Resolution

The concept of *temporal resolution* is important in a space-borne remote sens-

ing system, since remote sensing satellite generally re-visit the same imaged area after one entire orbit cycle. Indeed, the absolute *temporal resolution* of a remote sensing system is expressed in terms of the period of time needed to re-image the exact same area at the same viewing angle. Contemporary high resolution satellite missions of Earth obsefrvation usually provide a pointing functionality, i.e., the viewed area can be pointed (to some extent) on basis of user/operator's choice. In this case, temporal resolution and revisit time become non-periodic and generally irregular.

1.6.2 Passive vs. Active Sensing

As mentioned in section 1.1, remote sensing systems which measure energy that is naturally available from the sun are called *passive sensors*. Passive sensors can only be used to detect energy when the naturally occurring energy is available. For all reflected energy, this can only take place during the time when the sun is illuminating the Earth. *Active sensors*, on the other hand, provide their own energy source for illumination. The sensor emits radiation which is directed toward the target to be investigated. The radiation reflected from that target is detected and measured by the sensor. Forthwith, an increasing amount and variety of last-generation very-high-resolution satellite missions are available. Figure 1.8 shows the rapid growth of the number of observation systems operating in both optical and microwave spectrum as well as their performance in term of resolutions.

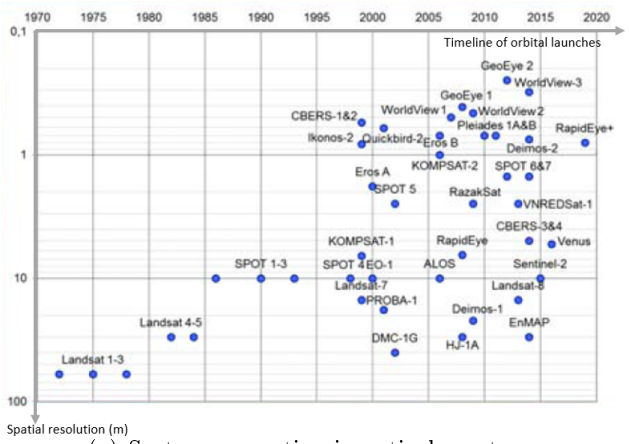
1.6.3 Passive sensors operating in the optical spectrum

This section focuses on satellite systems that operate within the optical spectrum. Optical remote sensing makes use of visible, near infrared and short-wave infrared sensors to form images of the Earth's surface by detecting the solar radiation reflected from targets on the ground. Different materials reflect and absorb differently at different wavelengths. Thus, the targets can be differentiated by their spectral reflectance signatures in the remotely sensed images. Optical remote sensing systems are classified into the following types, depending on the number of spectral bands used in the imaging process.

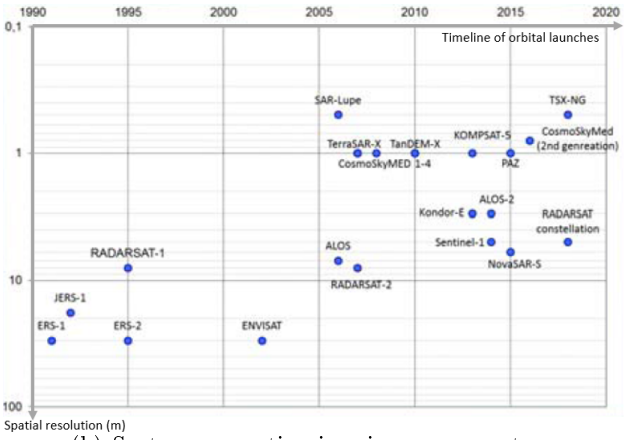
Panchromatic imaging system : The sensor is a single channel detector sensitive to radiation within a broad wavelength range usually encompassing the visible range and possibly part of the near-infrared range.

Multi-spectral imaging system : The sensor is a multichannel detector with a few spectral bands. Each channel is sensitive to radiation within a relatively broad wavelength band. The resulting image is a multilayer image.

Hyper-spectral imaging system : A hyper-spectral imaging system is also known as an "imaging spectrometer". it acquires images in about a hundred or more very narrow spectral bands.



(a) Systems operating in optical spectrum



(b) Systems operating in microwave spectrum

Figure 1.8: The growth of the number of observation systems operating in both optical and microwave spectrum [http://rsde.fbk.eu/]

As shown in Figure 1.8, Earth observation data are becoming available at increasingly finer resolutions. Very High Resolution (VHR) satellite imagery already in existence or due to be launched in the near future offers sub-meter resolution. In the remaining of this section we present the main characteristics of the optical sensors whose images were used during the Ph.D.

Pléiades

The Pléiades system was designed under the French-Italian ORFEO program (Optical & Radar Federated Earth Observation) between 2001 and 2003. The Pléiades constellation is composed of two very-high-resolution optical satellites that provide the coverage of Earth's surface with a repeat cycle of 26 days. The panchromatic and multispectral datasets are available separately or can be purchased as fused (pan-sharpened) at 0.5m of resolution. The main characteristics of the Pléiades system are reported in Table 1.1

GeoEye

GeoEye-1 is a very high resolution satellite that was launched in 2008 from Vandenberg Airbase and is owned and operated by DigitalGlobe Inc (CO, USA). The satellite collects image data down to a pixel resolution of 0.41 m. The panchromatic and multispectral datasets are available separately or can be purchased as fused (pan-sharpened) at 0.41m. The main characteristics of the GeoEye system are reported in Table 1.2

1.6.4 Active sensors operating in the microwave spectrum

An increasing amount of valuable resource information is being acquired by sensors that operate in the microwave portion of the electromagnetic spectrum. Active imaging RADAR systems are generally considered to include microwaves with wavelength from 1 mm to 1 m and are the subject of this section (for more details refer to [Henderson et al., 1998]). From a remote sensing standpoint, there are tree operational characteristic that distinct microwave radiation.

1. Active sensors based on microwaves are independent of the natural source of illumination. Therefore, microwave satellites acquire images during daytime as well at night.

2. Based on the scattering phenomena detailed in section 1.3, microwaves can penetrate the atmosphere in all weather conditions, including haze light rain, snow, cloud and smoke, depending on the wavelengths involved.

3. There is no direct relationship between microwaves scattered bye the Earth surface to their counterparts in the visible portion of the spectrum. For instance, surface that appear "rough" in the visible portion of the spectrum may be "smooth" as seen by microwaves.

Number of Satellites	Twins (Pléiades 1A and Pléiades 1B)
Launch	• Pléiades 1A: 16th December 2012; • Pléiades 1B: Q4 2012
Altitude	694km
Type	Sun-synchronous, 10:30 AM descending node
Period	98.79 minutes
Inclination	98.2°
Cycle	26 days
Spectral Bands	• Pan: 0.47-0.83 μm; • Blue = 0.43-0.55 μm, • Green = 0.50-0.62 μm, • Red = 0.59-0.71 μm, • Near Infrared (NIR) = 0.74-0.94 μm
Ground Sampling Distance	• Panchromatic: 0.7m; • Multispectral: 2.8m
Product Resolution	• Panchromatic: 0.5m; • Multispectral: 2.0m
Swath Width	20km at nadir
Radiometric resolution	12 bits per pixel

Table 1.1: Main characteristics of the Pléiades satellites. [http://www.geo-airbusds.com/pleiades/]

Number of Satellites	One
Launch	GeoEye, September 6, 2008
Altitude	694km
Type	Sun-synchronous, 10:30 AM descending node
Period	98 minutes
Inclination	98.2°
Cycle	26 days
Spectral Bands	• Pan: 0.45-0.80 μm; • Blue = 0.45-0.51 μm, • Green = 0.51-0.58 μm, • Red = 0.65-0.69 μm, • Near Infrared (NIR) = 0.78-0.92 μm
Product Resolution	• Panchromatic: 0.41m; • Multispectral: 1.65m
Swath Width	15.3km at nadir
Radiometric resolution	11 bits per pixel

Table 1.2: Main characteristics of the GeoEye satellite. [http://global.digitalglobe.com/]

As a further comparison, the visible part of the electromagnetic spectrum can be said to include the red, green, and blue spectral regions. Similarly, the active microwaves region includes X, C, L and K among others, that refer ti specific segments of the microwave portion of the electromagnetic spectrum. To illustrate, an X band system would be a radar that operates at a single wavelength within this band (e.g., 3.2 cm.). A specific RADAR technique is used for remote sensing: Synthetic Aperture Radar (SAR) imaging, in which microwave pulses are transmitted by an antenna towards the Earth surface. The microwave energy scattered back to the spacecraft is measured. The SAR makes use of the radar principle to form an image with acceptably fine spatial resolution by utilizing the motion of the satellite on the orbit. In the following of this section the basics of SAR are recalled (for a more detailed explanation of SAR the reader could refer to [Armin and Fred, 2004]). A SAR system produces a two-dimensional image. One dimension in the image is called *range* and is related to of the distance from the radar to the target. The *range* is determined by measuring the time from the transmission of a pulse to the reception of the corresponding echo from a target. The range resolution is determined by the transmitted pulse width and waveform, (i.e., narrow pulses yield fine range resolution) and by suitably filtering to echo signal (pulse compression). The second dimension is called *azimuth* and is perpendicular to the range direction. To obtain fine *azimuth* resolution, a physically large antenna would be needed to focus the transmitted and received energy into a sharp beam. The sharpness of the beam defines the *azimuth* resolution. The SAR techniques exploit the theory of antenna arrays and of the Doppler effect to improve azimuth resolution.
When discussing microwave energy, the polarization of the radiation is also important. Polarization refers to the orientation of the electric field with respect to an orthonormal basis perpendicular to the direction of irradiation. Most radar sensors are designed to transmit microwave radiation either horizontally polarized (H) or vertically polarized (V). Similarly, the antenna receives either the horizontally or vertically polarized component of the backscattered field, and some radars can receive both.

COSMO-SkyMed

COSMO-SkyMed is an Italian satellite constellation commissioned and operated by the Italian space agency (ASI). It consists of 4 medium-size satellites, each one equipped with a microwave high-resolution synthetic aperture radar (SAR) operating in X-band (wavelength of 3.2 cm), having 600 km single side access ground area, orbiting in a sun-synchronous orbit at 620km height over the Earth surface, with the capability to change attitude in order to acquire images at both right and left sides of the satellite ground track. As shown in Figure 1.9 and Table 1.3 several combinations between image size and spatial resolution have been chosen:

- A Spotlight mode, for $1m$ resolution over small images;
- Two Stripmap modes, for approximately $3m$ resolution over tenth of km im-

ages;

- Two ScanSAR for medium to coarse (100 m) resolution over large swath.

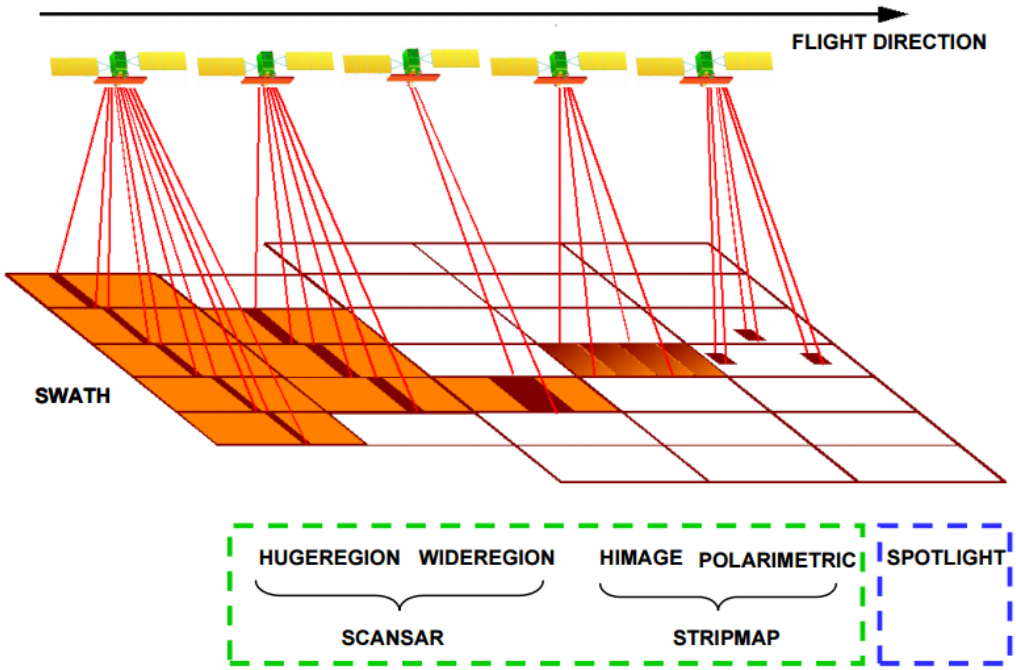


Figure 1.9: The 3 acquisition modes of COSMO-SkyMed Satellite (© ASI)

	Spotlight	Stripmap		ScanSAR	
		HIMAGE	Ping Pong	Wide Region	Huge Region
Polarisation	Single	Single	Dual	Single	Single
Swath [kmxkm]	10 x 10	40 x 40	30 x 30	100 x 100	200 x 200
Resolution [m]	1	3	15	30	100

Table 1.3: Fundamental characteristics of CSK operational modes. [http://www.cosmo-skymed.it/en/index.htm]

RADARSAT-2

Radarsat-2 is a Canadan commercial Earth observation satellite operated by the Canadian Space Agency (CSA). It utilizes a synthetic aperture radar (SAR) sensor to image the Earth at a single microwave wavelength, in the C-band (wavelength of 5.6 cm). Figure 1.10 gives a visual representation of the coverage area of

RADARSAT-2 imaging mode along the sub-satellite ground track, as well as the satellite velocity vector. The beam modes are shown in order of decreasing coverage area:

- Low Resolution Mode (100m);
- Medium Resolution Mode and Low Noise / Ship Detection Mode (50m);
- Medium Resolution Mode (30m);
- Medium Resolution Mode and Quad-Pol Mode (16m);
- High Resolution Mode (5m);
- Very High Resolution Mode (3m);
- Spotlight Mode (1m).

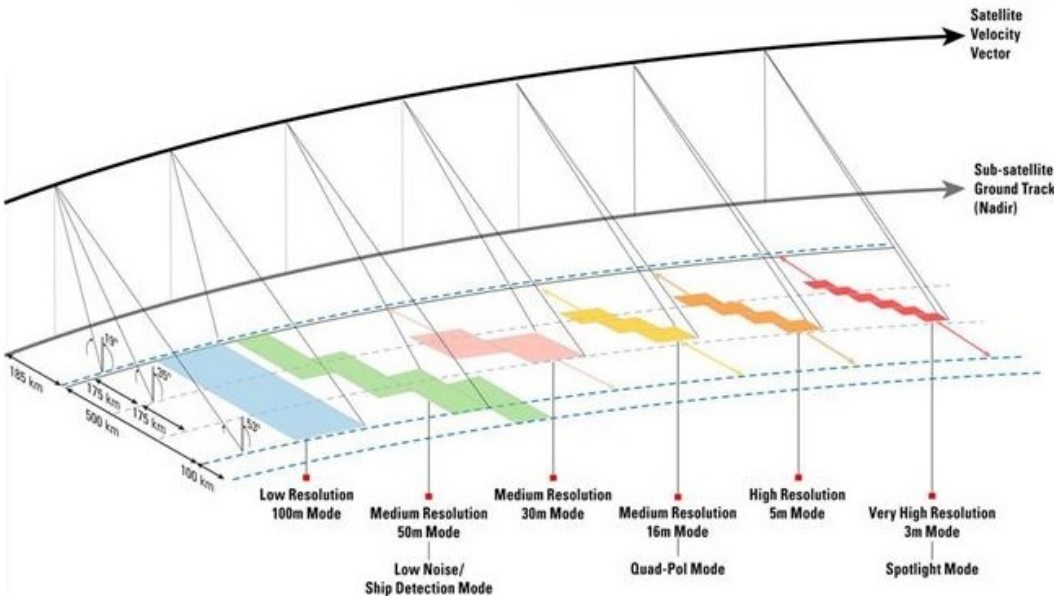


Figure 1.10: The acquisition modes of Radarsat-2 Satellite (© CSA)

1.7 Remote sensing applications

Satellite information is fundamentally important to solve some of the major challenges of our time. For issues ranging from defense, weather prediction, and mineral exploration to applications as diverse as civil protection, risk management, urban mapping, precision agriculture, commercial fishing and infrastructure management. Remote sensing provides a wealth of information at a global scale. Accordingly, an informed decision can be made by involving many processing tasks including:

Otho-rectification Warp an image to its location on the Earth.

Image Classification Categorization of pixels based on reflectance into different land cover classes.

Spectral Analysis by employing a small number of carefully chosen spectral bands spread across the visible and infrared regions of the electromagnetic spectrum. For example, using non-visible parts of the electromagnetic spectrum to determine if a forest is healthy.

Multi-resolution Analysis To enhance results and to use all available data at their different spatial resolution

Multi-temporal Analysis by using the valuable spatio-temporal information resulting from multi-temporal image acquisitions in order to analyze phenomena happened on the Earth surface during a given time period.

Change Detection To determine the changes from images taken at different times on the same area.

Chapter 2

Data fusion in Remote sensing

Contents

2.1 Data fusion in Remote sensing

The literature in data fusion is extensive, indicating the intense interest in this topic (highlighted by the sharp increase in the number of papers published in the major remote sensing journals, and the increasing number of related sessions in international conferences). In fact, data fusion gave rise to a continuing tradition in remote sensing, since Earth observation is by definition dynamic (thus implying the multi-temporal capability of remote sensing sensors), multi-resolution (multiple spatial and spectral resolutions) and related to different physical quantities (thus requiring multi-view/multi-sensor capability) [Clark and Yuille, 2013, Waltz et al., 1990].

Data fusion is defined differently depending on the final goal of the user. In fact, [Pohl and van Genderen, 2014, Li et al., 1995] considered data fusion in remote sensing as the combination of two some algorithms. This may include, *multi-resolution fusion and pansharpening techniques* with the aim is to obtain multi-spectral images of increased spatial resolution [Vivone et al., 2015], *resolution blending* that consists in providing time series of data at their maximum spatial and spectral resolutions (referred to as parallel pansharpening in a multi-temporal domain) [Huang and Song, 2012] and *data fusion for missing information reconstruction* by using complementary data [Wang and Liang, 2014].
An alternative perspective is to define data fusion in remote sensing as a decision fusion technique which combines the information that is obtained from different data sets and provides sufficient generalization capability [Wald, 1999]. According to this definition, any type of image processing that combines two or more data sets, such as image classification or atmospheric correction or application of vegetation indices, should be considered data fusion.

Within the former definition, various procedures of data fusion techniques have been proposed in the literature with regard to, on the one hand, the *application requirements* such as the availability of ground reference data, the collected prior information and/or some ancillary data that can be used in the development of the system according to a multi-source processing architecture. On the other hand, it is important to properly understand the *user needs* with respect to economic costs, processing time and performance. Figure 2.1 summarizes the general architecture of a data fusion techniques.

2.2 Multi-sensor and multi-resolution approaches

2.2.1 Multi-resolution approaches

As exposed before, the availability of remote sensing imagery of varying resolution has increased. Merging images of differing spatial resolution has become a significant operation in the field of digital remote sensing. A variety of different multi-scale fusion approaches have been developed since the late 80s. In the following we give an

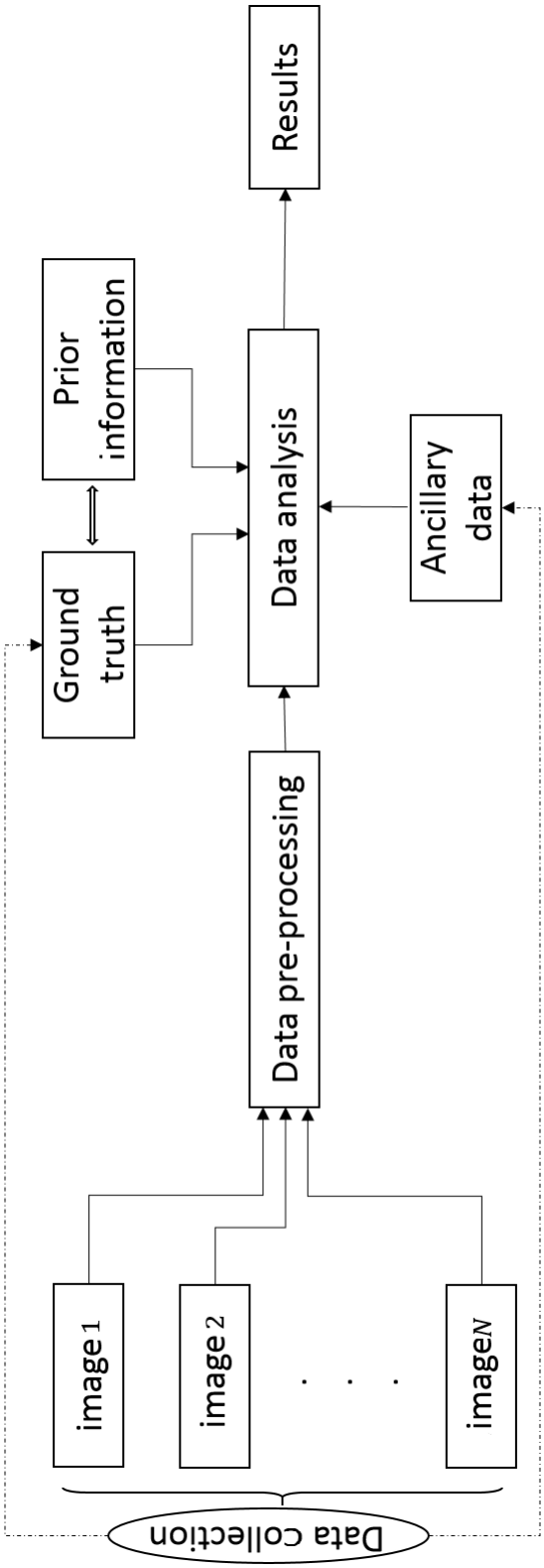


Figure 2.1: General data fusion architecture.

excerpt on the most common approaches found in the literature. We divided them into the two groups, (i) transformation techniques and (ii) modeling techniques.
Methods in (i) consist in replacing the entire set of multi-scale information by a single composite representation which incorporates all relevant data. The multiscale transformations usually employ Pyramid Transforms [Burt, 1984], the Discrete Wavelet Transform (DWT) [Piella, 2003, Forster et al., 2004, Zhang and Hong, 2005], the Undecimated Wavelet Transform (UWT) [Rockinger, 1996, Chibani and Houacine, 2003], the Dual-Tree Complex Wavelet Transform (DTCWT) [Demirel and Anbarjafari, 2010, Iqbal et al., 2013], the Curvelet Transform (CVT) [Nencini et al., 2007, Choi et al., 2005], the Contourlet Transform (ConT)[ALEjaily et al., 2008, Shah et al., 2008] and the Nonsubsampled Contourlet Transform (NSCT) [Yang et al., 2007].
(ii) includes multi-scale approaches with a focus on the use of coarser resolution of the data set in order to obtain fast computational algorithms. In a seminal papers [Basseville et al., 1992a, Basseville et al., 1992b] the basis for multi-scale autoregressive modeling in dyadic trees was introduced. Since, straightforward approaches were performed to deal with multi-resolution images using trees [Pérez, 1993, Kato, 1995, Laferté et al., 2000, Chardin, 2000, Voisin, 2012, Hedhli et al., 2014b]. Due to their importance, the subsequent chapters provide a more detailed view on the use of multi-scale modeling on trees. However, a detailed review of some of these methods could be found in [Graffigne et al., 1995, Willsky, 2002].

2.2.2 Multi-sensor approaches

Multi-sensor analysis is a process dealing with data and information from multiple sensors to achieve refined/improved information as compared to the result that could be obtained by using data from only one individual source (see [Hall and Llinas, 2001, Waltz et al., 1990, Pohl, 1998]). The classification accuracy of remote sensing images, for instance, is improved when multiple source image data are introduced to the processing (e.g., [Nguyen et al., 2011, Gamba et al., 2011, Dousset and Gourmelon, 2003, Hedhli et al., 2015]). Images from microwave and optical sensors provide complementary information that helps in discriminating the different classes. Several procedures have been introduced in the literature including, on one hand, post-classification techniques in which, first, the two data sets are separately segmented, and then the joint classification is produced by using, for example, random forest (RF) (e.g., [Waske and van der Linden, 2008]), support vector machine(SVM) with ad-hoc kernels [Muñoz-Marí et al., 2010], artificial neural networks (ANN) [Mas and Flores, 2008]. On the other hand, other methods directly classify the combined multi-sensor data by using for instance, statistical mixture models (e.g., [Dousset and Gourmelon, 2003, Voisin et al., 2012, Prendes, 2015]), entropy based techniques (e.g., [Roberts et al., 2008]), fuzzy analysis (e.g., [Benz, 1999, Stroppiana et al., 2015]). Further more, for complex data, especially when dealing with urban areas, radar images can contribute to different reflectance due to differences in surface roughness, shape and moisture content of the observed

ground cover(e.g., [Brunner et al., 2010]). The use of multi-sensor data in image classification becomes more and more popular with the increased availability of sophisticated software and hardware facilities to handle the increasing volumes of data. The decision on which technique is the most suitable is very much driven by the application and the typology of input remote sensing data.

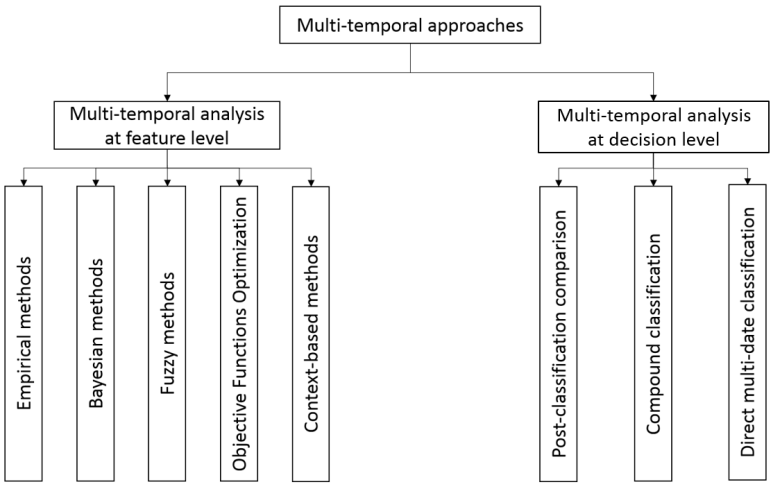


Figure 2.2: Multi-temporal method categorization.

2.3 Multi-temporal analysis approaches

As discussed in Chapter 1, satellite remote sensors for Earth observation have the capability to visit the same geographic area several times. Several survey papers provide thorough reviews of multi-temporal remote sensing image analysis methods, [Singh, 1989] classifies these approaches into two categories: (1) comparative analysis of independently produced classifications for different dates [Johnson and Kasischke, 1998, Sziranyi and Shadaydeh, 2014] and (2) simultaneous analysis of multi-temporal data [Bazi et al., 2005, Hedhli et al., 2014b, Singh et al., 2014]. Whereas, [Malila, 1980] points out two other basic approaches for multi-temporal analysis: (1) classification approaches [Hoberg et al., 2015, Hedhli et al., 2014b, Melgani and Serpico, 2003] and (2) change measurement (stratification) [Benedek et al., 2015, Gamba et al., 2006]. [Petitjean et al., 2011] slightly changes the focus by centering the definitions more on a temporal scale and brings out three main usages of the time dimension: (1) time as identifier: time is only used to identify the information, i.e., there is no ordering between the images of the series [Melgani and Serpico, 2003, Hoberg et al., 2015]; (2) pairwise time ordering: time is used to structure the

images, pairwise [Liu et al., 2012, Bovolo and Bruzzone, 2008]; (3) time ordering the sequence: time is used to structure the image series [Petitjean et al., 2011, Weber et al., 2012, Hedhli et al., 2014a, Hedhli et al., 2014b]. Other scientists [Nelson, 1983], [Milne, 1988] and [Coppin et al., 2004] have employed more sophisticate multi-class schemes. In the following of this section, the categorization in [Bovolo and Bruzzone, 2015] is adopted. Hereafter, we present the most often used methods in the literature that deal with multi-temporal data by classifying them into two main categories as shown in Figure 2.2

2.3.1 Multi-temporal fusion at feature level

the aim of these methods is to generate new features (i.e., change indices in a bi-temporal image pairs) that highlight multi-temporal information, and then to extract changes occurred in the temporal series or to perform more sophisticated analysis such as assessing, monitoring, and predicting the dynamics of natural land covers. Further more, fusion at feature level is characterized by several particular factors that render ineffective some of the multi-temporal image analysis techniques typically used in other application domains. Nonetheless, pre-processing steps are mandatory. As shown in the general schema of data pre-processing illustrated in Figure 2.3 these steps may include:

- **Radiometric Correction**: Multi-temporal image series are generally impacted by differences in light conditions, sensor calibration, and ground moisture. These last items might be considered a legitimate ground change, though differences in light and atmospheric conditions between images, for instance, can be mitigated by applying radiometric calibration to the temporal series. In the literature there are two major calibration approaches depending on the particular application considered and on the specific information available. The first one is the *Absolute Calibration* which consists in transforming the quantized pixel intensities (digital numbers) into the corresponding ground reflectance values by employing radiometric transfer models [Kurum et al., 2011, Biswas et al., 2013, Montopoli et al., 2013] or regression algorithms applied to ground-reflectance measurements collected during the data acquisition phase [Henderson et al., 1998, Gordon and Morel, 2012, Bernstein et al., 2012]. The second category widely used is the *Relative Calibration* which directly modifies the histograms, so that the same gray levels values in images can represent similar reflectance values [Gadallah et al., 2000, Tokola et al., 1999]
- **Geometric Corrections & Image Registration**: Non-perfect alignment (registration Moise) and different acquisition conditions (view angle, shadows, ...) are common problems within the considered image time series. In fact, it is not possible to obtain perfect alignment between multi-temporal images. This is mainly due to local defects in the geometries of the images which result in a very critical source of noise called *registration noise* [Maes et al., 1997, Le Moigne et al., 2011].

- **Image Filtering**: This step is not mandatory and depends on the specific sensor and on the quality of the images. A wide variety of filtering techniques are used in remote sensing (see for instance [Schowengerdt, 2006]). In the case of optical images one way to create a smoothed image time series is to use multi-point smoothing functions, such as moving averages or polynomial filters([Udelhoven, 2011, Savitzky and Golay, 1964, Chen et al., 2004]. The task is more delicate with SAR images since they are inherently affected by multiplicative speckle noise. Therefore, many adaptive filters for speckle reduction have been proposed. Using for instance, Frost filter which models the scene reflectivity [Frost et al., 1982], Kuan Filter based on the minimum mean-square error (MMSE) criterion [Kuan et al., 1985], two-dimensional Kalman filter satisfying a causal Auto-Regressive model [Azimi-Sadjadi and Bannour, 1991], Gamma MAP filter based on a Bayesian analysis of the image statistics where both signal and speckle noise follow a Gamma distribution [Baraldi and Parmiggiani, 1995], and wavelet transform as a powerful tool for recovering SAR images from noisy data [Fukuda and Hirosawa, 1999].

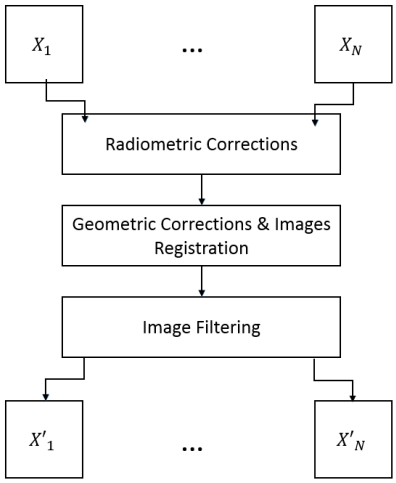


Figure 2.3: General schema of data pre-processing.

Then, after the pre-processing operations a broad range of methods can be applied in order to extract temporal information at feature level that depend essentially on the type of the sensor used. indeed, the noise model in optical images acquired by passive sensors is usually considered additive and the natural classes tends to follow a Gaussian distribution [Richards, 2013]. The speckle in SAR amplitude and intensity images, which is generated within the scattering phenomenon, acts through a multiplicative model. Table 2.1 and Table 2.2 summarize feature level methods using optical and SAR images, respectively.

Technique		Descriptions & Examples
Empirical Techniques	Univariate Image Differencing	Fusion is mainly performed by the difference operator applied to one spectral band [Sowell, 1992].
	Change Vector Analysis (CVA)	Fusion is mainly performed by the difference operator applied to multiple spectral bands. A variety of CVA techniques have been developed, (e.g., Change Vector Analysis [Lambin and Strahlers, 1994, Melgani et al., 2002], Compressed Change Vector Analysis [Marchesi et al., 2010], Polar Change Vector Analysis [Bovolo and Bruzzone, 2007] ...).
	Regression	Establishes relationships between bi-temporal images, then estimates pixel values of the second-date image by use of a regression function, and subtracts the regressed image from the first-date image [JHA and Unni, 1994, Esbensen et al., 1992].
	Transformation	Reduces data redundancy between bands and emphasizes different information in the derived components using for instance Principal component analysis (PCA) [Deng et al., 2008], Tasseled Cap Transformation [Healey et al., 2005], or Gram-Schmidt Transformation [Rosa et al., 2015].
Multi-scale/resolution features		For example, a multi-resolution analysis with wavelet transforms applied to image differencing results enabled the extraction of changed sites according to size classes. [Carvalho et al., 2001, Moser et al., 2011]
Unsupervised Bayesian framework		Use automatic techniques based on both implicit or explicit parameter estimation in an unsupervised framework, and perform the multi-temporal analysis with either a context sensitive or insensitive method (e.g., [Hame et al., 1998, Melgani et al., 2002]).
Fuzzy theory		The threshold selection is performed using fuzzy theory by assuming that the gray tone image processes some ambiguity (e.g., [Du et al., 2013, Madanian et al., 2014, Gamba and Dell'Acqua, 2003])

Table 2.1: Multi-temporal techniques at the feature level using optical images.

Technique	Descriptions & Examples
Image (log-)ratio	Calculates the ratio of registered images in order to reduce the multiplicative distortions common within SAR images due to speckle noise (e.g., [Bazi et al., 2005]).
Information theoretical & similarity measures	A wide variety of distance functions and similarity measures have been used for multi-temporal analysis using SAR images, such as, Kullback-Leibler distance [Aspert et al., 2007, Inglada and Mercier, 2007], mutual information [Cui and Datcu, 2012], variational information [Gueguen et al., 2010], relative entropy [Martinis and Twele, 2010], etc.
Additional specific fusion operators	Several features are specific for SAR imagery such as : Backscattering coefficient, cloud decomposition (or H-α decomposition), polarimetric signatures, etc. Diverse method arise from the analysis of these features : difference of scattering matrices [Touzi et al., 2009], change indices [Hachicha et al., 2011], contrast ratio (or Rayleigh quotient) [Molinier and Rauste, 2007], Bartlett test [Kersten et al., 2005], etc.

Table 2.2: Multi-temporal techniques at the feature level using SAR images.

2.3.2 Multi-temporal fusion at decision level

Decision or interpretation level analysis induces methods that use value-added data where the input images are processed individually for information extraction. the obtained information is then combined applying decision rules to reinforce common interpretation, resolve differences and provide a better understanding of the observed data. Most prior studies of multi-temporal analysis have involved classification techniques which actually combine the change detection and change identification aspects of using information to update resource surveys. Other aspects of classification are the requirements of training and additional ground truth for supervised or unsupervised classification of images. Image transformation, vegetation indices, advanced classification methods and integration of different data sources are often used to improve classification results. Table 2.3 summarizes the methods using multi-temporal information at the decision level based on classification approaches.

In the present manuscript, we will focus on supervised classification methods for multi-source data. These methods have the advantage of being applicable to two or

Technique	Descriptions & Examples
Post-classification comparison	To minimize impacts of atmospheric, sensor and environmental differences between images in time series, these techniques separately classify the multi-temporal data into thematic maps, then implement comparison of the classified images, pixel by pixel [Serra et al., 2003, Van de Voorde et al., 2007, Manandhar et al., 2009].
Combined classification	Such methods are based on a single analysis of a combined data set of two or more dates. Two Bayesian approaches can generally be adopted for this purpose. The 'cascade' classifier (e.g., [Swain, 1978, Hedhli et al., 2014b]) removes the coupling between the spectral and temporal dimensions and classifies each image in the input series on the basis of itself and of the previous images. The 'mutual' approach classifies each image on the basis of the previous and the subsequent images in the series, (e.g., [Hoberg et al., 2015, Melgani and Serpico, 2003]). The two approaches are basically complementary in terms of applicability: online processing can be feasible within the cascade domain, whereas the mutual domain can be generally more appropriate for batch processing.
Artificial neural networks (ANN)	ANN is a nonparametric supervised method and has the ability to estimate the properties of data based on the training samples. A back-propagation algorithm is often used to train the Multi-Layer Perceptron (MLP) neural network model (e.g., [Miller et al., 1995, Bruzzone et al., 1999]).
Other change multi-temporal techniques	Include spatial statistics-based method, Biophysical parameter method, Hybrid change detection, Unsupervised methods, Data mining approaches, multi-resolution techniques.

Table 2.3: Multi-temporal techniques at the decision level images.

more images taken by single or multiple sensors at the same or at different spatial resolutions. The available images are temporally and spatially correlated. In fact, temporal and spatial contextual constraints are unavoidable in multi-temporal data interpretation. Within this framework, Markovian Models provide a convenient and consistent way of modeling context-dependent spatio-temporal entities originated from multiple information sources, such as images in multi-temporal, multi-sensor, and multi-resolution context. In the next chapter, we introduce the basis of the contextual classification using Markovian models.

Chapter 3

Markovian image models

Contents

3.1 Inverse problems and contextual information

Change detection, image classification, image restoration are many examples that show the number and diversity of inverse problem in image processing [Neelamani, 2003]. These problems can be regarded as the process that estimates *hidden informations* (i.e., latent variables) x (e.g., land cover class labels) from *observations* y (e.g., satellite data) attached to a set of nodes (e.g., pixels) S. Solutions to inverse problems are based on models that link labels to the given observation. These problems are often *ill-posed* in the sense of Hadamard[1], due to the loss of information that occurs when forming the observations. To overcome this ill-posedness, a major approach is to promote certain types of solutions by imposing prior knowledge on the labels, which may include, for instance, contextual priors that favor a spatial regularization of the considered signals. To stress this aspect, many issues of image analysis can be modeled and coped with by designing suitable *energy* functions $U(x, y)$ which capture the interactions between the unknown variables $x = \{x_s\}_{s\in S}$ to be estimated, and the observed variables $y = \{y_s\}_{s\in S}$, while incorporating the desired prior characterization.

3.1.1 The Ising model

Inspired from ferromagnetism, the simplest theoretical description of the interaction between variables is called the *Ising model*. Our starting point is a lattice, which is a finite set of regularly spaced points in a space of dimension $d = 1, 2$ or 3. In dimension 1 we simply have a string of points on line, which can be enumerated from 1 to N. At each point there is a spin which is either *up* or *down* at any given configuration $\omega = \{\omega_i\}_{1\leq i\leq N}$ as shown in Figure 3.1. In this context each spin is a function:

$$\delta_i(\omega) = \begin{cases} 1 & \text{if} \quad \omega_i \; is \; up \\ -1 & \text{if} \quad \omega_i \; is \; down \end{cases}$$

An energy U is then assigned to each configuration:

$$U(\omega) = -\alpha \sum_i \delta_i(\omega) - \beta \sum_{i,j} \delta_i(\omega)\delta_j(\omega). \tag{3.1}$$

The first term represents the influence of each spin on the energy, regardless of the other spins. The second term represents the energy contribution caused by the spin interactions. If $\beta > 0$, the interaction tend to keep neighboring spins in the same direction (*attractive case*) otherwise, neighboring spins with opposite orientations are favored (*repulsive case*) [Parisi, 1988, Baxter, 2007].

[1] Well-posedness of a problem stems from a definition given by [Hadamard, 1923].A problem, involving some mathematical model of physical phenomena, is Hadamard well-posed if a solution exists, is unique and its behavior changes continuously with the initial conditions ; otherwise, it is ill-posed.

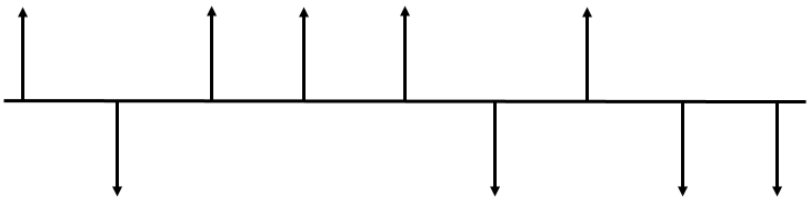


Figure 3.1: One dimensional Ising model.

3.1.2 The Potts model

The Potts model [Li, 2009] is a generalization of the Ising model to more than two states per point. The Potts model consists of N spins that are placed on a lattice; at each point i the spin takes one of q possible values, distributed uniformly about the circle, at angles $\theta_i = \frac{2\pi N}{q}$, exemples are shown in Figure 3.2. Then, the interaction between two neighboring spins i and j is characterized by the angle Θ_{ij} between them. Thus, a generalization of (3.1) is given by:

$$U(\omega) = -\alpha \sum_i \delta_i(\omega) - \beta \sum_{i,j} \Theta_{ij}(\omega). \tag{3.2}$$

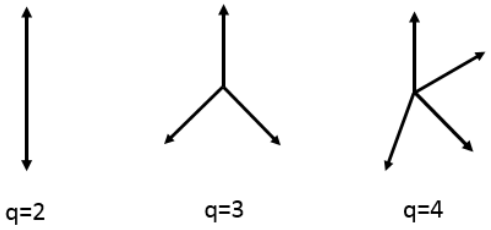


Figure 3.2: System of spins in the Potts model.

This energy model is encountered in various fields (e.g., statistical physics, multivariate statistics, combinatorial optimization, artificial intelligence). We are here interested in its use in Markovian priors as a regularization term.

3.2 General characteristics of Markovian image models

In probability theory and statistics, Markovian models refer to the finite memory stochastic processes. A Markov chain is the simplest Markovian model and is widely used in one-dimensional signal processing, queuing theory, and image processing as well.

Definition 3.2.1 (Markov chain) *A Markov chain is a sequence of random variables $X = (X_i)_{i=1}^{N}$ that satisfy the Markov property, i.e., such that the distribution of the present state given the entire past can be restricted to the distribution conditioned only to the most recent past sample. Formally:*

$$\forall i > 1, p(X_i \mid X_{i-1}, \cdots, X_1) = p(X_i \mid X_{i-1}). \tag{3.3}$$

Spatial Markov Random Fields (MRFs) were formulated by [Besag, 1974] and have become popular image processing tools with [Geman and Geman, 1984]. These models are based on the definition of a local neighborhood system.

Definition 3.2.2 (Neighborhood system and clique) *Let us denote as S a finite set. The elements of S are named sites. Each site $s \in S$ is associated with a set of neighbors $\mathcal{V}_s \subset S$ such as:*

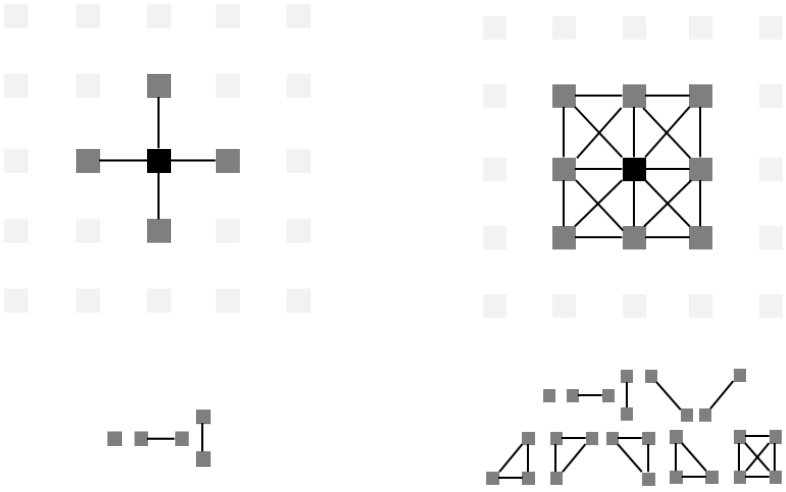
- $s \notin \mathcal{V}_s$, $\forall s \in S$
- $s \in \mathcal{V}_t \Leftrightarrow t \in \mathcal{V}_s$,$\forall s, t \in S$

$\mathcal{N} = \{\mathcal{V}_s, s \in S\}$ is called a neighborhood system.

*A subset $c \subseteq S$ is a **clique** if every pair of distinct sites in c are neighbors. $\mathcal{C}$ denotes the set of cliques.*
In image processing applications, the set S of sites often coincides with the pixel lattice, and each site is consequently a pixel. In the case of region-based image analysis, the set S may include regions obtained by a segmentation algorithm. For example, in the former case, the set of neighbors of site s in a regular lattice S is commonly defined as

$$\mathcal{V}_s = \{r,\ d(s,\ r) \leqslant o\},$$

where o is the order of the neighborhood system, and $d()$ is the Euclidean distance. Figure 3.3 illustrates the neighborhood systems for $o = 1$ and 2 and their corresponding cliques.



(a) (b)

Figure 3.3: Neighborhood systems: (a) first-order; (b) second-order.

Definition 3.2.3 (Markov random field) *Let X be a collection of random variables X indexed by the elements of a set of sites S. X is a Markov random field (MRF) with respect to a neighborhood system $\mathcal{N}$ on S if:*

$$\forall s \in S,\ p(X_s \mid X_t,\ t \neq s) \ = \ p(X_s \mid X_t,\ t \in \mathcal{V}_s). \tag{3.4}$$

Local specifications are given by the probabilities $p(X_s = x_s \mid X_t = x_t,\ t \in \mathcal{V}_s)$.

The definition of Markov fields is very local. Nevertheless, the success of Markov fields is largely due to their capability to also provide an expression of the global distribution $p(X)$. A theoretical result about the equivalence between Markov random fields and Gibbs distributions is given by the *Hammersley-Clifford* theorem [Besag, 1974].

Theorem 3.2.1 (Hammersley-Clifford) *Let us denote as X a random field indexed by a set S with respect to the neighborhood system $\mathcal{N}$*

$$\begin{gathered} X \text{ is a Markov random field with respect to } \mathcal{N} \\ \Leftrightarrow \\ X \text{ is a Gibbs field with respect to } \mathcal{N} \end{gathered}$$

Definition 3.2.4 (Gibbs fields) *X is a Gibbs field (GF) indexed by a set S with respect to the neighborhood system $\mathcal{N}$ if:*

$$p(X = x) \ = \ \frac{1}{Z} e^{-U(x)}, \tag{3.5}$$

where Z is a normalization constant (called partition function), and U n energy function that can be decomposed into a sum of potentials associated with the set of cliques $\mathcal{C}$:

$$U(x) = \sum_{c \in \mathcal{C}} V_c(x). \tag{3.6}$$

Equivalently, the joint distribution of x factorizes into a product of positive factor potentials:

$$p(x) \;=\; \frac{1}{Z} \prod_{c \in \mathcal{C}} g_c(x), \tag{3.7}$$

where $g_c = e^{-V_c}$.

3.3 Inference and Bayesian estimation

We now focus on the mathematical formulation of MRF-based image classification. The aim of the classification is to estimate a set of hidden labels X (e.g., land cover class labels) given a set of observations Y (e.g., satellite data) attached to the set of nodes (pixels) S . Each label occupies a value in the set $\Lambda = \{0, 1, \cdots, M-1\}$. The configuration space $\Omega = \Lambda^{|S|}$ is the set of all global discrete labeling.

$x = \{x_s\}_{s \in S}$ and $y = \{y_s\}_{s \in S}$ denote realizations of the random fields of the class labels and observations of all nodes. In this context, we consider the problem of inferring the "best" configuration $x \in \Omega$ given the observation y. This leads naturally to an inference problem in which the *posterior* distribution for the possible realization x, given the observations y, is computed via *Bayes*'s formula as :

$$p(X = x|Y = y) \propto p(Y = y|X = x)p(X = x). \tag{3.8}$$

In (B.1), $p(X = x)$ is the prior distribution over configurations. In other words, the distribution of the possible realizations of X in the absence of any observations. For convenience, (B.1) is abbreviated to:

$$p(x|y) \propto p(y|x)p(x). \tag{3.9}$$

Similar less cumbersome notations will be used in the following whenever possible. According to (3.9), when both the prior distribution $p(x)$ and the likelihood function $p(y|x)$ are given, the optimal solution that can be estimated from these sources of knowledge is the Bayes labeling. Bayes's theorem can only be applied when all these distributions are suitable. However, it is not uncommon for the resulting *posterior* to be a valid probability distribution. In this case, the posterior expected loss (i.e., Bayes risk) is typically well defined and finite:

$$R(x^\star) = \sum_{x \in \Omega} C(x^\star, x)p(x|y), \tag{3.10}$$

where C is the cost function penalizing the discrepancy between the estimated configuration and the "ideal" random configuration. Indeed, The standard Bayesian

formulation of the inference problem consists of minimizing the Bayes risk which is referred to as the *generalized Bayes estimator*:

$$x_{opt} = arg \min_{x^\star \in \Omega} R(x^\star) = arg \min_{x^\star \in \Omega} \sum_{x \in \Omega} C(x^\star, x) p(x|y). \tag{3.11}$$

3.3.1 Bayesian estimators

Among the different cost functions employed, three have been widely used in the literature:

Maximum A Posteriori (MAP) estimate

This algorithm aims to exactly estimate the maximum a posteriori (MAP) configuration. The cost function of this algorithm is defined by the following:

$$C_{MAP}(x^\star, x) = 1 - \delta(x^\star, x), \tag{3.12}$$

where δ is the Kronecker delta. This function has the identical cost for all configurations of $x^\star$ different from x.
From Equations (3.10) and (3.12), the Bayes's risk is:

$$R(x) = 1 - p(x|y). \tag{3.13}$$

Minimizing (3.13) is equivalent to maximizing the posterior probability. Therefore, the minimal risk estimate is:

$$x^{MAP} = arg \max_{x \in \Omega} p(x|y) = arg \max_{x \in \Omega} p(y|x) p(x), \tag{3.14}$$

In addition, if X is a Markovian process and according to the Markov-Gibbs equivalence evoked in theorem (3.2.1), the MAP estimate results in the minimization of an energy function as:

$$x^{MAP} = arg \min_{x \in \Omega} [U(y|x) + U(x)], \tag{3.15}$$

where $U(x)$ is the energy function associated with the MRF X, and $U(y|x) = -lnp(y|x)$.

Marginal a Posteriori Modes (MPM) estimate

The cost function associated with the MPM is the following:

$$C_{MPM}(x^\star, x) = \sum_{s \in S} [1 - \delta(x_s^\star, x_s)], \tag{3.16}$$

which is related to the number of sites . Thus, the MPM criterion offers the possibility of penalizing errors according to their number, a property that makes it especially appealing for image segmentation and classification. The resulting Bayesian estimator is given by the following:

$$\forall s \in S, x_s^{MPM} = arg \max_{x_s \in \Lambda} p(x_s|y), \tag{3.17}$$

Mean Field (MF) estimate

A third cost function that may be considered is the MF estimator using the following cost function:

$$C_{MF}(x^\star, x) = \sum_{s \in S} (x_s^\star - x_s)^2, \tag{3.18}$$

The resulting Bayesian estimator is given by the following:

$$\forall s \in S, x_s^{MF} = \sum_{x \in \Omega} x_s p(x|y) = E[x_s|Y = y], \tag{3.19}$$

which is nothing else but the conditional expectation of X given $Y = y$ which explains the terminology "mean field". Note that, each label takes values in the set $\Lambda = \{1, 2, \ldots, M\}$. In most usual applications to image classification and segmentation, these M possible labels are meant as symbols that indicate M possible classes. In the case of the MF criterion, they are also meant in their numeric sense as M integer numbers.

3.3.2 Optimization approaches

Given the Bayesian justification about the form of the cost function that was presented in the previous section, we now turn into the matter of how to obtain the optimal solution in a Markovian framework. For general MRFs, this problem is NP-hard and so finding the optimal solution is far from being an easy task to achieve. As the full list of MRF optimization techniques is vast, we refer here only to a selected and very small subset of these methods. A thorough review is reported in [Li, 2009, Kato and Zerubia, 2012]. In the following, we will briefly describe the main optimization techniques for MRF models.

Simulated annealing and Markov Chain Monte Carlo (MCMC) methods

Actually, MCMC techniques which can be used for sampling any type of distribution. However, thanks to a procedure named simulated annealing ([Černỳ, 1985, Goffe et al., 1994]), sampling can be also used for optimization purposes with MRF models. Simulated annealing is inspired by an analogy with metallurgy, in which slow cooling (annealing) is used to produce metal that is tougher than that which is produced from rapid cooling. When a Markov chain simulation (typically a Metropolis sampler [Metropolis et al., 1953]) is used to sample a Gibbs distribution, the analogous procedure is to introduce a temperature parameter T into the Gibbs distribution. The temperature is gradually reduced from a very high initial value to a value close to zero. On one hand, when T is large the sampler can move freely through the state space, thus escaping poor local minima. On the other hand, when the temperature is close to zero the distribution is concentrated in the region near the global minimum and so sampling becomes equivalent to optimization. The rationale is that, as T is gradually decreased, one will be able to approach the global

minimum. Asymptotically, simulated annealing guarantees to extract the global optimum provided that the temperature follows an appropriate decreasing schedule. This, however, requires infinite time in practice and so only local minima can be usually estimated.

Iterated Conditional Modes (ICM)

Iterated Conditional Modes (ICM) is another deterministic optimization technique which has been proposed by Besag [Besag, 1974, Besag, 1986]. ICM tries to maximize the joint probability of an MRF by maximizing local conditional probabilities sequentially. More specifically, at each step of the algorithm, a MRF node is chosen (under raster scanning) and that node is then assigned the label which minimizes the energy of the MRF under the condition that all other nodes keep their labels fixed. This is repeated for all nodes of the MRF until convergence i.e. until the energy cannot decrease further. The main disadvantage of ICM is that it is greedy and very sensitive to initialization. Therefore, unless a good initial estimate is given to ICM, it can easily get trapped to poor local minima. Highest Confidence First (HCF) is a deterministic algorithm which tries to improve ICM and has been proposed by Chou and Brown [Chou and Brown, 1990]. Its feature is that it processes the nodes of the MRF in a specific order. To this end, it introduces an uncommitted label and then uses a certain strategy for choosing which MRF node "to committ" next. Experimental results show that HCF is, on the whole, better than ICM with respect to the task of minimizing the energy of an MRF.

Graph-cuts

Graph cuts are considered effective approaches for solving the energy minimization problems in computer vision. [Boykov and Jolly, 2001] proposed methods to solve MAP-MRF using graph-cut algorithms and they showed that MAP-MRF estimate is equivalent to min-cut problems on a graph [Boykov and Kolmogorov, 2001]. This equivalence makes graph cuts extremely important. They showed that using max-flow min-cut graph algorithms it is possible to solve some class of energy functions with MAP-MRF framework.

Other MRF optimization techniques

Before finishing this section, we mention some other techniques, which have also been used in computer vision for the optimization of MRF methods.
Another class of techniques, resembling the way simulated annealing works, are the so-called continuation methods with graduated non convexity (GNC) [Blake and Zisserman, 1987] being one such example. In these methods the role of the temperature is played by another parameter γ. In this case, the intractable non-convex energy function is approximated by a sequence of energy functions parameterized by this parameter γ. When γ is large the energy function becomes strictly convex and so locating the global optimum is easy. However, as γ decreases,

the energy function becomes non-convex and local minima are starting to appear. The hope is that if γ is gradually decreased, then by tracking the sequence of local minima we will be able to locate the global optimum at the end. Unfortunately, however, continuation methods cannot provide any optimality guarantees about their solutions, except for certain special cases.

Another technique, which has also been widely used is the relaxation labeling method ([Ishikawa, 2003, Faugeras and Berthod, 1981, Rosenfeld et al., 1976]). The idea of relaxation labeling is to replace the discrete labels with continuous ones that must lie on a high dimensional simplex. In this way, the problem of optimizing a discrete MRF is converted into a constrained continuous optimization problem, which can then be solved using standard gradient descent techniques.

Finally, we should note that another class of optimization techniques are the so-called genetic algorithms ([Goldberg and Holland, 1988, Holland, 1975]). These are optimization methods which are inspired by the principle of natural evolution in the biological world and try to simulate the evolutionary process: in a population of individuals, those who possess the highest goodness-of-fit values are the ones who finally survive. Although genetic algorithms are general purpose optimization methods and were found effective in many applications, their main disadvantage is that they are mostly based on heuristic procedures.

3.4 Graphical representation of Markovian models

The theory of Markovian models, as a part of the theory of random processes, shows a strong bonding with *graphs*. Formally, a graph G is a pair of sets (S, E), where S is the set of vertices (nodes) and E is the set of edges, formed by pairs of vertices. In this framework, it is natural to describe a family of random variables by a set of sites that constitute the nodes of the graph that represent the random process. One of the advantages of using graphs in a probabilistic framework is the possibility of immediate visual interpretation of the relationships between variables that expresses different dependencies among the nodes; that is why we can talk about dependency graphs. The underlying probability distributions of Markovian models can be represented in a graphical form, this is why these models are often called probabilistic graphical models [Bishop, 2006]. A probabilistic graphical model is a diagrammatic representation of a probability distribution. In such a graph there is a node for each random variable. The absence of an edge between two variables represents conditional independence between those variables.

Definition 3.4.1 (Conditional independence) *Conditional independence means that two random variables X and Y are independent given a third random variable Z (denoted as, $X \perp Y \mid Z$), i.e., if they are independent in their conditional probability distribution, formally :*

$$p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$$

Conditional independence is an important concept as it can be used to decompose complex probability distributions into a product of factors, each consisting of the subset of corresponding random variables.

In the case of a regular lattice, the graph takes a particularly simple form called *Markov blanket* as shown in Figure 3.4: a node is conditionally independent of all other nodes when conditioned only on the neighboring nodes. Thus, the Markov blanket of a node simply consists of the set of all neighboring nodes.

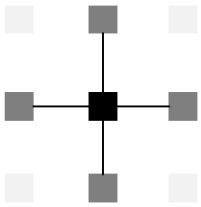


Figure 3.4: First order Markov blanket.

3.4.1 Dependency graph

Let us denote X a Markovian process. To reflect the interaction structure induced by the energy function $U(X = x)$, one could define a dependency graph as follows:

Definition 3.4.2 (Dependency graph) *With the same notation as in Section 3.2 the dependency graph associated with the energy decomposition* $U(x) = \sum_{c \in \mathcal{C}} V_c(x)$ *is the graph* $G = (S,\ E)$ *such as:*

$$\forall \{s,\ t\} \subset S$$
$$(s,\ t) \in E \ \Leftrightarrow\ \exists c \in \mathcal{C} \colon (s,\ t) \in c.$$

In other word, given a neighborhood system $\mathcal{N} = \{\mathcal{V}_s, s \in S\}$*:*

$$t \in \mathcal{V}_s \ \Leftrightarrow\ (s,\ t) \in E.$$

See [Lauritzen, 1996, Lauritzen et al., 1990, Whittaker, 2009] for more discussion about dependency graphs and their semantics.

3.4.2 Graphical interpretation

In this section, we examine the structural transformations generated on independence graphs by the following basic operations.

Freezing

The independences can be read off from the graph by freezing a subset of variables. As shown in Figure 3.5, an alternative way to view the conditional independence

between to subsets of vertices A and B conditioned to C ($A \perp B \mid C$), is to imagine removing all nodes in set C from the graph together with any links that connect to those nodes. We then ask if there exists a path that connects any node in A to any node in B. If there are no such path, then the conditional independence property holds.

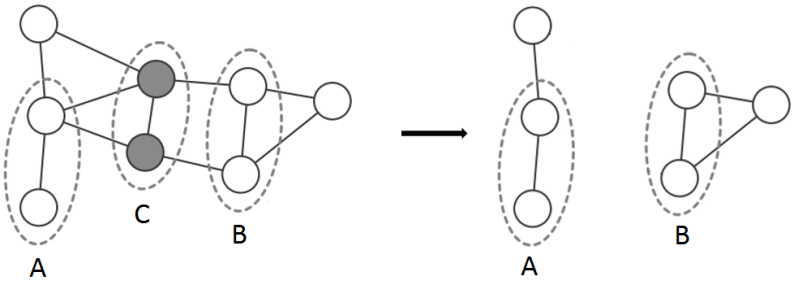


Figure 3.5: Graphical consequence of the conditional independence.

Summing and maximizing

Summing and maximizing act similar way as freezing on the graph. Indeed, It has been established in [Perez and Heitz, 1996] that, given a Markovian process, X associated to a graph $G(S, E)$, the marginal $p(x_A)$ for some subset $A \subset S$ has an independence graph G', in which two sites are neighbors if they are neighbors in G, or if they belong to the neighborhood of a same connected component $\overline{A} = S \setminus A$. This results from the summation of $p(x) = p(x_A, x_{\overline{A}})$ with respect to $x_{\overline{A}}$ which provides the related marginal distribution [Fieguth, 2010].
As a consequence of this result, from a graphical viewpoint, a dependency graph of the marginal distribution $p(x_A)$ is derived from the dependency graph associated with the joint distribution $p(x)$ by removing all nodes $a \in A$ and then by making all neighbors of a mutually neighbor as shown in Figure3.6.

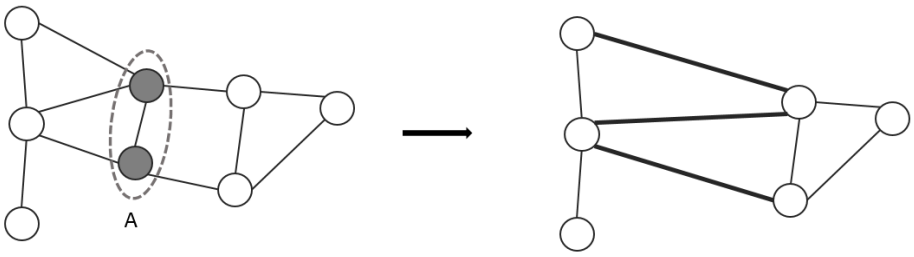


Figure 3.6: Graphical consequence of summing/maximizing out the variables in A.

Using the graphical representation, *probabilistic causal models* have

been already thoroughly studied [Luettgen et al., 1994, Pérez et al., 2000, Laferté et al., 2000, Willsky, 2002]. The classes of causal autoregressive fields [Luettgen et al., 1994], unilateral MRFs [Devijver, 1988], mesh MRFs [Razlighi et al., 2009, Yousefi and Kehtarnavaz, 2011], and mutually compatible MRFs on bidimensional lattices [Laferté et al., 2000, Voisin et al., 2014] have thus been introduced. These models rely on a probabilistic causality concept captured by the factorization of $p(x)$ in terms of causal transition kernels [Whittaker, 2009]. In the following of this chapter, we shall recall the basics of causality that could be applied to Markovian Processes.

3.5 Causality

To introduce the concept of causality, it is necessary to define an order over the set of sites S. In such a way, we characterize the past of a site s, as the set of nodes preceding s. For example, is S is a regular two-dimensional lattice, the past of a site (pixel) s located on the rth row and c column can be defined as the set of all the pixels located from the first and the $(r-1)^{th}$ rows and of all the pixels located in the rth row from the first to the $(c-1)^{th}$ column [Devijver, 1988] . More examples will be mentioned later with regard to hierarchical MRFs.
Given the ordered set $S = \{1,\ \cdots,\ s-1,\ s,\ s+1,\ \cdots,\ N\}$, the past of the site s will be denoted as $pa(s) = \{i \in S\colon i < s\}$. Thus, the causality is used to define a subset of sites $\partial_s \subseteq pa(s)$ that satisfies the following property ([Lauritzen, 1996, Whittaker, 2009, Lauritzen et al., 1990]):

$$\forall s > 1,\ p(x_s \mid x_{pa(s)}) \ = \ p(x_s \mid x_{\partial_s}). \tag{3.20}$$

If (3.20) holds, one can break the joint distribution $p(x = x_1, \cdots, x_N)$ as a product of the conditional probabilities over the past neighborhood. Formally:

$$p(x) = p(x_1) \prod_{s=2}^{N} p(x_s \mid x_{pa(s)}) \ = \ p(x_1) \prod_{s=2}^{N} p(x_s \mid x_{\partial_s}) \tag{3.21}$$

The most remarkable implication of (3.21) is the absence of a normalization constant, which helps to easily perform efficient and non-iterative forward recursive sampling of this Markov chain-type distribution. By "non-iterative", it is meant here that a solution x is obtained directly, i.e., without generating a sequence $\{x^k\}_{k=0}^{\infty}$ of solutions that converge to x.

In the case of a Markovian process indexed by an ordered set S, denoting as $U(x) = \sum_{c \in \mathcal{C}} V_c(x)$ the energy function with the dependency graph $G = (S,\ \mathcal{N})$ and given the nice properties offered by causality, the question that arises is to whether we can determine if a Markov process is causal on the sole basis of a small past neighborhood ∂_s defined as:

$$\forall s,\ \partial_s \ \subseteq \ \mathcal{V}_s \cap \ pa(s).$$

Before continuing, let us rearrange the potential product $\prod_{c\in\mathcal{C}} g_c$ such as:

$$\prod_{c\in\mathcal{C}} g_c = \prod_{1\leq i\leq N} g_i,$$

where g_i is the product of g_c's for all c containing i and no further site j (i.e., $j > i$). the joint distribution (3.21) becomes:

$$p(x) = p(x_1) \prod_{i=2}^{N} g_i(x_i, x_{\partial_i}), \tag{3.22}$$

where by definition of g_i, $\partial_i \subseteq \mathcal{V}_i \cap pa(i)$. However, by successively considering marginals of vectors $x_{pa(N)}$, $x_{pa(N-1)}$, $\cdots$, x_1), [Chardin, 2000] established functional cases where causal representation turns out to be at least as local as the original non-causal one (i.e., $\partial_i = \mathcal{V}_i \cap pa(i)$).

3.5.1 Functional characterization of the causality

As shown in the Section 3.4.2, the construction of the dependency graph associated with a marginal distribution $p(x_1, \cdots, x_i)$ from the dependency graph associated with the joint distribution $p(x_1, \cdots, x_N)$ involves the creation of edges between nodes neighboring the set of removed sites. However, in certain cases depending on the expression of the factors under concern, simplifications might occur in the marginal $p(x_1, \cdots, x_i)$, leading to a simpler independence graph [Pérez, 1993].

Proposition 3.5.1 (Functional characterization) *We denote by X a Markovian process with the joint distribution $p(x) = p(x_1) \prod_{i=2}^{N} g_i(x_i, x_{\partial_i})$, where $\partial_i = \mathcal{V}_i \cap pa(i)$.*

X is causal according to ∂_i, if $\forall i, \prod_{i=2}^{N} g_i(x_i, x_{\partial_i})$ is a constant with refer to x_i .

3.5.2 Graphical characterization of the causality

Graphical considerations allow to point out an important class of interaction models for which the functional characterization of the causality systematically holds. In fact, the graphical viewpoint allows in some cases to identify at first glance (without need of any computational or probabilistic argument) interaction structures that support causal models.

Proposition 3.5.2 (Graphical characterization) *We denote by X a Markovian process admitting G as an independence graph with the joint distribution $p(X = x) = p(x_1) \prod_{i=2}^{N} g_i(x_i, x_{\partial_i})$, where $\partial_i = \mathcal{V}_i \cap pa(i)$.*

X is causal according to ∂_i, if $\forall i, \partial_i = \mathcal{V}_i \cap pa(i)$ is a complete graph[2].

[2] A complete graph is a simple undirected graph in which every pair of distinct vertices is connected by a unique edge.

Again, the benefit of dependency graph is highlighted because it allows for a visual immediate identification of certain causal models. In fact, it has been shown in [Whittaker, 2009] that graphs satisfying the graphical characterization in Proposition 3.5.2 with respect to an ordered set are *triangulated* (or chordal), i.e., they contain no cycles of length 4 without a chord. Examples of triangulated graphs are shown in Figure 3.7.

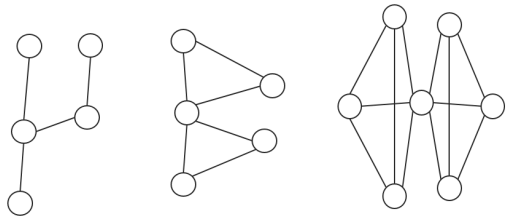


Figure 3.7: Examples of triangulated graphs.

3.5.3 Causal Optimization techniques on Markov chain

Let us denote $\{(X_t, Y_t)\}_{t\in[1,n]}$ a hidden Markov chain that respects the graph in Figure 3.8

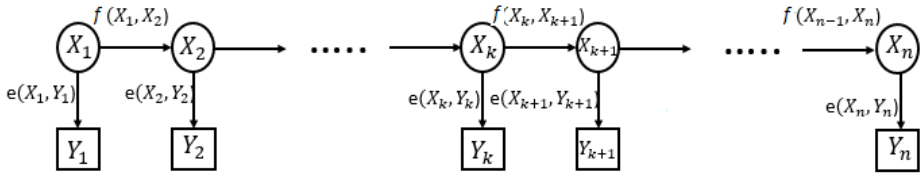


Figure 3.8: Hidden Markov chain.

where $f(.,.)$ denote the transition probabilities and the $e(.,.)$ are emission probabilities. Formally :

$$f(i,j) = p(X_t = i|X_{t-1} = j); \quad e(k,l) = p(X_t = k|Y_t = l) \tag{3.23}$$

For the sake of argument, all random variables in (3.23) are assumed discrete but extensions to the continuous observation variables y are rather straightforward. Because of its causal nature, the graph generated by Markov chain allows obtaining the posterior distribution of the hidden variables efficiently using a two stage message passing algorithm. In the particular context of the hidden Markov model, this is known as the forward-backward (F-B) algorithm which is an example of *dynamic programing*[3] [Rabiner et al., 1989]. There are several variants of the basic

[3]Dynamic programing refers to optimization methods that allows an exact inference technique [Bellman, 1956].

F-B technique [Jordan, 2003]. In the following, we shall focus on the most widely used algorithms.

Viterbi algorithm

The general form of the F-B algorithm can be used to do inference on individual states, which represent hidden variables, of the Markov chain given a sequence of observations. However, in many applications, the hidden variables have some meaningful interpretations, so it is often of interest to find the most probable sequence of hidden states for a given observation set. This is what the Viterbi algorithm does ([Viterbi, 1967, David-Forney, 1973]). Formally, the goal of this algorithm is to compute recursively the best sequence x of the hidden variables that maximize the posterior distribution:

$$\begin{aligned} x^\star &= arg\max_x p(X = x|Y = y) \\ &= arg\max_x p(X = x, Y = y) \end{aligned} \tag{3.24}$$

First let us consider the following proposition:

Proposition 3.5.3

$$\begin{gathered} \forall a, b \text{ and } \forall \text{ } u, \text{ } v \text{ bounded above functions:} \\ \text{if } u(a) \geqslant 0 \quad \text{and} \quad v(a, b) \geqslant 0 \text{ then:} \\ \max_{a,b}\left[u(a)v(a, b)\right] = \max_a \left[u(a) \max_b v(a, b)\right] \end{gathered}$$

At time t, let us consider all the paths in the sequence of states that recognize the observations until time t, $(Y_1 = y_1, \ldots, Y_t = y_t)$ and where the last state is equal to x_t $(X_t = x_t)$. Then, one could introduce th following quantity:

$$\delta_t(x_t) = \max_{x_1,\ldots,x_{t-1}} p(Y_1 = y_1, \ldots, Y_t = y_t, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x_t), \tag{3.25}$$

For convenience, (3.25) is abbreviated to:

$$\delta_t(x_t) = \max_{x_1,\ldots,x_{t-1}} p(y_1, \ldots, y_{t-1}, x_1, \ldots, x_{t-1}, x_t), \tag{3.26}$$

which represents the probability of the best path from those previously considered. Next, under conditional independence assumptions derived from the graph in Figure 3.8 and using Proposition 3.5.3 one could compute (3.25) recursively as follows:

$$\begin{aligned} \delta_t(x_t) &= \max_{x_1,\ldots,x_{t-1}} p(y_1, \ldots, y_t, x_1, \ldots, x_{t-1}, x_t) \\ &= \max_{x_1,\ldots,x_{t-1}} [p(y_t|x_t)p(x_t|x_{t-1})p(y_1, \ldots, y_t, x_1, \ldots, x_{t-1})] \\ &= \max_{x_{t-1}}[p(y_t|x_t)p(x_t|x_{t-1}) \max_{x_1,\ldots,x_{t-2}} p(y_1, \ldots, y_t, x_1, \ldots, x_{t-1})] \\ &= \max_{x_{t-1}}[e(y_t, x_t)f(x_t, x_{t-1})\delta_{t-1}(x_{t-1})], \end{aligned} \tag{3.27}$$

which gives the following recursion:

$$\delta_t(x_t) = e(y_t, x_t) \max_{x_{t-1}} [f(x_t, x_{t-1}) \delta_{t-1}(x_{t-1})], \tag{3.28}$$

where

$$\delta_1(x_1) = p(y_1, x_1) = p(x_1) e(y_1, x_1) \tag{3.29}$$

which gives

$$\begin{aligned} x_t^\star &= arg(\delta_t(x_t)) \\ x_t^\star &= arg \max_{x_{t-1}} [f(x_t, x_{t-1}) \delta_{t-1}(x_{t-1})] \end{aligned} \tag{3.30}$$

Algorithm 3.5.1 (Viterbi Algorithm)

↻ **Initialization**

$\delta_1(x_1) = p(x_1) e(y_1, x_1)$.

⇒ **Forward pass**

$\forall t \in \{2, \ldots, n\}$:

$$\delta_t(x_t) = e(y_t, x_t) \max_{x_{t-1}} [f(x_t, x_{t-1}) \delta_{t-1}(x_{t-1})]$$
$$x_t^\star = arg \max_{x_{t-1}} [f(x_t, x_{t-1}) \delta_{t-1}(x_{t-1})]$$

⇐ **Backward pass**

Construction of the optimal path x as a sequence of the optimal states computed during the forward step.

Baum-Welch algorithm

[Baum et al., 1970] tried to solve the problem of estimating the optimal sequence based on a likelihood criterion :

$$\begin{aligned} \mathcal{L}_t &= p(x_t, y) \\ x_t^\star &= arg \max_{x_t} \mathcal{L}_t \end{aligned} \tag{3.31}$$

Then, *Forward* and *Backward* probabilities have been introduced via the decomposition of (3.32):

$$\begin{aligned} \mathcal{L}_t &= p(x_t) p(y|x_t) \\ &= p(x_t, y_1, \ldots, y_t) \times p(y_{t+1}, \ldots, y_n | x_t) \\ &= \mathcal{F}_t \times \mathcal{B}_t \end{aligned} \tag{3.32}$$

where $\mathcal{F}_t$ and $\mathcal{B}_t$ are computed recursively as follows:

$$\begin{aligned}
\mathcal{F}_t &= p(x_t, y_1, \ldots, y_t) \\
&= \sum_{x_{t-1}} p(x_t, x_{t-1}, y_1, \ldots, y_t) \\
&= \sum_{x_{t-1}} p(x_t, y_t | x_{t-1}) p(x_{t-1}, y_1, \ldots, y_{t-1}) \\
&= \sum_{x_{t-1}} p(y_t | x_t) p(x_t | x_{t-1}) \mathcal{F}_{t-1} \\
&= e(y_t | x_t) \sum_{x_{t-1}} f(x_t, x_{t-1}) \mathcal{F}_{t-1}
\end{aligned} \tag{3.33}$$

$$\begin{aligned}
\mathcal{B}_t &= p(y_{t+1}, \ldots, y_n | x_t) \\
&= \sum_{x_{t+1}} p(x_{t+1}, y_{t+1}, \ldots, y_n | x_t) \\
&= \sum_{x_{t+1}} p(x_{t+1} | x_t) p(y_{t+1}, \ldots, y_n | x_{t+1}) \\
&= \sum_{x_{t+1}} f(x_{t+1}, x_t) p(y_{t+1} | x_{t+1}) p(y_{t+2}, \ldots, y_n | x_{t+1}) \\
&= \sum_{x_{t+1}} f(x_{t+1}, x_t) e(y_{t+1}, x_{t+1}) \mathcal{B}_{t+1}
\end{aligned} \tag{3.34}$$

Algorithm 3.5.2 (Baum-Welch algorithm)

$\Rightarrow$ **Forward pass**

Initialization

$$\mathcal{F}_1 = p(x_1, y_1) = e(x_1, y_1)p(x_1)$$

Recursion $\forall t \in \{2, \dots, n\}$:

$$\mathcal{F}_t = e(y_t, x_t) \sum_{x_{t-1}} f(x_t, x_{t-1})\mathcal{F}_{t-1}$$

$\Leftarrow$ **Backward pass**

Initialization

$$\mathcal{B}_n = 1$$
$$x_n^\star = arg \max_{x_n} \mathcal{F}_n$$

Recursion $\forall t \in \{n-1, \dots, 2\}$:

$$\mathcal{B}_t = \sum_{x_{t+1}} f(x_{t+1}, x_t)e(y_{t+1}, x_{t+1})\mathcal{B}_{t+1}$$
$$x_t^\star = arg \max_{x_n} \mathcal{B}_t \times \mathcal{F}_n$$

In the following chapter, we will focus on causal models on Markov fields via the particular case of tree structures as they are triangulated by definition and do not include cycles.

Chapter 4

Hierarchical Markov Models on quad-trees

Contents

4.1 Basic model structure

The objective of this chapter is to describe multi-resolution classification techniques based on hierarchical Markovian modeling. These methods have two requirements for accomplishing this task:

(i) the method should be able to handle the data acquired at different resolutions

(ii) the method should provide a structure that can easily handle the interactions between different images in the input data set.

Parallel multi-grid (or pyramidal) schemes are one of the possible approaches satisfying requirement **(i)**. The pyramid structure is a type of signal representation in which images are organized according to their resolutions [Jolion and Rosenfeld, 2012] (see Figure 4.1), i.e., a pyramid P is a stack of images I_n for which the scale $n \in [0, R]$ and R is the height of the pyramid. An element of this pyramid is called a node and may correspond to a pixel or a group of pixels in the image domain.

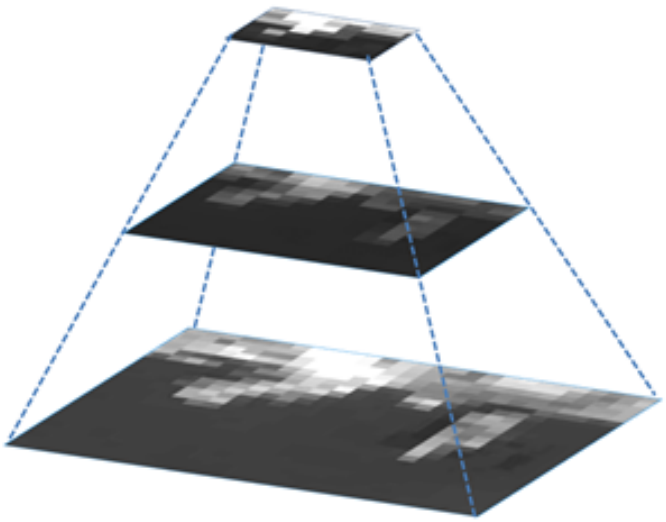
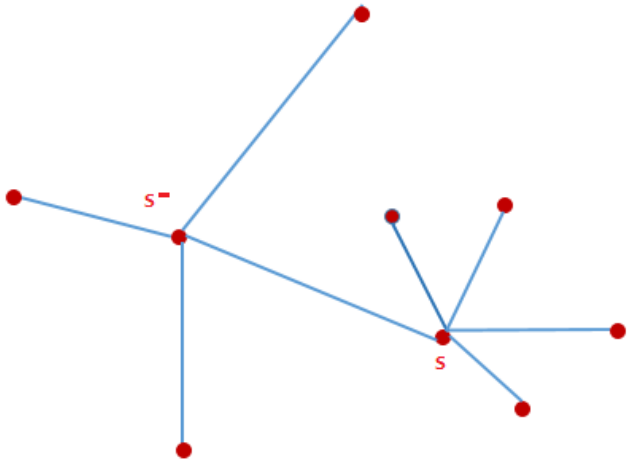
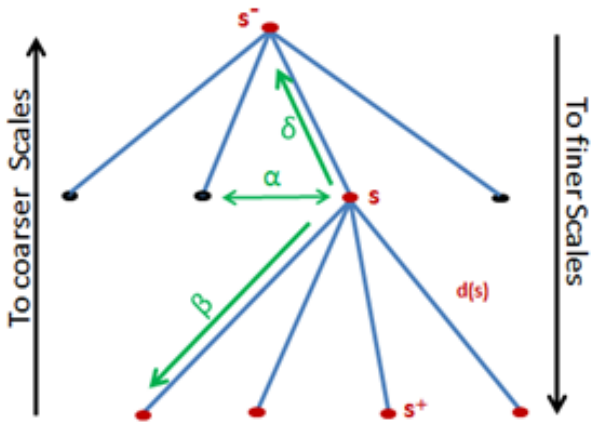


Figure 4.1: Images are organized according to their resolutions in a pyramid structure.

To handle requirement **(ii)**, we define for each node of the pyramid a set of links to other nodes to model scale-to-scale interactions. The theory of multi-scale signals has been widely studied, and their representations naturally lead to models of signals on trees. Among others, dyadic trees (e.g., [Basseville et al., 1992a, Basseville et al., 1992b]) and quad-trees (e.g., [Pérez et al., 2000, Laferté et al., 2000, Chardin, 2000, Voisin et al., 2014]) have been proposed as attractive candidates for modeling these scale-to-scale interactions



(a)



(b)

Figure 4.2: (a) A typical example of a tree; (b) tree structure that models the interaction between images in the pyramid structure.

in mono-dimensional and bi-dimensional signals, respectively. As discussed in the section 3.5.2 of the previous chapter, the selection of these structures is justified by their causality properties over scale and by the possibility of employing a fast optimization method. Furthermore, the inference technique using tree structures is much simpler than the techniques for other triangulated graphs because of the acyclic (i.e., loop-free) property. A useful interpretation makes use of the relationship between directed graphical models and undirected ones. In a directed graphical model, the quantities that must be specified include the conditional distribution at each node given the values of all of its parents (where s^- is a parent of s if there is a directed edge from s^- to s). It is straightforward to convert a directed graphical model into an undirected one, but the construction of a directed graphical model equivalent to an undirected one is generally very complex and, in fact, requires defining new node and edge sets where the nodes consist of entire cliques of nodes of the original undirected graph. For a tree, however, the construction of a directed graphical model from an undirected one is trivial and in fact does not change the nodes of the graph nor the graphical structure (except that edges become directed rather than undirected). Specifically, consider an undirected graphical model over a tree and choose any node to designate as the "root" node. Consider then "hanging" the tree from this node i.e., redraw the graph with the root node at the top level, with its neighbors at the next level, etc. For example, in Figure 4.2 (a), we have labeled one node, s^-, in Figure 4.2 (b) as the root node of the considered tree.

Let us denote a generic node on the specific case of a quad-tree as s and the finite set of all nodes as $S(s \in S)$. Each node is a pixel in one of the levels of the tree. The set of nodes is then hierarchically partitioned, (i.e., $S = S^0 \cup S^1 \cup \ldots \cup S^R$) where S^n indicates the subset of nodes associated with the n^{th} level of the tree ($n = 0, 1, \ldots, R$), $n = R$ denotes the root of the tree (coarsest resolution) and $n = 0$ indicates its leaves (finest resolution).
In the considered structure, a parent-child relationship can be defined: an upward shift operator δ such that $s^- = \delta(s)$ is the parent of node s. The operator δ is not one-to-one, but four-to-one because each parent has four offspring (because of the quad-tree structure). We define the forward shift operator β such that $s^+ = \beta(s)$ is the set of all the descendants of s, the interchange operator α is defined as between the nodes in the identical scale, and $d(s)$ is the set including s and all its descendants in the tree as illustrated in Figure 4.2 (b). This framework allows data at different resolutions and different spectral bands to be fused.

4.2 Estimation and inference algorithm on quad-tree

4.2.1 Problem statement

The described hierarchical structure allows, in a natural way, the use of an explicit statistical model through a hierarchical Markov random field formulation using a

set of random fields, which are associated with the different scales, and exploiting the operators defined above on the quad-tree structure. Therefore, let us denote the class label of site s as a discrete random variable x_s and its value as ω_s ($s \in S$). If there are M classes in the considered scene, then each label occupies a value in the set $\Lambda = \{0, 1, \ldots, M-1\}$ (i.e., $x_s, \omega_s \in \Lambda$). The class labels of all pixels can be collected in a set $X = \{x_s\}_{s \in S}$ of random fields $X^n = \{x_s\}_{s \in S^n}$ associated with each scale n, where S_n is the related set of lattice points. The corresponding configuration at scale n and can be represented as $\omega^n = (\omega_s)_{s \in S^n}$. The configuration space $\Omega = \Lambda^{|S|}$ is the set of all global discrete labelings (i.e., $X \in \Omega$).
One could then assume the following to fit a MRF model to the aforementioned hierarchical structure:

(i) The fundamental assumption of the model is that the sequence of random fields from coarse to fine scales forms a Markov chain over scale :

$$p(X^n = \omega^n | X^q = \omega^q, q > n) = p(X^n = \omega^n | X^{n+1} = \omega^{n+1}). \tag{4.1}$$

For convenience (4.1) is abbreviated to:

$$p(x^n | x^q, q > n) = p(x^n | x^{n+1}). \tag{4.2}$$

(ii) The transition probabilities of this Markov chain factorize so that the components (pixels or node s) of X^n are mutually independent given the components of X^{n+1} :

$$p(X^n = \omega^n | X^{n+1} = \omega^{n+1}) = \prod_{s \in S^n} p(x_s = \omega_s | x_{s^-} = \omega_{s^-}). \tag{4.3}$$

Again for convenience purpose (5.2) is abbreviated to:

$$p(x^n | x^{n+1}) = \prod_{s \in S^n} p(x_s | x_{s^-}). \tag{4.4}$$

Using the quad-tree structure allows benefiting from the good properties discussed in the Section 4.1 (i.e., causality) and applying non-iterative algorithms, thus resulting in a decrease in computational time compared to iterative optimization procedures over graphs. As shown in Chapter 3, several Bayesian approaches are offered to solve this inference problem by minimizing the Bayes risk (see 3.10).

Among the different classification algorithms employed on a quad-tree structure in the literature, two have been widely used and match the MAP and MPM criteria, respectively. The first algorithm aims to exactly estimate the MAP configuration, the resulting estimator is given by 3.15. The second algorithm produces the configuration that maximizes at each site s the a posteriori marginal, the resulting estimator is given by 3.17.

To better highlight the difference between these two criteria and their implications in the application to quad-trees, let us recall the formulation of the cost functions associated with the MAP and MPM (see equations 3.12 and 3.16 in Chapter 3). The cost function of MAP is defined by the following:

$$C_{MAP}(x^\star, x) = 1 - \delta(x^\star, x) = 1 - \prod_{s \in S} \delta(x_s^\star, x_s). \tag{4.5}$$

This function implies identical cost for all configurations of $x^\star$ (different from x). the MAP estimator of the label field X is given by the following:

$$x^{MAP} = arg \max_{x \in \Omega} p(x|y), \tag{4.6}$$

This combinatorial optimization problem can be resolved by using a Kalman-like filter [Willsky, 2002], owing to the formal similarity between MRF models and the Viterbi algorithm [David-Forney, 1973]. The extension of the Viterbi algorithm, which computes the exact MAP estimate of X given $Y = y$ on the quad-tree has been introduced by [Dawid, 1992] (see Section 3.3.1) in the context of probabilistic expert systems, and [Laferté et al., 2000] in the context of image classification by proposing a non-iterative algorithm on the quad-tree. However, in the application to quad-trees, these algorithms exhibit two main shortcomings. First, computationally, they are known to be affected by underflow problems because of the small probabilities involved. Second, according to (4.5), the MAP cost function penalizes the discrepancies between configurations regardless of their corresponding scales, an undesirable property from the viewpoint of segmentation, labeling, and classification. Specifically, an error at a coarser scale will be paid the same cost as an error at a finer scale whereas it is desirable to have a higher cost for errors at coarser levels because they may generally lead to the misclassification of groups of pixels at level 0 (e.g., one pixel at the root corresponds to 4^R pixels at the finest scale).

On the contrary, the MPM criterion is based on a criterion function that aims at segmentation accuracy and allows errors on distinct scales to be penalized differently [Laferté et al., 2000, Kato and Zerubia, 2012]. The cost function is:

$$C_{MPM}(x^\star, x) = \sum_{s \in S} [1 - \delta(x_s^\star, x_s)], \tag{4.7}$$

which is related to the number of sites $s \in S$ such that $x_s^\star \neq x_s$. The MPM criterion penalizes errors according to their number, at the scale at which they occur. The Bayesian estimator resulting from (4.7) is given by the following:

$$\forall s \in S, x_s^{MPM} = arg \max_{x_s \in \Lambda} p(x_s|y), \tag{4.8}$$

Furthermore, as shown in [Marroquin et al., 1987, Bouman, 1991], MPM well adapts the estimator to the quad-tree topology. Indeed, because the tree is acyclic, the labels are estimated recursively by MPM through a forward-backward algorithm similar to the classical Baum and Welch algorithm for Markov chains [Baum et al., 1970].

In the following of this thesis MPM criterion is chosen to address inference problems on the quad-tree structures.

4.2.2 MPM inference

The starting point of the original procedure introduced in [Laferté et al., 2000] lies in the expression of the posterior marginal $p(x_s|y)$ as a function of the posterior marginal at the parent node $p(x_{s^-}|y)$. Formally it has been shown that under some independence conditional assumptions detailed in [Laferté et al., 2000]:

$$p(x_s|y) = \sum_{x_{s^-}} \frac{p(x_s, x_{s^-}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}|y_{d(s)})} \; p(x_{s^-}|y). \tag{4.9}$$

Where $y_{d(s)}$ is the vector collecting the observations of all the descendants of site $s \in S^n, n = 0, 1, \ldots, R-1$. This yields a top-down recursion whereas the posterior marginal is initialized at the root node as well as the probabilities $p(x_s, x_{s^-}|y_{d(s)})$ are made available. Moreover, these probabilities can be computed as:

$$p(x_s, x_{s^-}|y_{d(s)}) = \frac{p(x_s|x_{s^-})p(x_{s^-})}{p(x_s)} p(x_s|y_{d(s)}), \tag{4.10}$$

where another conditional independence assumption has been used, $p(x_s|x_{s^-})$ is the transition probability over scale, and $p(x_s)$, $p(x_{s^-})$ are the prior marginals computed using a simple top down recursion:

$$p(x_s) = \sum_{x_{s^-}} p(x_s|x_{s^-}) \; p(x_{s^-}). \tag{4.11}$$

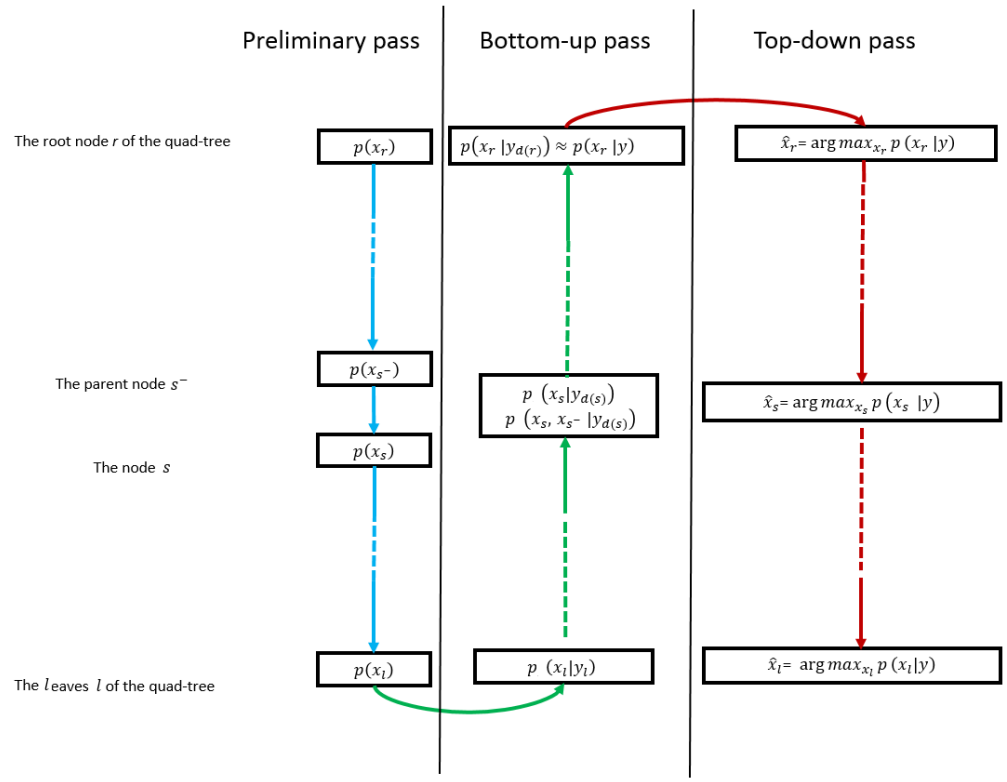
A bottom-up pass allows to compute the partial posterior marginals $p(x_s|y_{d(s)})$ in 4.10. [Laferté et al., 2000] proved that these marginals could be expressed as a recursion over scales ($s \in S^n, n = 1, 2, \ldots, R$):

$$p(x_s|y_{d(s)}) \propto p(y_s|x_s)p(x_s) \prod_{t \in s^+} \sum_{x_t} \frac{p(x_t|y_{d(t)})}{p(x_t)} p(x_t|x_s), \tag{4.12}$$

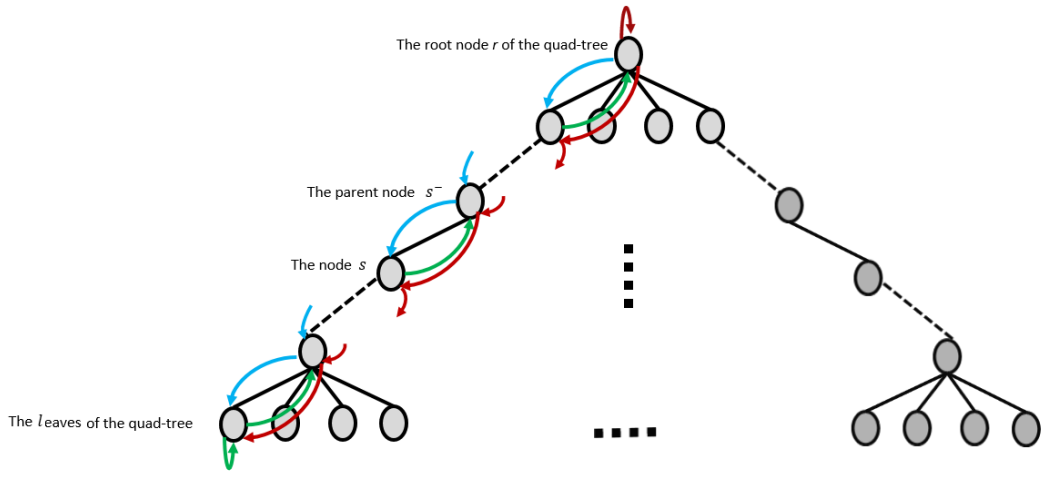
where $p(y_s|x_s)$ are the pixelwise class-conditional PDFs of the image data at each node of the quad-tree. Note that the product over the children set is actually absent at the leaves of the tree (i.e., $s \in S^0$) in which the partial posterior marginals are computed directly:

$$p(x_s|y_s) \propto p(y_s|x_s) \; p(x_s). \tag{4.13}$$

The whole procedure is summarized in Figure 4.3 and Algorithm 4.2.1.



(a)



(b)

Figure 4.3: The recursive algorithm on the quad-tree via the MPM criterion

Algorithm 4.2.1 (MPM estimation on the quad-tree)

↻ **Preliminary pass**
A preliminary top-down recursion is performed to compute the prior probabilities using (4.11).

⇑ **Bottom-up pass**
This step provides the distributions $p(x_s|y_{d(s)})$ and $p(x_s, x_{s^-}|y_{d(s)})$ via (4.10) and (4.12) respectively.

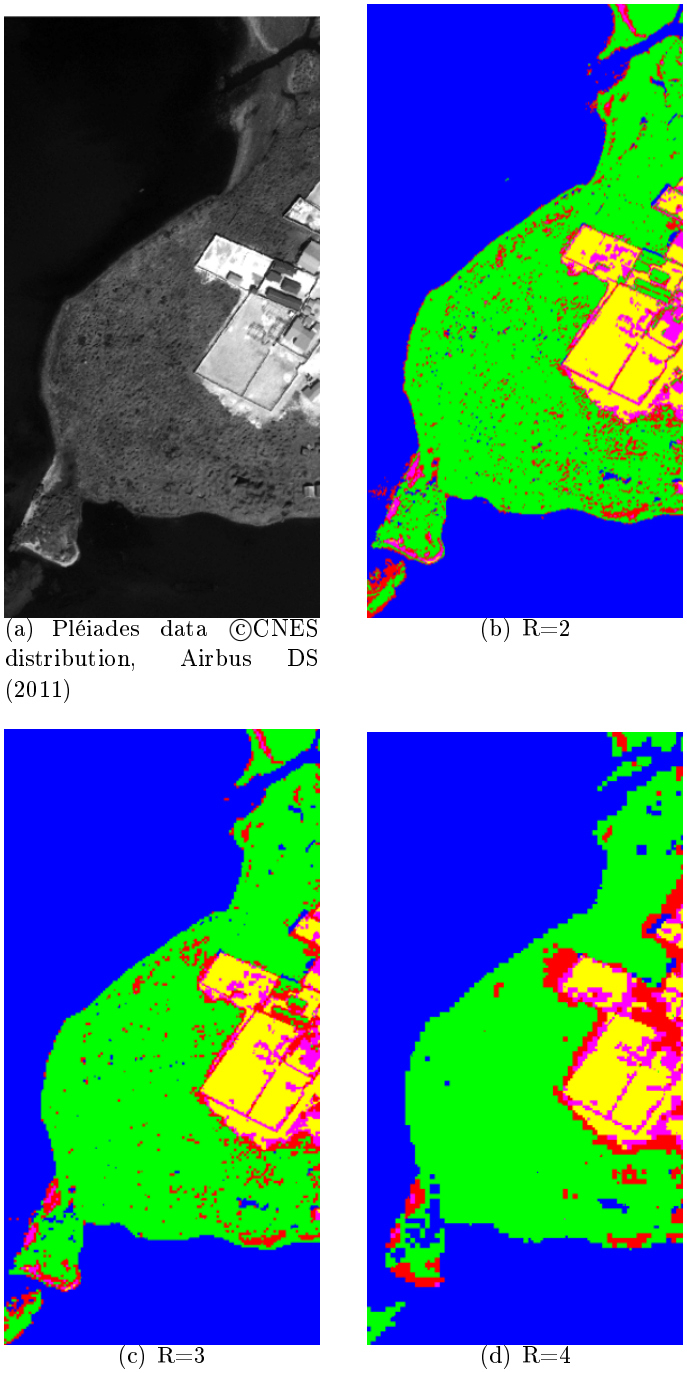
⇓ **Top-down pass**
This downward recursion reassembles the complete posterior marginals using (4.9) and then deduces the MPM solution via (4.8).

4.3 Blocky artifacts and further extensions of the hierarchical model

4.3.1 Blocky artifacts

The tree structure induces non-stationarity in space: the distribution at the leaves is not invariant with respect to the δ operator, since the correlation between two variables depends on the "distance" to their common ancestor in the tree. This may result in "blocky" artifacts in the final classification map. This phenomenon can be explained by the fact that two neighboring sites at a given scale may not have the same parent. In this case, a boundary is more likely to appear than when they are linked by a parent node. This artifact has been reported in several papers [Bouman, 1991, Chou et al., 1993, Luettgen et al., 1994, Fieguth et al., 1998, Laferté et al., 2000]. Figure 4.4 shows the blocky artifacts where the method in [Laferté et al., 2000] was applied using the MPM criterion at varying levels of the quad-tree ($R = 1, 2, 3$) using Pléiades images.

Several techniques have been proposed to circumvent such undesired effects. For example, a posteriori smoothing was developed in [Luettgen et al., 1994] using a multi-scale regularization (MR) algorithm [Chou, 1991]. Another way to alleviate these effects was to define a tree structure with overlapping data leaves [Irving et al., 1997]. Further methods consists in using hierarchical graph structures that are more complex than usual trees [Bouman, 1991, Kato et al., 1996, Comer and Delp, 1999]. Unfortunately, in these cases, the practical advantages of the tree structure are then partly or completely lost. Finally, other scientists avoided the block artifacts by using semi-iterative algorithms in which the prior distribution was computed iteratively on the quad-tree (prior smoothing) [Chardin, 2000, Voisin, 2012]. Some of these algorithms are described in the next section.



(a) Pléiades data ©CNES distribution, Airbus DS (2011)

(b) R=2

(c) R=3

(d) R=4

Figure 4.4: Blocky artifacts using the method proposed in [Laferté et al., 2000] using the MPM criterion at varying hight of the quad-tree.

4.3.2 Other techniques using trees

4.3.2.1 Kato et al.'s method

[Kato et al., 1996] consider a more complex graph (called 3D pyramidal MRF) in which the original quad-tree is combined with a spatial lattice neighborhood at each level as shown in Figure 4.5.

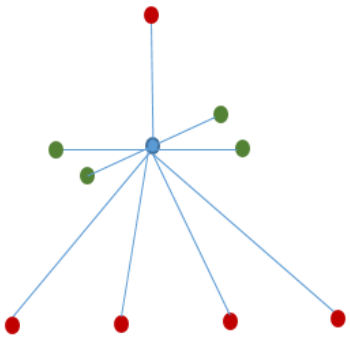
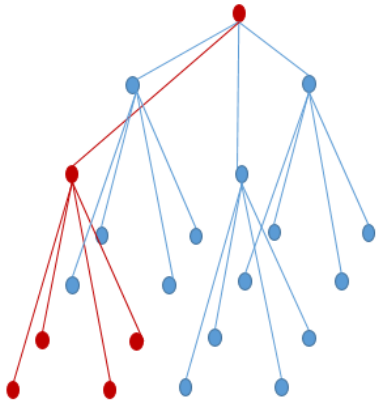


Figure 4.5: Neighborhood system in [Kato et al., 1996].

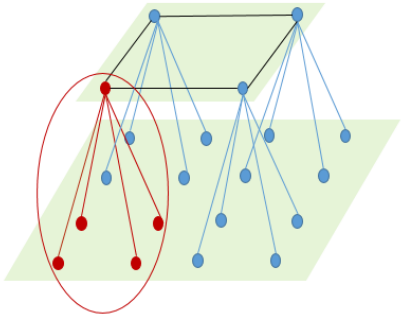
This results in an interleaved model whose manipulation is iterative. The inference is conducted using a parallel relaxation scheme [Besag, 1986] which is based on a modified annealing procedure (multi-temperature annealing [Kato et al., 1996]) where temperature is kept high at the coarse levels of the structure, thus, enabling the algorithm to be less sensitive to local minima. This approach provides accurate classification results but is computationally demanding.

4.3.2.2 Chardin et al.'s method

[Chardin, 2000, Chardin and Pérez, 1999] investigated an extension of the hierarchical stochastic model described in [Laferté et al., 2000] by taking benefit from both the spatial and the hierarchical prior model, combine a causal hierarchical prior on trees with a non-causal spatial prior at the coarsest level of the hierarchy: it is a Markov random field attached to the nodes of a truncated tree as shown in Figure 4.6. The proposed hybrid structure leads to semi-iterative algorithms which mix exact non-iterative procedures on sub-trees and iterative procedures on the grid at the top of the structure.



(a) Quad-tree



(b) Hybrid structure

Figure 4.6: Two hierarchical structures: (a) quad-tree with three levels; (b) truncated tree with two levels proposed by [Chardin, 2000, Chardin and Pérez, 1999].

4.3.2.3 Voisin et al.'s method

[Voisin, 2012, Voisin et al., 2014] enhanced the classification technique in [Laferté et al., 2000] by better estimating the prior at the coarsest scale through a case-specific prior update algorithm. Thus, a MPM estimation on a R-scale tree is performed first by choosing an uniform prior, and then by using the preliminary results as a new prior. Next, a smaller tree of scale $R-1$ is considered to which the MPM algorithm is applied to estimate a new prior. This step is repeated iteratively until scale 0 is reached. The scheme of this algorithm is given in Figure 4.7. This method improves robustness with respect to blocky artifacts as compared to the original method in [Laferté et al., 2000], thus increasing the accuracy when compared to the basic MPM tree as experimentally demonstrated in [Voisin et al., 2014].

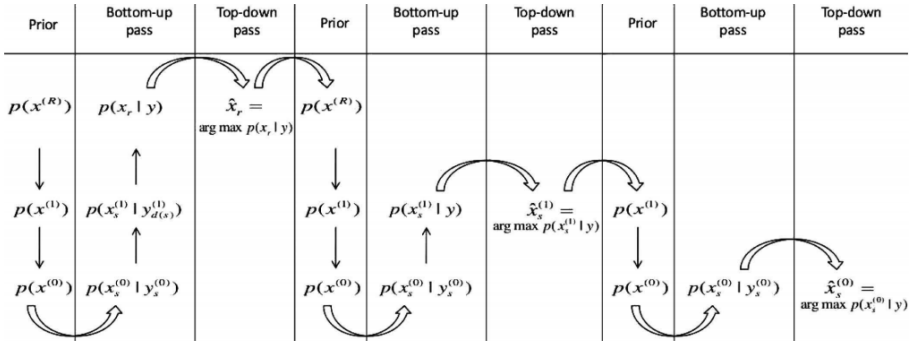


Figure 4.7: MPM estimation on the quad-tree as proposed in [Voisin, 2012, Voisin et al., 2014] In this representation, R = 2.

The main advantage of algorithms which use quad-tree structures is their capabilities to built non-iterative procedures when the causality property holds. In particular, a recursive procedure on a quad-tree is feasible for the marginal posterior modes (MPM) criterion. Then the aim is to maximize the posterior marginal at each site s. Because the tree is acyclic, the labels are estimated recursively through a forward-backward algorithm similar to the classical Baum and Welch technique for Markov chain described in algorithm 3.5.3. In practice, the use of one quad-tree structure with the MPM criterion often yields "blocky" effects as shown in Figure 4.4. This phenomenon can be explained by the fact that two neighboring sites at a given scale may not have the same parent. This undesirable effect can be reduced using some modification as in the algorithms described in the previous section. In the next chapters, several other hierarchical MRF models based on quad-tree will be introduced to support multi-resolution, multi-temporal, multi-sensor data, and to incorporate spatial contextual info either at the root or in all levels of the trees.

Chapter 5

Multi-temporal hierarchical Markov model on quad-tree

Contents

5.1 Multi-temporal structure

In the previous chapter (section 4.1) the basic quad-tree model was described. A generic node was denoted as s and the finite set of all nodes was denoted as $S(s \in S)$. In such a structure (see Figure 4.2), a parent-child relationship was defined:

δ: is the upward shift operator such that $s^- = \delta(s)$ is the parent of node s.

β: is the forward shift operator such that $s^+ = \beta(s)$ is the set of the descendants of s.

α: is the interchange operator between the nodes at the same scale.

$d(s)$: is the set including s and all its descendants in the tree.

This framework allows data at different resolutions and different spectral bands to be fused. A novel element of the first proposed approach which is methodologically described and experimentally validated in this chapter, is the multi-temporal aspect. Therefore, multiple pyramids and quad-trees are employed in a cascade, each pyramid being associated with the set of images available at a different date to characterize the temporal correlations associated with distinct images in the input time series, while still supporting the classification of data acquired at multiple spatial resolutions. For this purpose, this extended multi-temporal quad-tree structure is endowed with new operators to link between the nodes across different dates, and is combined with multi-scale wavelet feature extraction. We define an upward shift operator ϖ such that $s^= = \varpi(s)$ is the parent of node s in the previous date of the time series. Furthermore, we define an interchange operator σ between the nodes at the identical scale and identical position but from consecutive dates to characterize the temporal correlation between images given at different dates (see Figure 5.1). This multi-temporal hierarchical structure is aimed at supporting the joint classification of both multi-temporal and multi-resolution input images. This implies that, if only one image is available on a certain acquisition time, then it will be included at the finest resolution layer (level 0) of the corresponding quad-tree. Hence, the intrinsic resolution of the image will be the finest resolution of the quad-tree of that date. All other levels of the tree will be filled in using wavelet transforms [Mallat, 2008] of the input image.

If images corresponding to multiple resolutions are available on a certain time, a scenario that occurs, for example, when there are both higher resolution panchromatic and coarser resolution multi-spectral data, then each single-resolution input image is included in a separate level of the quad-tree. In this case, an assumption implicit in the quad-tree topology is that the spatial resolutions of the input images are related by a power-of-2 relationship. This condition is satisfied with minor approximations by most current multi-resolution space-borne optical sensors (see Chapter 1), so it is currently only a mild restriction from an operational viewpoint.

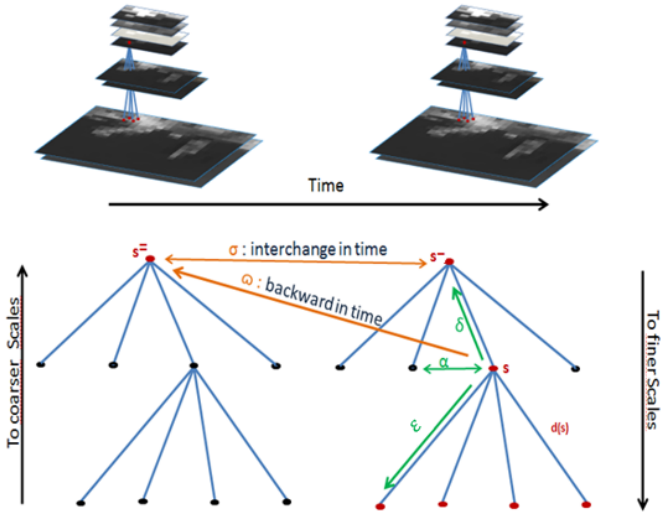


Figure 5.1: Multi-temporal hierarchical structure.

After the input images are included in the layers corresponding to their spatial resolutions, level 0 of the quad-tree corresponds to the finest-resolution input image, while some intermediate levels generally remain empty and are filled in using wavelet transforms of the images from the lower (finer resolution) layers. For example, if an IKONOS acquisition composed of a panchromatic component at 1-m resolution and a multi-spectral component at 4-m resolution is used on a certain date, then the finest resolution of the quad-tree is 1 m, the panchromatic and multi-spectral images are included at levels 0 and 2, respectively, and level 1 is computed as a wavelet transform of the panchromatic data at level 0.
The proposed hierarchical structure allows, in a natural way, the use of an explicit statistical model through a hierarchical Markov random field formulation using a series of random fields at varying scales and times, using the operators defined above, on the consecutive quad-trees.
Let us denote the class label of site s as a discrete random variable x_s and its value as $\omega_s(s \in S = \cup S_t^n$, where t and n refer to the corresponding time and scale respectively). If there are M classes in the considered scene, then each label takes a value in the set $\Lambda = \{0, 1, \ldots, M-1\}$. The class labels of all pixels can be collected in a set $\mathcal{X} = \{x_s\}_{s \in S}$ of random fields $\mathcal{X}_t^n = \{x_s\}_{s \in S_t^n}$ associated with each scale n and date t, where S_t^n is the related set of lattice points. The corresponding configuration at scale n and date t can be represented as $\omega_t^n = \{\omega_s\}_{s \in S_t^n}$. The configuration space $\Omega = \Lambda^{|S|}$ is the set of all global discrete labelings (i.e., $\mathcal{X} \in \Omega$).
We then assume the following to fit an MRF model to the aforementioned hierarchical structure:

(i) The fundamental assumption of the model is that the sequence of random fields

from coarse to fine scales form a Markov chain over scale and time:

$$p(\mathcal{X}_t^n = \omega_t^n | \mathcal{X}_p^q = \omega^q, p < t, q > n) = p(\mathcal{X}_t^n = \omega_t^n | \mathcal{X}_t^{n+1} = \omega_t^{n+1}, \mathcal{X}_{t-1}^{n+1} = \omega_{t-1}^{n+1}). \tag{5.1}$$

(ii) The transition probabilities of this Markov chain factorize so that the components (pixels or nodes s) of $\mathcal{X}_t^n$ are mutually independent given the components of $\mathcal{X}_t^{n+1}$ and $\mathcal{X}_{t-1}^{n+1}$:

$$p(\mathcal{X}_t^n = \omega_t^n | \mathcal{X}_t^{n+1} = \omega_t^{n+1}, \mathcal{X}_{t-1}^{n+1} = \omega_{t-1}^{n+1}) = \prod_{s \in S^n} p(x_s = \omega_s | x_{s^-} = \omega_{s^-}, x_{s^=} = \omega_{s^=}). \tag{5.2}$$

In the next section, an extended version of the MPM estimator will be discussed using the proposed multi-temporal hierarchical structure.

5.2 Multitemporal MPM inference

In the proposed approach, the posterior marginal $p(x_s|y)$ of the label of each spatio-temporal node s is expressed as a function of the posterior marginal $p(x_s|y)$ of the parent node s^- in the corresponding quad-tree and the posterior marginal $p(x_s|y)$ of the parent node $s^=$ in the quad-tree associated with the previous date to characterize the temporal correlations associated, at different scales, with distinct images in the input time series. The posterior marginal of each spatio-temporal site s can be written as follows:

$$\mathbf{p}(\mathbf{x_s}|\mathbf{y}) = \sum_{x_{s^-}, x_{s^=}} \frac{p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})} \, \mathbf{p}(\mathbf{x_{s^-}}|\mathbf{y}) \mathbf{p}(\mathbf{x_{s^=}}|\mathbf{y}) \tag{5.3}$$

where bold fonts denote the marginal posteriors of interest to the MPM.
Proof of equation (5.3):

$$\begin{aligned}
p(x_s|y) &= \sum_{x_{s^-}, x_{s^=}} p(x_s|x_{s^-}, x_{s^=}, y) p(x_{s^-}, x_{s^=}|y) \\
&= \sum_{x_{s^-}, x_{s^=}} p(x_s|x_{s^-}, x_{s^=}, y_{d(s)}) p(x_{s^-}, x_{s^=}|y) \\
&= \sum_{x_{s^-}, x_{s^=}} \frac{p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})} \, p(x_{s^-}, x_{s^=}|y) \\
&= \sum_{x_{s^-}, x_{s^=}} \frac{p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})} \, p(x_{s^-}|y) p(x_{s^=}|y),
\end{aligned}$$

The equalities across rows 1 and 2, and across rows 3 and 4, derive from two conditional independence assumptions:

A1. The label x_s, given the labels of the parents x_{s^-} and $x_{s^=}$ at the same and the previous dates, depends only on the observations $y_{d(s)}$ of site s and of its descendants and not on the observations of the other sites, i.e.,

$$p(x_s|x_{s^-}, x_{s^=}, y) = p(x_s|x_{s^-}, x_{s^=}, y_{d(s)})$$

A2. Given the observations y, the label of the parent s^- of a site s at the same date is independent on the label of the parent $s^=$ at the previous date, i.e.,

$$p(x_{s^-}, x_{s^=}|y) = p(x_{s^-}|y)p(x_{s^=}|y)$$

These assumptions are analogous to the conditional independence assumptions that are commonly accepted when dealing with (hierarchical or single-scale) MRF-based image analysis, and have been mentioned in the previous chapters. They are used within the proposed method for analytical convenience.

The formulation in (5.3) allows calculating recursively the posterior marginal $p(x_s|y)$ at each spatio-temporal node s while the probabilities $p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})$ are produced. Thus, this simplifies to the determination of the other probabilities based on the following equation:

$$p(x_s, x_{s^-}, x_{s^=}|y_{d(s)}) = p(x_s|x_{s^-}, x_{s^=}).\frac{p(x_{s^-}|x_{s^=})p(x_{s^=})}{p(x_s)}p(x_s|y_{d(s)}) \tag{5.4}$$

In (5.4), the first factor $p(x_s|x_{s^-}, x_{s^=})$ corresponds to the child-parent transition probability; $p(x_s)$ is the prior probability; $p(x_{s^-}|x_{s^=})$ is the temporal transition probability at the same scale; and $p(x_s|y_{d(s)})$ is the partial posterior marginal probability.

Proof of equation (5.4):

$$\begin{aligned}
p(x_s, x_{s^-}, x_{s^=}|y_{d(s)}) &= p(x_{s^-}, x_{s^=}|x_s, y_{d(s)})\ p(x_s|y_{d(s)}) \\
&= p(x_{s^-}, x_{s^=}|x_s)\ p(x_s|y_{d(s)}) \\
&= \frac{p(x_s|x_{s^-}, x_{s^=})\ p(x_{s^-}, x_{s^=})}{p(x_s)}\ p(x_s|y_{d(s)}) \\
&= p(x_s|x_{s^-}, x_{s^=})\frac{p(x_{s^-}|x_{s^=})p(x_{s^=})}{p(x_s)}p(x_s|y_{d(s)})
\end{aligned}$$

where the equality across rows 1 and 2 derives from the following assumption:

A3. The distribution of the labels x_{s^-} and $x_{s^=}$ of the parents of a site s are independent on the observations $y_{d(s)}$ of the descendants of s, when conditioned to the label x_s of s, i.e.,

$$p(x_{s^-}, x_{s=}|x_s, y_{d(s)}) = p(x_{s^-}, x_{s=}|x_s)$$

To compute these probabilities, we benefit from the hierarchical structure defined above and use three recursive passes on the quad-tree, including one *bottom-up* and two *top-down* passes. For the sake of brevity, only the steps associated with a pair of images acquired at two different times ($t = 0$ and $t = 1$) are explained in the following (see Figure 5.2). The recursive extension to more than two acquisition times is straightforward.

5.2.1 Time t = 0: single-time MPM.

According to the cascade approach, first, classification is performed at time t=0 using a single-date MPM as shown in chapter 4, in which the labeling is obtained recursively over scales through a top-down stage and a bottom-up stage. Details of this single-date formulation can be found in [Laferté et al., 2000] and has been recalled in Section 4.2. We only recall that the process is initialized by predefining the pixelwise prior probability distribution at the root of the corresponding quad-tree, i.e., $p(x_s)$, $s \in S_0^R$. This initialization is required to begin a top-down recursion and compute the priors at all levels of the quad-tree at time 0. A simple initialization strategy is to use a uniform prior distribution on Λ. Here, to incorporate spatial contextual information and mitigate possible blocky artifacts, a case-specific initialization strategy is applied that makes use of a spatial MRF model: a neighborhood system is defined on the lattice S_0^R in the root at time 0, and for each pixel $s \in S_0^R$, the unconditional prior $p(x_s)$ is replaced by the local conditional prior $p(x_s|x_{s'}, s' \sim s, s' \in S_0^R)$, where $s \sim s'$ denotes that the sites s and s' are neighbors. This choice generally provides a biased prior-probability estimate but favors spatial adaptivity, a desired property when working with high resolution images in which spatial details are common.
The well-known Potts MRF model, which favors the same labeling in homogeneous image regions, is used [Li, 2009, Kato and Zerubia, 2012], i.e.:

$$p(x_s|x_{s'}, s' \sim s, s' \in S_0^R) \propto exp(\beta \sum_{s \sim s'} \delta(x_s, x_{s'})) \tag{5.5}$$

where β is a positive spatial smoothness parameter. Several methods have been proposed to optimize the value of this parameter including the maximization of the pseudo-likelihood function over the training set [Geman and Geman, 1984]. In [Chardin, 2000] also combined a hierarchical structure and the Potts model, leading to a semi-iterative technique in which the Potts component is used to compute a unique prior distribution for each scale. On the contrary, here, we use the Potts model to define a local characteristic for each node of the root level to maximize spatial adaptivity.

As a result of single-time processing at time $t = 0$, the posterior marginal $p(x_s|y)$ is known for each pixel of the corresponding quad-tree; $p(x_s|y_{d(s)})$, in which $y_{d(s)}$

denotes the collection of the observations of all quad-tree sites that are descendants of site s, is also derived as a by-product ($s \in S_0^n, n = 0, 1, \ldots, R$). Details can be found in [Laferté et al., 2000] and in Section 4.2.

5.2.2 Time t = 1: first top-down pass.

In the proposed method, the recursive top-down/bottom-up formulation used for the single-time case in Chapter 4 is extended to the multi-temporal classification at time $t = 1$. In this case as well, first, the prior distribution on the root lattice, i.e., $p(x_s), s \in S_1^R$, has to be defined to initialize a top-down pass. Following the cascade approach, at time $t = 1$, we take benefit of the inference conducted at time $t = 0$: for each pixel $s \in S_1^R$ on the root lattice at $t = 1$, the unconditional prior $p(x_s)$ is initialized as the posterior marginal $p(x_{\sigma(s)}|y_{d[\sigma(s)]})$, which corresponds to the same pixel $\sigma(s) \in S_0^R$ in the root lattice S_0^R at $t = 0$ (blue arrow labeled with the number one in Figure 5.2) and has been computed as a by-product of the single-date MPM application at time $t = 0$.

After initializing the prior in the root, a top-down pass (blue arrow labeled with the number two in Figure 5.2) is performed for each finer level $n < R$ at time $t = 1$. The prior-probability distribution is derived as a function of the prior-probability distribution at the parent level and of the transition probabilities from the parent to the current level ($s \in S_1^n, n = 0, 1, \ldots, R-1$):

$$p(x_s) = \sum_{x_{s^-}} p(x_s|x_{s^-})\ p(x_{s^-}). \tag{5.6}$$

This derivation favors an identical parent-child labeling and models the statistical interactions between consecutive levels of the quad-tree. We model the transition probability in the form introduced by [Bouman, 1991], i.e., ($s \in S_1^n, n = 0, 1, \ldots, R-1$):

$$p(x_s|x_{s^-}) = \begin{cases} \theta & x_s = x_{s^-} \\ \dfrac{1-\theta}{M-1} & x_s \neq x_{s^-} \end{cases} \tag{5.7}$$

where θ is a parameter ranging in $[\frac{1}{M}, 1]$. As a result of the first top-down pass, the prior distribution $p(x_s)$ is derived for each pixel s ($s \in S_1^n, n = 0, 1, \ldots, R$) of each level of the quad-tree at time $t = 1$.

5.2.3 Time t = 1: bottom-up pass.

A bottom-up pass recursion is then performed to estimate the joint probabilities $p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})$ starting from the leaves of the quad-tree at time $t = 1$ and proceeding until the root is reached based on the factorization in equation (5.4). In addition to priors, which have been computed in the previous top-down pass, three sets of probabilities are required to compute this factorization:

(i) the set of temporal transition probabilities at the same scale $p(x_{s^-}|x_{s^=})$;

(ii) the child-parent transition probability $p(x_s|x_{s^-}, x_{s^=})$;

(iii) the partial posterior marginals $p(x_s|y_{d(s)})$.

Details of the calculation of (i) and (ii) are shown in Section 5.3.2. Concerning (iii), [Laferté et al., 2000] proved the following ($s \in S_1^n, n = 1, 2, \ldots, R$):

$$p(x_s|y_{d(s)}) \propto p(y_s|x_s)p(x_s) \prod_{u \in s^+} \sum_{x_u} \frac{p(x_u|y_{d(u)})}{p(x_u)} p(x_u|x_s), \tag{5.8}$$

Thus, the bottom-up pass is a recursion that estimates $p(x_s|y_{d(s)})$. It starts from the leaves of the quad-tree in which the partial posterior marginals are computed via (green arrow labeled with the number 1 in Figure 5.2):

$$p(x_s|y_s) \propto p(y_s|x_s)p(x_s), \tag{5.9}$$

and then proceeds until the root is reached using equation (5.8) (green arrow labeled with the number two in Figure 5.2). Equation (5.8) involves the pixelwise class-conditional PDFs $p(y_s|x_s)$ of the image data at each node of each quad-tree (see Section 5.3.1). As a result of the bottom-up pass, we now have all needed probabilities to compute $p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})$ at each level of the quad-tree.

5.2.4 Time t = 1: second top-down pass.

According to equation (5.3), first, the posterior marginal is initialized at the root of time $t = 1$ (red arrow labeled with the number one in Figure 5.2). For this purpose, we initialize $p(x_s|y)$ as $p(x_s|y_{d(s)})$ for $s \in S_1^R$, as in the usual single-date formulations of MPM. Then, the posterior $p(x_s|y)$ at each pixel s for all other tree levels at time $t = 1$ ($s \in S_1^n, n = 0, 1, \ldots, R-1$) can be easily computed recursively in a top-down pass (red arrow labeled with the number two in Figure 5.2) using the formulation in equation (5.3).

5.2.5 Both times: combination with MMD.

At each time $t \in \{0, 1\}$, the aforementioned steps lead to the computation of the posterior marginal $p(x_s|y)$ on each pixel ($s \in S_t^n, n = 0, 1, \ldots, R$). In principle, the class label x_s that maximizes $p(x_s|y)$ over the finite set Λ of classes could be selected and assigned to s. This is a feasible procedure but is often avoided in the literature of hierarchical MRFs because of its computational burden (linear with respect to the number of classes and the number of sites for all scales and times) and of possible blocky artifacts. As an alternate approach, here, a case-specific formulation of the modified Metropolis dynamics (MMD) [Kato, 1995] is applied separately for each scale and time. Specifically, in the case of the root layer of the quad-tree corresponding to each time t, MMD is used to minimize the following energy with respect to the label configuration $\mathcal{X}_t^R = \{x_s\}_{s \in S_t^R}$:

$$U(\mathcal{X}_t^R|y) = -\sum_{s \in S_t^R} log\ p(x_s|y) - \beta \sum_{s \sim s' \in S_t^R} \delta(x_s, x_{s'}) \tag{5.10}$$

where the first term is expressed in terms of the pixel-wise posteriors computed by MPM and the second contribution is due to the Potts model on the root of the tree. In the case of each other layer $n = 0, 1, \ldots, R-1$, no Potts model is used and MMD is applied to minimize:

$$U(\mathcal{X}_t^R|y) = -\sum_{s \in S_t^R} log\ p(x_s|y) \tag{5.11}$$

This means that, in this case, MMD is equivalent to iteratively selecting a random subset of pixels for which random replacements in class membership are attempted. In all cases, the iterative procedure of MMD is repeated until the difference in energy on consecutive iterations goes below a predefined threshold (which was set to 10^{-4} in the experiments). In the case of the root layer, the solutions obtained using MMD and maximizing $p(x_s|y)$ directly intrinsically differ because the former takes into account spatial context through the Potts model while the latter does not. In the case of the other layers, MMD basically acts as a randomized version of the maximization of $p(x_s|y)$ on every pixel. Computationally, the number of iterations of MMD that suffices to reach convergence is usually significantly smaller than the number of individual operations leading to the maximization of $p(x_s|y)$ on every pixel with respect to the class label. Accordingly, MMD is expected to be advantageous from a computational viewpoint. This is consistent with various previous works using MPM on hierarchical MRF models (see, e.g., [Pérez et al., 2000] that combines MPM and ICM; and [Voisin et al., 2014] for MPM and MMD).

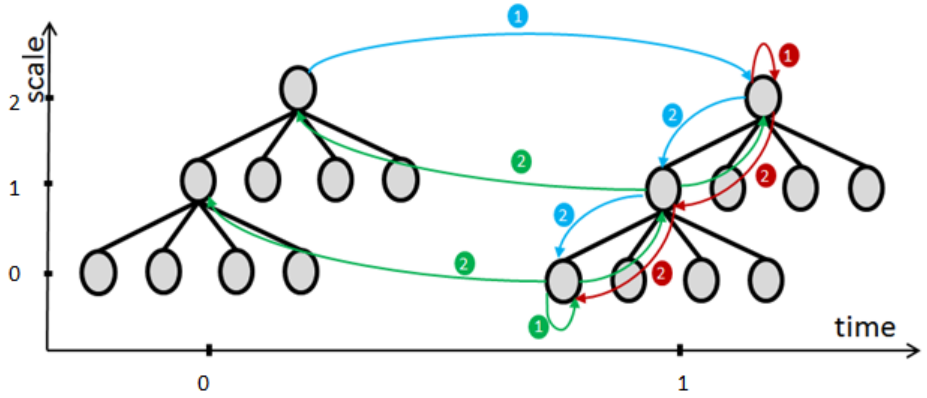


Figure 5.2: Multi-date MPM estimation on the quad-tree: R=2 and two dates.

5.3 Pixelwise class-conditional PDFs and transition Probabilities

5.3.1 Pixelwise class-conditional PDFs

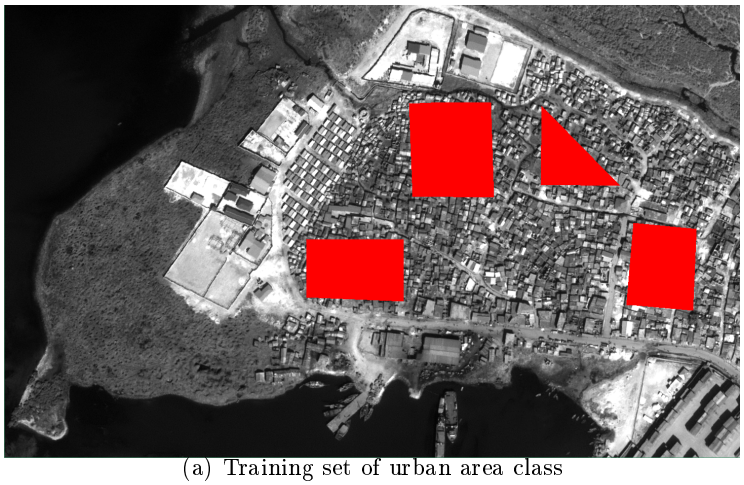
Given a training set for each input date, for each class m, scale n and acquisition time t we model the corresponding class-conditional marginal PDF $p(y_s|x_s = m)$ using finite mixtures of independent distributions:

$$p(y_s|x_s = m) = \sum_{i=1}^{K^{mnt}} \pi_i^{mnt} \mathcal{F}_i^{mnt}(y_s|\theta_i^{mnt}), \qquad \forall s \in S_t^n \tag{5.12}$$

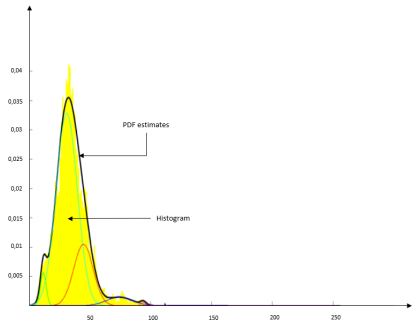
where π_i^{mnt} are the mixing proportions, θ_i^{mnt} is the set of the parameters of the i^{th} PDF mixture component of class m at scale level n and time t, and $\mathcal{F}_i^{mnt}$ is the corresponding parametric family $(n = 0, 1, \ldots, R; m = 0, 1, \ldots, M-1; t = 0, 1)$. When the data at scale level n and time t is an optical image, the class-conditional marginal PDF $p^{mnt}(y|m)$ related to each class m is modeled by a multivariate Gaussian mixture [Figueiredo and Jain, 2002] with a set of parameters $\theta_i^{mnt} = \{\mu_i^{mnt}, \Sigma_i^{mnt}\}$ associated with the corresponding means and covariance matrices. Formally the component i of the multivariate Gaussian mixture is given by (for ease of notation the subscripts m, n, and t are dropped):

$$p_i(y|m) = (2\pi)^{-\frac{d}{2}} |\Sigma_i|^{-\frac{1}{2}} \, exp\,[-\frac{1}{2}(y - \mu_i)^T \Sigma_i^{-1}(y - \mu_i)] \tag{5.13}$$

This Gaussian assumption, especially when combined with a finite mixture is a well-known and widely accepted model for the statistics of optical data. Thanks to the linearity of the wavelet operator, the same assumption also holds for the resulting transformed levels of the quad-tree. The use of finite mixtures instead of single PDFs offers the possibility to consider heterogeneous PDFs, usually reflecting the contributions of different materials present in each class. This class heterogeneity is relevant when we address VHR images. The parameters of the mixture model θ_i^{mnt} in the Gaussian mixture are estimated through the stochastic expectation maximization (SEM) algorithm [Celeux et al., 1996], which is an iterative stochastic parameter estimation algorithm developed for problems characterized by data incompleteness and approaching, under suitable assumptions, maximum likelihood estimates. For each scale and time, SEM is separately applied to the training samples of each class to estimate the related parameters as shown in Figure 5.3 using the algorithm 5.3.1.



(a) Training set of urban area class



(b) Urban area modeling

Figure 5.3: Example of PDF modeling using the SEM algorithm on a panchromatic Pléiades image acquired on Port-au-Prince, Haiti, ©CNES, distribution Airbus DS.

Algorithm 5.3.1 (Stochastic Expectation Maximization (SEM) algorithm)

For each class $m \in \{0, 1, \ldots, M\}$, time $t \in \{0, 1\}$, and scale $n \in \{0, 1, \ldots, R\}$ and at each iteration $j \in [1, j_{max}]$, dropping the subscripts m, n, and t for ease of notation:

E-step:

For each observation y_k, $k \in [1, \mathcal{N}]$ and the i^{th} mixture component, $i \in [1, K]$, posterior probability estimation:

$$\tau_i^j(y_k) = \frac{\pi_i^j \, p(y_k|\mu_i^j, \Sigma_i^j)}{\sum_{l=1}^{K} \pi_l^j \, p(y_k|\mu_l^j, \Sigma_l^j)};$$

S-step:

A label $s^j(y_k)$ is associated to each training y_k of the considered class by using the previous posterior estimation $\{\tau_i^j(y_k) : i \in [1, K]\}$;

M-step:

For each mixture component i, the proportions and the parameters at the iteration $j+1$ are given by:

$$\pi_i^{j+1} = \frac{Q_i^j}{\mathcal{N}}$$

$$\mu_i^{j+1} = \frac{1}{Q_i^j} \sum_{k=1}^{\mathcal{N}} y \tau_i^j(y_k)$$

$$\Sigma_i^{j+1} = \frac{1}{Q_i^j} \sum_{k=1}^{\mathcal{N}} \tau_i^j(y_k)(y - \mu_i^{j+1})(y - \mu_i^{j+1})^T$$

where Q_i^j is the set of samples assigned to the i^{th} mixture component in the S-step.

5.3.2 Transition probabilities

The transition probabilities between consecutive scales and consecutive dates determine the properties of the hierarchical MRF because they formalize the causality of the statistical interactions involved. Therefore, they must be carefully defined. In the proposed method, two types of probabilities involve time:

(i) The First is the set of temporal transition probabilities at the identical scale $p(x_{s^-}|x_{s^=})$, which are estimated using a specific formulation of the expectation-maximization (EM) algorithm [Figueiredo and Jain, 2002]. An iterative fixed-point EM-like algorithm is performed to estimate the prior

joint probabilities $p(x_{s^-}, x_{s^=})$ for each scale n, and the temporal transition probabilities are then derived [Bruzzone et al., 1999]. The probabilities $p(x_{s^-} = m, x_{s^=} = m')$, where m and m' range in $\Lambda = \{0, 1, \ldots, M-1\}$, are regarded as the elements of an $M \times M$ matrix J, which is computed by maximizing the following pseudo-likelihood ($n = 0, 1, \ldots, R$):

$$L(J) = \prod_{s \in S_1^n} \sum_{x_{s^-}} \sum_{x_{s^=}} p(x_{s^-}, x_{s^=}) p(y_{s^-}, y_{s^=} | x_{s^-}, x_{s^=}) \tag{5.14}$$

The recursive equation to be used to maximize the pseudo-likelihood in (5.14) is the following:

$$p_{k+1}(x_{s^-}, x_{s^=}) \propto \sum_{s \in S_1^n} \frac{p_k(x_{s^-}, x_{s^=}) p(y_{s^-} | x_{s^-}) p(y_{s^=} | x_{s^=})}{\sum_{x_{s^-}} \sum_{x_{s^=}} p_k(x_{s^-}, x_{s^=}) p(y_{s^-} | x_{s^-}) p(y_{s^=} | x_{s^=})} \tag{5.15}$$

where $p_k(x_{s^-}, x_{s^=})$ is the iterative joint probability estimate at the k^{th} EM iteration, and N is the total number of pixels. These estimates are initialized by assigning equal probabilities to each pair of classes:

$$p_0(x_{s^-}, x_{s^=}) = \frac{1}{M^2} \tag{5.16}$$

(ii) The second type of transition probabilities that involve time is the child-parent transition probability $p(x_s | x_{s^-}, x_{s^=})$. To our knowledge, a case-specific formulation of EM is not available for inter-scale transition probabilities. However, parametrically modeling these probabilities have demonstrated an effective choice in the case of single-date classification as it yielded accurate results to be obtained [Laferté et al., 2000, Voisin et al., 2012]. Indeed, we extend here the algorithm proposed in [Baum et al., 1970], which favors the identity between the children and parents (in the current and previous dates), all other transitions being unlikely:

$$p(x_s | x_{s^-}) = \begin{cases} \theta & x_s = x_{s^-} = x_{s^=} \\ \varphi & ((x_s = x_{s^-}) \, or \, (x_s = x_{s^-})) \, and \, x_{s^=} \neq x_{s^-} \\ \dfrac{1-\theta}{M-1} & x_s \neq x_{s^-} \, and \quad x_s \neq x_{s^=} \, and \quad x_{s^-} = x_{s^=} \\ \dfrac{1-2\varphi}{M-2} & x_s \neq x_{s^-} and \quad x_s \neq x_{s^=} and \quad x_{s^-} \neq x_{s^=} \end{cases} \tag{5.17}$$

with the parameters $\theta > \frac{1}{M}$ and $\frac{1}{M} < \varphi < 1/2$. Here, θ has the same meaning as in equation (5.6), and the same parameter value is used in both transition probabilities.

5.4 Experimental results

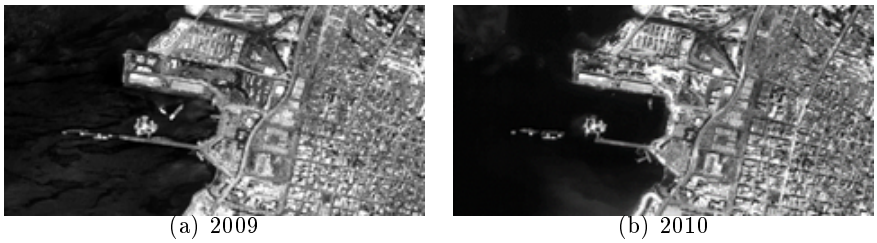
5.4.1 Data sets and experimental setup

In this section, we discuss the results of the experimental validation of the developed multi-temporal and multi-resolution classifier on two datasets (examples are shown in Figures 5.4 and 5.5):

- A three-date series of panchromatic and multispectral Pléiades images acquired over Port-au-Prince (Haiti) in 2011, 2012, and 2013.
- Two pan-sharpened GeoEye acquisitions acquired over Port-au-Prince (Haiti) in 2009 and 2010.

(a) 2011 (b) 2012 (c) 2013

Figure 5.4: Examples of images time series acquired over Port-au-Prince (Haiti) by Pléiades satellite ©CNES, distribution Airbus DS.

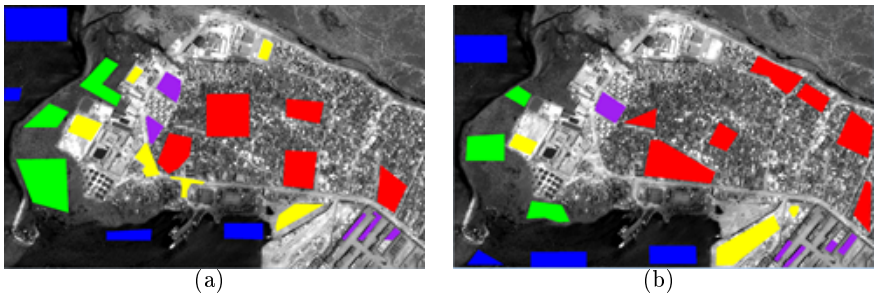


(a) 2009 (b) 2010

Figure 5.5: Examples of images time series acquired over Port-au-Prince (Haiti) by GeoEye satellite ©GeoEye.

Five land cover classes have been considered for both data sets: urban (red), water (blue), vegetation (green), bare soil (yellow), and containers (purple). We note that these classes represent semantically high level land covers. However, a classification map associated with more detailed classes can be produced when a sophisticated ground truth is available. In the present work, manually annotated non-overlapping training and test sets were selected in homogeneous areas (as shown in Figure 5.6 and Table 5.4.1). Spatially disjoint training and test areas were used in all experiments

to minimize correlation between training and test samples and to prevent possible optimistic biases in accuracy assessment.



(a) (b)

Figure 5.6: Ground truth for the Pléiades image acquired in 2013 (a) training set, (b) test set.

class name and color	♯ of pixels in training set	♯ of pixels in test set
Water	49 057	45 790
Urban	75 508	72 327
Vegetation	50 688	27 086
Bare soil	29 333	25 541
Container	16 064	14 652

Table 5.1: Number of training and test samples on the panchromatic pixel lattice of the Pléiades image (1600 x1000 pixels) acquired in 2013

In the case of the Pléiades images, the finest resolution of the multi-resolution pyramid (level 0) was set equal to the finest resolution of the input panchromatic images (i.e., 0.5 m). Co-registered multi-spectral images (at 2 m) were integrated in level 2 of the pyramid. To fill level 1 of each quad-tree, a wavelet decomposition of each panchromatic image was used. As a preliminary experiment, the combination of the proposed method with numerous wavelet operators, including Daubechies, biorthogonal, and reverse biorthogonal wavelets, symlets, and coiflets of various orders [Mallat, 2008], was examined. The results were similar, and the main difference relied on the level of smoothness of the final classification map. On one hand, as shown in Figure 5.7, the average of the overall accuracies obtained on the test sets of all individual dates was remarkably stable as a function of the selection of the wavelet operator, suggesting that this selection is not critical in the application of the proposed approach. On the other hand, an exception was represented by the Daubechies wavelets of order 10 (db10) whose combination with the proposed multi-resolution method resulted in higher accuracies than the other considered wavelet transforms. This wavelet operator will be used in all other experiments discussed in this dissertation (see Figure 5.7).

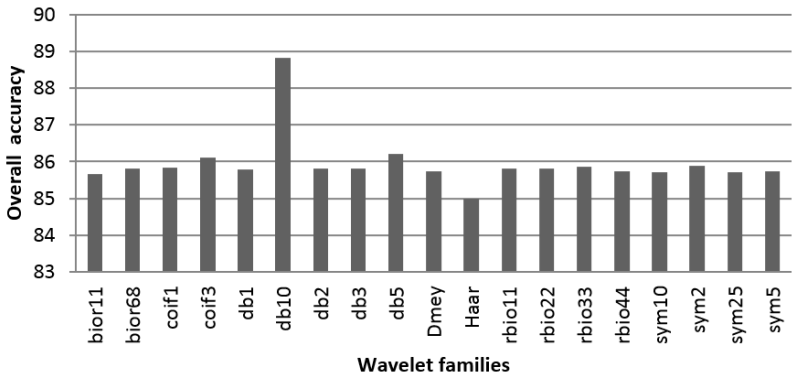
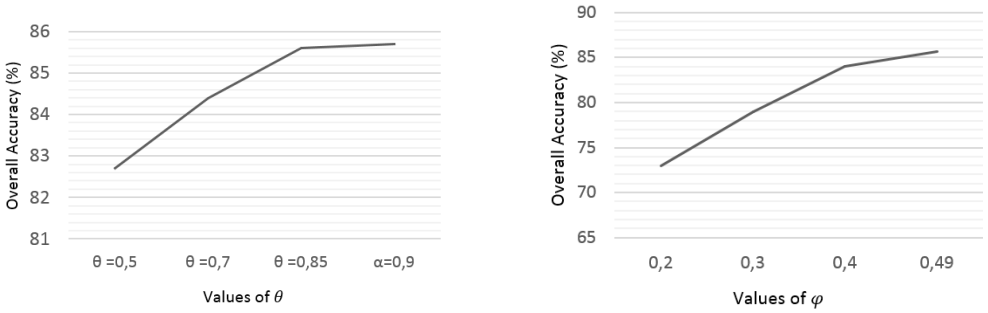


Figure 5.7: Average of the overall accuracies obtained on the test sets of all individual dates results using several wavelet families.

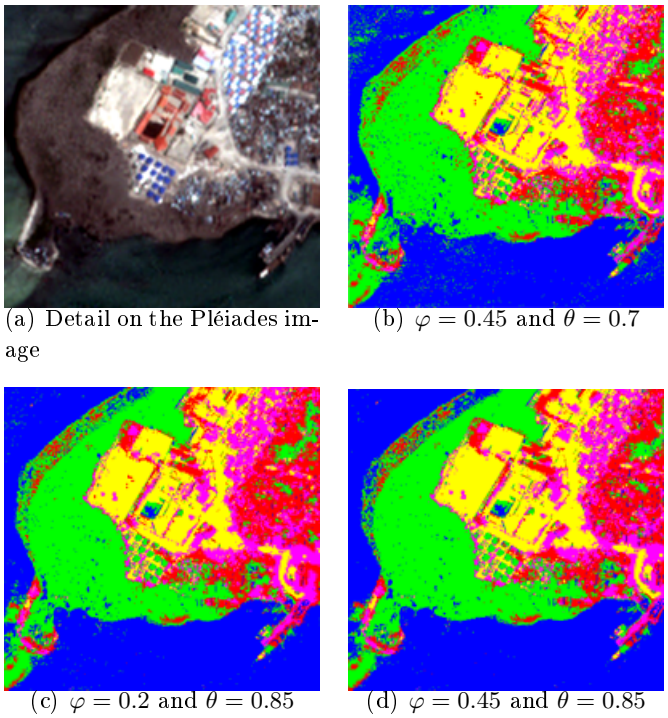
The GeoEye image resolution is 0.5 meter (the finest resolution of the multi-resolution pyramid). Comments similar to those reported for the experiments with Pléiades images regarding the selection of the wavelet operator hold here as well, and db10 was used to fill level 1 and level 2 of the pyramid.

As discussed in Section 5.2, the proposed method depends on four parameters, i.e., β in (B.9), θ in (5.7) and (5.17), φ in (5.17), and R. the classification results included in this manuscript were obtained using the following parameter values: $\beta = 0.8$, $\theta = 0.85$, $\varphi = 0.48$, and $R = 2$, i.e., three levels in each quad-tree. Including more decomposition levels generally helps in discriminating homogeneous land covers but might result in the removal of small-size image details (see Figure 4.4). Indeed, with $R = 2$, the classification map is generated at 50-cm spatial resolution for both data sets, while the coarsest scale corresponds to 2-m resolution. Including coarser-resolution features would generally favor spatial smoothness but may progressively hinder the capability to discriminate classes characterized by spatial details such as "buildings" and "containers." The value of β was automatically optimized by applying the well-known pseudo-likelihood method [Besag, 1974] to the training samples. Accordingly, a user/operator does not have to perform a trial-and-error procedure to set β. In general, a lower value of β prevents spatial over-smoothing at the price of an accuracy decrease on test samples located inside homogeneous regions associated with the same thematic class. As β is automatically optimized, the only free parameters are θ and φ. According to (5.17), θ is the probability that a site, its parent at the same date and its parent at the previous date share the same class label; φ is the probability that a site shares the same label of one of the two parents while the parents disagree. For (5.17) to define a probability distribution, θ and φ can take values in $[0, 1]$ and $[0, 0.5]$, respectively, in the case of $M = 5$ classes. Figures 5.8 and 5.9 show the behavior of the overall accuracy (OA) of the proposed method on the test set as each one of these two parameters range in these intervals while the other parameter is fixed to the aforementioned reference values (i.e., $\theta = 0.85$ and $\varphi = 0.48$).



(a) (b)

Figure 5.8: Overall accuracy of the proposed method on the test set of the Pléiades data set as a function of the parameters (a) φ and (b) θ.



(a) Detail on the Pléiades image (b) $\varphi = 0.45$ and $\theta = 0.7$

(c) $\varphi = 0.2$ and $\theta = 0.85$ (d) $\varphi = 0.45$ and $\theta = 0.85$

Figure 5.9: Details of the the classification maps obtained by the proposed method when applied to the Pléiades data set with different values φ and θ.

On one hand, these plots suggest that the method is sensitive to the values of θ and φ. This is an expected result because they involve the causality of the model. On the other hand, limited sensitivity was observed and OA remained higher than approximately 85% as long as φ and θ were larger than 0.4 and 0.85, respectively. On the contrary, it was basically for relatively extreme and not

meaningful values of θ or φ that poor values of OA were obtained. For example, $\varphi = 0.2$ implies that there is a 15% probability that the labels of a site and of its two parents are all different (see (5.17) with $M = 5$). Because of the very high inter-scale and temporal correlations associated with the multi-resolution data sets, this event is highly unlikely, and $\varphi = 0.2$ yields to significantly overestimating its chances of occurring in the classification process, thus affecting OA. Similarly, $\theta = 0.5$ implies a 12.5% probability that the parents of a site at the two dates agree on a certain class membership but the site disagrees, another outcome that is very unlikely because of inter-scale correlation and whose probability is overestimated. Therefore, although the proposed method is overall sensitive to θ and φ, the experimental analysis suggests that the meaning of the two parameters in relation to inter-scale and temporal correlation allows a user/operator to rather easily determine values of or ranges on these parameters that lead to accurate classification maps. Preliminary experiments also pointed out that the use of the MMD optimization technique resulted in a significant reduction in the number of iterations needed to estimate the class label of each pixel as compared to the direct maximization of the posterior marginals. To reach convergence in the case of one single level of a quad-tree including 1600×1000 pixels and 5 classes, MMD required fewer than 1000 iterations. We recall that, in each MMD iteration, one individual pixel is examined. On the contrary, the direct choice of the class label that maximizes the posterior marginal on each pixel required to iterate over all 1600×1000 pixels and 5 classes, thus taking a much longer time.

The results obtained by the proposed method were compared to those generated by:

(i) the technique in [Laferté et al., 2000], used as a multi-resolution single-date benchmark in both its **(a)** MPM- and **(b)** MAP-based formulations;

(ii) the MRF-based algorithm proposed in [Melgani and Serpico, 2003], used as single-resolution multi-temporal benchmark;

(iii) a contextual combination of the K nearest neighbor and MRF-based approaches, used as a single-resolution single-date benchmark;

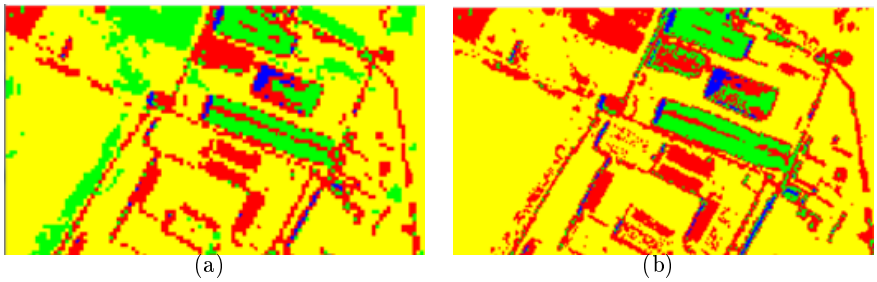
(iv) the well-known K-means clustering technique, used as a basic unsupervised benchmark.

The results of the proposed and previous techniques are reported in the following subsection along with further details on their applications.

5.4.2 Experimental Results and comparisons

In this section, we present the classification maps and discuss the corresponding classification accuracies that were obtained on the test set. Figure 5.11 and Table 5.2 refer to the results obtained using Pléiades images, and Figure 5.12 and

Table 5.3 regard those obtained from the GeoEye acquisitions. All computation times reported in the tables refer to a C++ implementation on an Intel i7 quad-core (2.40 GHz) 8-GB-RAM 64-bit Linux system. The analysis of the classification maps has suggested that the proposed hierarchical method leads to accurate results. In particular, several experimental comparisons were performed with methods exploiting multi- or single-resolution, multi- or single-date, supervised or unsupervised approaches. First, the results of the proposed technique were compared to the separate hierarchical classification results obtained at individual dates using the multi-resolution single-time method in [Laferté et al., 2000], in both its MPM (Figure 5.11(c) and Figure 5.12(c)) and MAP (see Figure 5.11(d) and Figure 5.12(d)) formulations and using a 3 level pyramid with the following parameters: $\beta = 0.8$ and $\theta = 0.85$. We recall that several extensions of the method in [Laferté et al., 2000] have been developed including the approach presented in [Voisin et al., 2012] for the specific case of multi-sensor classification and based on the integration of the hierarchical MRF model of [Laferté et al., 2000] with copula functions for merging data from both optical and SAR sensors within the same pyramid. In the present thesis, the focus is on multi-temporal classification with optical images and not on multi-sensor fusion. Accordingly, we used the original method in [Laferté et al., 2000] for comparison purposes. The results of the comparison show the effectiveness of the proposed multi-temporal hierarchical model in fusing the temporal, spatial, and multi-resolution information associated with the input data (see Table **??**). In practice, the use of one quad-tree structure with the MPM criterion yields "blocky" segmentation (see Figure 5.10 (a)). This phenomenon can be explained by the fact that two neighboring sites at a given scale may not have the same parent. In this case, a boundary appears more easily than when they are linked by a parent node. These blocky artifacts are avoided by the use of the multi-temporal hierarchical structure proposed in this thesis in which causal relationships between parents and offspring in the same quad-tree are relaxed by the introduction of other causal relationships over time and scale (see Figure 5.10(b)). One of the main sources of misclassification in the single-date results is the confusion between the "urban" and "vegetation" classes. This misclassification is reduced in the multi-temporal classification obtained by the proposed method because of the modeling of the temporal relationships among the input multi-resolution data. Furthermore, as expected, the MAP criterion was poorly effective when applied to the considered hierarchical structure because errors were propagated from the root to the leaves and led to severe misclassifications, especially regarding the classes that most strongly overlap in the feature space (e.g., "urban" and "containers"; see Figures 5.11(d) and 5.12(d)).

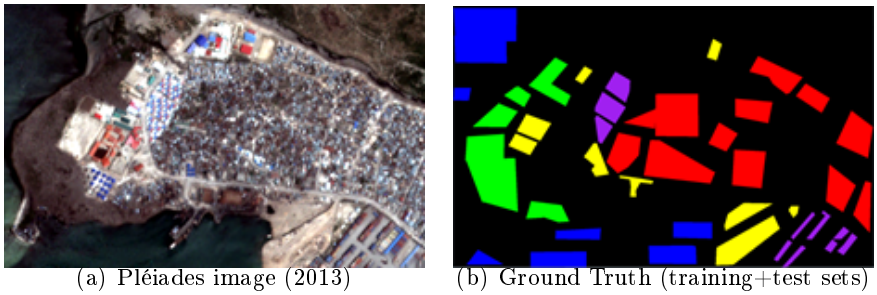


(a) (b)

Figure 5.10: (a) Blocky artifacts obtained using quad-tree structure via MPM formulation, (b) reduction of these blocky artifacts using the proposed method.

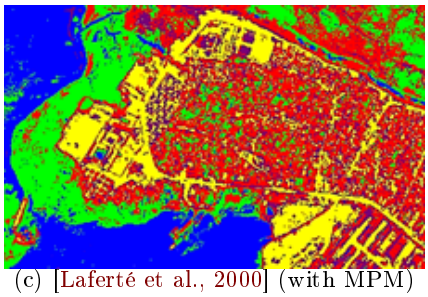
Second, in the context of multi-temporal classification, the proposed classifier was compared to the multi-temporal single-resolution MRF-based method proposed in [Melgani and Serpico, 2003]. It uses the mutual approach and consists in performing a bidirectional exchange of the temporal information between the (non-hierarchical) single-time MRF models associated with consecutive images in the sequence. In the form of an appropriate energy function, each single-time MRF model integrates three types of information (spectral, spatial contextual, and temporal contextual) using a multilayer perceptron (MLP) neural network to extract the spectral information. The results reported in Table 5.3 show that a better exploitation of the spatio-temporal information allowed the proposed cascade multi-resolution approach to provide more accurate results than the previous mutual single-resolution approach in [Melgani and Serpico, 2003]. More generally, the mutual approach reduces the risk of propagating the classification error between consecutive dates, while the use of the hierarchical schema provided more accurate classification maps, at least, on the considered data sets. Furthermore, because of the hierarchical aspects and the non-iterative algorithm, only few minutes were necessary to obtain satisfactory results using the proposed approach compared to those obtained by the mutual approach that required a much longer computation time (several hours). According to the formulation of the method in [Melgani and Serpico, 2003], this time included the times required to compute the texture features from the given image time series, to train and apply an MLP neural network for the image of each date using the back-propagation algorithm, and to estimate the parameters of the corresponding MRF model using the case-specific parameter optimization procedure in [Ibáñez and Simó, 2003]. The classification maps obtained using the well-known K-nearest-neighbors (K-NN) method are also shown in Figures 5.11(h) and 5.12(h). K-NN was used as a benchmark non-parametric classifier. It is non-contextual, so to perform a fair comparison between the proposed method and a spatial-contextual technique, it was combined with an MRF model. A hidden MRF whose unary term was expressed in terms of the pixelwise posterior probabilities estimated by K-NN and whose contextual term was represented by an isotropic Potts model was used. K = 30 was estimated by cross validation on the training set, and the smooth-

ing parameter of the Potts model was optimized using the automatic method in [Serpico and Moser, 2006], which is based on the Ho-Kashyap algorithm. The numerical results on the test sets suggest that this single-scale MRF-based method (see Tables 5.2 and 5.3) leads to rather poor accuracy and severe spatial oversmoothing as shown in Figures 5.11(g) and 5.12(g). This is consistent with the fact that this combined K-NN + MRF classifier is intrinsically single-resolution and single-date, and can exploit neither the multi-resolution nor the multi-temporal structure of the input data set. In the map in Figure 5.12(g), obtained from GeoEye data, the combined K-NN + MRF well discriminated the "water" and "urban" classes but almost did not identify the other thematic classes due to the strong spectral overlapping and the imbalance between the training sample sizes of these classes. Finally, a further comparison was performed between the results of the proposed method and those of an unsupervised algorithm. K-means was used for this benchmark comparison as a well-known consolidated approach, and was applied with $K = 5$. This number of clusters was used to match the number of classes in each data set. The clusters obtained by K-means generally do not coincide with the thematic classes of a supervised classification problem. An alternate strategy could be to, first, apply K-means using a significantly larger number of clusters, and then, perform a cluster-to-class assignment either manually or on the basis of the training set. In either case, this assignment would incorporate prior knowledge. This experiment was meant as a benchmark comparison with an unsupervised method using no prior knowledge. Accordingly, the simple choice $K = 5$ was accepted. As expected due to its unsupervised, non-contextual, and single-resolution formulation, K-means performed the worst in terms of classification accuracy, while it exhibited the lowest computation time (see Figures 5.11(h), 5.12(h)).

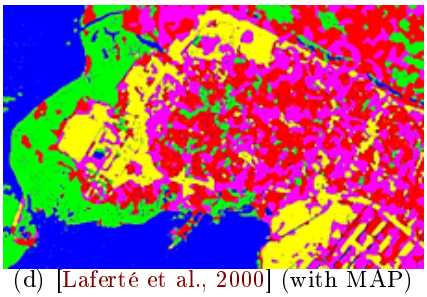


(a) Pléiades image (2013)

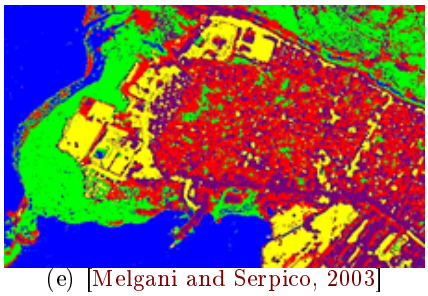
(b) Ground Truth (training+test sets)



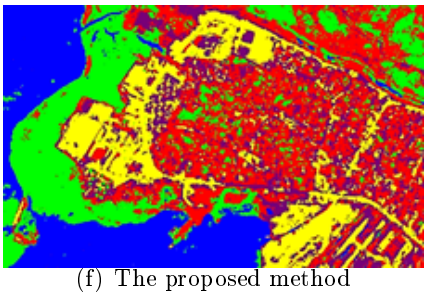
(c) [Laferté et al., 2000] (with MPM)



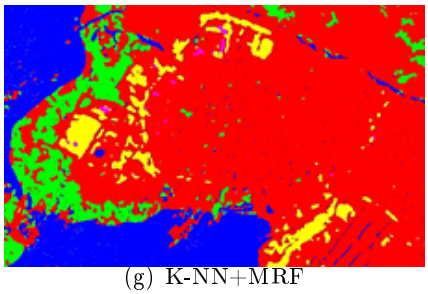
(d) [Laferté et al., 2000] (with MAP)



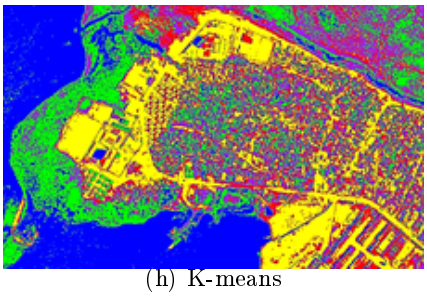
(e) [Melgani and Serpico, 2003]



(f) The proposed method

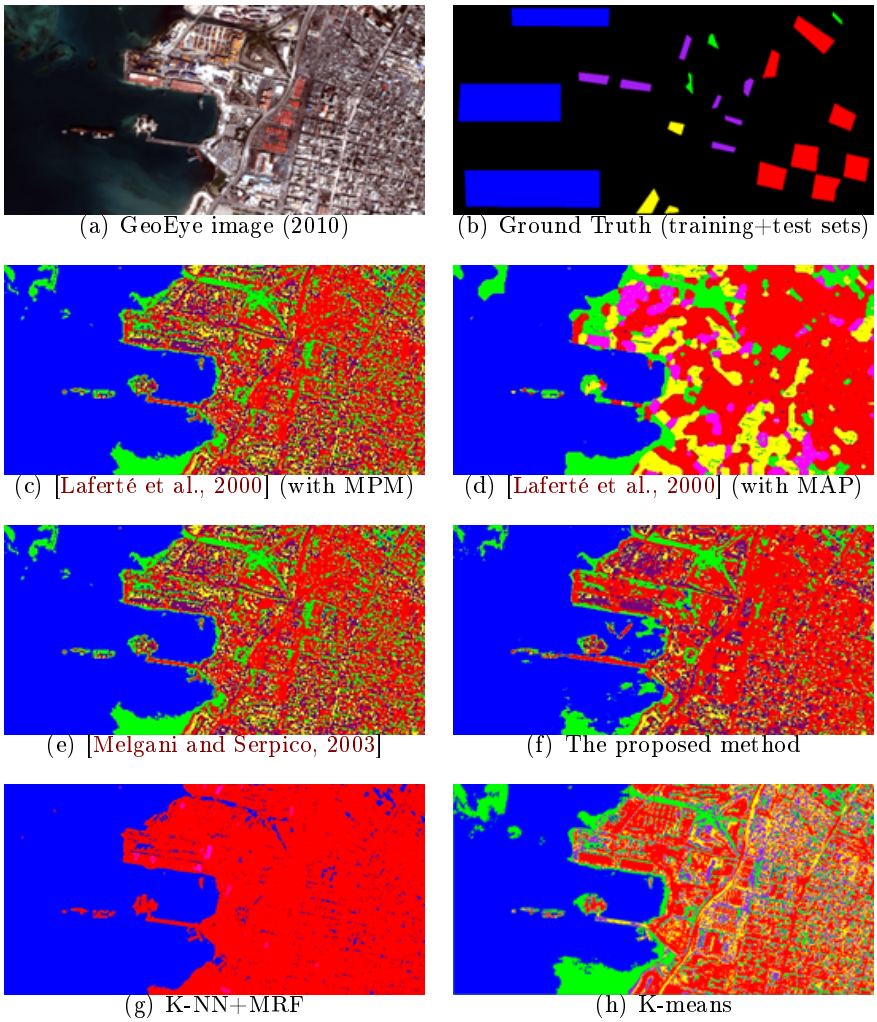


(g) K-NN+MRF



(h) K-means

Figure 5.11: Classification maps obtained from Pléiades data set, (©CNES distribution Airbus DS).



(a) GeoEye image (2010) (b) Ground Truth (training+test sets)

(c) [Laferté et al., 2000] (with MPM) (d) [Laferté et al., 2000] (with MAP)

(e) [Melgani and Serpico, 2003] (f) The proposed method

(g) K-NN+MRF (h) K-means

Figure 5.12: Classification maps obtained from GeoEye data set (©GeoEye).

	urban	water	vegetation	bare soil	containers	over all	computation time
Proposed method	81.62	100	90.69	92.82	62.82	85,59	480 seconds
method in [Laferté et al., 2000] using MPM criterion	77.45	88.62	72.59	86.02	57.02	76.34	160 seconds
method in [Laferté et al., 2000] using MAP criterion	56.14	100	81.90	87.02	73.21	79.65	220 seconds
method in [Melgani and Serpico, 2003]	80.63	100	86.33	87.61	69.61	84.83	$\approx$ 1 hour
K-NN + MRF	96.84	92.42	47.15	71.83	16.75	64.99	90 seconds
K-means	12.37	98.63	59.18	91.66	29.42	58.25	20 seconds

Table 5.2: Classification accuracies on the test set of the Pléiades dataset: class accuracies, overall accuracy, and computation time. Experiments were conducted using one (1600x1000) image at level 0, one (800x500) image at level 1 and four (400x250) bands at level 2 on an Intel i7 quad-core (2.40 GHz) 8-GB-RAM 64-bit Linux system.

	urban	water	vegetation	bare soil	containers	over all	computation time
Proposed method	87.59	100	98.12	72.82	82.27	88,16	345 seconds
method in [Laferté et al., 2000] using MPM criterion	77.45	100	88.34	66.22	67.87	79.97	160 seconds
method in [Laferté et al., 2000] using MAP criterion	64.52	100	92.15	85.62	49.47	78.35	140 seconds
method in [Melgani and Serpico, 2003]	80.63	100	89.79	70.54	74.29	83,05	$\approx$ 1 hour
K-NN + MRF	100	100	0	0	12.28	42.45	40 seconds
K-means	88.97	100	88.14	45.6	36.96	71.93	15 seconds

Table 5.3: Classification accuracies on the test set of the GeoEye dataset: class accuracies, overall accuracy, and computation time. Experiments were conducted using one (1600x800) image at level 0, one (800x400) image at level 1 and one (400x200) bands at level 2 on an Intel i7 quad-core (2.40 GHz) 8-GB-RAM 64-bit Linux system.

In the proposed method, multi-date and multi-resolution fusion is based on explicit statistical modeling. The method combines a joint statistical model of the considered input optical images through hierarchical Markov random field modeling, leading to a statistical supervised classification approach. We have developed a novel MPM-based hierarchical Markov random field model that considers multi-temporal information and, thus, supports the joint supervised classification of multiple images taken over the same area at different times and different spatial resolutions. We analyzed the results obtained with the proposed method through experiments with multi-temporal Pléiades and GeoEye data sets. The experimental results show that the method is able to provide accurate classification maps. The proposed algorithm was compared to a previous single-date multi-resolution method and a previous multi-date single-resolution one, both based on MRF models associated with suitable (hierarchical or single-scale) pixel lattices, and a couple of well-known classifiers including a contextual combination of the K nearest neighbor and MRF-based approaches, used as a single-resolution single-date benchmark; and the K-means clustering technique, used as a basic unsupervised benchmark. The proposed technique has been advantageous in terms of the classification accuracy on the test set, the spatial regularity of the classification maps, the minimization of spatial artifacts, and the tradeoff with respect to computation time. These results show the effectiveness of the algorithm in fusing both multi-temporal and multi-resolution information for supervised classification purposes and confirm that MRF models represent a powerful fusion tool in remote sensing. The computational advantages of the hierarchical MRFs, for which exact recursive formulations of the MPM decision rule are feasible with no need for time-expensive Metropolis or Gibbs sampling procedures, has also been confirmed by the experimental results of the proposed method and those of the benchmark single-resolution or multi-resolution classifiers. The proposed method is based on an MRF model on a case-specific topology that comprises multiple hierarchical quad-trees, each associated with an acquisition date. Wavelet transforms are used to fill in those levels of each quad-tree that are associated with input remote sensing imagery. The selection of the wavelet operator among a large family of possible transforms was not critical because most transforms lead to classification results with similar accuracies. Nevertheless, Daubechies wavelets of order 10 yielded higher accuracies than the other considered transforms. One main advantage of the proposed classifier is that it can be extended to be used for optical data, synthetic aperture radar (e.g., COSMO-SkyMed, or RADARSAT-2) or multi-sensor data. The extension to the multi-sensor case will be discussed in the next chapter.

Chapter 6

Multi-sensor hierarchical Markov model on quad-tree

Contents

6.1 Multi-sensor PDF modeling

As discussed in Chapter 1, the availability of different kind of sensors is very advantageous for land cover mapping applications. It allows us to capture a wide variety of properties of the objects contained in a scene as measured by each sensor. These properties can be exploited to extract richer information about the imaged area. In particular, as previously discussed, the opportunity of joint availability of SAR and optical images can possibly offer high resolution, all-weather, day/night, short revisit time data, polarimetric, and, multi-frequency acquisition capabilities. This potential is especially emphasized by current and forthcoming satellite missions for EO; e.g., Sentinel, Pléiades, GeoEye, COSMO-SkyMed (CSK), RADARSAT-2 (RS2) and TerraSAR-X, which convey a huge potential for multi-sensor optical and SAR observation. They allow a spatially distributed and temporally repetitive view of the monitored area at the multiple spatial scales. However, the use of multi-source image analysis for land cover classification purposes has been mostly addressed so far by focusing on single-resolution multi-sensor optical-SAR imagery whereas the joint use of multi-sensor and multi-resolution capabilities has been more scarcely investigated. A general discussion of the classical literature on multi-sensor optical-SAR image classification has been presented in Section 2.2. This approach bears the obvious advantage of simplicity but is, in general, suboptimal. From a methodological viewpoint, when multi-sensor (optical and SAR) or multi-resolution images of a given scene are available, using them separately discards part of the correlations among these multiple data sources and, most importantly, their complementarity. As illustrated in Figure 6.1, SAR and multi-spectral images exhibit complementary

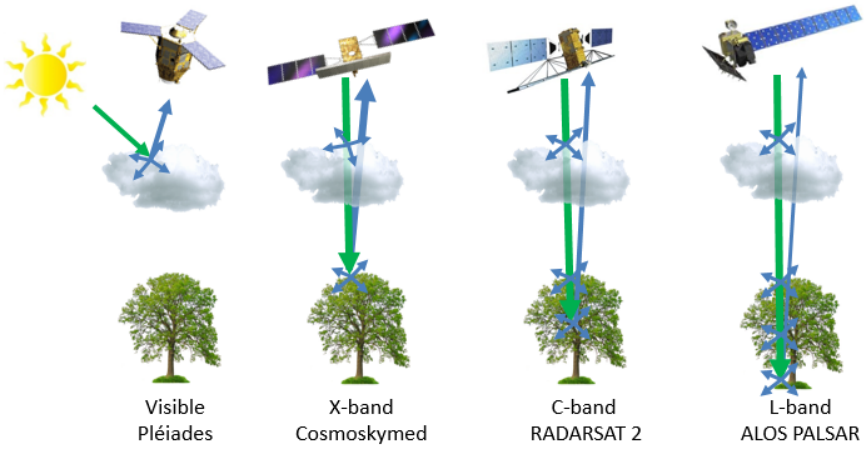


Figure 6.1: Sensitivity to cloud cover and object size using different wavelength range.

properties in terms of wavelength range (active microwave vs. passive visible and infrared), noisy behavior (often strong in SAR due to speckle, usually less critical in

optical imagery), feasibility of photo-interpretation (usually easier with optical than with SAR data), impact of atmospheric conditions and cloud cover (strong for optical acquisitions, and almost negligible for SAR) and sensitivity to Sun-illumination (strong for optical imagery and negligible for SAR). This makes the joint use of HR optical and SAR imagery particularly interesting for many applications related to environmental monitoring and risk management.
Within this framework, there is a definite need for classification methods that automatically correlate different sets of images taken on the same area from different sensors and at different resolutions. One way to address this problem is to resort to an explicit statistical modeling by finding a joint probability distribution given the class-conditional marginal PDFs of the data collected by each sensor (see Figure 6.2). The joint statistics can be designed by resorting to meta-Gaussian distributions [Storvik et al., 2009], multivariate statistics such as multivariate copulas [Voisin et al., 2014], or non-parametric density estimators [Fukunaga, 2013]. However, employing heterogeneous data (SAR-optical in our case) makes the task of finding an appropriate multivariate statistical model complex, time demanding, and possibly prone to over-fitting.

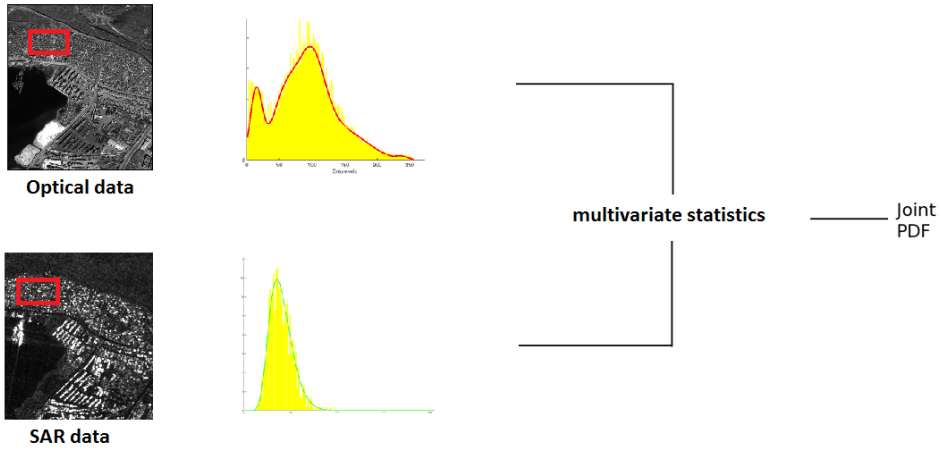


Figure 6.2: Multivariate statistical modeling.

In this chapter, two methods are proposed for the joint supervised classification of multi-sensor images including SAR and optical components, acquired at multiple spatial resolutions. The rationale of both approaches is to take benefit of the data fusion capabilities of hierarchical MRFs and to avoid the computation of joint statistics. A novel approach, based on multiple quad-trees in cascade, applied to multi-sensor and multi-resolution fusion is described. In the first proposed method, for each sensor, the input image is associated with a separate quad-tree structure on the basis of its resolution. The goal is to jointly classify a SAR and an optical image acquired over the same area. The proposed approach formalizes a supervised Bayesian classifier within this multiple quad-tree topology that combines a class-

conditional statistical model for pixel-wise information and a hierarchical MRF for multi-sensor and multi-resolution contextual information.
The second proposed method focuses on the specific case of the fusion of multi-frequency SAR data collected by the COSMO-SkyMed (X band) and RADARSAT-2 (C band) sensors together with optical Pléiades data. A multiple quad-tree structure is used again but optical and SAR data are jointly included in each quad-tree to take benefit of the finest resolution available from each sensor. Compared to the first proposed method, which considers the fusion of data from generally arbitrary SAR and optical sensors, this second method focuses on a specific combination of space-borne SAR and optical sensors to investigate the synergy among the multi-frequency and multi-resolution information they provide.
Given a training set for each input data, for each class m, scale n and sensor, we model the corresponding class-conditional marginal using finite mixtures of independent gray level distributions as detailed in Section 5.3.1. The mixture modeling is performed depending on the different types of remote sensing imagery used in the study. Indeed, when the input data at the n^{th} scale level is an optical image, class-conditional marginal PDF related to each class m can be modeled by a Gaussian mixture model. On the opposite, SAR acquisitions are known to be affected by speckle. For this reason, we use appropriate SAR specific models for such images, such as the generalized Gamma distribution [Li et al., 2011]:

$$p(y|\theta_{mn}) = p(y|\nu_{mn}, \sigma_{mn}, \kappa_{mn}) = \frac{\nu_{mn}}{\sigma_{mn}\Gamma(\kappa_{mn})} \left(\frac{y}{\sigma_{mn}}\right)^{\kappa_{mn}\nu_{mn}-1} exp\left\{-\left(\frac{y}{\sigma_{mn}}\right)^{\nu_{mn}}, \right\} \quad (6.1)$$

where ν_{mn}, σ_{mn} and κ_{mn} are the related parameters and $\Gamma(.)$ is the Gamma function. A wide variety of PDF families can be deduced from (6.1) by varying the parameters. This includes: Rayleigh ($\nu_{mn} = 2$, $\kappa_{mn} = 1$), Nakagami ($\nu_{mn} = 1$), log-normal ($\kappa_{mn} \longmapsto \infty$) and ($\kappa_{mn} = 1$). The evolution of the probability density of the generalized Gamma distribution as function of each individual variable, the two other being fixed is given in Figure 6.3.

The parameters of the mixture model for both SAR and optical images are estimated through the SEM algorithm (see Algorithm 5.3.1).

6.2 The first proposed multi-sensor hierarchical model

6.2.1 methodology

In this section, the first proposed method for SAR-optical multi-resolution classification based on a hierarchical Markovian model is presented. Let an optical image and a SAR image, acquired at the same time over the same area, be given, together with a training map characterizing M thematic classes. The multi-temporal structure developed in the previous chapter (see Section 5.1) is adapted in order to handle the heterogeneous data acquired by different sensors with different physical properties. The novelty in the proposed approach is the use of multiple quad-trees in cascade

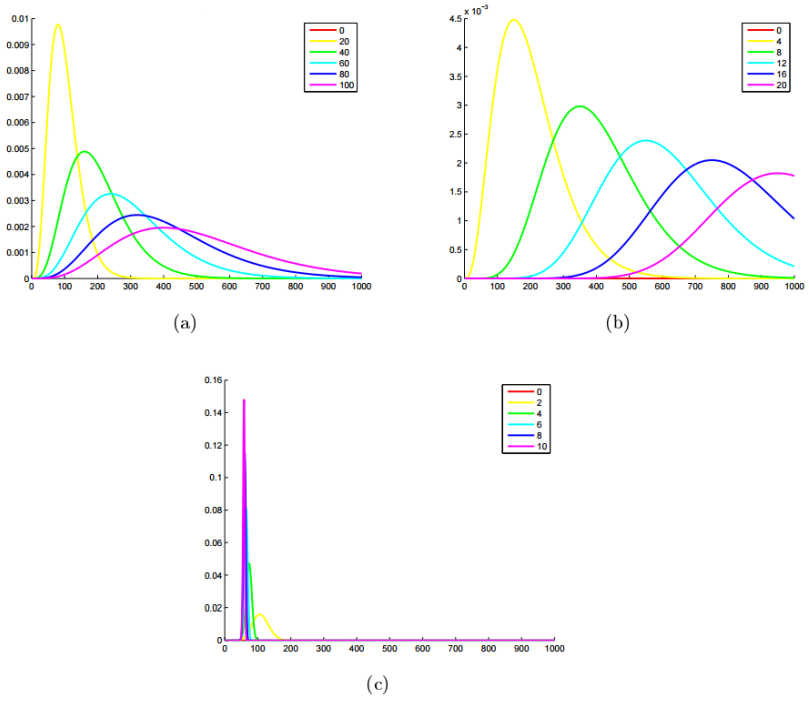


Figure 6.3: The evolution of the probability density of the generalized Gamma distribution as a function of σ_{mn} (a), ν_{mn} (b), and κ_{mn} (c).

(see Figure 6.4), each associated with the set of images given by one specific sensor. In this respect, this method takes benefit of the capability of a multi-quad-tree structure to fuse the information associated with correlated data sources, a capability that has been exploited for multi-temporal image classification in the previous chapter and is investigated in the multi-sensor case here. This approach also aims at exploiting the multi-scale information that is typically associated with either SAR or optical HR imagery by using a hierarchical MRF.
It is worth recalling that, here as in the case of multi-temporal images, the proposed hierarchical structure implies a constraint among the resolutions of the input multi-sensor images at the various levels of the pyramid. The choice of quad-tree topology imposes a factor of 2 between the spatial resolutions of the images at two successive levels. The "empty" levels are again filled in through a wavelet decomposition of the images associated with the finer levels.
The quad-tree structure allows in a very natural way the use of an explicit statistical modeling through a hierarchical Markov random field formulation using a series of random fields at varying scales, resolutions, and sensors, on the basis of the transitions defined on the quad-trees as shown in Figure 6.4.

Accordingly, the formulation of MPM developed in Section 5.2 is used through multiple quad-trees in cascade, while here each quad-tree is biunivocally associated

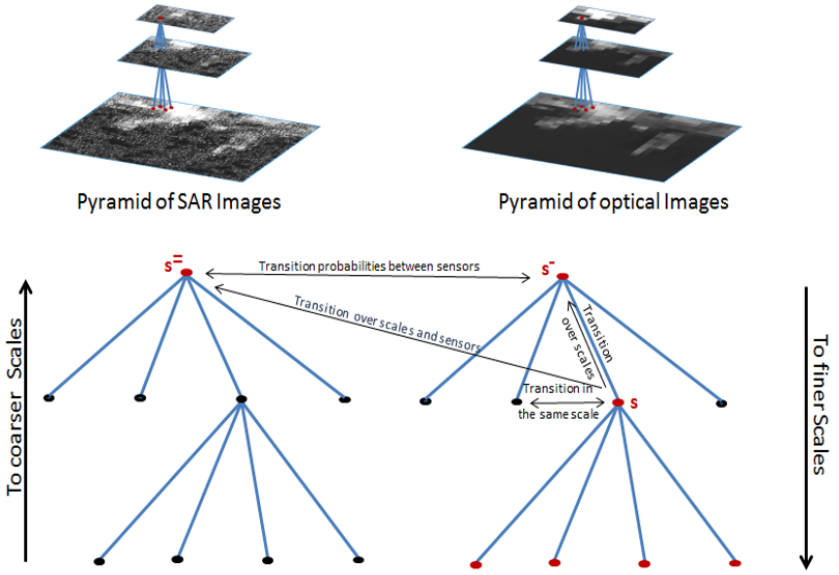


Figure 6.4: Multi-sensor hierarchical structure.

with the data acquired by one individual sensor. Specifically, the posterior marginal $p(x_s|y)$ of the label of each site s in the quad-tree related to the optical sensor is expressed as a function not only of the posterior marginal of the parent node $p(x_{s^-}|y)$ in the same quad-tree but also of the posterior marginal of the parent node in the quad-tree associated with SAR images $p(x_{s^=}|y)$ as shown in (6.2), with the aim to characterize the SAR-optical correlations associated, at different scales, with distinct images in the input multi-source data i.e.:

$$\mathbf{p}(\mathbf{x_s}|\mathbf{y}) = \sum_{x_{s^-}, x_{s^=}} \frac{p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})} \mathbf{p}(\mathbf{x_{s^-}}|\mathbf{y})\mathbf{p}(\mathbf{x_{s^=}}|\mathbf{y}) \tag{6.2}$$

To compute these probabilities, we take benefit from the hierarchical structure defined above and we use two recursive passes on the quad-tree, referred to as "bottom-up" and "top-down" passes as substantiated in Algorithm 6.2.1.

Algorithm 6.2.1 (Multi-sensor hierarchical technique)

Quad-tree associated with SAR images

The classification is done at the first quad-tree composed by the SAR image and of the corresponding wavelet transforms, using the classical MPM using the Algorithm 4.2.1

Quad-tree associated with optical images

↻ **Initialization,** $(\mathbf{s} \in \mathbf{S_{opt}^{R}})$**:**

To estimate the prior in the root given the resulting classification map, we use a spatial Markovian model which takes into account the contextual information given by SAR images (as detailed in the case of multi-temporal, see Section 5.2.1) in the estimation of the prior probability distribution associated with the quad-tree of the optical images.

⇓ **Top-down pass I,** $(\mathbf{s} \in \mathbf{S_{opt}^{n}}, \mathbf{n = 0, 1, \ldots, R-1})$**:**

$$p(x_s) = \sum_{x_{s^-}} p(x_s|x_{s^-})\ p(x_{s^-})$$

↻ **Initialization,** $(\mathbf{s} \in \mathbf{S_{opt}^{0}})$**:** $p(x_s|y_{d(s)}) \propto p(y_s|x_s)p(x_s)$

⇑ **Bottom-up pass,** $(\mathbf{s} \in \mathbf{S_{opt}^{n}}, \mathbf{n = 1, \ldots, R})$**:**

$$p(x_s|y_{d(s)}) \propto p(y_s|x_s)p(x_s) \prod_{u \in s^+} \sum_{x_u} \frac{p(x_u|y_{d(u)})}{p(x_u)} p(x_u|x_s)$$

↻ **Initialization,** $(\mathbf{s} \in \mathbf{S_{opt}^{R}})$**:** $p(x_s|y) = p(x_s|y_{d(s)})$

⇓ **Top-down pass II,** $(\mathbf{s} \in \mathbf{S_{opt}^{n}}, \mathbf{n = 0, 1, \ldots, R-1})$**:**

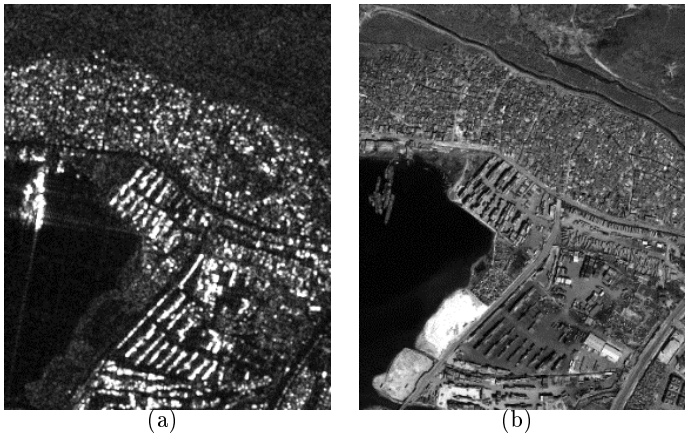
$$p(x_s|y) = \sum_{x_{s^-}, x_{s^=}} \frac{p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})}\ p(x_{s^-}|y) p(x_{s^=}|y)$$

6.2.2 Experimental results

Experiments are discussed regarding two datasets collected over Port-au-Prince (Haiti) including:

(i) A single polarization single-look CSK image, 2010 (©ASI; Figure 6.5(a)) with HH polarization, stripmap acquisition modality, and 2.5-m pixel spacing;

(ii) a coregistered GeoEye RGB acquisition, 2010 (©GeoEye; Figure 6.5(a)(b)).



(a) (b)

Figure 6.5: (a) SAR image (©ASI), (b) one channel from the optical image (©GeoEye)

Multi-scale features are extracted through 2D discrete wavelet transform. As in the case of multi-temporal image classification, preliminary experiments suggested Daubechies wavelets of order 10 as an appropriate choice. Five land cover classes have been considered: urban (red), water (blue), vegetation (green), soil (yellow) and containers (purple).

We compare the proposed method with the classification results obtained using the following techniques:

(1) Separate hierarchical classifications of the images provided by the two sensors (Figure 6.6(b) for optical and Figure 6.6(c) for SAR). In this case, the classification is obtained recursively over the scales by combining the approach in [Laferté et al., 2000] with the same PDF estimation algorithms described in Section 5.3.1 for optical and SAR data. We recall that the use of quad-tree structure in MPM scheme may yield "blockly" segmentations (details can be found in [Laferté et al., 2000] and Section 4.2).

(2) The multi-sensor multiscale method proposed in [Voisin et al., 2014], in which a model for the multivariate joint class-conditional statistics of the co-registered input images at each resolution is designed by resorting to multivariate copulas. The estimated joint probability density function is plugged into a hierarchical Markovian model based on a quad-tree structure (see Figure 6.6(d)).

(3) The multisensor single-scale approach in [Storvik et al., 2009] (see Figure 6.6(e)), in which the likelihood term is constructed by merging generalized Gamma (for SAR) and Gaussian (for optical) marginals into a meta-Gaussian distribution. The classification is obtained by the maximum likelihood rule.

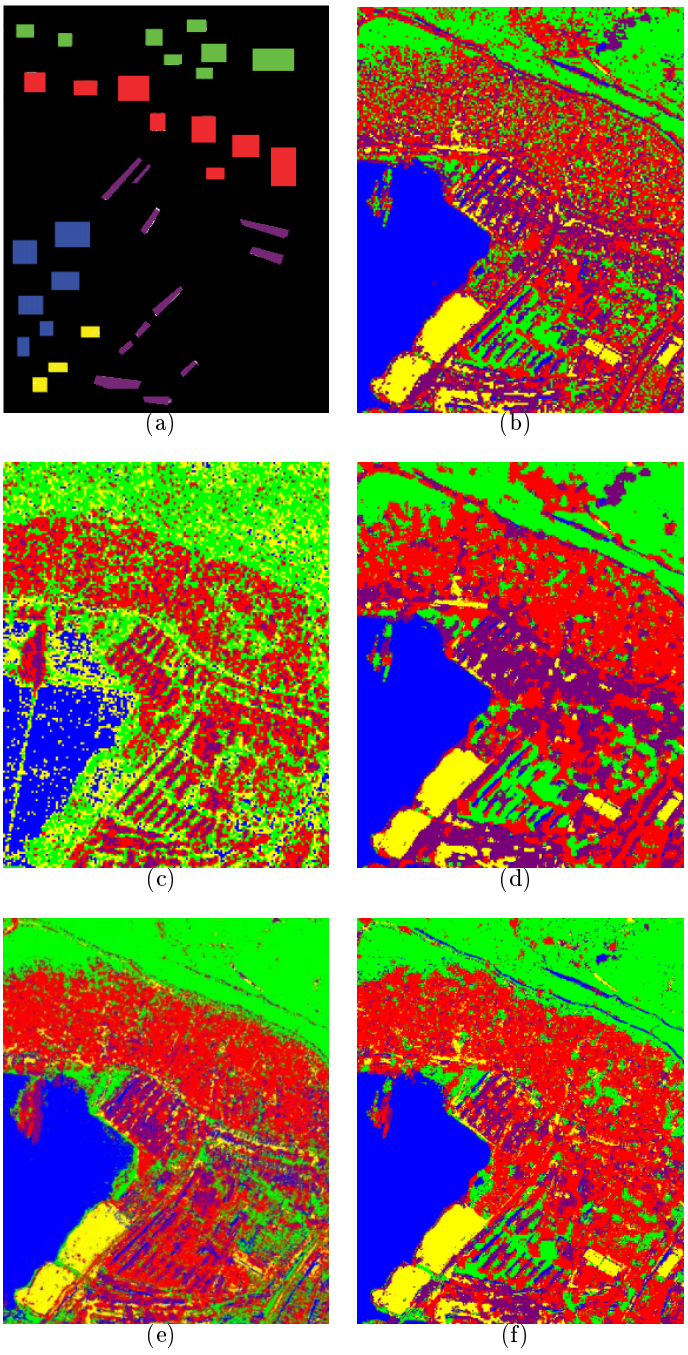


Figure 6.6: (a) ground truth map (including training and test areas), (b) hierarchical MRF-based classification obtained for the optical image, using the method in [Laferté et al., 2000], (c) hierarchical MRF-based classification obtained for the SAR image, using method in [Laferté et al., 2000] with generalized Gamma class-conditional PDFs, (d) result of the multi-sensor and multi-resolution method in [Voisin et al., 2014], (e) result of the multi-sensor method in [Storvik et al., 2009], (f) hierarchical MRF-based classification obtained by the proposed multi-sensor and multi-resolution method.

	water	urban	vegetation	bare soil	containers	over all	computation time
Proposed method	100	78.12	89.46	98.78	47.12	82.69	254 seconds
method in [Storvik et al., 2009]	99.95	97.32	90.81	96.22	37.25	79.44	298 seconds
method in [Voisin et al., 2014]	100	75.24	87.16	98.89	49.31	82.12	668 seconds

Table 6.1: Classification accuracies for the Port-au-Prince dataset. Experiments were conducted on an Intel i7 quad-core (2.40 GHz) 8-GB-RAM 64-bit Linux system.

The resulting classification maps show that the proposed hierarchical method leads to accurate results, especially as compared to separate hierarchical classifications of the images provided by the two sensors (see Figures 6.6(b),(c)). Indeed, experimental results obtained by using only SAR data accurately detect roads and containers, while the results generated by using only optical data better discriminate classes that are spatially homogenous. The proposed method effectively takes benefit from both SAR and optical imagery, and allows generating a classification result that visually well discriminates all classes in the considered HR data set. It is interesting to notice that the detection of the "containers" class, which is generally very overlapping with other urban or built-up classes in the feature space, improves significantly compared to the method in [Voisin et al., 2014], in which the classification map is over smoothed (see Figure 6.6(d)). Compared to the previous multi-sensor classification method in [Storvik et al., 2009] (see Figure 6.6(e) and Table 6.1), which is based on transformations of the input features to a common jointly Gaussian domain, the proposed algorithm provides a spatially more regular classification result, thanks to contextual MRF modeling and wavelet feature extraction. There is an improvement using the proposed method compared to the hierarchical multi-sensor method in [Voisin et al., 2014] (see Figure 6.6(d)) and Table 6.1) in terms of accuracy and computation time. For this method, the main source of misclassification is the container area, where asphalt is erroneously classified as vegetation. Moreover, with the proposed method, we avoid the computation of joint PDFs, while copulas are used in [Storvik et al., 2009] and meta-Gaussian distributions are used in [Voisin et al., 2014]. This choice results in reducing the computation time (see Table 6.1).

6.3 The second proposed multi-sensor hierarchical model

6.3.1 Methodology

The second proposed method addresses a specific problem of multi-sensor optical-SAR fusion focused on exploiting the synergy between current space missions for EO carrying HR optical and SAR payloads. The joint use of COSMO-SkyMed, RADARSAT-2, and Pléiades images is investigated with the aim of taking benefit from the related multi-frequency information (radar X band, radar C band, optical VNIR band). The interest of this specific study is further enforced by the role of these sensors in the area of HR remote sensing. All the three sensors support multiple spatial resolutions up to 0.5 m for Pléiades, approximately 1 m for COSMO-SkyMed (in its spotlight mode), and around 1.5 m of pixel spacing for RADARSAT-2 (in its ultrafine mode). To both benefit from the multi-sensor and multi-frequency information they convey and ensure the capability of mapping land cover at the finest of the available spatial resolutions, a case-specific hierarchical MRF based on a double quad-tree is proposed The approach extends the recent method proposed in

the previous chapter and section, which focused on multi-resolution/multi-temporal image classification to a multi-scale and multi-sensor model that fuses the spatial, multi-resolution, multi-sensor, and multi-frequency information conveyed by input images collected by CSK, RS2, and one optical sensor (Pléiades) at multiple spatial resolutions. For each of the two SAR sensors, the input images are inserted in a separate quad-tree on the basis of their spatial resolutions. These resolutions are expected to be coarser than the finest resolution that can be achieved using Pléiades data. Therefore, the input Pléiades image is embedded in the nest-resolution levels (i.e., the leaves) of both quad-trees. Empty levels of each quad-tree are filled in using wavelet transforms of the Pléiades data (see Figure 6.7). Then, Algorithm 6.2.1 is applied to compute recursively the posterior marginal of each site, while Gaussian and generalized Gamma PDFs are used to estimate the class-conditional statistics of the optical and SAR channels, respectively (and of the wavelet features derived from them).

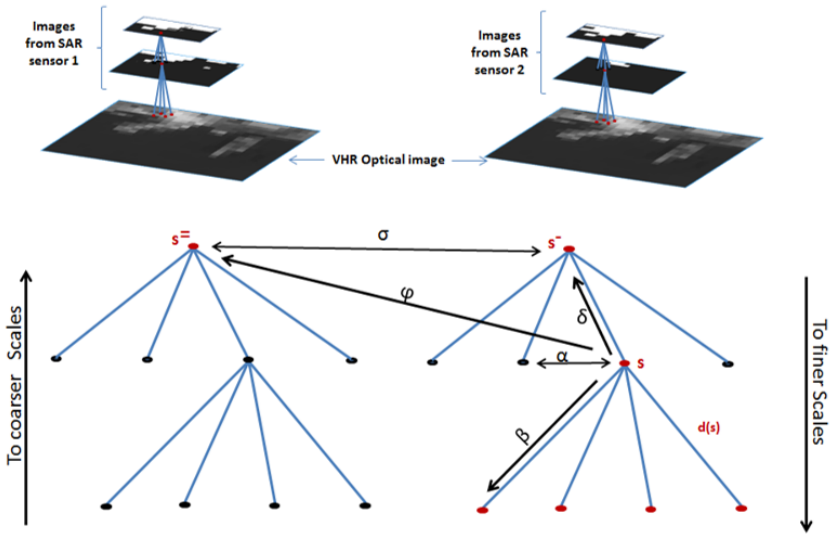


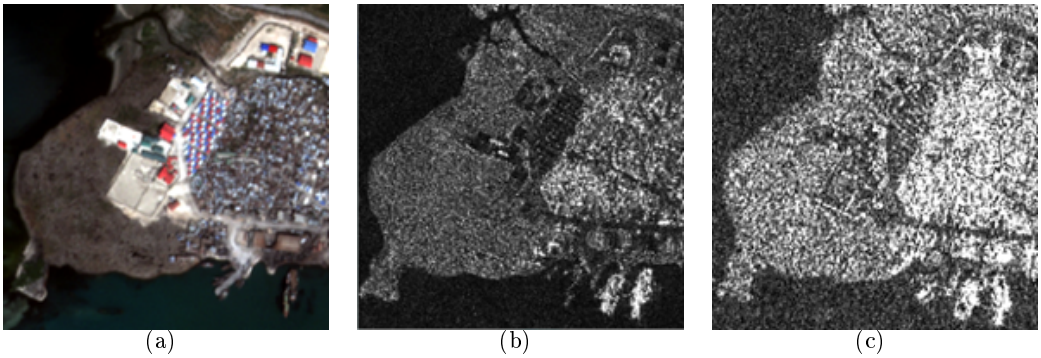
Figure 6.7: Multi-sensor hierarchical structure (case 2).

6.3.2 Experimental results

In this section, we discuss the results of the experimental validation of the developed model on two datasets acquired over Port-au-Prince (Haiti) using:

(i) a panchromatic Pléiades acquisition at 0.5m resolution (Pléiades, ©CNES distribution Airbus DS, 2011), shown in Figures 6.8(a) and 6.9(a).

(ii) a CSK image (©ASI, 2011), X band, HH polarization, Spotlight mode (1m resolution), geocoded, singlelook, shown in Figures 6.8(b) and 6.9(b).

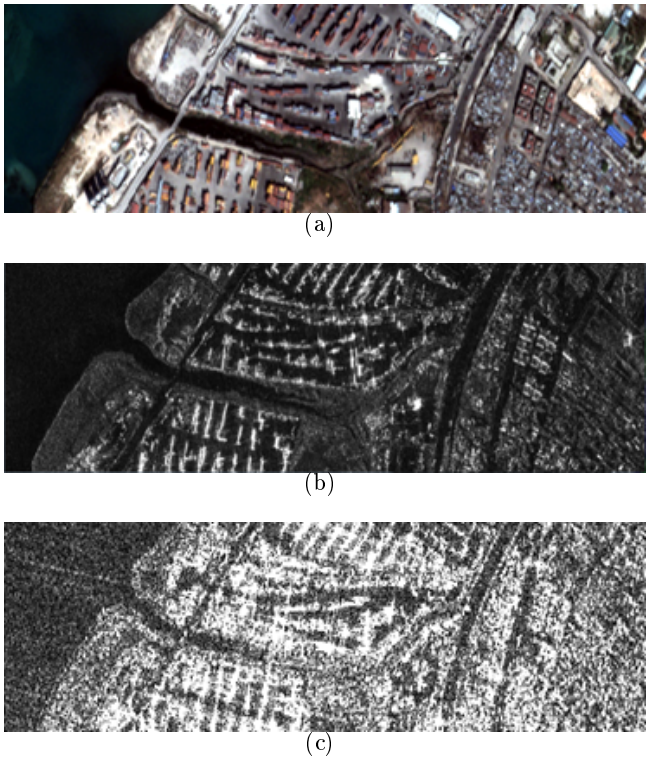


(a) (b) (c)

Figure 6.8: First site used for experiments: (a) Pléiades (©CNES distribution Airbus DS), (b) CSK (©ASI), and (c) RS2 images (©CSA).

(iii) a RS2 image (©CSA, 2011), C band, HH polarization, Ultra-Fine mode (1.56 m pixel spacing), geocoded, single-look, shown in Figures 6.8(c) and 6.9(c).

The RS2 image comes at 1.56 m pixel spacing. To fit with the dyadic decomposition imposed by the quad-tree, we slightly resampled the data to obtain the $0.5 = 2\ /4$ m resolution. Down sampling from 1.56 to 2 m is expected to have a minor impact on the classification map, because the resampling ratio is quite close to unity. Five classes are considered: urban (red), water (blue), low vegetation (green), sand (yellow) and containers (purple). We present the final classification maps in Figures 6.10 and 6.11 and the corresponding classification accuracies on the test samples in Table 6.2. To the best of our knowledge, none of the previously developed application-specific methods allows a direct integration of multi-resolution, multi-frequency, and multi-sensor optical-SAR data without major pre- or post-processing. The results obtained by the proposed multisensor hierarchical method, see Figures 6.10(d) and 6.11(d), lead to a detailed classification with a remarkable level of classification map regularity. The main source of misclassification is the container area, where containers are partly classified as urban. This is consistent with the fact that no texture features are used as input to the considered hierarchical MRF. In Table 6.2 we compare numerically the results obtained with the proposed hierarchical method when considering either only Pléiades, or both SAR (CSK or RS2) and optical images. We observe an improvement related to the combination of the two SAR images, in particular in the urban areas for which the joint use of CSK and RS2 acquisition represents a significant source of discriminative information. More specifically, we have observed that the optical image has a relevant effect in the sand and vegetation discrimination, and the SAR acquisitions are jointly very helpful to detect the urban area. This confirms the potential of exploiting the synergy of the data provided by the CSK and RS2 missions with respect to each other and to the imagery collected by space-borne optical HR cameras.



(a) (b) (c)

Figure 6.9: Second site used for experiments: (a) Pléiades (©CNES distribution Airbus DS), (b) CSK (©ASI), and (c) RS2 images (©CSA).

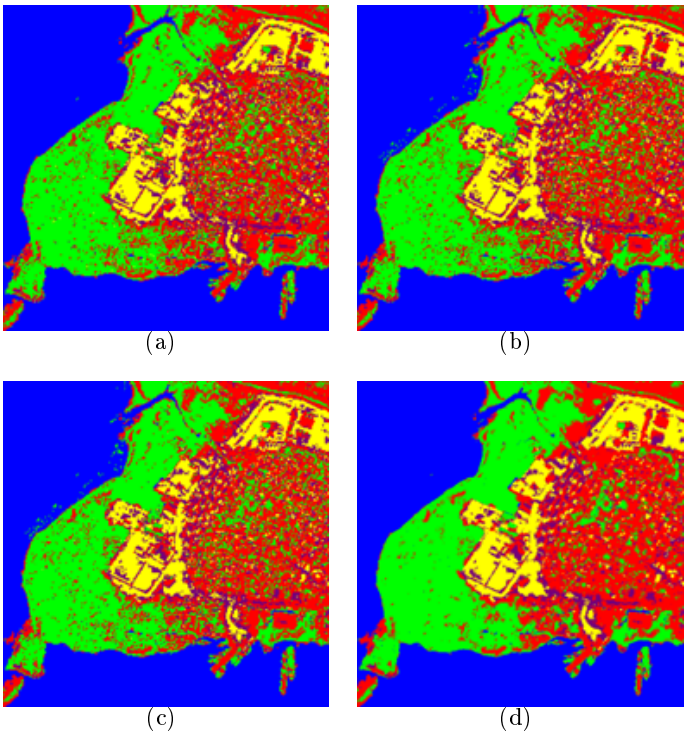


Figure 6.10: (a) Result using a only optical Pléiades images, (b) Result using both Pléiades and CSK acquisitions, (c) Result using both Pléiades and RS2 acquisitions (d) Result using all sensors.

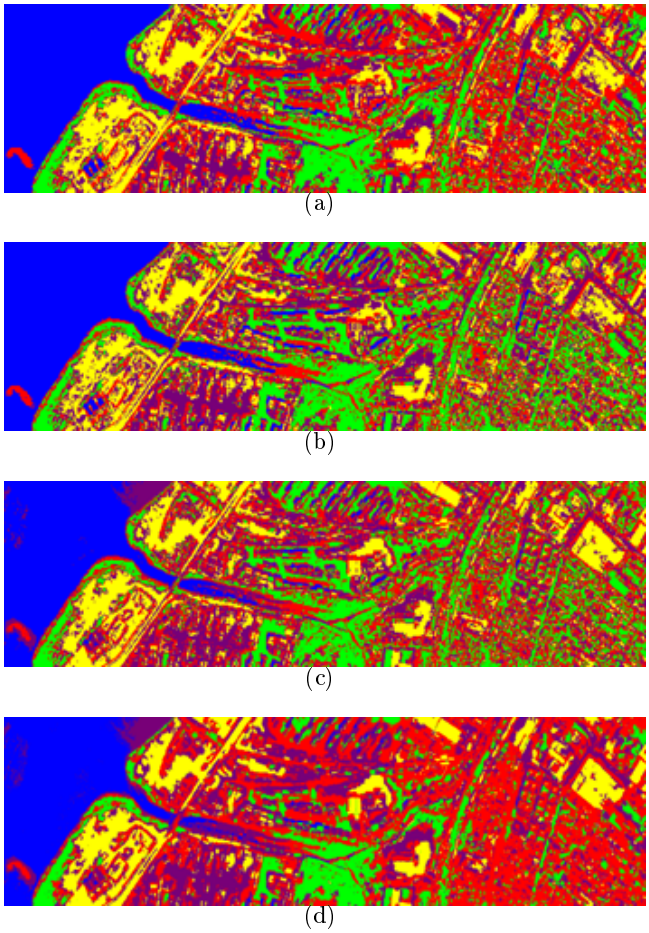


Figure 6.11: (a) Result using a only optical Pléiades images, (b) Result using both Pléiades and CSK acquisitions, (c) Result using both Pléiades and RS2 acquisitions (d) Result using all sensors.

	water	urban	vegetation	bare soil	containers	over all
Only Pléiades	100	61.66	81.69	82.82	56.72	76.57
Pléiades + CSK	100	44.32	83.54	74.75	49.12	70.34
Pléiades + RS2	92.56	44.85	79.85	78.62	42.15	67.60
Pléiades + RS2 + CSK	90.79	91.45	82.59	81.02	54.85	80.14

Table 6.2: Classification accuracies of results. Experiments were conducted on an Intel i7 quad-core (2.40 GHz) 8-GB-RAM 64-bit Linux system.

In the proposed method, multi-sensor and multi-resolution fusion is based on explicit statistical modeling. It combines a marginal statistical model of the considered input optical and SAR images, through hierarchical Markov random field modeling based on quad-trees in cascade, leading to a statistical supervised classification approach. We have developed a novel multi-source MPM-based hierarchical Markov random field model that takes into account both SAR and optical information and leads to improved robustness of the classifier. When applied to a several challenging high-resolution data sets associated with urban and semi-urban test sites, the proposed method gives high overall classification accuracy with a small computation time (a few minutes).

Chapter 7

Contextual multi-scale image classification on quad-tree

Contents

7.1 Markov mesh random field

As discussed in Chapter 4 probabilistic causal image models have been thoroughly studied since the early 90's through hierarchical MRFs on quad-trees. These models rely on a causality concept captured by the factorization of the prior distribution in terms of causal transition probabilities. In practice, this structure tends to generate "blocky" effects in the final classification map. Several techniques have been introduced in Section 4.3.2 that could mitigate such undesired effects. However, in these methods, the practical advantages of the tree structure (i.e., causality) are then partially (e.g., [Chardin and Pérez, 1999]) or completely lost (e.g., [Kato et al., 1996], because of the spatial interactions introduced in the respective models). In this chapter, the quad-tree approach is extended and a novel contextual multi-scale technique is proposed in order to classify multi-resolution remote sensing data that incorporate spatial contextual information and mitigate possible blocky artifacts while keeping the causality of the hierarchical model. Here, the focus is on the fusion of multi-resolution and spatial-contextual information for the supervised classification of single-date imagery. For this purpose, let us mention another important class of random fields. Markov mesh random fields (MMRFs), or causal Markov Random Fields that are also known as Unilateral MRFs (UMRFs) were first recalled in [Abend et al., 1965, Besag, 1972, Pickard, 1980]. As indicated in Chapter 3, images are usually modeled on a finite rectangular lattice with each site $s \in S$ being associated with one or more random variables.

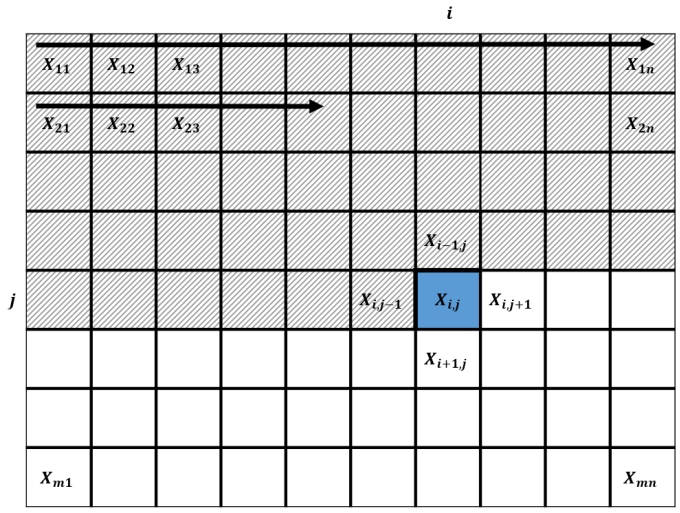
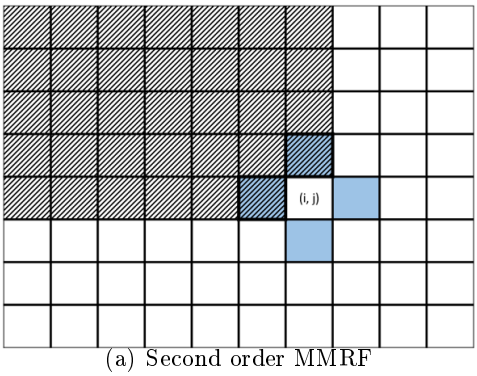
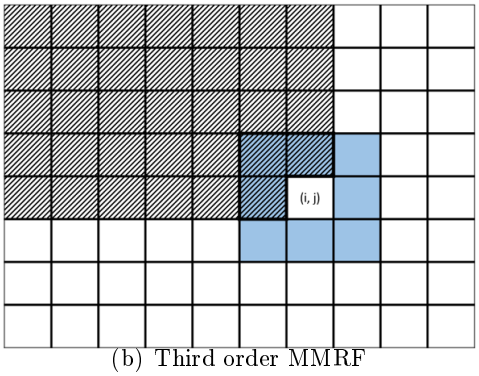


Figure 7.1: Regular rectangular lattice X of size $m \times n$

To build a causal process on this lattice, an order over the set of sites S, i.e., a well-defined characterization of the "past" of a site s as a set of nodes preceding s (see Section 3.5) is implicitly required. As shown in 7.1, $X_{i,j}$ is a random



(a) Second order MMRF



(b) Third order MMRF

Figure 7.2: Causal neighborhood.

variable associated with a site $s_{i,j}$ located at the coordinates $(i,j) \in [1,n] \times [1,m]$. $X = X_{11}, X_{12}, \ldots, X_{mn}$ represents the corresponding random field.

The probability that the random variable $X_{i,j}$ takes on the value $x_{i,j}$ is denoted by $p(X_{i,j} = x_{i,j})$, for ease of notation this probability is abbreviated as $p(x_{i,j})$ and the global probability is denoted by $p(X = x) = p(X_{11} = x_{11}, X_{12} = x_{12}, \ldots, X_{mn} = x_{mn})$ and abbreviated as $p(x)$. In this context, X is said to be an MMRF if:

$$p(x_{i,j}|pa(s_{i,j})) = p(x_{i,j}|\partial(s_{i,j})) \tag{7.1}$$

where, $pa(s_{i,j}) = \{X_{a,b}|a < i \, and \, b < j\}$ is considered to be the past of $X_{i,j}$ and $\partial(s_{i,j})$ is a subset of $pa(s_{i,j})$ that presents an unilateral neighborhood of $s_{i,j}$
Thus, X is a MMRF with the neighborhood size of 2 if (see Figure 7.2):

$$p(x_{i,j}|pa(s_{i,j})) = p(x_{i,j}|x_{i-1,j}, x_{i,j-1}) \tag{7.2}$$

As shown previously in Section 3.5, if this property is assumed, then the joint probability is equal to the product of the probabilities of each pixel in the lattice, each probability being conditional to the small subset of the past of the corresponding site. Formally:

$$p(x) = \prod_{i=1}^{m} \prod_{j=1}^{n} p(x_{i,j}|x_{i-1,j}, x_{i,j-1}) \tag{7.3}$$

where the required boundary adjustment is fulfilled by assuming the sites outside of the finite lattice to be zero. This factorization is what makes this model attractive. [Abend et al., 1965] proved that a MMRF is also a MRF while the inverse does not necessary holds, i.e., MMRFs form a causal subclass of MRFs.

In this chapter, a novel classification method is proposed by integrating an explicit hierarchical graph-based classifier, which uses a quad-tree structure to model multi-scale interactions, and an MMRF to deal with pixel wise contextual information in the same scale. The choices of a quad-tree and an MMRF allow taking benefit from their good analytical properties (especially causality) and consequently apply non-iterative algorithms that fuse multi-resolution and spatial-contextual information for classification purposes.

7.2 Methodology

7.2.1 Combined Structure

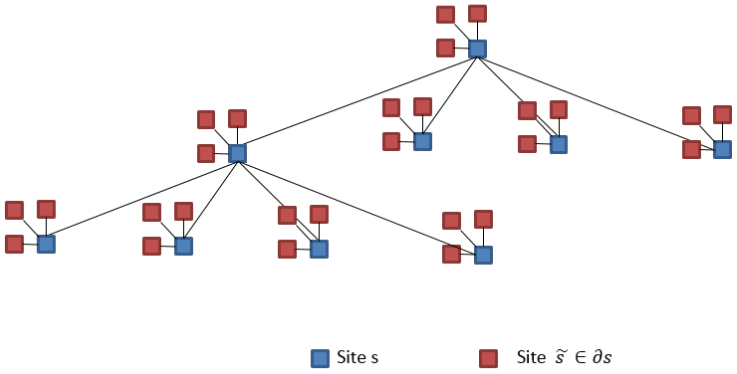
In this dissertation, a novel hybrid structure that combines a spatial grid using a causal MMRF and a hierarchical MRF via quad-tree is employed to circumvents the blocky artifacts of quad-tree-based methods and incorporate spatial contextual information in each scale. The starting point is to give an order on the set of all nodes S. For each scale of the quad-tree, a causal MMRF is integrated into the hierarchical structure. Accordingly, a node s at each level of the quad-tree except the root depends on one parent (at the level just above) and three neighbors (at the same level). For each pixel in the root level, there is no parent node and only the neighbors remains. The shapes of the neighborhoods of the pixels in the top and left borders of each lattice at each scale of the pyramid are obviously adapted to the image borders. The resulting dependence graph is shown in Figure 7.3.

7.2.2 MPM inference

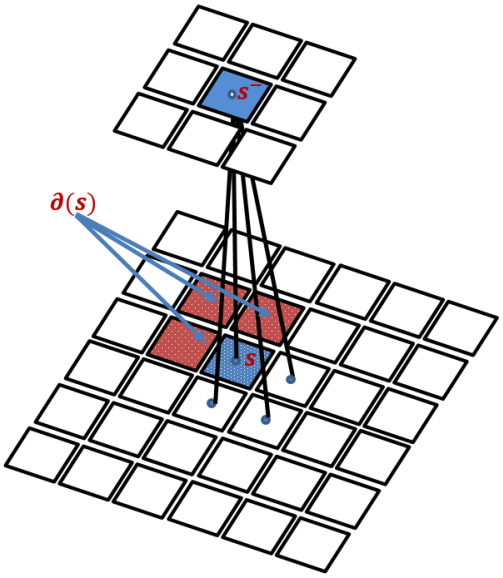
Again, when the causality property holds, non-iterative classification algorithms can be applied. In particular, a recursive procedure on a quad-tree is feasible for the MPM criterion, in which the aim is to maximize the posterior marginal at each site s:

$$\forall s \in S, x_s^\star = arg \max_{x_s \in \Lambda} p(x_s|y), \tag{7.4}$$

which produces the configuration that maximizes the a posteriori marginal $p(x_s|y)$ over the finite set Λ of thematic classes.



(a) Graphical model for the hierarchical Markov Mesh



(b) hierarchical structure

Figure 7.3: Hybrid structure that combines a spatial grid using a causal MMRF and a hierarchical MRF via quad-tree.

To resolve recursively the optimization problem in (7.4), one could use the same conditional independence assumptions developed in Section 5.2 for the multi-temporal model, the posterior marginal $p(x_s|y)$ of the label of each node s could be expressed as a function of the posterior marginal $p(x_{s^-}|y)$ of the parent node s^- in the quad-tree and the posterior marginals $p(x_{\tilde{s}}|y)$ where $\tilde{s}$ is each one of the sites in the set $\partial(s)$ of the three causal neighbors of s. Formally:

$$\mathbf{p}(\mathbf{x_s}|\mathbf{y}) = \sum_{x_{s^-},\partial(s)} \frac{p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})} \, \mathbf{p}(\mathbf{x_{s^-}}|\mathbf{y}) \prod_{\tilde{\mathbf{s}}\in\partial(\mathbf{s})} \mathbf{p}(\mathbf{x_{\tilde{s}}}|\mathbf{y}) \tag{7.5}$$

where bold fonts denote the marginal posteriors of interest to the MPM, as usual.

Proof of equation (7.5):

$$\begin{aligned}
p(x_s|y) &= \sum_{x_{s^-}} p(x_s|x_{s^-}, y) p(x_{s^-}|y) \\
&= \sum_{x_{s^-}} \left[\sum_{\partial(s)} p(x_s|x_{s^-}, x_{\partial(s)}, y) p(x_{\partial(s)}|y) \right] p(x_{s^-}|y) \\
&= \sum_{x_{s^-}} \left[\sum_{\partial(s)} \frac{p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})} p(x_{\partial(s)}|y) \right] p(x_{s^-}|y) \\
&= \sum_{x_{s^-}} \left[\sum_{\partial(s)} \frac{p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})} \prod_{\tilde{s}\in\partial(s)} p(x_{\tilde{s}}|y) \right] p(x_{s^-}|y) \\
&= \sum_{x_{s^-},\partial(s)} \frac{p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})} \, p(x_{s^-}|y) \prod_{\tilde{s}\in\partial(s)} p(x_{\tilde{s}}|y)
\end{aligned}$$

The equalities across rows 1 and 2 and across rows 3 and 4 derive from two conditional independence assumptions:

A1. the distribution of the labels $(x_s, x_{s^-}, x_{\partial(s)})$ given all the observations y can be restricted to the distribution conditioned only to the descendants of site s;

A2. the labels $\tilde{s}$ of the causal neighbors $\partial(s)$ are independent when conditioned to the data y.

These assumptions are analogous to the conditional independence assumptions that are commonly accepted when dealing with (hierarchical or single-scale) MRF-based image analysis, and have been mentioned in the previous chapters. They are used within the proposed method for analytical convenience.

The formulation in (7.5) allows calculating recursively the posterior marginal $p(x_s|y)$ at each node s while the probabilities $p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})$ are produced.

Thus, this simplifies to the determination of the other probabilities based on:

$$p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)}) = \frac{p(x_s|x_{s^-})p(x_{s^-})}{p(x_s)} \left[\prod_{\tilde{s}\in\partial(s)} \frac{p(x_s|x_{\tilde{s}})p(x_{\tilde{s}})}{p(x_s)} \right] p(x_s|y_{d(s)}) \quad (7.6)$$

In (7.6), the first factor $p(x_s|x_{s^-})$ corresponds to the child-parent transition probability; $p(x_s)$ is the prior probability; $p(x_s|x_{\tilde{s}})$ is the causal neighborhood transition probability in the same scale; and $p(x_s|y_{d(s)})$ is the partial posterior marginal probability.

Proof of equation (7.6):

$$\begin{aligned}
p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)}) &= p(x_{s^-}, x_{\partial(s)}|x_s)p(x_s|y_{d(s)}) \\
&= p(x_{s^-}|x_s)p(x_{\partial(s)}|x_s)p(x_s|y_{d(s)}) \\
&= p(x_{s^-}|x_s) \left[\prod_{\tilde{s}\in\partial(s)} p(x_{\tilde{s}}|x_s) \right] p(x_s|y_{d(s)}) \\
&= \frac{p(x_s|x_{s^-})p(x_{s^-})}{p(x_s)} \left[\prod_{\tilde{s}\in\partial(s)} \frac{p(x_s|x_{\tilde{s}})p(x_{\tilde{s}})}{p(x_s)} \right] p(x_s|y_{d(s)})
\end{aligned}$$

where the equality across rows 1 and 2 and across rows 2 and 3 derives from the following assumptions:

A3. The distribution of the labels x_{s^-} and $x_{\partial s}$ are independent on the observations $y_{d(s)}$ of the descendants of s, when conditioned to the label x_s of s i.e.,

$$p(x_{s^-}, x_{\partial s}|x_s, y_{d(s)}) = p(x_{s^-}, x_{\partial s}|x_s) \text{ (row 1)}$$

A4. The distribution of the labels x_{s^-} and $x_{\partial s}$ are independent when conditioned to x_s i.e.,

$$p(x_{s^-}, x_{\partial(s)}|x_s) = p(x_{s^-}|x_s)p(x_{\partial(s)}|x_s) \text{ (from row 1 to row 2)}$$

A5. The distribution of the labels x_{s^-} and $x_{\tilde{s}}$ are independent when conditioned to x_s, i.e.,

$$p(x_{s^-}, x_{\partial(s)}|x_s) = p(x_{s^-}|x_s)p(x_{\partial(s)}|x_s) \text{ (from row 2 to row 3)}$$

These results allow the recursive technique in Algorithm 7.2.1 to be formulated. It provides a generalization of the hierarchical MRF-based classification method proposed in [Laferté et al., 2000] and of the multi-temporal technique in Chapter 3 that allows spatial-contextual information to be incorporated through the MMRF component, in addition to multi-resolution information through the quad-tree hierarchy. For transition probabilities and class-conditional PDFs, the same models and estimates as in Chapter 4, which are based on the model described in [Bouman, 1991], on finite Gaussian mixtures, and on SEM, are extended here as well.

Algorithm 7.2.1 (Contextual hierarchical technique)

↻ **Initialization,** $(\mathbf{s} \in \mathbf{S^R})$:

Definition of the inter-scale transition probability $p(x_s|x_{s^-})$ *using the formulation introduced in [Bouman, 1991], and extension of the same formulation to the inter-neighborhood transition probability* $p(x_s|x_{\tilde{s}})$.
Initialization of the prior $p(x_s)$ *at the root* $s \in S^R$ *using Potts model as discussed in Section 5.2.1.*

⇓ **Top-down pass I,** $(\mathbf{s} \in \mathbf{S^n}, \mathbf{n} = \mathbf{0}, \mathbf{1}, \ldots, \mathbf{R-1})$:

$$p(x_s) = \sum_{x_{s^-}} p(x_s|x_{s^-})\ p(x_{s^-})$$

↻ **Initialization,** $(\mathbf{s} \in \mathbf{S_0})$: $p(x_s|y_{d(s)}) \propto p(y_s|x_s)p(x_s)$

⇑ **Bottom-up pass,** $(\mathbf{s} \in \mathbf{S^n}, \mathbf{n} = \mathbf{1}, \ldots, \mathbf{R})$:

$$p(x_s|y_{d(s)}) \propto p(y_s|x_s)p(x_s) \prod_{u \in s^+} \sum_{x_u} \frac{p(x_u|y_{d(u)})}{p(x_u)} p(x_u|x_s)$$

↻ **Initialization,** $(\mathbf{s} \in \mathbf{S^R})$: $p(x_s|y) = p(x_s|y_{d(s)})$

The likelihood term $p(y_s|x_s)$ *is estimated using a finite Gaussian mixture as in Section 5.3.1*

⇓ **Top-down pass II,** $(\mathbf{s} \in \mathbf{S^n}, \mathbf{n} = \mathbf{0}, \mathbf{1}, \ldots, \mathbf{R-1})$:

$$\mathbf{p(x_s|y)} = \sum_{x_{s^-}, x_{\partial(s)}} \frac{p(x_s, x_{s^-}, x_{\partial(s)}|y_{d(s)})}{\sum_{x_s} p(x_s, x_{s^-}, x_{s=}|y_{d(s)})}\ \mathbf{p(x_{s^-}|y)} \prod_{\tilde{\mathbf{s}} \in \partial(\mathbf{s})} \mathbf{p(x_{\tilde{s}}|y)}$$

7.3 Experimental results

In this section, we discuss preliminary results of the experimental validation of the developed contextual hierarchical classifier based on MMRF, using panchromatic and multi-spectral Pléiades images acquired over Port-au-Prince (Haiti). The finest resolution of the multi-resolution pyramid (level 0) was set equal to the finest resolution of the input panchromatic images (i.e. 0.5 m). Co-registered multi-spectral images (at 2 m) were integrated at level 2 of the pyramid. To fill

level 1, a wavelet decomposition of the panchromatic image was used. Five land cover classes have been considered: urban (red), water (blue), vegetation (green), soil (yellow) and containers (purple). In the present work, manually annotated (provided by an expert) non-overlapping training and test sets were selected in homogeneous areas. Spatially disjoint training and test areas were used. As shown in Figures 7.5, 7.4 and Table 7.1, the resulting classification map shows that the proposed hierarchical method leads to accurate results, especially as compared to the original hierarchical classification technique based on the MPM criterion in [Laferté et al., 2000] which yields to "blocky" classification (see Figures 7.5(b) and 7.4 (b)). These blocky artifacts are mitigated by incorporating spatial contextual information. The proposed method was also compared to an extended version of the method in [Laferté et al., 2000] in which the blocky artifacts were reduced by employing a prior update technique in the top-down step of the hierarchical algorithm as described in [Voisin et al., 2014]. Comparing to this method, the proposed technique obtains higher accuracies especially over the urban area included in the imaged scene as shown in Figures 7.4, 7.5 and Table 7.1.

The proposed method is aimed at performing hierarchical classification using input multi-resolution imagery. It combines a causal hierarchical MRF model using a quad-tree and a Markov mesh to preserve contextual information at each scale by applying a non-iterative classification algorithm using the MPM criterion. Experimental results with HR satellite imagery suggest that the method allows to effectively incorporates spatial information in the hierarchical classification process and provides higher accuracies than previous benchmark techniques. However, Considering the above, it is noted that MMRFs and their lattice models have the following weakness: MMRFs are corner-dependent and consequently the proposed hierarchical MRF model based on them are not necessarily symmetric (there exists an irregularity issue with this MMRF based method due to the fact that each pixel depend on a non-symmetric neighborhood, as highlighted with black arrows in Figures 7.4, 7.5). To circumvent this drawback several techniques have been introduced recently in the literature. Quadrilateral Markov Random Field (QMRF) was introduced in [Razlighi et al., 2009], in which, the non-regularity problem is avoided by enforcing all four MMRF with respect to different corners of the lattice into a new field definition; however, the model was non-symmetric. To overcome these limitations from both mathematical and practical points of view, [Yousefi et al., 2013] established a new random field, a symmetric, corner-independent, and isotropic image model which incorporates the dependency of a pixel on all its neighbors using a Symmetric Markov Mesh Random Field (SMMRF). This can be reached by scanning the lattice based on a specific scheme as detailed in [Yousefi et al., 2013]. The integration of SMMRF with an hierarchical MRF could be an interesting direction of further research.

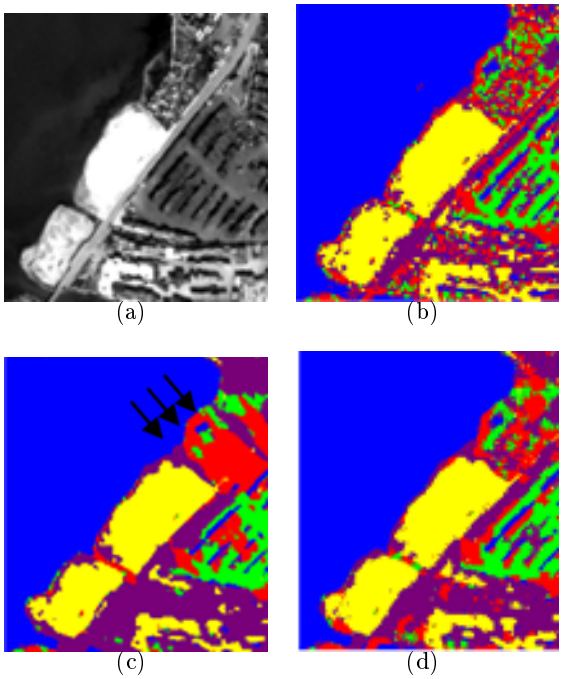


Figure 7.4: classification maps of optical(pléiades) image (a) using the original method proposed in [Laferté et al., 2000] (b), the proposed method (c) and the method in [Voisin et al., 2014] (d).

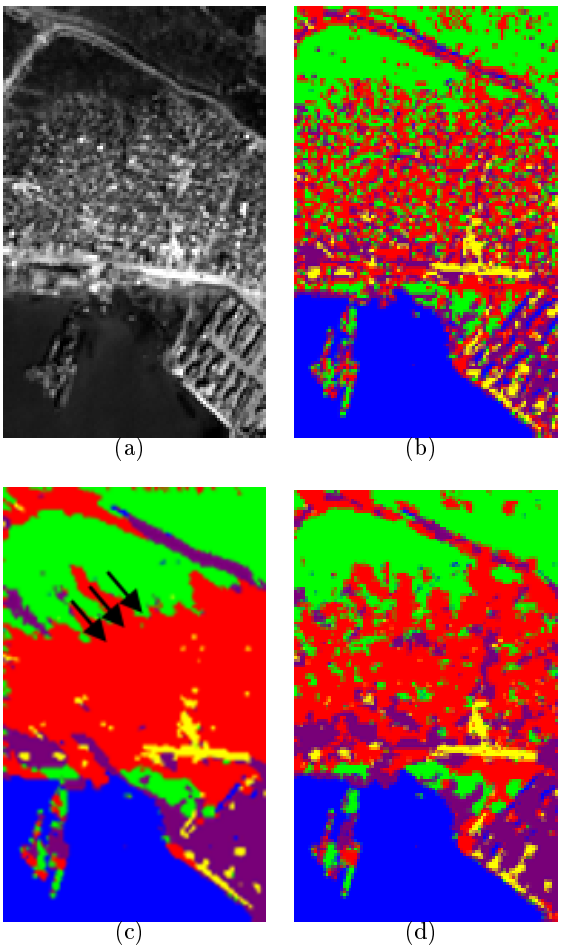


Figure 7.5: classification maps of optical(pléiades) image (a) using the original method proposed in [Laferté et al., 2000] (b), the proposed method (c) and method in [Voisin et al., 2014] (d).

	water	urban	vegetation	containers	soil	over all	computation time
Proposed method	100	92	89	81	94	91	147 seconds
method in [Laferté et al., 2000]	100	62	76	72	91	80	120 seconds
method in [Voisin et al., 2014]	100	74	83	86	92	87	154 seconds

Table 7.1: Results obtained using the Pléiades dataset: class accuracies, overall accuracy, and computation time.

Conclusion and perspectives

Conclusion

Contemporary satellite missions for Earth observation through optical (panchromatic and multi-spectral) and SAR sensors currently offer unprecedented observation capabilities, thanks to the attained very high spatial resolutions (up to a few dozens of centimeters), the short revisit times, and the joint availability of data with different spatial resolutions and coverages. In this framework, the present Ph.D. thesis focused on the development of novel multi-modal data fusion techniques for the joint supervised classification of images collected on the same geographical area at different times by distinct sensors at multiple spatial resolutions in different wavelength ranges. Several challenging methodological problems have been addressed to support the joint exploitation of multiple remote sensing data modalities at once for the generation of thematic classification results. The mathematical formalizations of the proposed techniques are rooted in the area of hierarchical multi-scale and causal random fields. They aim at taking benefit from the complementary information conveyed by remote sensing observations with diverse resolutions, sensors, frequencies, and acquisition times, at formulating their fusion in a flexible and analytically rigorous way.

Joint multi-date and multi-resolution fusion of optical data was addressed on basis of explicit statistical modeling. The method combined a joint model of the input images through hierarchical MRF modeling, leading to a statistical supervised classification approach. A novel MPM-based hierarchical Markov random field model was developed by considering multi-temporal information and, thus, supporting the classification of multiple images taken over the same area at different times and different spatial resolutions, a problem that has been very scarcely addressed so far in the data fusion literature. The results obtained by the proposed method were analyzed through experiments with multi-temporal Pléiades and GeoEye data sets. The experimental results showed that the method was able to provide accurate classification maps on both data sets. Indeed, the technique was demonstrated to be advantageous in terms of the classification accuracy on the test set, the spatial regularity of the classification maps, the minimization of spatial artifacts, and the tradeoff with respect to computation time as compared to previous benchmark methods that model multi-resolution or multi-temporal or spatial-contextual information. These results suggest the effectiveness of the algorithm in fusing both multi-temporal and multi-resolution information for supervised classification purposes and confirm that MRF models represent powerful fusion tools in remote sensing.
The methodological rationale of this multi-temporal analysis method has been extended to the case of multi-sensor image classification. The proposed method

addressed the problem of joint supervised classification of multi-sensor images including SAR and optical components acquired at multiple spatial resolutions. A novel approach, based on multiple quad-trees in cascade, to multi-sensor and multi-resolution fusion was developed. This approach formalizes a supervised Bayesian classifier within this multiple quad-tree topology that combines a class-conditional statistical model for pixel-wise information and a hierarchical MRF for multi-sensor and multi-resolution contextual information. Moreover, a focus on the specific case of the fusion of multi-frequency SAR data collected by the COSMO-SkyMed (X band) and RADARSAT-2 (C band) sensors together with multi-spectral Pléiades data was relevant to investigate the synergy among the multi-frequency SAR and multi-resolution information provided by these sensors.
Furthermore, another causal technique has been developed by combining a causal hierarchical MRF model using a quad-tree and a Markov mesh to preserve contextual information at each scale, classify multi-resolution remote sensing data, incorporate spatial contextual information, and mitigate possible blocky artifacts while keeping the causality of the hierarchical model and applying a non-iterative algorithm using the MPM criterion. Experimental results with HR satellite imagery showed that the method allows to effectively incorporate spatial information in the hierarchical classification process and provides higher accuracies than previous benchmark techniques.

The computational advantages of the hierarchical MRFs, for which exact recursive formulations of the MPM decision rule are feasible with no need for time-expensive Metropolis or Gibbs sampling procedures, have been confirmed by the experimental results of the proposed methods. All the techniques developed in this thesis are based on MRF models on a specific topology that comprises multiple hierarchical quad-trees, each associated with an acquisition date, a sensor, or a combination of sensors. This topology is flexible enough to natively incorporate images related to distinct spatial resolutions, acquisition times, and sensors. It can also accommodate the use of further multi-scale features. Wavelet transforms are used to fill in those levels of each quad-tree that are not natively associated with input remote sensing data. The selection of the wavelet operator among a large family of possible transforms is not critical because most transforms lead to classification results with similar accuracies. Nevertheless, Daubechies wavelets of order 10 yields higher accuracies than the other considered transforms.

Perspectives

Automation of the selection of the wavelet operator

As a future extension of the proposed methods, the automation of the selection of the wavelet operator using, for example, a dictionary of multiple transforms [Moser et al., 2011] could be incorporated in the developed models. Similarly, the pseudo-likelihood method was used to optimize the smoothing pa-

rameter of the Potts spatial component of the proposed MRF model. Alternate parameter estimation algorithms based, for example, on mean-square error [Serpico and Moser, 2006], stochastic gradient [Younes, 1988], or Monte-Carlo techniques [Ibáñez and Simó, 2003] could be integrated in the proposed method to address the optimization of this parameter and of the parameters involved in the transition probabilities across multiple scales, times, or sensors. The accuracy of the proposed method was found sensitive to these parameters. However, the experimental results suggested that, based on the meaning of these parameters in relation to temporal and inter-scale correlations, it is not difficult for a user/operator to identify ranges on their values that lead to meaningful models of the transition probabilities and yield accurate results.

SMMRF and quad-tree

Markov mesh random fields and their lattice models exhibit a well-known weakness, i.e., they are "corner-dependent". Consequently, the proposed integration of a hierarchical MRF and a spatial Markov mesh model does not necessarily exhibit an anisotropic behavior and can be affected by the use of a non-regular (non-symmetric) neighborhood. To mitigate this drawback several techniques have been recently introduced in the literature. Quadrilateral MRFs were introduced in [Razlighi et al., 2009], in which the non-regularity problem is avoided by using four Markov meshes related to the different corners of the lattice and enforcing them into a unique field definition; however, the model is still non-symmetric. To overcome these limitations from both mathematical and practical points of view, [Yousefi et al., 2013] established a new random field: a symmetric, corner-independent, and isotropic model that incorporates the dependency of a pixel on all its neighbors using a symmetric Markov mesh random field. The integration of this approach with the proposed hierarchical techniques could be an interesting direction of further research.

Semantic classification

In this thesis, the number of classes for the proposed methods has been kept fixed for all levels of the hierarchical structure. On one hand, this is common practice in supervised image classification. On the other hand, especially when VHR data are considered, different types of land cover classes may be appreciated at different resolutions. Therefore, a further extension of this work would be to define different sets of classes at distinct levels of the pyramid and introduce a hierarchical link between these classes according to their semantic meanings. A pixel-level classification may not exceed the spectral signature or backscattering features resulting from the biophysical and geophysical properties of each individual pixel, but the neighborhood of the pixel brings substantial information that can be used to reconstruct landscape units and functional areas. Then, a vertical strategy could be applied: each pixel could be assigned to an elementary class, corresponding to the

related ground material, and each group of pixels constitutes classes as a mixture of these material classes. Therefore, a pixel may belong to different classes according to the spatial resolution of the image in which it belongs, and a semantic relationship between classes may be defined and would critically involve the availability of multi-resolution ground truth data. The extension of the proposed hierarchical MRF model to this case-specific typology of soft classification, for example based on fuzzy reasoning, will be a further challenging generalization of the work described in this thesis.

Appendix A

Scientific activity

The following papers have been published or submitted on works contained in this dissertation:

A.1 Peer-reviewed papers for international conferences

- I. Hedhli, G. Moser, J. Zerubia et S. B. Serpico, "Contextual multi-scale image classification on quad-tree", IEEE International Conference on Image Processing (ICIP), 2016, (soumis).
- I. Hedhli, G. Moser, J. Zerubia, and S. B. Serpico, "New hierarchical joint classification method of SAR-optical multiresolution remote sensing data", IEEE/EURASIP European Signal Processing Conference (EUSIPCO), Nice, France, Aug 2015.
- I. Hedhli, G. Moser, J. Zerubia, and S. B. Serpico, New cascade model for hierarchical joint classification of multisensor and multiresolution remote sensing data. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Jul 2015, Milan, Italy. 2015.
- I. Hedhli, G. Moser, J. Zerubia, and S. B. Serpico, "Fusion of multitemporal and multiresolution remote sensing data and application to natural disasters", in IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Québec, Canada, July 2014.
- I. Hedhli, G. Moser, J. Zerubia, and S. B. Serpico, "New cascade model for hierarchical joint classification of multitemporal, multiresolution and multisensor remote sensing data", in IEEE International Conference on Image Processing (ICIP), Paris, France, October 2014.

A.2 Peer-reviewed journals

- I. Hedhli, G. Moser, J. Zerubia, and S. B. Serpico, "A New Cascade Model for the Hierarchical Joint Classification of Multitemporal and Multiresolution Remote Sensing Data", IEEE Transactions on Geoscience and Remote Sensing (TGRS) (under revision).
- I. Hedhli, G. Moser, J. Zerubia, "Nouvelle méthode en cascade pour la classification hiérarchique multi-temporelle ou multi-capteur d'images satellitaires

haute résolution", Revue Française de Photogrammétrie et de Télédétection (RFPT) (submitted).

A.3 International Summer School

- SPLIT SUMMER SCHOOL (May 2015, THESSALONIKI, GREECE) Environmental Security: Remote Sensing of Natural Resources
- GR4S SUMMER SCHOOL (Jul 2015, PAVIA, ITALY) Data Fusion for Risk Mapping

Appendix B

Résumé étoffé

Les moyens mis en oeuvre pour surveiller la surface de la Terre, notamment les zones urbaines, en cas de catastrophes naturelles telles que les inondations ou les tremblements de terre, et pour évaluer l'impact de ces événements, jouent un rôle primordial du point de vue sociétal, économique et humain. De très nombreux usages (scientifiques, civils et militaires) sont faits de ces données. Les satellites ont eu une contribution importante et unique dans certains domaines dont l'évaluation et la surveillance de l'environnement (déforestation, pollution, pollution lumineuse, urbanisation et périurbanisation, fragmentation écologique, érosion, etc.). Dans ce cadre, des méthodes de classification précises et efficaces sont des outils particulièrement importants pour aider à l'évaluation rapide et fiable des changements au sol et des dommages provoqués.
Étant données l'énorme quantité et la variété des données disponibles grâce aux missions satellitaires de dernière génération et de différents types, la principale difficulté est de trouver un classifieur qui puisse prendre en compte des données multi-bande, multi-résolution, multi-date, et éventuellement multi-capteur, tout en gardant un temps de calcul raisonnable.
Considérant une série d'images satellitaires, un modèle multi-source est proposé afin de fusionner les informations spatiales, temporelles et multi-résolution. A cette fin, deux approches bayésiennes peuvent être adoptées : un classifieur en cascade obtenu en supprimant le couplage entre les dimensions spatiales et temporelles, et en classifiant chaque image de la série sur la base d'elle-même et des images précédentes ou l'approche mutuelle qui classifie chaque image sur la base des images précédentes et ultérieures. L'approche en cascade est adoptée dans cet thése afin de préserver l'ordre temporel des images et d'améliorer la classification à chaque nouvelle donnée disponible. La méthode proposée aborde le problème de la classification multi-temporelle et multi-résolution dans laquelle, la fusion multi-date et multi-résolution est fondée sur une modélisation statistique explicite au travers d'un modèle hiérarchique de champs de Markov (MRF) fusionnant l'information spatio-temporelle et multi-résolution.

Classifieur hiérarchique proposé

L'objectif de la classification est d'estimer un ensemble d'étiquettes cachées $\mathcal{X}$ sachant un ensemble d'observations $\mathcal{Y}$ attachées à l'ensemble des noeuds (pixels) $\mathcal{S}$. Chaque étiquette a une valeur dans l'ensemble des étiquettes, $\Lambda = \{0, 1, ..., M-1\}$.

L'espace des configurations $\Omega = \Lambda^{|\mathcal{S}|}$ est l'ensemble de toutes les configurations possibles.
$\mathcal{X} = \{x_s\}_{s\in\mathcal{S}}, x_s \in \Lambda$ et $\mathcal{Y} = \{y_s\}_{s\in\mathcal{S}}$ sont considérés comme des processus aléatoires. Dans ce contexte, nous étudions le problème d'inférence de la meilleure configuration $\hat{x}$ de $\mathcal{X}$ dans Ω. La formulation bayésienne de ce problème d'inférence consiste à minimiser le coût d'erreur de segmentation selon un critère choisi :

$$\hat{x} = \arg\min_{x'\in\Omega} \sum_{x\in\Omega} \mathcal{C}(x, x')p(x|y) \tag{B.1}$$

où $\mathcal{C}(.,.)$ est le coût pénalisant l'écart entre la configuration estimée et la configuration aléatoire idéale.
Parmi les différents algorithmes de classification bayésienne à minimum d'erreur, deux critères ont été largement utilisés. Le premier vise à estimer le maximum a posteriori (MAP) où la configuration optimale est obtenue par :

$$\hat{x}_{MAP} = \arg\min_{x\in\Omega} p(x|y) \tag{B.2}$$

Cela en remplaçant dans (B.1) la fonction du coût suivante :

$$\begin{aligned} \mathcal{C}(x, x') &= 1 - \delta(x, x') \\ &= 1 - \prod_{n=0}^{R} \delta(x^n, x'^n) \\ &\quad \text{(dans le cadre hiérarchique)} \end{aligned} \tag{B.3}$$

où R est le nombre de niveaux de la pyramide et δ est le symbole de Kronecker.
Cette fonction de coût est égale à 1 si une erreur de classification apparaît dans un niveau quelconque de la pyramide. Par conséquence, cette fonction pénalise les erreurs indépendamment de l'échelle à laquelle elles se produisent, ce qui n'est pas très intéressant pour des méthodes de classification hiérarchiques et ne compte pas le nombre total d'erreurs.
Le second critère très utilisé dans la littérature est celui des Modes Marginales a Posteriori (MPM en anglais) dont la fonction du coût est définie par:

$$\mathcal{C}(x, x') = \sum_{s\in S} 1 - \delta(x_s, x'_s) \tag{B.4}$$

L'utilisation de cette fonction offre l'avantage de pénaliser les erreurs en fonction de leur nombre et de l'échelle à laquelle elles se produisent: par exemple, une erreur à une échelle grossière est plus fortement pénalisée qu'une erreur à une échelle plus fine. L'estimateur bayésien résultant est défini pour chaque pixel s par:

$$\forall s \in \mathcal{S}, \hat{x}_s = \arg\max_{x_s\in\Lambda} p(x_s|y) \tag{B.5}$$

Parmis les méthodes utilisées pour résoudre ce problème d'optimisation, on distingue la méthode fondée sur la structure quad-arbre employée par [Laferté et al., 2000]

qui vise à affecter à chaque pixel s l'étiquette optimale $\hat{x}_s$, en maximisant récursivement la probabilité marginale a posteriori $p(x_s|y)$, cela en exprimant la probabilité marginale du pixel s à travers la probabilité marginale $p(x_{s^-}|y)$ du parent s^-. Malgré ses performances en temps de calcul et en précision, l'utilisation de cet algorithme conduit à des artéfacts de blocs [Laferté et al., 2000]. Ce phénoméne peut être expliqué par le fait que, à une échelle donnée, deux pixels voisins, peuvent ne pas avoir le même parent. Dans ce cas, une limite apparaît beaucoup plus facilement que s'ils étaient liés au même parent ou s'il existait d'autres liaisons entre les pixels à des échelles consécutives. Afin d'éviter ces artéfacts, nous pouvons utiliser une structure hiérarchique plus connectée dans laquelle la liaison causale ascendant-descendant dans la même pyramide n'est pas l'unique liaison entre les niveaux consécutifs (cf. par exemple [Bouman, 1991]).
Dans ce contexte, une nouvelle formulation de l'algorithme fondé sur le critère MPM est proposée en utilisant la structure hiérarchique en cascade définie dans la partie **??**. La probabilité marginale a posteriori $p(x_s|y)$ de chaque pixel s, est exprimée non seulement à travers la probabilité marginale a posteriori $p(x_{s^-}|y)$ du parent s^-, mais aussi à travers la probabilité marginale a posteriori $p(x_{s^=}|y)$ du parent à une source antérieure $s^=$ afin de caractériser les corrélations entre des images consécutives de la structure en cascade définie et éviter ainsi les artéfacts de blocs :

$$\begin{aligned} p(x_s|y) &= \sum_{s^-,s^=} p(x_s|x_{s^-},x_{s^=},y).p(x_{s^-},x_{s^=}|y) \\ &= \sum_{s^-,s^=} p(x_s|x_{s^-},x_{s^=},y_{d(s)}).p(x_{s^-},x_{s^=}|y) \\ &= \sum_{s^-,s^=} \mathcal{A}\ .\ p(x_{s^-},x_{s^=}|y) \\ &= \sum_{s^-,s^=} \mathcal{A}\ .\ p(x_{s^-}|y).p(x_{s^=}|y) \end{aligned} \tag{B.6}$$

où

$$\begin{aligned} \mathcal{A} &= p(x_s|x_{s^-},x_{s^=},y_{d(s)}) \\ &= \frac{p(x_s,x_{s^-},x_{s^=}|y_{d(s)})}{\sum_s p(x_s,x_{s^-},x_{s^=}|y_{d(s)})} \end{aligned} \tag{B.7}$$

Il est à noter que les équations dans (B.6) impliquent deux hypothèses d'indépendance conditionnelle:

- pour le passage de la ligne 1 à la ligne 2 dans (B.6) : étant donné les étiquettes x_{s^-} et $x_{s^=}$, l'étiquette x_s ne dépend pas de tout l'ensemble d'observations y, elle dépend seulement des observations attachées à s et à ses descendants ce qui implique que :

$$p(x_s|x_{s^-},x_{s^=},y) = p(x_s|x_{s^-},x_{s^=},y_{d(s)})$$

- pour le passage de la ligne 3 à la ligne 4 dans (B.6) : les étiquettes x_{s^-} et $x_{s^=}$ sont indépendantes, étant donné l'ensemble d'observation y, ce qui implique :

$$p(x_{s^-}, x_{s^=}|y) = p(x_{s^-}|y).p(x_{s^=}|y)$$

Il reste à calculer les probabilités jointes $p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})$ évoquées dans (B.7) qui peuvent être déterminées en estimant les probabilités introduites dans la factorisation suivante :

$$\begin{aligned} &p(x_s, x_{s^-}, x_{s^=}|y_{d(s)}) = \\ &\quad p(x_s|x_{s^-}, x_{s^=}).\frac{p(x_{s^-}|x_{s^=}).p(x_{s^=})}{p(x_s)}.p(x_s|y_{d(s)}) \end{aligned} \tag{B.8}$$

incluant :

- des probabilités a priori ($p(x_{s^=})$ et $p(x_s)$),
- des probabilités inter-échelle et/ou multi-source ($p(x_s|x_{s^-})$ et $p(x_s|x_{s^-}, x_{s^=})$),
- des probabilités a posteriori partielles $p(x_s|y_{d(s)})$.

Par la suite, à chaque nouvelle donnée t (nouvelle image à une date différente ou acquise par un capteur différent), pour estimer la probabilité marginale a posteriori $p(x_s|y)$, il suffit de calculer les probabilités évoqueés dans (B.8). A cet effet, l'approche proposée se déroule en trois passes récursives, descendante (1), ascendante (1) et descendante (2)

Passe descendante (1) : estimation des probabilités a priori

Afin de calculer les probabilités a priori dans la pyramide $\mathcal{P}$ (à une nouvelle date t dans le cadre multi-temporel ou avec un nouveau capteur c dans le cadre multi-capteur), nous procédons à une passe descendante, en commençant par estimer les probabilités a priori à la résolution la plus grossière. Pour ce faire, nous utilisons la carte de classification obtenue en utilisant la structure hiérarchique en cascade jusqu'à la pyramide $\mathcal{P}-1$ en considérant un modèle markovien spatial qui conduit à une meilleure estimation de l'a priori. Ainsi, en utilisant le théorème de Hammersley-Clifford, nous pouvons définir un a priori local pour chaque pixel s à la résolution la plus grossière comme étant un modèle de Potts décrit ci-dessous :

$$p(x_s|x_{s'}, s' \sim s) \propto \exp\left(-\beta \sum_{s \sim s'} \delta\left(x_s, s_{s'}\right)\right) \tag{B.9}$$

Où $\delta(.)$ est le symbole de Kronecker, $s \sim s'$ signifie que s et s' sont voisins par rapport à un système de voisinage donné et β est un paramètre de lissage, de signe positif pour faire décroitre l'énergie lorsque les étiquettes sont identiques.

Puis, une passe descendante est effectuée pour calculer les probabilités a priori à chaque niveau du quad-arbre en utilisant la formulation récursive suivante :

$$p(x_s) = \sum_{s^-} p(x_s|x_{s^-}) \cdot p(x_{s^-}) \tag{B.10}$$

Les probabilités de transition inter-échelle au sein de l'arbre $p(x_s|x_{s^-})$ peuvent être calculées indépendamment de telle sorte que x_s appartienne à la même classe que son ascendant x_{s^-}.

Passe ascendante (1) : estimation des probabilités jointes

En se basant sur (B.8), les probabilités jointes $p(x_s, x_{s^-}, x_{s^=}|y_{d(s)})$ sont déterminées par :

- les probabilités a priori $p(x_s)$, déjà estimées lors de la passe descendante précédente,
- les probabilités inter-échelle et/ou inter-pyramide $p(x_s|x_{s^-})$ et $p(x_s|x_{s^-}, x_{s^=})$, qui peuvent être calculées en favorisant l'appartenance des pixels concernés à une même classe;
- les probabilités a posteriori partielles $p(x_s|y_{d(s)})$ qui peuvent être calculées moyennant une passe ascendante en utilisant la formulation introduite par [Laferté et al., 2000]:

$$p(x_s|y_{d(s)}) \propto p(y_s|x_s) \cdot p(x_s) \cdot \prod_{\tilde{s} \in s^+} \sum_{\tilde{s}} \frac{p(x_{\tilde{s}}|y_{d(\tilde{s})})}{p(x_{\tilde{s}})} \cdot p(x_{\tilde{s}}, x_s) \tag{B.11}$$

Les probabilités a posteriori partielles nécessitent l'estimation d'un modèle statistique conditionnel par classe combinant des informations pixel par pixel, à la même résolution pour le calcul de $p(y_s|x_s)$. Cela peut se faire en traitant la modélisation des probabilités marginales des images acquises à chaque résolution, conditionnées à chaque classe.

Passe descendante (2) : Probabilité a posteriori et optimisation

Connaissant les probabilités jointes calculées dans la partie précédente, nous procédons à une passe descendante dans laquelle les probabilités marginales a posteriori $p(x_s|y)$ sont estimées en utilisant (B.6), puis maximisées par un algorithme de Metropolis modifié [Kato et al., 1996].

Dans les méthodes proposées, la fusion multi-résolution, multi-temporelle ou multi-capteur est fondée sur une modélisation statistique explicite au travers d'un modèle hiérarchique de champs de Markov. La nouveauté principale de l'approche est l'utilisation en cascade de plusieurs quad-arbres, chacun étant associé à une

nouvelle image disponible, en vue de caractériser les corrélations temporelles ou spatiales associées à des images distinctes. Lorsqu'elle est appliquée à des images à haute résolution, la méthode proposée donne une précision globale de la classification de valeur élevée avec un temps de calcul raisonnable grâce à la structure hiérarchique utilisée.

Bibliography

[Abend et al., 1965] Abend, K., Harley, T. J., and Kanal, L. (1965). Classification of binary random patterns. *IEEE Transactions on Information Theory*, 11(4):538–544. (Cited on pages 118 and 120.)

[ALEjaily et al., 2008] ALEjaily, A., El Rube, I., and Mangoud, M. (2008). Fusion of remote sensing images using contourlet transform. *Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering*, pages 213–218. (Cited on page 30.)

[Armin and Fred, 2004] Armin, W. and Fred, M. (2004). Synthetic aperture radar. *Opt. Photon. News*, 15(11):28–33. (Cited on page 23.)

[Aspert et al., 2007] Aspert, F., Bach-Cuadra, M., Thiran, J., Cantone, A., and Holecz, F. (2007). Time-varying segmentation for mapping of land cover changes. In *Proceeding of ESA Symposium, Montreux*. (Cited on page 35.)

[Azimi-Sadjadi and Bannour, 1991] Azimi-Sadjadi, M. and Bannour, S. (1991). Two-dimensional adaptive block Kalman filtering of SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 29(5):742–753. (Cited on page 33.)

[Baraldi and Parmiggiani, 1995] Baraldi, A. and Parmiggiani, F. (1995). A refined Gamma map SAR speckle filter with improved geometrical adaptivity. *IEEE Transactions on Geoscience and Remote Sensing*, 33(5):1245–1257. (Cited on page 33.)

[Basseville et al., 1992a] Basseville, M., Benveniste, A., and Willsky, A. (1992a). Multiscale autoregressive processes. I. Schur-Levinson parametrizations. *IEEE Transactions on Signal Processing*, 40(8):1915–1934. (Cited on pages 30 and 60.)

[Basseville et al., 1992b] Basseville, M., Benveniste, A., and Willsky, A. (1992b). Multiscale autoregressive processes. II. lattice structures for whitening and modeling. *IEEE Transactions on Signal Processing*, 40(8):1935–1954. (Cited on pages 30 and 60.)

[Baum et al., 1970] Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Snnals of Sathematical Statistics*, pages 164–171. (Cited on pages 55, 64 and 85.)

[Baxter, 2007] Baxter, R. (2007). *Exactly solved models in statistical mechanics*. Dover pub. (Cited on page 40.)

[Bazi et al., 2005] Bazi, Y., Bruzzone, L., and Melgani, F. (2005). An unsupervised approach based on the generalized Gaussian model to automatic change detection

in multitemporal SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):874–887. (Cited on pages 31 and 35.)

[Bellman, 1956] Bellman, R. (1956). Dynamic programming and Lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767. (Cited on page 53.)

[Benedek et al., 2015] Benedek, C., Shadaydeh, M., Kato, Z., Szirányi, T., and Zerubia, J. (2015). Multilayer Markov random field models for change detection in optical remote sensing images. *Journal of Photogrammetry and Remote Sensing*, 107:22–37. (Cited on page 31.)

[Benz, 1999] Benz, U. (1999). Supervised fuzzy analysis of single-and multichannel SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):1023–1037. (Cited on page 30.)

[Bernstein et al., 2012] Bernstein, L., Jin, X., Gregor, B., and Adler-Golden, S. (2012). Quick atmospheric correction code: algorithm description and recent upgrades. *Optical engineering*, 51(11). (Cited on page 32.)

[Besag, 1972] Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society (B)*, pages 75–83. (Cited on page 118.)

[Besag, 1974] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236. (Cited on pages 42, 43, 47 and 88.)

[Besag, 1986] Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259–302. (Cited on pages 47 and 69.)

[Bishop, 2006] Bishop, C. (2006). *Pattern recognition and machine learning*. Springer. (Cited on page 48.)

[Biswas et al., 2013] Biswas, S., Farrar, S., Gopalan, K., Santos-Garcia, A., Jones, W., and Bilanow, S. (2013). Intercalibration of microwave radiometer brightness temperatures for the global precipitation measurement mission. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3):1465–1477. (Cited on page 32.)

[Blake and Zisserman, 1987] Blake, A. and Zisserman, A. (1987). *Visual reconstruction*, volume 2. MIT press Cambridge. (Cited on page 47.)

[Bouman, 1991] Bouman, C. (1991). *A multiscale image model for Bayesian image segmentation*. Purdue University, School of Electrical Engineering. (Cited on pages 64, 67, 79, 123, 124 and 137.)

[Bovolo and Bruzzone, 2007] Bovolo, F. and Bruzzone, L. (2007). 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. (Cited on page 34.)

[Bovolo and Bruzzone, 2008] Bovolo, F. and Bruzzone, L. (2008). An adaptive technique based on similarity measures for change detection in very high resolution SAR images. In *IEEE Geoscience and Remote Sensing Symposium*, volume 3. (Cited on page 32.)

[Bovolo and Bruzzone, 2015] Bovolo, F. and Bruzzone, L. (2015). The time variable in data fusion: A change detection perspective. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):8–26. (Cited on page 32.)

[Boykov and Jolly, 2001] Boykov, Y. and Jolly, M.-P. (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *IEEE International Conference on Computer Vision*, volume 1, pages 105–112. (Cited on page 47.)

[Boykov and Kolmogorov, 2001] Boykov, Y. and Kolmogorov, V. (2001). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In *Energy Minimization methods in Computer Vision and Pattern Recognition*, pages 359–374. Springer. (Cited on page 47.)

[Brunner et al., 2010] Brunner, D., Lemoine, G., and Bruzzone, L. (2010). Earthquake damage assessment of buildings using VHR optical and SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2403–2420. (Cited on page 31.)

[Bruzzone et al., 1999] Bruzzone, L., Prieto, D. F., and Serpico, S. (1999). A neural-statistical approach to multitemporal and multisource remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3):1350–1359. (Cited on pages 36 and 85.)

[Burt, 1984] Burt, P. (1984). *The pyramid as a structure for efficient computation*. Springer. (Cited on page 30.)

[Carvalho et al., 2001] Carvalho, L., Fonseca, L., Murtagh, F., and Clevers, J. (2001). Digital change detection with the aid of multiresolution wavelet analysis. *International Journal of Remote Sensing*, 22(18):3871–3876. (Cited on page 34.)

[Celeux et al., 1996] Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4):287–314. (Cited on page 82.)

[Černỳ, 1985] Černỳ, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51. (Cited on page 46.)

[Chardin, 2000] Chardin, A. (2000). *Modeles energetiques hierarchiques pour la resolution des problemes inverses en analyse d'images, application a la teledetection.* PhD thesis, University of Rennes 1, France. (Cited on pages 30, 52, 60, 67, 69, 70 and 78.)

[Chardin and Pérez, 1999] Chardin, A. and Pérez, P. (1999). Semi-iterative inferences with hierarchical energy-based models for image analysis. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 83–98. Springer. (Cited on pages 69, 70 and 118.)

[Chen et al., 2004] Chen, J., Jönsson, P., Tamura, M., Gu, Z., Matsushita, B., and Eklundh, L. (2004). A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote sensing of Environment*, 91(3):332–344. (Cited on page 33.)

[Chibani and Houacine, 2003] Chibani, Y. and Houacine, A. (2003). Redundant versus orthogonal wavelet decomposition for multisensor image fusion. *Pattern Recognition*, 36(4):879–887. (Cited on page 30.)

[Choi et al., 2005] Choi, M., Kim, R., Nam, M.-R., and Kim, H. (2005). Fusion of multispectral and panchromatic satellite images using the curvelet transform. *IEEE Geoscience and Remote Sensing Letters*, 2(2):136–140. (Cited on page 30.)

[Chou, 1991] Chou, K. (1991). *A stochastic modeling approach to multiscale signal processing.* PhD thesis, Citeseer. (Cited on page 67.)

[Chou et al., 1993] Chou, K., Golden, S., and Willsky, A. (1993). Multiresolution stochastic models, data fusion, and wavelet transforms. *Signal Processing*, 34(3):257–282. (Cited on page 67.)

[Chou and Brown, 1990] Chou, P. and Brown, C. (1990). The theory and practice of Bayesian image labeling. *International Journal of Computer Vision*, 4(3):185–210. (Cited on page 47.)

[Clark and Yuille, 2013] Clark, J. and Yuille, A. (2013). *Data fusion for sensory information processing systems*, volume 105. Springer Science & Business Media. (Cited on page 28.)

[Comer and Delp, 1999] Comer, M. and Delp, E. (1999). Segmentation of textured images using a multiresolution Gaussian autoregressive model. *IEEE Transactions on Image Processing*, 8(3):408–420. (Cited on page 67.)

[Coppin et al., 2004] Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., and Lambin, E. (2004). Review articledigital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, 25(9):1565–1596. (Cited on page 32.)

[Cui and Datcu, 2012] Cui, S. and Datcu, M. (2012). Statistical wavelet subband modeling for multi-temporal SAR change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(4):1095–1109. (Cited on page 35.)

[David-Forney, 1973] David-Forney, G. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278. (Cited on pages 54 and 64.)

[Dawid, 1992] Dawid, A. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2(1):25–36. (Cited on page 64.)

[Demirel and Anbarjafari, 2010] Demirel, H. and Anbarjafari, G. (2010). Satellite image resolution enhancement using complex wavelet transform. *IEEE Geoscience and Remote Sensing Letters*, 7(1):123–126. (Cited on page 30.)

[Deng et al., 2008] Deng, J., Wang, K., Deng, Y., and Qi, G. (2008). PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing*, 29(16):4823–4838. (Cited on page 34.)

[Devijver, 1988] Devijver, P. (1988). Image segmentation using causal Markov random field models. In *Pattern Recognition*, pages 131–143. Springer. (Cited on page 51.)

[Dousset and Gourmelon, 2003] Dousset, B. and Gourmelon, F. (2003). Satellite multi-sensor data analysis of urban surface temperatures and landcover. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(1):43–54. (Cited on page 30.)

[Du et al., 2013] Du, P., Liu, S., Xia, J., and Zhao, Y. (2013). Information fusion techniques for change detection from multi-temporal remote sensing images. *Information Fusion*, 14(1):19–27. (Cited on page 34.)

[Esbensen et al., 1992] Esbensen, K., Geladi, P., and Grahn, H. (1992). Strategies for multivariate image regression. *Chemometrics and Intelligent Laboratory Systems*, 14(1):357–374. (Cited on page 34.)

[Faugeras and Berthod, 1981] Faugeras, O. and Berthod, M. (1981). Improving consistency and reducing ambiguity in stochastic labeling: An optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4):412–424. (Cited on page 48.)

[Fieguth, 2010] Fieguth, P. (2010). *Statistical image processing and multidimensional modeling*. Springer Science & Business Media. (Cited on page 50.)

[Fieguth et al., 1998] Fieguth, P., Karl, W., and Willsky, A. (1998). Efficient multiresolution counterparts to variational methods for surface reconstruction. *Computer Vision and Image Understanding*, 70(2):157–176. (Cited on page 67.)

[Figueiredo and Jain, 2002] Figueiredo, M. and Jain, A. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396. (Cited on pages 82 and 84.)

[Forster et al., 2004] Forster, B., Van De Ville, D., Berent, J., Sage, D., and Unser, M. (2004). Complex wavelets for extended depth-of-field: A new method for the fusion of multichannel microscopy images. *Microscopy Research and Technique*, 65(1-2):33–42. (Cited on page 30.)

[Frost et al., 1982] Frost, V., Stiles, J., Shanmugan, K., and Holtzman, J. (1982). A model for radar images and its application to adaptive digital filtering of multiplicative noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):157–166. (Cited on page 33.)

[Fukuda and Hirosawa, 1999] Fukuda, S. and Hirosawa, H. (1999). Smoothing effect of wavelet-based speckle filtering: The haar basis case. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):1168–1172. (Cited on page 33.)

[Fukunaga, 2013] Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Academic press. (Cited on page 101.)

[Gadallah et al., 2000] Gadallah, F., Csillag, F., and Smith, E. (2000). Destriping multisensor imagery with moment matching. *International Journal of Remote Sensing*, 21(12):2505–2511. (Cited on page 32.)

[Gamba and Dell'Acqua, 2003] Gamba, P. and Dell'Acqua, F. (2003). Increased accuracy multiband urban classification using a neuro-fuzzy classifier. *International Journal of Remote Sensing*, 24(4):827–834. (Cited on page 34.)

[Gamba et al., 2006] Gamba, P., Dell'Acqua, F., and Lisini, G. (2006). Change detection of multitemporal SAR data in urban areas combining feature-based and pixel-based techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 44(10):2820–2827. (Cited on page 31.)

[Gamba et al., 2011] Gamba, P., Lisini, G., Tomás, L., Almeida, C., and Fonseca, L. (2011). Joint VHR-LIDAR classification framework in urban areas using a priori knowledge and post processing shape optimization. In *IEEE Urban Remote Sensing Event (JURSE)*, pages 93–96. (Cited on page 30.)

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741. (Cited on pages 42 and 78.)

[Goffe et al., 1994] Goffe, W., Ferrier, G., and Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, 60(1):65–99. (Cited on page 46.)

[Goldberg and Holland, 1988] Goldberg, D. and Holland, J. (1988). Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99. (Cited on page 48.)

[Gordon and Morel, 2012] Gordon, H. and Morel, A. (2012). *Remote assessment of ocean color for interpretation of satellite visible imagery: A review*, volume 4. Springer Science & Business Media. (Cited on page 32.)

[Graffigne et al., 1995] Graffigne, C., Heitz, F., Perez, P., Preteux, F., Sigelle, M., and Zerubia, J. (1995). Hierarchical Markov random field models applied to image analysis: a review. In *International Symposium on Optical Science, Engineering, and Instrumentation (SPIE)*, pages 2–17. International Society for Optics and Photonics. (Cited on page 30.)

[Gueguen et al., 2010] Gueguen, L., Cui, S., Schwarz, G., and Datcu, M. (2010). Multitemporal analysis of multisensor data: Information theoretical approaches. In *IEEE Journal of Geoscience and Remote Sensing Symposium (IGARSS)*, pages 2559–2562. (Cited on page 35.)

[Hachicha et al., 2011] Hachicha, S., Deledalle, C., Chaabane, F., and Tupin, F. (2011). Multi-temporal SAR classification according to change detection operators. In *International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multi-Temp)*, pages 133–136. (Cited on page 35.)

[Hadamard, 1923] Hadamard, J. (1923). *Lectures on Cauchy's problem in linear partial differential equations*. Courier Corporation. (Cited on page 40.)

[Hall and Llinas, 2001] Hall, D. and Llinas, J. (2001). *Multisensor data fusion*. CRC press. (Cited on page 30.)

[Hame et al., 1998] Hame, T., Heiler, I., and San Miguel-Ayanz, J. (1998). An unsupervised change detection and recognition system for forestry. *International Journal of Remote Sensing*, 19(6):1079–1099. (Cited on page 34.)

[Healey et al., 2005] Healey, S., Cohen, W., Zhiqiang, Y., and Krankina, O. (2005). Comparison of Tasseled Cap-based Landsat data structures for use in forest disturbance detection. *Remote Sensing of Environment*, 97(3):301–310. (Cited on page 34.)

[Hedhli et al., 2015] Hedhli, I., Moser, G., Serpico, S., and Zerubia, J. (2015). New hierarchical joint classification method of SAR-optical multiresolution remote sensing data. In *IEEE European Signal Processing Conference*. (Cited on page 30.)

[Hedhli et al., 2014a] Hedhli, I., Moser, G., Zerubia, J., and Serpico, S. (2014a). Fusion of multitemporal and multiresolution remote sensing data and application to natural disasters. In *IEEE Geoscience and Remote Sensing Symposium*, pages 207–210. (Cited on page 32.)

[Hedhli et al., 2014b] Hedhli, I., Moser, G., Zerubia, J., and Serpico, S. (2014b). New cascade model for hierarchical joint classification of multitemporal, multiresolution and multisensor remote sensing data. In *IEEE International Conference on Image Processing (ICIP), 2014*, pages 5247–5251. (Cited on pages 30, 31, 32 and 36.)

[Henderson et al., 1998] Henderson, F., Lewis, A., et al. (1998). *Principles and applications of imaging radar. Manual of remote sensing, volume 2.* John Wiley and sons. (Cited on pages 20 and 32.)

[Hoberg et al., 2015] Hoberg, T., Rottensteiner, F., Queiroz Feitosa, R., and Heipke, C. (2015). Conditional random fields for multitemporal and multiscale classification of optical satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):659–673. (Cited on pages 31 and 36.)

[Holland, 1975] Holland, J. (1975). Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. (Cited on page 48.)

[Huang and Song, 2012] Huang, B. and Song, H. (2012). Spatiotemporal reflectance fusion via sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10):3707–3716. (Cited on page 28.)

[Ibáñez and Simó, 2003] Ibáñez, M. and Simó, A. (2003). Parameter estimation in Markov random field image modeling with imperfect observations. a comparative study. *Pattern recognition letters*, 24(14):2377–2389. (Cited on pages 92 and 131.)

[Inglada and Mercier, 2007] Inglada, J. and Mercier, G. (2007). A new statistical similarity measure for change detection in multitemporal SAR images and its extension to multiscale change analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 45(5):1432–1445. (Cited on page 35.)

[Iqbal et al., 2013] Iqbal, M., Ghafoor, A., and Siddiqui, A. (2013). Satellite image resolution enhancement using dual-tree complex wavelet transform and nonlocal means. *IEEE Geoscience and Remote Sensing Letters*, 10(3):451–455. (Cited on page 30.)

[Irving et al., 1997] Irving, W., Fieguth, P., and Willsky, A. (1997). An overlapping tree approach to multiscale stochastic modeling and estimation. *IEEE Transactions on Image Processing*, 6(11):1517–1529. (Cited on page 67.)

[Ishikawa, 2003] Ishikawa, H. (2003). Exact optimization for Markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1333–1336. (Cited on page 48.)

[JHA and Unni, 1994] JHA, S. and Unni, N. (1994). Digital change detection of forest conversion of a dry tropical indian forest region. *International Journal of Remote Sensing*, 15(13):2543–2552. (Cited on page 34.)

[Johnson and Kasischke, 1998] Johnson, R. and Kasischke, E. (1998). Change vector analysis: a technique for the multispectral monitoring of land cover and condition. *International Journal of Remote Sensing*, 19(3):411–426. (Cited on page 31.)

[Jolion and Rosenfeld, 2012] Jolion, J.-M. and Rosenfeld, A. (2012). *A pyramid framework for early vision: multiresolutional computer vision*, volume 251. Springer Science & Business Media. (Cited on page 60.)

[Jordan, 2003] Jordan, M. (2003). *An introduction to probabilistic graphical models*. (Cited on page 54.)

[Kato, 1995] Kato, J. (1995). *Modelisations markoviennes multiresolutions en vision par ordinateur. Application a la segmentation d'images SPOT*. PhD thesis, University of Nice Sophia Antipolis, France. (Cited on pages 30 and 80.)

[Kato et al., 1996] Kato, Z., Berthod, M., and Zerubia, J. (1996). A hierarchical Markov random field model and multitemperature annealing for parallel image classification. *Graphical Models and Image Processing*, 58(1):18–37. (Cited on pages 3, 67, 69, 118 and 139.)

[Kato and Zerubia, 2012] Kato, Z. and Zerubia, J. (2012). *Markov random fields in image segmentation*. Now Pub. (Cited on pages 46, 64 and 78.)

[Kersten et al., 2005] Kersten, P., Lee, J., and Ainsworth, T. (2005). A comparison of change detection statistics in POLSAR images. Technical report, DTIC Document. (Cited on page 35.)

[Kuan et al., 1985] Kuan, D., Sawchuk, A., Strand, T., and Chavel, P. (1985). Adaptive noise smoothing filter for images with signal-dependent noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):165–177. (Cited on page 33.)

[Kurum et al., 2011] Kurum, M., Lang, R., O'Neill, P., Joseph, A., Jackson, T., and Cosh, M. (2011). A first-order radiative transfer model for microwave radiometry of forest canopies at l-band. *IEEE Transactions on Geoscience and Remote Sensing*, 49(9):3167–3179. (Cited on page 32.)

[Laferté et al., 2000] Laferté, J.-M., Pérez, P., and Heitz, F. (2000). Discrete Markov image modeling and inference on the quadtree. *IEEE Transactions on Image Processing*, 9(3):390–404. (Cited on pages 3, 30, 51, 60, 64, 65, 67, 68, 69, 70, 78, 79, 80, 85, 90, 91, 94, 95, 96, 106, 107, 123, 125, 126, 127, 128, 136, 137 and 139.)

[Lambin and Strahlers, 1994] Lambin, E. and Strahlers, A. (1994). Change-vector analysis in multitemporal space: a tool to detect and categorize land-cover change processes using high temporal-resolution satellite data. *Remote Sensing of Environment*, 48(2):231–244. (Cited on page 34.)

[Landgrebe, 2005] Landgrebe, D. (2005). *Signal theory methods in multispectral remote sensing*, volume 29. John Wiley & Sons. (Cited on page 10.)

[Lauritzen, 1996] Lauritzen, S. (1996). *Graphical models*. Oxford University Press. (Cited on pages 49 and 51.)

[Lauritzen et al., 1990] Lauritzen, S., Dawid, A., Larsen, B., and Leimer, H. (1990). Independence properties of directed Markov fields. *Networks*, 20(5):491–505. (Cited on pages 49 and 51.)

[Le Moigne et al., 2011] Le Moigne, J., Netanyahu, N., and Eastman, R. (2011). *Image registration for remote sensing*. Cambridge University Press. (Cited on page 32.)

[Li et al., 1995] Li, H., Manjunath, B., and Mitra, S. (1995). Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing*, 57(3):235–245. (Cited on page 28.)

[Li et al., 2011] Li, H.-C., Hong, W., Wu, Y.-R., and Fan, P.-Z. (2011). On the empirical-statistical modeling of SAR images with generalized Gamma distribution. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):386–397. (Cited on page 102.)

[Li, 2009] Li, S. (2009). *Markov random field modeling in image analysis*. Springer Science & Business Media. (Cited on pages 41, 46 and 78.)

[Lillesand et al., 2014] Lillesand, T., Kiefer, R., and Chipman, J. (2014). *Remote sensing and image interpretation*. John Wiley & Sons. (Cited on pages 10 and 13.)

[Liu et al., 2012] Liu, Z.-G., Dezert, J., Mercier, G., and Pan, Q. (2012). Dynamic evidential reasoning for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5):1955–1967. (Cited on page 32.)

[Luettgen et al., 1994] Luettgen, M., Karl, W., and Willsky, A. (1994). Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Transactions on Image Processing*, 3(1):41–64. (Cited on pages 51 and 67.)

[Madanian et al., 2014] Madanian, M., Soffianian, A., and Hajian, A. (2014). Change detection through four techniques using multi-temporal Landsat thematic mapper data: A case study on Falavarjan area, Isfahan, Iran. *Journal of Environmental Informatics*, 23(2):58–66. (Cited on page 34.)

[Maes et al., 1997] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198. (Cited on page 32.)

[Malila, 1980] Malila, W. (1980). Change vector analysis: an approach for detecting forest changes with Landsat. In *LARS Symposia*, page 385. (Cited on page 31.)

[Mallat, 2008] Mallat, S. (2008). *A wavelet tour of signal processing*. Academic press, 3 edition. (Cited on pages 74 and 87.)

[Manandhar et al., 2009] Manandhar, R., Odeh, I. O., and Ancev, T. (2009). Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement. *Remote Sensing*, 1(3):330–344. (Cited on page 36.)

[Marchesi et al., 2010] Marchesi, S., Bovolo, F., and Bruzzone, L. (2010). A context-sensitive technique robust to registration noise for change detection in VHR multispectral images. *IEEE Transactions on Image Processing*, 19(7):1877–1889. (Cited on page 34.)

[Marroquin et al., 1987] Marroquin, J., Mitter, S., and Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89. (Cited on page 64.)

[Martinis and Twele, 2010] Martinis, S. and Twele, A. (2010). A hierarchical spatio-temporal Markov model for improved flood mapping using multi-temporal X-band SAR data. *Remote Sensing*, 2(9):2240–2258. (Cited on page 35.)

[Mas and Flores, 2008] Mas, J. and Flores, J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*, 29(3):617–663. (Cited on page 30.)

[Melgani et al., 2002] Melgani, F., Moser, G., and Serpico, S. (2002). Unsupervised change-detection methods for remote-sensing images. *Optical Engineering*, 41(12):3288–3297. (Cited on page 34.)

[Melgani and Serpico, 2003] Melgani, F. and Serpico, S. (2003). A Markov random field approach to spatio-temporal contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 41(11):2478–2487. (Cited on pages 31, 36, 90, 92, 94, 95 and 96.)

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092. (Cited on page 46.)

[Miller et al., 1995] Miller, D., Kaminsky, E., and Rana, S. (1995). Neural network classification of remote-sensing data. *Computers & Geosciences*, 21(3):377–386. (Cited on page 36.)

[Milne, 1988] Milne, A. (1988). Change direction analysis using Landsat imagery: a review of methodology. In *Geoscience and Remote Sensing Symposium, 1988*, volume 1, pages 541–544. (Cited on page 32.)

[Molinier and Rauste, 2007] Molinier, M. and Rauste, Y. (2007). Comparison and evaluation of polarimetric change detection techniques in aerial SAR data. In *IGARSS*, pages 2386–2389. (Cited on page 35.)

[Montopoli et al., 2013] Montopoli, M., Cimini, D., Lamantea, M., Herzog, M., Graf, H., and Marzano, F. (2013). Microwave radiometric remote sensing of volcanic ash clouds from space: Model and data analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 51(9):4678–4691. (Cited on page 32.)

[Moser et al., 2011] Moser, G., Angiati, E., and Serpico, S. (2011). Multiscale unsupervised change detection on optical images by Markov random fields and wavelets. *IEEE Geoscience and Remote Sensing Letters*, 8(4):725–729. (Cited on pages 34 and 130.)

[Muñoz-Marí et al., 2010] Muñoz-Marí, J., Bovolo, F., Gómez-Chova, L., Bruzzone, L., and Camp-Valls, G. (2010). Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(8):3188–3197. (Cited on page 30.)

[Neelamani, 2003] Neelamani, R. (2003). Inverse problems in image processing. *Rice University, Texas*. (Cited on page 40.)

[Nelson, 1983] Nelson, R. (1983). Detecting forest canopy change due to insect activity using Landsat MSS. *Photogrammetric Engineering and Remote Sensing*, 49(9):1303–1314. (Cited on page 32.)

[Nencini et al., 2007] Nencini, F., Garzelli, A., Baronti, S., and Alparone, L. (2007). Remote sensing image fusion using the curvelet transform. *Information Fusion*, 8(2):143–156. (Cited on page 30.)

[Nguyen et al., 2011] Nguyen, N., Nasrabadi, N., and Tran, T. (2011). Robust multi-sensor classification via joint sparse representation. In *International Conference on Information Fusion*, pages 1–8. (Cited on page 30.)

[Parisi, 1988] Parisi, G. (1988). *Statistical field theory*. Addison-Wesley pub. (Cited on page 40.)

[Pérez, 1993] Pérez, P. (1993). *Champs markoviens et analyse multirésolution de l'image: application à l'analyse du mouvement*. PhD thesis. (Cited on pages 30 and 52.)

[Pérez et al., 2000] Pérez, P., Chardin, A., and Laferté, J.-M. (2000). Noniterative manipulation of discrete energy-based models for image analysis. *Pattern Recognition*, 33(4):573–586. (Cited on pages 51, 60 and 81.)

[Perez and Heitz, 1996] Perez, P. and Heitz, F. (1996). Restriction of a Markov random field on a graph and multiresolution statistical image modeling. *IEEE Transactions on Information Theory*, 42(1):180–190. (Cited on page 50.)

[Petitjean et al., 2011] Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693. (Cited on pages 31 and 32.)

[Pickard, 1980] Pickard, D. (1980). Unilateral markov fields. *Advances in Applied Probability*, pages 655–671. (Cited on page 118.)

[Piella, 2003] Piella, G. (2003). Adaptive wavelets and their applications to image fusion and compression. (Cited on page 30.)

[Pohl and van Genderen, 2014] Pohl, C. and van Genderen, J. (2014). Remote sensing image fusion: an update in the context of digital earth. *International Journal of Digital Earth*, 7(2):158–172. (Cited on page 28.)

[Pohl, 1998] Pohl, C.and Van Genderen, J. (1998). Review article multisensor image fusion in remote sensing: concepts, methods and applications. *International Journal of Remote Sensing*, 19(5):823–854. (Cited on page 30.)

[Prendes, 2015] Prendes, J. (2015). *New statistical modeling of multi-sensor images with application to change detection*. PhD thesis, Université Paris-Sud, France. (Cited on page 30.)

[Rabiner et al., 1989] Rabiner, L., Lee, C., Juang, B., and Wilpon, J. (1989). HMM clustering for connected word recognition. In *International Conference on Acoustics, Speech, and Signal Processing, 1989.*, pages 405–408. (Cited on page 53.)

[Razlighi et al., 2009] Razlighi, Q., Kehtarnavaz, N., and Nosratinia, A. (2009). Computation of image spatial entropy using quadrilateral Markov random field. *IEEE Transactions on Image Processing*, 18(12):2629–2639. (Cited on pages 51, 125 and 131.)

[Richards, 2013] Richards, J. (2013). *Remote sensing digital image analysis*. Springer, 5 edition. (Cited on page 33.)

[Roberts et al., 2008] Roberts, J., Van Aardt, J., and Ahmed, F. (2008). Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1). (Cited on page 30.)

[Rockinger, 1996] Rockinger, O. (1996). Pixel-level fusion of image sequences using wavelet frames. In *Proceedings of the 16th Leeds applied shape research workshop, Leeds University Press*. Citeseer. (Cited on page 30.)

[Rosa et al., 2015] Rosa, R., Fernandes, D., Nogueira, J. B., and Wimmer, C. (2015). Automatic change detection in multitemporal X-and P-band SAR images using Gram-Schmidt process. In *Geoscience and Remote Sensing Symposium*, pages 2797–2800. (Cited on page 34.)

[Rosenfeld et al., 1976] Rosenfeld, A., Hummel, R., and Zucker, S. (1976). Scene labeling by relaxation operations. *IEEE Transactions on Systems, Man and Cybernetics*, (6):420–433. (Cited on page 48.)

[Savitzky and Golay, 1964] Savitzky, A. and Golay, M. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639. (Cited on page 33.)

[Schowengerdt, 2006] Schowengerdt, R. (2006). *Remote sensing: models and methods for image processing*. Academic press. (Cited on page 33.)

[Serpico et al., 2012] Serpico, S., Dellepiane, S., Boni, G., Moser, G., Angiati, E., and Rudari, R. (2012). Information extraction from remote sensing images for flood monitoring and damage evaluation. *Proceedings of the IEEE*, 100(10):2946–2970. (Cited on page 5.)

[Serpico and Moser, 2006] Serpico, S. and Moser, G. (2006). Weight parameter optimization by the Ho–Kashyap algorithm in mrf models for supervised image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 44(12):3695–3705. (Cited on pages 93 and 131.)

[Serra et al., 2003] Serra, P., Pons, X., and Sauri, D. (2003). Post-classification change detection with data from different sensors: some accuracy considerations. *International Journal of Remote Sensing*, 24(16):3311–3340. (Cited on page 36.)

[Shah et al., 2008] Shah, V., Younan, N., and King, R. (2008). An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1323–1335. (Cited on page 30.)

[Singh, 1989] Singh, A. (1989). Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003. (Cited on page 31.)

[Singh et al., 2014] Singh, P., Kato, Z., and Zerubia, J. (2014). A multilayer markovian model for change detection in aerial image pairs with large time differences. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 924–929. (Cited on page 31.)

[Sowell, 1992] Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of econometrics*, 53(1):165–188. (Cited on page 34.)

[Storvik et al., 2009] Storvik, B., Storvik, G., and Fjortoft, R. (2009). On the combination of multisensor data using meta-Gaussian distributions. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2372–2379. (Cited on pages 101, 106, 107, 108 and 109.)

[Stroppiana et al., 2015] Stroppiana, D., Azar, R., Calo, F., Pepe, A., Imperatore, P., Boschetti, M., Silva, J., Brivio, P., and Lanari, R. (2015). Remote sensing

of burned area: A fuzzy-based framework for joint processing of optical and microwave data. In *IEEE, Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1409–1412. (Cited on page 30.)

[Swain, 1978] Swain, P. (1978). Bayesian classification in a time-varying environment. (Cited on page 36.)

[Sziranyi and Shadaydeh, 2014] Sziranyi, T. and Shadaydeh, M. (2014). Segmentation of remote sensing images using similarity-measure-based fusion-mrf model. *IEEE Geoscience and Remote Sensing Letters*, 11(9):1544–1548. (Cited on page 31.)

[Tokola et al., 1999] Tokola, T., Löfman, S., and Erkkilä, A. (1999). Relative calibration of multitemporal Landsat data for forest cover change detection. *Remote Sensing of Environment*, 68(1):1–11. (Cited on page 32.)

[Touzi et al., 2009] Touzi, R., Deschamps, A., and Rother, G. (2009). Phase of target scattering for wetland characterization using polarimetric C-band SAR. *IEEE Transactions on Geoscience and Remote Sensing*, 47(9):3241–3261. (Cited on page 35.)

[Udelhoven, 2011] Udelhoven, T. (2011). Timestats: A software tool for the retrieval of temporal patterns from global satellite archives. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(2):310–317. (Cited on page 33.)

[Van de Voorde et al., 2007] Van de Voorde, T., De Genst, W., and Canters, F. (2007). Improving pixel-based VHR land-cover classifications of urban areas with post-classification techniques. *Photogrammetric Engineering and Remote Sensing*, 73(9):1017. (Cited on page 36.)

[Viterbi, 1967] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269. (Cited on page 54.)

[Vivone et al., 2015] Vivone, G., Alparone, L., Chanussot, J., Dalla M., M., G., A., Licciardi, G., Restaino, R., and Wald, L. (2015). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586. (Cited on page 28.)

[Voisin, 2012] Voisin, A. (2012). *Classification supervisée d'images d'observation de la Terre à haute résolution par utilisation de méthodes markoviennes*. PhD thesis, University of Nice Sophia Antipolis, France. (Cited on pages 3, 30, 67, 70 and 71.)

[Voisin et al., 2012] Voisin, A., Krylov, V., Moser, G., Serpico, S., and Zerubia, J. (2012). Multichannel hierarchical image classification using multivariate copulas. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics. (Cited on pages 30, 85 and 91.)

[Voisin et al., 2014] Voisin, A., Krylov, V., Moser, G., Serpico, S., and Zerubia, J. (2014). Supervised Classification of Multi-sensor and Multi-resolution Remote Sensing Images with a Hierarchical Copula-based Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3346–3358. (Cited on pages 51, 60, 70, 71, 81, 101, 106, 107, 108, 109, 125, 126, 127 and 128.)

[Wald, 1999] Wald, L. (1999). Some terms of reference in data fusion. *IEEE Transactions on geoscience and remote sensing*, 37(3):1190–1193. (Cited on page 28.)

[Waltz et al., 1990] Waltz, E., Llinas, J., et al. (1990). *Multisensor data fusion*, volume 685. Artech house Boston. (Cited on pages 28 and 30.)

[Wang and Liang, 2014] Wang, D. and Liang, S. (2014). Improving LAI mapping by integrating modis and cyclopes lai products using optimal interpolation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(2):445–457. (Cited on page 28.)

[Waske and van der Linden, 2008] Waske, B. and van der Linden, S. (2008). Classifying multilevel imagery from SAR and optical sensors by decision fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1457–1466. (Cited on page 30.)

[Weber et al., 2012] Weber, J., Petitjean, F., and Gancarski, P. (2012). Towards efficient satellite image time series analysis: Combination of dynamic time warping and quasi-flat zones. In *IEEE Geoscience and Remote Sensing Symposium*, pages 4387–4390. (Cited on page 32.)

[Whittaker, 2009] Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Pub. (Cited on pages 49, 51 and 53.)

[Willsky, 2002] Willsky, A. (2002). Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458. (Cited on pages 30, 51 and 64.)

[Yang et al., 2007] Yang, B., Li, S., and Sun, F. (2007). Image fusion using nonsubsampled contourlet transform. In *IEEE International Conference on Image and Graphics*, pages 719–724. (Cited on page 30.)

[Younes, 1988] Younes, L. (1988). Estimation and annealing for Gibbsian fields. In *Annales de l'IHP Probabilités et statistiques*, volume 24, pages 269–294. (Cited on page 131.)

[Yousefi and Kehtarnavaz, 2011] Yousefi, S. and Kehtarnavaz, N. (2011). Generating symmetric causal Markov random fields. *Electronics letters*, 47(22):1. (Cited on page 51.)

[Yousefi et al., 2013] Yousefi, S., Kehtarnavaz, N., and Cao, Y. (2013). Computationally tractable stochastic image modeling based on symmetric markov mesh

random fields. *IEEE Transactions on Image Processing*, 22(6):2192–2206. (Cited on pages 125 and 131.)

[Zhang and Hong, 2005] Zhang, Y. and Hong, G. (2005). An IHS and wavelet integrated approach to improve pan-sharpening visual quality of natural colour IKONOS and QuickBird images. *Information Fusion*, 6(3):225–234. (Cited on page 30.)