

TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Mensch-Maschine-Kommunikation

Heterogeneous Data Mining of Earth Observation Archives: Integration and Fusion of Images, Maps, and In-situ Data

Kevin Alonso Gonzalez

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr.-Ing. Th. Eibert

Prüfer der Dissertation: 1. Prof. Dr.-Ing. habil. G. Rigoll
2. Prof. Dr.-Ing. habil. M. Datcu

Die Dissertation wurde am 28.09.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 13.02.2017 angenommen.

Abstract

The amount of Earth Observation (EO) data is in constant growth due to the proliferation of Earth Observation (EO) missions in space and the continuous evolution of their instruments. The catalog of EO products is enriched by the high diversity of the imaging sensors. Along with the imagery data, EO products accommodate different metadata containing several parameters related to the image, the satellite and the instrument. In addition, we can also consider as EO information the data derived from third party systems not directly related with satellite EO products. An example are the widely spread Geographical Information Systems (GIS), which store map information that can be used for different purposes during EO image analysis. In this heterogeneous Big Data scenario, the main challenge is not only to provide better and more efficient algorithms, but also to design and implement tools that allow a greater exploitation of the available information.

In line with the challenge, this thesis focuses on the integration, mining and exploitation of a wide range of EO heterogeneous data in order to efficiently extract valuable information for a better understanding of EO image content. The presented Heterogeneous Data Mining (HDM) system prototype overcomes the limitations of previous systems by including multispectral images, Synthetic Aperture Radar (SAR) images, and digital maps in an accelerated active learning algorithm. The learning stage of the algorithm is based on naive Bayes Classifiers which make use of posterior probabilities of a user-defined semantic label given a query image. This accelerated algorithm opens new ways for knowledge-driven information mining systems to Big Data scenarios. In conjunction with the learning algorithm, the Heterogeneous Data Mining (HDM) concept also contains a probabilistic search method based on the distances between the elements being used for the calculation of the posterior probabilities and image Bag of Words (BoW) in the database.

Additionally, a multilayer system for heterogeneous geospatial data analytics is introduced. The system manages data from the source and performs several transformations in order to enable the integration of remote sensing, cartographic and in-situ data. Specifically, we use as in-situ data the results from the Land

Use/Cover Area frame Survey (LUCAS). This survey monitors the state and change dynamics in land use and cover in the European Union. The system is tested in different scenarios and used for the development of a data mining methodology to filter and validate land cover changes recorded in multitemporal in-situ surveys.

Our final effort focuses on the visual exploitation of the integrated heterogeneous EO data. By combining the results obtained from automatic analysis methodologies with interactive visualization tools, one can navigate and understand the EO data more efficiently.

Zusammenfassung

Aufgrund der Zunahme von Erdbeobachtungsmissionen im Weltraum und der ständigen Weiterentwicklung der zugehörigen Instrumente wächst die Menge an verfügbaren Erdbeobachtungsdaten ständig. Das Spektrum der Erderkundungsprodukte wird dabei durch die große Vielfalt von bilderzeugenden Sensoren bereichert. Neben den eigentlichen Bilddaten umfassen die Datenprodukte der Erderkundung verschiedene Metadaten mit mehreren Parametern bezüglich Aufnahme, Satellit und Instrument. Darüber hinaus können wir als Informationen der Erdbeobachtung auch Daten von Drittsystemen betrachten, die nicht direkt zu Erdbeobachtungsprodukten von Satelliten gehören. Ein Beispiel dafür sind die weitverbreiteten Geographischen Informationssysteme (GIS), die Kartierungen aufsammeln, die während der Analyse von Erdbeobachtungsbildern für verschiedenen Zwecke genutzt werden können. In diesem heterogenen Big-Data-Umfeld besteht die hauptsächliche Herausforderung nicht nur darin, bessere und effizientere Algorithmen bereitzustellen, sondern auch Werkzeuge zu entwerfen und zu installieren, die eine breitere Nutzung der verfügbaren Daten erlauben.

Entsprechend dieser Herausforderung liegt der Schwerpunkt dieser Dissertation auf der Integration, dem Mining und der Nutzung eines breiten Spektrums von unterschiedlichen Erdbeobachtungsdaten, um für ein besseres Verständnis des Inhalts von Erdbeobachtungsbildern wertvolle Informationen effizient zu extrahieren. Der hier vorgestellte Systemprototyp für Heterogenes Data-Mining (HDM) überwindet die Einschränkungen früherer Systeme durch die Integration von multispektralen Bildern, von Radarbildern mit synthetischer Apertur (SAR) sowie von digitalen Karten in einem beschleunigten Algorithmus für Aktives Lernen. Die Lernphase des Algorithmus beruht auf Naiven Bayes-Klassifikatoren, die A-posteriori-Wahrscheinlichkeiten eines nutzerdefinierten semantischen Labels nach einer Bildabfrage nutzen. Dieser beschleunigte Algorithmus eröffnet neue Optionen von wissensbasierten Information-Mining-Systeme für Big-Data-Szenarien. Zusammen mit seinem Lern-Algorithmus beinhaltet das HDM-Konzept eine probabilistische Suchmethode, beruhend auf den Abständen zwischen den Elementen, die für die Berech-

nung der A-posteriori-Wahrscheinlichkeiten sowie der Bild-Bags-of-Words in der Datenbasis benutzt werden.

Zusätzlich führen wir ein Mehrebenen-System zur Analyse von heterogenen raumbezogenen Geo-Daten ein. Das System verwaltet Daten ab deren Quelle und führt mehrere Transformationen durch, um die Integration von Fernerkundungs-, von kartographischen und von In-situ-Daten zu ermöglichen. Insbesondere nutzen wir als In-situ-Daten die Ergebnisse aus der LUCAS-Untersuchung. Diese Untersuchung überprüft den Stand und die Änderungsdynamik der Landnutzung und Landbedeckung in der Europäischen Union. Unser System wurde für verschiedenen Szenarien getestet und für die Entwicklung einer Big-Data-Strategie genutzt, um Änderungen der Landbedeckung aus aufgezeichneten multitemporalen In-situ-Untersuchungen zu identifizieren und zu verifizieren.

Letztlich möchten wir die visuelle Nutzbarmachung von integrierten heterogenen Erdbeobachtungsdaten erreichen. Durch die Kombination von Ergebnissen aus automatisierten Analysemethoden mit denen von interaktiven Visualisierungswerkzeugen kann man Erdbeobachtungsdaten effizienter durchsuchen und verstehen.

Acknowledgments

I would like to thank to Prof. Mihai Datcu for the opportunity to conduct my research at the German Aerospace Center (DLR) and his clear guidance. I would also thank Prof. Gerhard Rigoll for his help and the thorough review of the thesis. I am also grateful to Prof. Thomas Eibert for chairing the examination board.

My sincere gratitude to my colleagues from the Photogrammetry and Image Analysis department, specially from the image analysis group including Daniela Espinoza-Molina, Ambar Murillo Montes de Oca, Wei Yao, Deniz Cagatay, Reza Bahmanyar, Gottfried Schwarz, Octavian Dumitru, and Shiyong Cui for their help during the preparation for the PhD defense and their support throughout these years. Also thanks to the Science Slam troupe for the provided warm welcome and necessary distraction during the initial PhD steps. My gratitude to Esteban Aguilera, Lotte Aguilera, Kuba Bienartz, Kasia Aszajkowska, Jagmal Singh, Russel Que, and Claas Grohnfeldt for their friendship, and above all for helping me meet and better know Muriel Pinheiro whose unconditional support "dramatically" help me during this endeavor. My appreciation also to the colleges and friends from Spain and in particular from Vicomtech-IK4 because even though far, they were always close. Lastly, my deepest thanks to my family for their unconditional support in every possible aspect.

Finally, I would like to extent my acknowledgment to the German Aerospace Center (DLR) and the German Academic Exchange Service (DAAD) for their financial aid to my PhD studies.

Oberpfaffenhofen, May 2017
Kevin Alonso Gonzalez

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Goals	2
1.3	Scope	2
1.4	Contributions	3
1.5	Thesis Overview	4
2	Heterogeneity in EO Data and Information Retrieval Systems	5
2.1	Heterogeneity in EO Data	5
2.1.1	Optical	5
2.1.2	Panchromatic	6
2.1.3	Multispectral	6
2.1.4	Hyperspectral	8
2.1.5	Synthetic Aperture Radar	8
2.1.6	Cartographic	9
2.1.7	LUCAS In-situ Survey	10
2.2	Information Retrieval Systems	11
2.3	Content-Based Image Retrieval	12
2.4	Semantic Gap	14
2.5	User Feedback and Interaction	15
2.6	Image Retrieval in EO	16
2.6.1	Colour Composites and Spectral Indexes	16
2.6.1.1	Color Composites	17
2.6.1.2	Spectral Indexes	18
2.6.2	Texture Features	20
2.6.2.1	Haralick Features	21
2.6.2.2	Gabor Feature Descriptor	21
2.6.2.3	Weber Local Descriptor	22

2.6.3	Scale-Invariant Feature Transform	22
2.6.4	Semantic Level Descriptors	23
2.6.4.1	Bag of Words	23
2.6.4.2	Latent Dirichlet Allocation	24
2.7	Image Information Mining Systems	24
2.8	Visual Analytics	25
3	Knowledge-driven Heterogeneous Data Mining	27
3.1	Knowledge-driven Information Mining	28
3.2	Heterogeneous Data Mining Concept	28
3.3	Feature Extraction	30
3.4	Clustering and Bag-of-Words	31
3.5	Machine Learning	32
3.5.1	Assumption of Feature Probability Independence	34
3.5.2	Assumption of Posterior Probability Independence	34
3.6	Probabilistic Retrieval	35
3.6.1	Retrieval Based on the posterior Probability Value	36
3.6.2	Retrieval Based on Similarity Metrics	36
3.7	User Interface	37
3.8	User Interaction	39
3.9	Application Scenarios	39
3.9.1	Case: Multispectral-SAR Fusion	39
3.9.2	Case: Multitemporal-SAR Fusion	42
3.9.3	Case: Multispectral-SAR-Map Fusion	43
3.10	Performance Evaluation and Discussion	47
3.11	Conclusions	50
4	Multilayer Architecture for Heterogeneous Geospatial Data Analytics	53
4.1	Heterogeneous EO Data Integration	54
4.2	Architecture	54
4.2.1	Data Sources	55
4.2.2	Data Ingestion	55
4.2.3	Database Management System	55
4.2.4	User Oriented Web Functionalities	57
4.2.5	Graphic User Interface	58
4.3	Performance Evaluation	60
4.4	Case Studies	60
4.4.1	Image Understanding	61
4.4.1.1	Munich	61
4.4.1.2	Karlsruhe	62
4.4.1.3	Stuttgart	64

4.4.2	Optimum Dataset Selection	64
4.5	Conclusions	68
5	Data Mining Methodology for the Assurance of in-situ Recorded Land Cover Changes	69
5.1	LUCAS Anomaly	70
5.2	Data Mining System Architecture	70
5.3	Data Mining Methodology	72
5.3.1	Database Query Refinement	73
5.3.2	In-situ Image Analysis and Similarity Measure Computation	74
5.3.3	On Map Data Visualization and Filtering	75
5.4	Query Refinement Evaluation	76
5.4.1	Case Study Germany	76
5.4.2	Case Study Spain	77
5.5	Image Analysis and Similarity Measure Evaluation	78
5.6	Data Visualization and Filtering Evaluation	79
5.7	Data Mining Methodology Evaluation	80
5.8	Conclusions	83
6	Visual Analytics for EO Archives and In-Situ Data	87
6.1	Data Categories	87
6.2	Data Models	88
6.2.1	TerraSAR-X Data Model	89
6.2.2	LUCAS Data Model	89
6.3	Analytics of the TerraSAR-X Archive	91
6.3.1	TerraSAR-X Archive	92
6.3.2	Visualizing EO Image Metadata	92
6.3.2.1	Geographical Distribution by Geospatial Metadata	93
6.3.2.2	Relationship among the Instrument Metadata	93
6.3.3	Visualizing Semantics and Geospatial Data	94
6.3.3.1	Main Land Covers by Continent	94
6.3.3.2	Land Cover Distribution Around the World	95
6.4	Analytics of the LUCAS Archive	98
6.5	Joint Analytics of LUCAS and TerraSAR-X	99
6.6	Conclusions	101
7	Summary, Conclusions and Future Works	103
7.1	Summary and Conclusions	103
7.2	Future Works	105

Contents

A Related Publications	107
A.1 Journals	107
A.2 Conferences	107
Acronyms	109
List of Symbols	113
Bibliography	115

Introduction

1.1 Motivation

The amount of EO data is constantly increasing. This is due to the growing number of EO missions in space and their instrument characteristics that have evolved continuously. In addition, a lot of the currently available EO instruments offer data with very high spatial resolution. As for imaging sensors, the well-known panchromatic and multispectral images have been complemented by hyperspectral images and a wide range of Synthetic Aperture Radar (SAR) images generated with different techniques such as Polarimetric SAR (PolSAR) or Interferometric SAR (InSAR). Not only the nature of the imagery should be taken into account. EO products also comprise metadata providing useful additional information such as satellite orbit state vectors, geographical coordinates and data acquisition times. In addition, the access to very accurate cartographic data has been widely extended due to their digitalization and posterior publication via Geographical Information Systems (GIS) [1, 2]. Moreover, with the proliferation of digital photography and related media it is possible to access diverse in-situ data that can be used for different purposes during EO image analysis.

In this data diversity context, the research community faces a heterogeneous Big Data scenario where the main challenges are not only to provide better and more efficient algorithms, but also to design and implement tools that allow a greater exploitation of the available information. In line with the challenge, this thesis focuses on the development of new tools, techniques, algorithms and concepts which rely on different research topics and disciplines to develop systems capable of quick extraction of valuable information for a better understanding of EO data.

1.2 Goals

The main goal of the thesis is to develop new methods and tools that combine data acquired from satellites with cartographic and in-situ resources in order to improve the analysis, understanding and exploitation of the abundant and diverse EO data. In the pursue of this general objective, it is possible to identify intermediate goals including:

- The development of an active learning concept which relies on Bayesian probabilities to fuse diverse EO images and maps in a performance environment suitable for Big Data analysis.
- The definition and implementation of a system architecture capable of handling the heterogeneity of EO, cartographic, and in-situ data in a seamless way.
- The implementation of visual representations of the information in order to improve the analysis and understanding of the EO archives.

1.3 Scope

In contemplation of the goals, the scope of this thesis extends to different research topics, represented in Fig. 1.1. We can wrap the presented work in two general research topics which are: Data Analytics and System Engineering. The union link between these disciplines is provided by Image and Data Processing which can be seen as the keystone of this thesis for two reasons. First, it sets the course for the System Engineering task. And second, it provides the means for the required data analysis processes.

Data Analytics comprises a set of processes that filter and model the available data with the purpose of discovering implicit information not easily identifiable. This information can then be used to improve the understanding of the data and support the decision-making. Under the wide umbrella of Data Analytics, this thesis specifically focuses on the development of Data Mining concepts and methodologies which depend on Machine Learning techniques to fuse and extract useful information from the data. Beforehand, the data requires to be processed in order to obtain a proper data integration, and useful analysis parameters. In addition, Visual Analytics exploit original data along with the results of the data mining processes to represent the information via interactive visual interfaces.

System Engineering covers all the aspects required to design and implement successfully a system capable of satisfying the needs of the user. In the context of this thesis, the requirements are set by the diversity of the data to be integrated, and by technical constraints, such as, data processing and visualization capabilities.

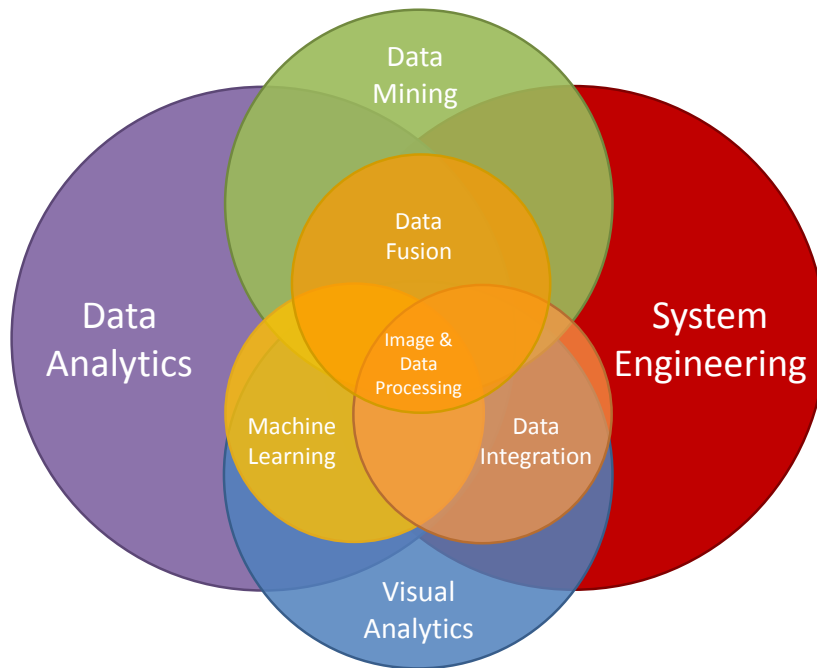


Figure 1.1: Research topics associated with the work presented in this thesis.

1.4 Contributions

The presented work focuses on the integration, analysis, and exploitation of wide range of heterogeneous data. Specifically, the main contributions of this dissertation are the following:

- Heterogeneous Data Mining Concept:** We present an Heterogeneous Data Mining (HDM) algorithm which is inspired by the main concept previously implemented in the Knowledge-driven Information Mining (KIM) system [3]. The HDM enhances the original KIM system overcoming the limitations related with the number of feature models used during the learning process. HDM introduces a faster active learning algorithm modifying the required statistical independence from the features to the posterior probabilities. The obtained speed-up factor allows the introduction of new feature models in the learning stage and the definition of more complex user semantics. The acceleration can also open new ways for knowledge-driven information mining systems to Big Data scenarios.
- Architecture for Heterogeneous Geospatial Data Analytics:** We introduce the architecture and a prototype of a multilayer system for heterogeneous geospatial data analytics. The system implements a server-client architecture, which integrates several web technologies. One of the benefits related to the

server-client approach is the simplicity of the client. The server is responsible for the most complex processing tasks making possible to offer lightweight clients for different devices. The presented architecture manages the data from the source. The initial layers read the original data and perform transformations to make viable the data integration. These heterogeneous data are linked and stored in a geographical database or in a system repository. The link among the data allows the User Oriented Web Functionality layer to exploit the database capabilities in order to perform geographical queries over the stored data. This layer also implements all the communication protocols to the linked third party services, and the server logic that interacts with the user via Graphic User Interface (GUI).

- **Methodology for Mining in-situ Land Cover Changes:** We describe a data mining methodology to filter and validate land cover change detections obtained from multitemporal Land Use/Cover Area frame Survey (LUCAS). Supported by the heterogeneous geospatial data analytics system, the data mining procedure obtains a clear reduction of the false detection of land cover changes. These results validate the proposed tools for the assurance of the in-situ recorded land cover changes.
- **Visual Analytics for EO Archives and In-Situ Data:** We present several interactive data visualizations that help end users to better understand the stored information in different archives. The data to be visualized is described by means of data models which are used for storing EO and in-situ information. The information provided by the data models is combined to generate interactive visualizations that make possible to analyze massive amounts of information in real-time.

1.5 Thesis Overview

The rest of the thesis is divided as follows. Chapter 2 deals with the heterogeneity of the EO data and presents the state of the art. Chapter 3 introduces the architecture and a prototype of a multilayer system for heterogeneous geospatial data analytics. Chapter 4 presents an architecture of a multilayer system for heterogeneous geospatial data analytics. Chapter 5 defines a data mining methodology to filter and validate land cover change detections obtained from multitemporal LUCAS in-situ surveys. Chapter 6 shows different data visualizations that summarize and help to better understand the content of EO archives. Finally, Chapter 7 sums up all the contributions presented in this dissertation.

Heterogeneity in EO Data and Information Retrieval Systems

This chapter introduces the different types of Earth Observation (EO) data, from satellite images to in-situ measurements. Moreover, state-of-the-art information retrieval systems are shortly presented, with special focus given to image retrieval systems in EO.

2.1 Heterogeneity in EO Data

One of the most remarkable properties of remote sensing imagery and EO data in general is the broad variety of products available that range from ancient cartographic data to modern satellite imagery and in-situ data, see Fig. 2.1. Focusing in remote sensing data, we can find imagery from two types of sensors: passive and active. Passive sensors are designed to receive and measure the radiation emitted or reflected by the observed objects. The intensity of the received radiation is dependent of the physical characteristics of the observed object or surface, e.g., temperature or roughness. On the other hand, active sensors firstly transmit a signal to the object or area to be observed and then they record the backscattered signals, i.e., signals reflected back to the emitter.

The rest of this section introduces some of the most relevant remote sensing imagery data along with cartographic and in-situ data used in the dissertation.

2.1.1 Optical

Optical data are obtained from passive sensors measuring the visible wavelengths of the spectrum, i.e., the wavelengths visible to the human eye. The visible spectrum goes from 390 to 700 nm and is represented in optical images by using three different channels or bands, Red spectral band (R), Green spectral band (G), and Blue

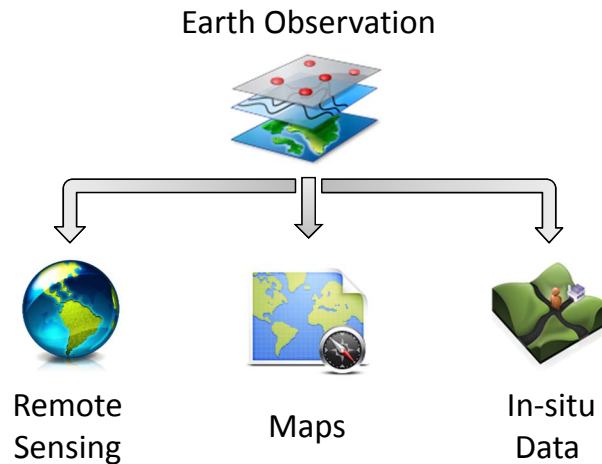


Figure 2.1: Representation of the three main EO data heterogeneity sources.

spectral band (B). Each band is recorded by independent sensors and represents the intensity values of the observed surfaces. The combination of the three bands will form an optical image. An image of Venice recorded by the satellite WorldView-2 (WV-2) with a 1.84 m spatial resolution is shown in Fig. 2.2.

2.1.2 Panchromatic

A panchromatic image is similar to an optical in the sense that it is generated only using the visible spectrum. The difference lies in the way it is represented. While in the optical image the visible spectrum is divided in three to match the human perception, a panchromatic image takes the visual spectrum as a whole, resulting in a gray level representation of it. Panchromatic sensors use of a wider wavelength range in comparison to the smaller range used in optical images. This factor increases the amount of energy received by the sensor which in consequence will offer a better spatial resolution. An example of panchromatic image is shown in Fig. 2.3 and corresponds to the panchromatic sensor of WV-2 with a spatial resolution of 0.46 m at nadir. The image shows the center of Washington D.C. with the presidential White House at the top-right, the Lincoln memorial at the bottom-left, and the Washington monument at the bottom-right.

2.1.3 Multispectral

Most of the in orbit satellites do not limit their sensors to panchromatic or R, G and B bands. On the contrary, the operational multispectral sensors may include a variety of other visual spectral bands (e.g., yellow or coastal blue) along with bands covering different parts of the InfraRed (IR) spectrum. We can see the multispectral



Figure 2.2: WorldView-2 optical image of Venice, Italy.



Figure 2.3: WorldView-2 panchromatic image of Washington D.C, USA.

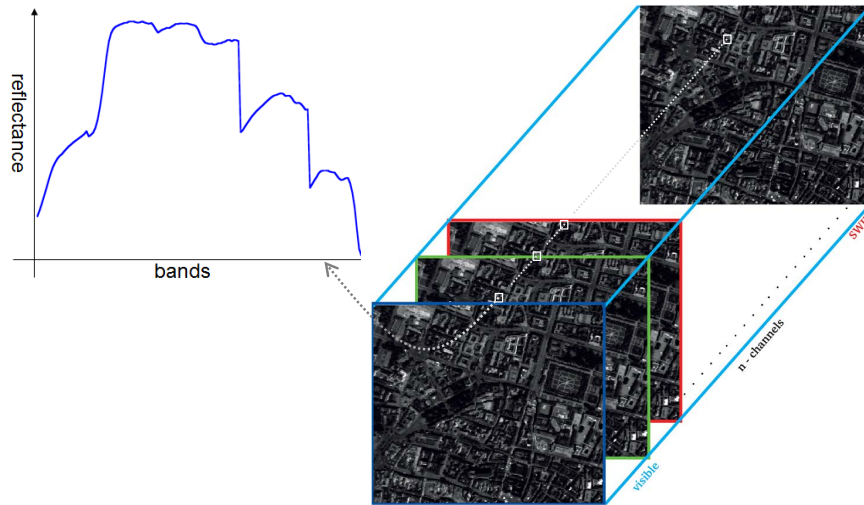


Figure 2.4: Hyperspectral cube characterization. A spectral signature of each pixel in the cube can be represented by using reflectance values in the spectral range. *Image courtesy of Jacub Bieniarz. Firstly published on [4].*

sensors as the opposite of panchromatic. While panchromatic sensors record the total radiation received in each pixel, the multispectral sensor uses spectral filters to divide the radiation in different bands and record them individually.

2.1.4 Hyperspectral

Hyperspectral follows the principles of the multispectral sensors but it records images using very narrow wavelength ranges over a wider spectral range. The output produces a hyperspectral cube, i.e., a stack of hundreds of images from narrow adjacent spectral bands. Subsequently, each spatial pixel of the hypercube image can be represented by their reflectance values in the spectrum, see Fig 2.4. In other words, the spectral reflectance plot shows the reflecting radiant energy of a pixel as a function of the wavelength. Each material has an unique reflectance signature which can be used in classification processes.

2.1.5 Synthetic Aperture Radar

A Synthetic Aperture Radar (SAR) usually operates in the microwave spectrum. The use of microwaves allows the sensors to be reasonably independent on the weather conditions. Moreover, as any active sensor, SAR systems can operate during day and night. SAR images are obtained transmitting a signal which is backscattered and received with a reduction in the intensity and a displacement in the phase. SAR sensors benefit from the movement of the platform (i.e. aircraft or satellite)

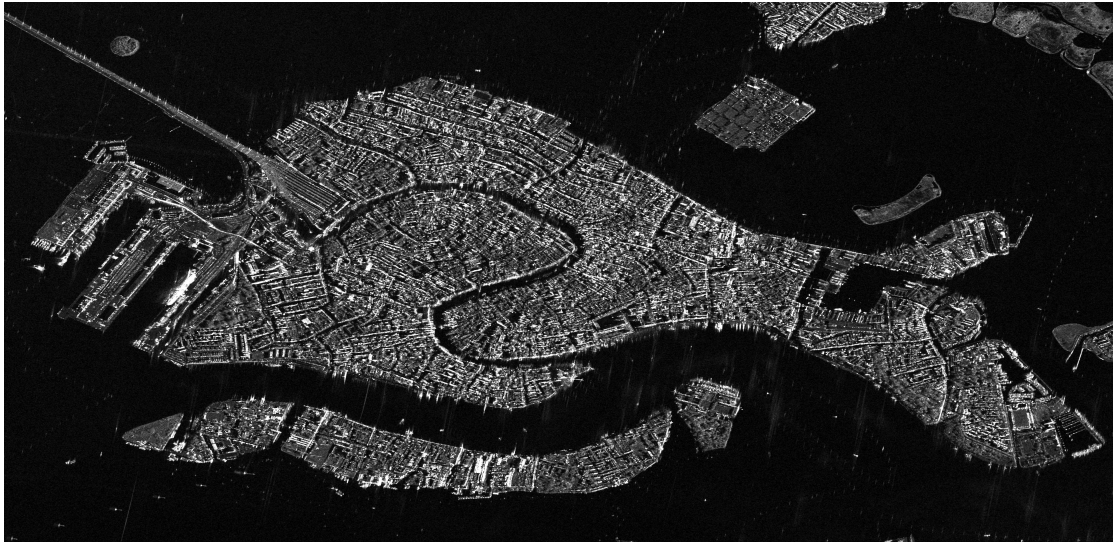


Figure 2.5: SAR image obtained by TerraSAR-X satellite over Venice, Italy.

to generate a large synthetic antenna aperture that makes possible to capture high resolution images with relatively small antennas. In Fig.2.5 an example of amplitude image generated from TerraSAR-X (TS-X) data is shown.

During the last decades several technologies have been developed for SAR data. Among them, we can remark Interferometric SAR (InSAR) [5]. Using an interferometric configuration, i.e., two or more SAR sensors observing the scene through different viewing geometries, it is possible to retrieve information about the terrain topography.

2.1.6 Cartographic

The use of cartography goes back to ancient times. Historically cartographic data have been treated as one of the most important secret documents among the countries due to their strategic relevance. In more modern times the access to very accurate cartographic data extended but it was not until the digitalization of cartographic data that the use of maps was tightly integrate in our daily life activities. We use precise maps and location services in every aspect of our life, from the classical travel planning to recording our workout paths. Nowadays, it is possible to access and reuse map information from two main different sources. The first one comes from official entities like national governments, government agencies, or supra national organisms. This kind of source is everyday more abundant due to Open Data and Open Government initiatives. In particular, the INSPIRE directive [6], promoted by the European Commission (EC) in 2007, established an infrastructure for spatial information in Europe to support policies or activities which may have an impact on the environment. The second source is made available by open collabora-

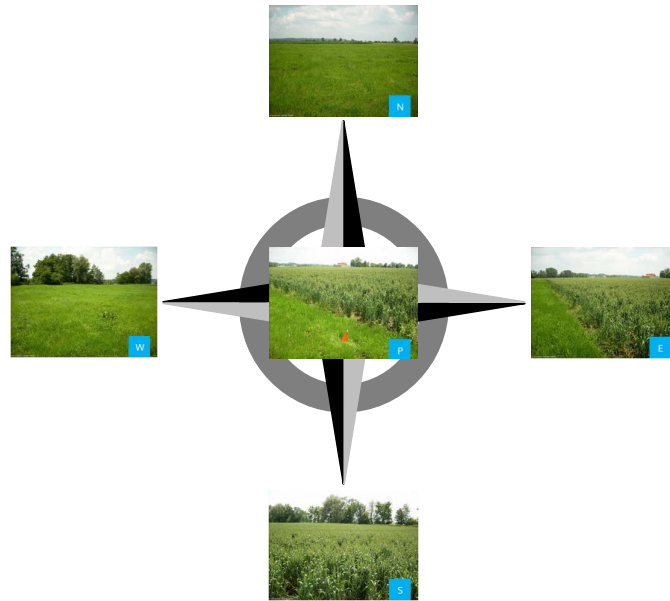


Figure 2.6: LUCAS in-situ images of a survey point in Germany. The images are composed of one photo of the exact survey geographical point of the acquisition and four images pointing to the main cardinal directions.

tive projects like OpenStreetMap [7], where the platform users upload geographical information every day and worldwide.

2.1.7 LUCAS In-situ Survey

Since 2006, EUROSTAT carries out a Land Use/Cover Area frame Survey (LUCAS) every three years to monitor the state and change dynamics in land use and cover in the European Union (EU). The survey comprises on ground observations that can be divided in three types: 1) micro data of the land cover, land use and environmental parameters associated to the single surveyed points; 2) in-situ photos of each point and landscape photos in the four cardinal directions, see Fig. 2.6; and 3) statistical tables with aggregated results by land cover/use at geographical level. LUCAS 2009 includes 234.561 points visited in-situ by 500 field surveyors on 23 countries, defining 77 different land cover classes. LUCAS 2012 survey includes 270.389 points visited in-situ by 594 field surveyors on 27 countries, defining 83 different land cover classes. In 2015 between March and October EUROSTAT carried out the LUCAS 2015 survey. Surveyors from 28 Member States visited a total of 273.401 points.

2.2 Information Retrieval Systems

Initial Information Retrieval Systems (IRS) were originally developed for structured collections as the ones stored in relational databases [8, 9]. Parallel to the growth of the Internet and its worldwide adoption as main communication medium, the amount of data and at the same time its variety increased. The available IRS which provided search and retrieval functions for structured text document collections were not designed for the new unstructured multimedia data. Thus, different research fields evolved or were born to develop solutions to classify, index and retrieve unstructured data. The developed system can be classified by their target data type:

- **Text-based Information Retrieval Systems:** In the nineties an estimated 90% of digital data was in the form of text documents [10]. Due to the predominant text quantity in the web, great effort and resources were focused on Text-based Information Retrieval. The effort put into the development and improvement of indexing and retrieval algorithms for unstructured text resulted in the modern Web search engines. One of the most famous and successful developments is the PageRank algorithm [11].
- **Image Information Retrieval Systems:** Large image collections require an effective system for image indexing and retrieval. Image data is composed of the digital image and the associated descriptive metadata. Therefore, image retrieval systems follow two different strategies in order to retrieve the visual documents [12]:
 - **Based on metadata:** Image metadata are divided in: 1) technical information about the device that captures the image along with date and location information; 2) annotations describing the content/context of the image which can be generated either manually, automatically or even by a semiautomatic procedure.
 - **Based on image content:** Commonly known as Content-Based Image Retrieval (CBIR), this approach is based on the statistical and mathematical analysis of the image. The objective of these analysis is to extract different low level features, e.g. color and textures, in order to compute the similarities of these features.
- **Multimedia Information Retrieval Systems:** The Multimedia Information Retrieval Systems (MIRS) are the natural evolution of the retrieval systems in environments where an integration of the newly available data types will improve the system response and usability. Thus, the initially independent retrieval systems converge merging the capabilities to query heterogeneous data, such as text, image, spatial data, audio or video [13]. Extensive surveys of this research area can be find in [14, 15].

The research work of this thesis is focused on the development of algorithms and tools to improve the navigation, retrieval and usability of EO image archives. Consequently, in the following Section 2.3 a more detailed explanation of CBIR is presented.

2.3 Content-Based Image Retrieval

CBIR as defined in [16] is the union of technologies which aim to help in the management of image and video digital archives by means of their visual content. Thus, anything ranging from a simple image similarity function to more complex image annotation engines can be seen as part of CBIR.

During the early 1990s, the CBIR term was coined in [17] and [18]. In these years different experiments to retrieve images from a image collection by different mechanisms based on image features were presented. Nowadays, CBIR community has increased and expanded to different fields of study. Good examples of these research fields are multimedia information retrieval, machine learning or computer vision which share common goals and challenges with CBIR [19].

Initial CBIR systems, summarized in [20, 21, 22], relied mostly in visual similarity. The main types of visual similarity measures were listed in [23]:

- **Color Similarity:** Humans use color as one of the main features to characterizes objects. In CBIR color was the first choices due to its simplicity and reliability. Different studies evaluated the color perception and proposed the use of different color spaces [24, 25, 26]. Red-Green-Blue (RGB) space is widely used because it simulates the input of the human eye and its performance is good under non changing conditions. A variation of RGB was presented in [27], where a combination of the RGB channels offered better retrieval results in images with shadows and brightness changes. Other color spaces like Munsel [28], the Lab color spaces [29], or Hue-Saturation-Lightness (HSV) [30] among others have been widely use for their invariant properties to different changes.

The color histogram has been the most widely used color representation technique which statistically describes the probability distribution of the color channels. In [27] the histogram intersection was proposed as metric to measure the sum of absolute distance between images. The authors of [31] proposed to use the mean-squared error to measure the similarities among not exactly identical colors. One common property of the color histograms is their sparsity. Sparse histograms have been shown to be more sensitive to noise and the use of cumulative histograms offer more robust results [32]. More recently, an evaluation of different color descriptors for object and scene recognition was presented in [33].

- **Texture Similarity:** An image texture is defined by a set of metrics via image processing which quantifies the perceived texture of an image and offers information about its spatial arrangement of color or intensities [34]. Initial image texture analysis focused on grey level spatial dependencies, i.e., orientation and distance, extracting meaningful statistic and constructing a co-occurrence matrix [35, 36]. Other popular texture representation methods rely on the modeling of the images via stochastic random processes. Remarkable examples of model based texture features are the Gaussian Markov Random Fields (GMRF) [37] and the Wold models [38]. In [39] a multiresolution gray scale rotation invariant method for representing the local image patterns via Local Binary Pattern (LBP) was presented.

In the beginning of the 90's Wavelet transform was introduced in the image retrieval context and became one of the most adopted image texture representation in a wide number of studies [40, 41]. In [42, 43] Wavelet transform was combined with co-occurrence matrix exploiting the statistical and transformation based analysis. Among the Wavelet transforms it is remarkable the performance of the Gabor filters [44, 45, 46]. A comparison study of several texture feature extraction methods was presented in [47].

- **Shape Similarity:** A shape descriptor is defined in [48] as a set of numbers which aims to quantify and represent a specific shape feature of a given object in ways that agree with the human perception. There are two main shape representation methods: 1) contour-based methods, and 2) region-based methods. If the shape representation method defines the shape as a whole, it follows a global representation approach. On the other hand, if the method describes the shapes by using segments or sections, it follows a structural approach. The methods based on contour make use only of the boundary information. Global contour-based methods include simple shape descriptors, such as perimeter, eccentricity or major axis orientation, convexity, compactness or elliptic variance [49, 50]. In the same category we find more complex descriptors like Hausdoff distance [51], the widely used Fourier descriptors [52], Wavelet descriptors [53], or elastic matching [54] among others. Some structural region-based methods are: chain codes [55, 56], polygon decompositions [57, 58], and invariant signatures [59]. The region-based methods describe shapes by using all the pixels inside the shape boundary rather than just the contour. Some global representation methods include the use of shape matrices [60] or grid methods [61]. Other methodologies rely on the calculation of different orthogonal moments. A comparison of the existing orthogonal moments can be found in [62]. In particular, Zernike moments [63] showed very promising results and were adopted by the MPEG-7 [64] standard as region-based shape descriptors. Generic Fourier Descriptors (GFD) were proposed aiming to deal with some of the problems related to the Zernike moments, such as computational

complexity, and some inconsistencies in the radial and circular features [65]. Among structural region-based methods, we highlight the medial axis transform [66] and core [67]. A wide survey on shape representation and description techniques is presented in [68].

- **Spatial Similarity:** Spatial information in the image can be used in conjunction with color or texture descriptors to improve the general retrieval accuracy. Initial approaches focused in the generation of layouts dividing the images in sub-blocks [69, 70]. A strategy based on quadtree data structure was presented in [71]. In [72] color correlograms were proposed for image indexing. Color correlograms demonstrated a certain tolerance to changes in the viewing point and zoom. More complex segmentation approaches focus on identifying meaningful regions or objects to improve the retrieval [73, 74].

We can find prominent examples of CBIR systems using a different combination of visual similarities. One is IBM's Query By Image Content (QBIC) system [75, 76], which was able to retrieve images from the catalogue based on color percentages, texture and shape. For the shape representation QBIC used circularity, eccentricity, major axis orientation and algebraic moment. The Virage image search engine [77, 78] offered an open framework for building a CBIR systems which rely on primitives, such as global and local color, different shape characterization techniques, and texture primitives very sensitive to high frequency features which were used to represent patterns within the image. The VisualSEEk image search engine allowed joint queries based on color information, region sizes and absolute and relative spatial locations [79].

By the second half of 90s, along with the World-Wide Web (WWW) blossoming, the image similarity search was implemented in several web image search engines. Examples of initial web image search engines are: WebSEEk [80, 81], Webseer [82], and PicToSeek [83]. With the appearance of new CBIR systems and the development of feature-based retrieval methods it became patent that the systems were not intuitive or user friendly for a non-expert users. This fact would certainly limit the adoption and usability of the CBIR systems unless a regular user could operate them in a more natural way. Thus, new research works focused in the development of new user friendly systems focusing in the use of semantics in the querying process which would bridge the so called "semantic gap".

2.4 Semantic Gap

In [16] the semantic gap is defined as "*the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation*". Fig. 2.7 represents the semantic gap in image analysis. The computers are able to analyze the digital data in the form of low level

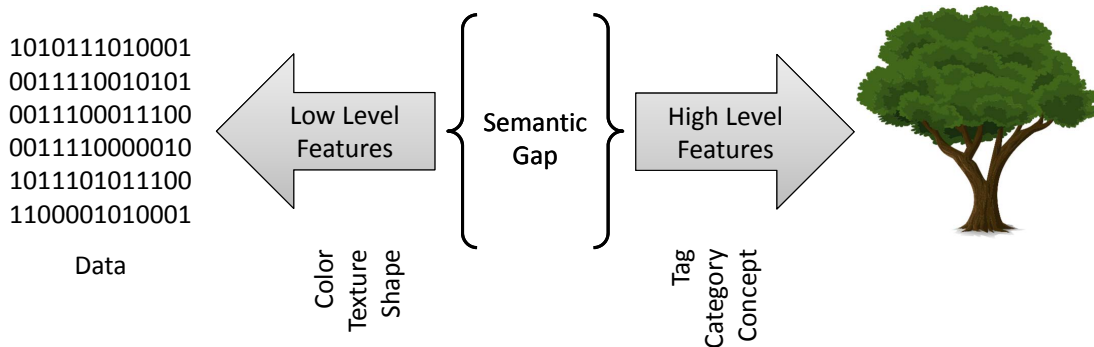


Figure 2.7: Semantic Gap between low level visual features and higher level semantics understood by humans.

features, such as color or texture, but human perceives only high level features in the form of concepts or keywords. The gap between the low level and high level features managed by humans is the semantic gap.

During the last decades, the research community has considered different approaches in order to bridge the semantic gap. One of the first approaches was introduced in the Multimedia Analysis and Retrieval System (MARS) [84] where they proposed a Query by Keyword (QbK) method which links the low level visual features with the high level semantic features. The use of the QbK was possible thanks to an automatically constructed Semantic Index (SI) which contained the concepts to query the multimedia content. More modern works [85] rely on machine learning algorithms like Latent Dirichlet Allocation (LDA) [86], Latent Semantic Indexing (LSI) [87] or probabilistic Latent Semantic Analysis (pLSA) [88]. The use of association rules has proven a valid tool to bridge the semantic gap in different domains [89]. Other approaches focus on the exploitation of the image visual content and the related available metadata. In [90] a methodology to fuse weighted graphs generated from the analysis of the visual features and the related tags is presented. One last way to effectively bridge the semantic gap is by including the user interaction [91].

2.5 User Feedback and Interaction

The user feedback via different interactions is a very important factor refining the retrieval data offered by CBIR search engines. Application wise, the search engines can be divided into three main categories according to their functionality and the role of the user in the querying process. The three main search engine categories are: Query by Example (QbE), Relevance Feedback (RF), and Active Learning (AL).

- **QbE:** In QbE engine the user introduces an image as a query, the system extracts the visual features and, based on them, returns similar images from the database ranked by their similarity. Examples of these engines are the previously mentioned QBIC [75, 76] or more modern systems focused on medical application like the one presented in [92].
- **RF:** In contrast, RF systems allow the user to refine a given query by iteratively specifying a set of relevant and a set of non-relevant images. A multimedia application of this type of search engines is IKONA [93].
- **AL:** AL engine requires interactive user action in order to refine the query parameters to obtain the desired results. A classical AL scenario contains an iterative loop composed of three main steps: user interaction, machine learning and retrieval. The loop starts with the user introducing a query to the system. The system based on the initial query parameters applies the machine learning processes and later returns initial results. Back on the user side, it is possible via user interface to point out positive examples (i.e., satisfactory results) and negative examples (i.e., results not matching the user requirements). This user feedback will update the machine learning algorithms which at the same time will improve the next computed results. If the results are not fully satisfactory, the user can refine the feedback and resume the loop once again. On the other hand, if the results are satisfactory, the interaction ends. More information about AL scenarios and a complete literature survey can be found in [94].

2.6 Image Retrieval in EO

So far the main general state of the art of CBIR has been presented. In this section we will extend its scope to EO. Thus, the following will describe the most commonly implemented features and descriptors in the EO field, along the existing systems and their capabilities.

2.6.1 Colour Composites and Spectral Indexes

The real difference between remote sensing color features and regular multimedia CBIR comes from the wide range of spectral bands available, see Section 2.1. The existence of multiple bands offers great flexibility regarding 1) representation of the images, and 2) generation of derived spectral indexes which can be also seen as new bands.

2.6.1.1 Color Composites

Also in the EO context, the RGB color space is widely used because it simulates the input of the human eye. The human visual system contains three different cone types that can discriminate three different wavelengths: R around 600-690 nm, G around 515-600 nm, and B around 450-515 nm. Mixing these three primary colors in different proportions, it is possible to represent any color in the visible spectrum. In EO it is a common practice to associate the available spectral bands to the primary color which results in color composites.

- **True Color Composite:** This composite is the normal composite produced when assigning the red band to R, green band to G and blue band to B. Fig. 2.8a shows a true color composite obtained from the Sentinel 2 multispectral satellite.
- **False Color Composite:** A false color composite is produced when the visual bands RGB are substituted by any combination of bands resulting in a different display of colors while comparing to the true color composition. While any band combination is certainly possible, certain combination schemes have shown remarkable results in improving the detection of certain objects. The use of the Near-InfraRed (NIR) is widely extended because of the high reflectance of the vegetation to this band. A classical false color composition is formed substituting RGB visual bands with NIR-R-G, Fig. 2.8b. This composite shows vegetation in different red tonalities, water can vary from dark-blue for clear water to cyan in the case of water containing sediments. In Fig. 2.8c we can see a composite using a Short-Wavelength Infrared (SWIR) in the band of $2.19\mu m$, the NIR band and the visual G band. This particular composite is very used in the fire management applications since it makes possible a clear differentiation between the burned and non burned forested areas. The composition highlights the fires with a bright red, shows vegetation in green, with can be particularly bright for healthy vegetation in growing seasons. Dry vegetation will appear in orange and it can discriminate different composition soils with several colors. Urban areas like the one in the image appear in different magenta tonalities. The composite shown in Fig. 2.8d shows no visible bands that are substituted by two SWIR bands (i.e., $2.19\mu m$ and $1.61\mu m$) and one NIR band. It provides very efficient atmospheric penetration and its main use is in geological studies to discriminate texture and moisture characteristics of the soil. Another possibility offered by false color composites is the introduction of spectral indexes, see the following Section 2.6.1.2. The Fig. 2.8e shows the grey scale representation of the Normalized Difference Vegetation Index (NVDI), while Fig. 2.8f shows a false color composite mixing NIR band, the NVDI and the G band. In there, high density of trees or vegetation canopy is bright green, yellow areas correspond to less

dense vegetation, grass is represented with golden yellow, and non-vegetated areas with dark blue and magenta.

- **Natural Color Composite:** This composite is used in EO products which lack one or more of the primary color bands and a true color style representation is required. Using different combinations of the available spectral bands, which might not be in the visible spectrum, it is possible to simulate a real photograph colors, i.e., water in blue, vegetation in green, etc.

There are many possible composite combinations with different outcomes depending the application purpose. In addition, in remote sensing as in general CBIR applications it is common the use of different colorimetric transforms (e.g., HSV, Cie Lab or Luv) because of their ability to enhance differentiation between some classes. Multiple research works have focused on the enhancement of the visualization of multiband EO products [95, 96]. A remarkable example is the work presented in [97] which tries to automatize the selection of optimum spectral features.

2.6.1.2 Spectral Indexes

The availability of bands allow to combine them to generate several spectral indexes which aim to convert the spectral reflectance into biophysical information. The following shows some of the most used spectral indexes:

- **RVI:** The Ratio Vegetation Index (RVI) or Simple Ratio (SR) was firstly introduced in [98, 99]. Healthy vegetation absorbs most of the visible R spectrum falling while reflecting a large portion of the NIR. On the other hand, unhealthy, dry or sparse vegetation reflects more R spectrum and less NIR. The RVI is defined as

$$RVI = \frac{NIR}{R}. \quad (2.1)$$

- **NDVI:** The Normalized Difference Vegetation Index (NDVI) was presented in [100, 101] as a transformation of the SR to simplify the computations and reduce the possible value range which for RVI is $[0, \infty)$. NDVI is defined as

$$NDVI = \frac{NIR - R}{NIR + R} \quad (2.2)$$

where NDVI values are in the range of $[-1, 1]$. Several research works have proven a direct relationship between the NDVI and the energy absorption of the plant canopies related to the photosynthetic processes [102] [103]. Thus, positive values over 0.3 show dense vegetation regions while lower positive values represent sparse vegetation or bare lands. Water lands present small negative values while clouds and snow areas have bigger negative values. NDVI has been used for monitoring the evolution of the vegetation growth [104].

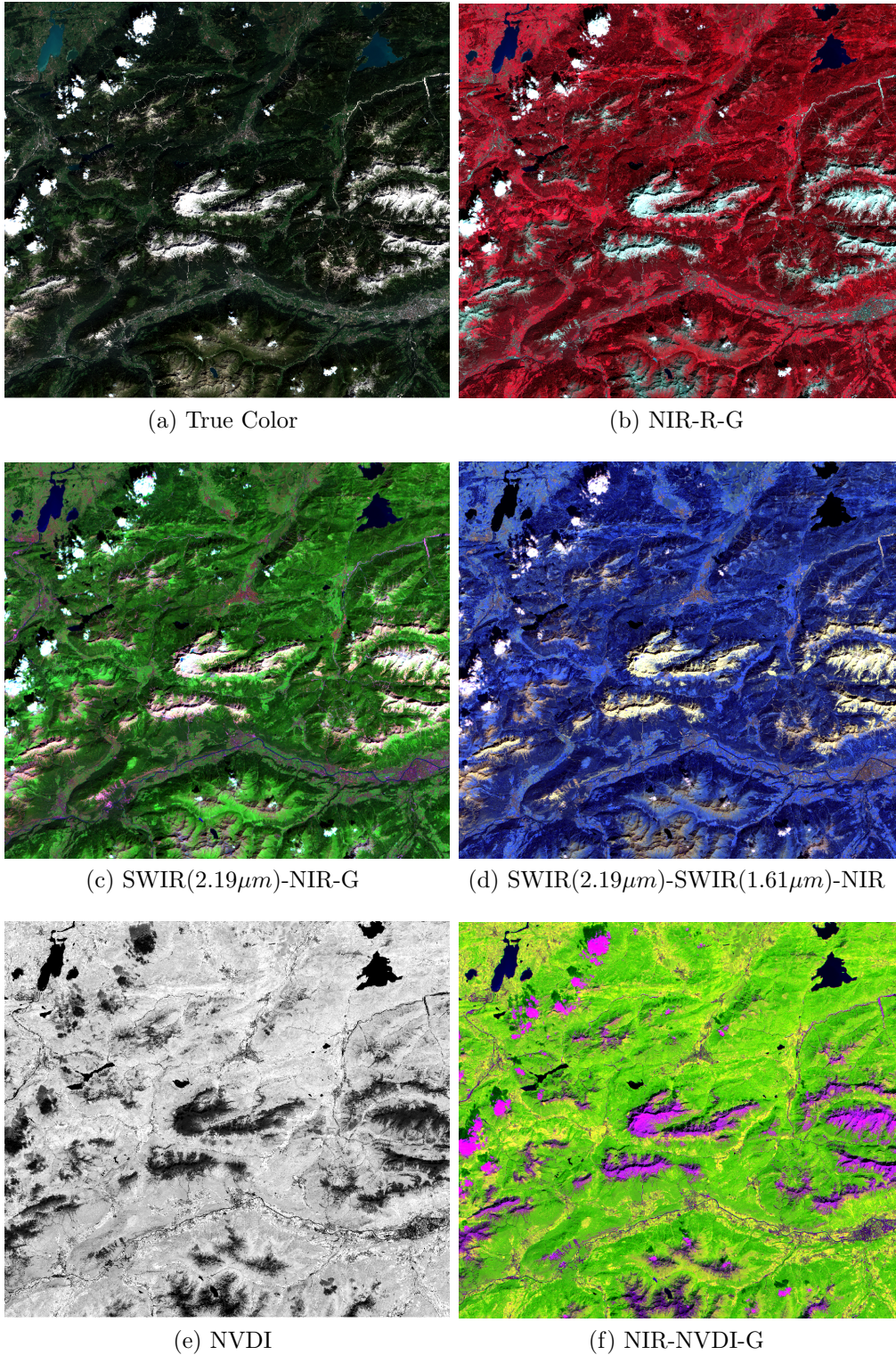


Figure 2.8: Color Composites using different Sentinel 2 multispectral bands.

- **SAVI:** The Soil Adjusted Vegetation Index (SAVI) was proposed to minimize the effect of the soil background [105] affecting the NVDI index. In this way a soil adjustment factor was introduced, reducing the impact of soil variations, background, and backscattering. The formula to compute the index is given below:

$$SAVI = \frac{NIR - R}{NIR + R + L_{cba}}(1 + L_{cba}) \quad (2.3)$$

being L_{cba} the canopy background adjustment factor which take usually value 0 for high vegetation cover, 0.5 for intermediate vegetation cover, and 1 for low vegetation cover.

- **EVI:** The Enhanced Vegetation Index (EVI) introduced in [106] as vegetation product of the Moderate Resolution Imaging Spectroradiometer (MODIS). EVI improves the measurement quality of the high biomass regions while minimizes the soil and atmosphere influence:

$$EVI = G \frac{N_{atm} - R}{N_{atm} + C_{aero1}R - C_{aero2}B + L_{soil}}(1 + L_{soil}) \quad (2.4)$$

where G is a gain factor, R and B are the red and blue bands; N_{atm} the atmospherically corrected surface reflectances in NIR; C_{aero1} and C_{aero2} are aerosol resistance coefficients; and L_{soil} is a soil-adjustment factor similar to the one defined in SAVI. The coefficients used in the EVI algorithm are, $L_{soil} = 1$, $C_1 = 6$, $C_2 = 7.5$, and $G = 2.5$ [107]. Great variety of studies have used EVI with different purposes from land cover change detection in time series [108] to Gross Primary Production (GPP) studies [109].

There are also many Atmospherically Resistant Vegetation Indexes (ARVI) [110] like the Green Atmospherically Resilient Index (GARI), Vegetation Index green (VIg) or Visible Atmospherically Resistant Index Green (VARIGreen) all described in [111], or Global Environmental Monitoring Index (GEMI) [112].

A very complete list of vegetation spectral indexes can be find in [113] and many other thematic spectral indexes can be find in [114].

2.6.2 Texture Features

A great variety of texture features have been introduced along the years for analyzing and classifying EO images. Basic texture models include the use of statistical moments over regions or windows. Due to the nature of SAR images, many research works have focused on the use of features from the transform domain. There are several feature descriptors based on Fourier transform such as, Fractional Fourier Transform (FrFT) [115, 116] or Short Time Fourier Transform (STFT) [117]. Also in the frequency domain the Quadrature Mirror Filter (QMF) has been used for EO image retrieval in [118].

Wavelet based texture descriptors are common for EO images. In [119] an evaluation of GMRF and Gibbs Random Fields (GRF) over TS-X images is presented.

In the following some of the most significant features in the context of this thesis are presented.

2.6.2.1 Haralick Features

The basis for the Haralick Features (HF) is the Gray Level Co-occurrence Matrix (GLCM) [35] which is a second-order texture measure, i.e., it considers the spatial linear relationship between groups of two neighboring pixels in four different directions. More specifically, GLCM generates matrices for pixels adjacent to each other horizontally (0°), vertically (90°) and for the two possible diagonals (45° , 135°). The co-occurrence matrices C are always square with a size of $N_G \times N_G$ being N_G the number of the gray levels defined in the image. Each element $C(i,j)$ of the matrix represents the number of times a pixel with value i is adjacent to a pixel with value j , divided by the total number of comparisons made. Four GLCMs are constructed one for each direction and since the elements in the matrices represent the amount of specific gray level relationships in the image and for each direction both possible senses are taken into account, the GLCMs are symmetric. The HF are the statistical features extracted from the GLCM and due to the high computation costs there are several practical approaches that try to improve the computation efficiency of the statistical features [120]. Another possibility to make the computation of the HF more efficient is to reduce the textural parameters used. GLCM originally proposed 14 textural parameters [35], but in [121] the six most relevant were considered to be: energy, contrast, variance, correlation, entropy and inverse difference moment. In [122] a study of those statistical parameters was presented in the EO scope and proposed a solution based on Gray Level Difference Histograms (GLDH) that offer a better compromise between texture accuracy and computation costs. The use of HF as texture parameter on SAR applications is extensive specially in sea ice studies [123, 124] and forestry [125, 126] but also as contextual descriptors in very high resolution SAR [127].

2.6.2.2 Gabor Feature Descriptor

Gabor Feature Descriptor (GFD) reduces the feature computation time compared with other texture features while maintaining a good retrieval performance. Furthermore, Gabor features are rotation and scale invariant. A successful implementation of Gabor wavelets for texture analysis and description in image retrieval applications was presented in [46], where the descriptors are formed using a mother Gabor wavelet which is tuned using different values of orientations K and scales S . The feature vector is then constructed concatenating the means and standard deviations

for every combination of orientations and scales,

$$GFD = [\mu_{11}\sigma_{11}\mu_{12}\sigma_{12}\dots\mu_{SK}\sigma_{SK}] \quad (2.5)$$

In the last year different publications have used Gabor descriptors in EO applications. In [128] Gabor filters are used for hyperspectral image classification. Regarding SAR images, [129] introduces an unsupervised change detection method of multitemporal SAR images based on Gabor features. Also working with SAR images, the work presented in [130, 131] introduces a feature descriptor modification based on second-kind moments, i.e., log-mean and log-variance, for performing image classification. The performance of GFD in very high resolution SAR patch classification was tested against other feature descriptors in [118] where it scored very high accuracy values.

2.6.2.3 Weber Local Descriptor

The Weber Local Descriptor (WLD) is developed in [132, 133] and it offers a robust edge extraction capabilities even in heavy noise images. WLD is based on Webber's Law which is a psychological law quantifying the perception of change in a given stimulus. It states that a human will only perceive the changes just in a constant ratio of the original stimulus [134]. Therefore, WLD uses the intensity difference in the image to detect the salient variations and simulate the human perception of patterns. This is performed by generating a series of histograms that represent the differential excitations ζ in various dominant orientations Φ . The WLD descriptor is formed afterwards by: 1) dividing the histograms in M segments, 2) constructing a new series of sub-histograms that contain the same M division of the initial histograms for the different Φ orientations, and 3) concatenating the reorganized sub-histograms. The length of the WLD descriptor is the product of ζ and Φ while M defines only the interleaving of the different Φ orientations M times. In remote sensing modifications of the WLD have been successfully implemented in [135] for indexing SAR image patches.

2.6.3 Scale-Invariant Feature Transform

The Scale-Invariant Feature Transform (SIFT) extracts the distinctive invariant features of the images which can be used later for object or scene matching tasks even with different view angles [136]. This is possible due to the ability of the features to be invariant to scale and rotation. The SIFT method comprises four different steps. First, the image is analyzed looking for the scale-invariant features using a Difference of Gaussians (DoG) functions [137]. DoG generates two different Gaussian blurs of the analysis image and subtracts them to highlight the image edges and propose potential interest points. Second, an analysis of the proposed interest points is made to select the most robust ones and discard the ones more

sensitive to noise. Third, based on the local image gradient directions, one or more orientations are assigned to the keypoints, this step removes the effects of rotation and scale. And forth, for each keypoint location the descriptor is generated by: 1) computing the magnitude and orientation of the gradient in a region around the keypoint and weighting it using a Gaussian window; 2) accumulating the results in $R \times R$ pixel subregions; and 3) generating histograms of the regions with bins equal to the number of selected θ_{SIFT} orientations. Consequently, the SIFT descriptor is a vector containing the values of all the orientation histogram entries with a length equal to the product of the subregions by the number of orientation used, i.e., $R \times R \times \theta_{SIFT}$. In [136] the most robust results were obtained using a region around the keypoint location of 16x16 pixel, a subregion of 4x4 and 8 different orientations. As result the descriptor length was a vector of 128 elements.

EO applications of SIFT can be found in urban-area and building detection [138], multispectral image registration [139, 140] or SAR image registration [141]. Also related to SAR image registration a modification of the original SIFT was proposed in [142].

2.6.4 Semantic Level Descriptors

The use of semantic level descriptors is one of the attempts from the research community to bridge the semantic gap. These descriptors link the low level features extracted directly from image analysis processes, with semantics/concepts used by humans by means of different machine learning approaches. The more relevant approaches for this thesis are Bag of Words (BoW) and Latent Dirichlet Allocation (LDA).

2.6.4.1 Bag of Words

BoW was originally proposed as a model for text document classification and initial references to the model go back to the 50s [143]. Basically, BoW just takes into account the multiplicity of words appearing in a document for which it constructs an histogram. BoW was posteriorly used in computer vision applications [144] where instead of words, the image visual feature vectors were used. Due to the high dimensionality of the visual features it is necessary to generate a codebook or dictionary. There are different alternatives for generating codebooks but all of them rely on Vector Quantization (VQ) techniques, from the simplest regular division of the dimensional space to more complex clustering methods, e.g., k-means [145]. Thus, the computed codebook, composed of a small subset of descriptors, is used to encode all the extracted visual features. There are different coding algorithms which define how the codewords are activated. A state of the art study of the existing coding techniques used in BoW applications can be found in [146]. The last step in the BoW generation encompass a pooling process [147], which typically

involves mean or max pooling, and results in the final signature representation of the image, i.e., the identification of the image by a BoW vector.

The use of BoW models in remote sensing can be found in applications for land use scene classification [148, 149], but also to annotate large satellite scenes [150], where the BoW is used as input for a LDA model.

2.6.4.2 Latent Dirichlet Allocation

Initially proposed for text corpora analysis, LDA has the goal to find short descriptions of the members of discrete data collections, enabling efficient processing of tasks like classification, summarization or novelty detection, among others. LDA makes use of generative probabilistic models to describe the underlying topics in a document as a probability distribution over a set of words in the existing vocabulary, which can be seen as an explicit representation of a document. Consequently, a corpus or collection can be defined as a finite set of latent set of topics. It is possible to extent the application scope of LDA by defining an analogy between the text corpora analysis and image collection analysis. Experiments of LDA in EO applications are numerous. In [150] LDA is used to classify QuickBird panchromatic image patches, using concepts defined by the user. Other example is presented in [151] where LDA is used to map heterogeneous pixels with similar intermediate-level semantic meaning into land cover classes of various mapping products. A complete description of LDA modeling can be found in [86].

2.7 Image Information Mining Systems

Image Information Mining (IIM) seeks solutions for automatizing the extraction of information from EO archives via interdisciplinary approaches comprehending computer vision, image retrieval, machine learning, data mining and database management among others [152]. The new information obtained from processes can result in a better image understanding or even in the discovery of knowledge [153]. Specifically, IIM systems offer the possibility to navigate and browse large archives and obtain implicit knowledge or patterns from images and/or between image and other alphanumeric data which are not explicitly stored [154].

A remarkable example of IIM system is the KIM system presented in [3]. KIM as any other AL system requires interaction of the users who provide semantic interpretation of the image content, which is internally linked to a hierarchical Bayesian network. The user can query the database for relevant images and obtains a probabilistic ranking of the entire image archive as an intuitive information representation. The KIM system is specially meaningful in the context of this thesis, and a more detailed description of it is introduced in Chapter 3.

During the last years, several IIM systems have been developed with different

technology approaches in order to handle EO image heterogeneity and its characteristics. Different QbE paradigms have been proposed to retrieve multispectral images like the region based image retrieval system developed at Oak Ridge National Laboratory in [155] and the Multi-sensor Evolution Analysis (MEA) [156]. We can also find in the literature systems like the Geospatial Information Retrieval and Indexing System (GeoIRIS) [157] and the Intelligent Interactive Knowledge Retrieval (I³KR) system [158] that try to retrieve images by means of semantics. In particular, GeoIRIS includes automatic feature preprocessing and indexing of EO images. Furthermore, it implements a complex query system that merges heterogeneous geospatial databases making possible to retrieve objects using different visual features which can be later semantically link to higher concept descriptors. There are also systems for the retrieval and analysis of SAR and corresponding image time series. Selected examples of these systems are Image Information Mining in Time Series (IIM-TS) [159], and the PicSOM system based on Self-Organizing Maps (SOM) [160]. Recently, different research projects like EOLib [161] or TELEIOS [162] have introduced the use of EO image metadata and linked data as query parameters in order to improve the results. Linked data can be seen as a collection of best practices for publishing semantically structured and interrelated datasets on the Web [163]. A review of current EO image information mining systems can be found in [164].

2.8 Visual Analytics

Image information retrieval systems have provided a wide variety of tools for interactive exploration of big image archives based on different metadata, keywords or visual descriptors. Regardless their outstanding performance, different sets of tools are required to successfully analyze and present the obtained results in a way that facilitates their understanding. In this sense, visual analytic techniques try to combine automatic analysis methodologies with interactive visualization tools in order to improve the understanding and analysis on Big Data scale datasets [165]. Visual analytic research goal is to provide the decision makers with tools which exploit the information abundance offering the opportunity to examine massive amounts of information in real-time situations [166]. Summing up, visual analytics is perceived as a conglomerate of interactive visualizations analysis techniques, which exploit different automatic analysis techniques, to facilitate the understanding, reasoning and decision making over large and complex datasets.

An example of the implementation of visual analytic techniques can be found in [167] where the available geographical tags and the underlying geographical context were exploited with image retrieval purposes. Moving the focus to EO, [168] successfully applied visual analytic techniques on large geospatial datasets with data mining purposes. Another example is LandEx GeoWeb tool [169] which provides a visual search engine to retrieve similar tiles based on pattern inputs and similar-

2. Heterogeneity in EO Data and Information Retrieval Systems

ity maps. Finally, the work in [170] introduces the Immersion Information Mining system that uses advanced visualization techniques to enable knowledge discovery from EO archives.

Knowledge-driven Heterogeneous Data Mining

This chapter presents an accelerated probabilistic learning concept and its prototype implementation for mining heterogeneous Earth Observation (EO) images, e.g., multispectral images, Synthetic Aperture Radar (SAR) images, image time series, or Geographical Information Systems (GIS) maps. The system prototype combines, at pixel level, the unsupervised clustering results of different features, extracted from heterogeneous satellite images and geographical information resources, with user defined semantic annotations in order to calculate the posterior probabilities that allow the final probabilistic searches. The system is able to learn different semantic labels based on a newly developed Bayesian network algorithm and allows different probabilistic retrieval methods of all semantically related images with only a few user interactions. The new algorithm reduces the computational cost, outperforming existing conventional systems, under certain conditions, by several orders of magnitude. The achieved speed-up allows the introduction of new feature models improving the learning capabilities of knowledge-driven image information mining systems and opening them to Big Data environments.

The chapter is organized as follows: Section 3.1 describes the main aspects behind a classical knowledge-driven information mining system. Section 3.2 introduces the Heterogeneous Data Mining (HDM) concept followed by the sections presenting the elements that composed the HDM system. Specifically, Section 3.3 describes the feature extraction processes and Section 3.4 the feature clustering and the generation of the Bag of Words (BoW). Continuing with the HDM modules, Section 3.5 introduces the machine learning methods implemented. Section 3.6 explains the available retrieval methods followed by the introduction of the user interface, Section

The content of this charter has been published in: K. Alonso and M. Datcu, "Accelerated Probabilistic Learning Concept for Mining Heterogeneous Earth Observation Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 7, pp. 3356-3371, July 2015.

3.7, and the description of the user interactions, Section 3.8. Section 3.9 demonstrates the system capabilities for different application scenarios, while Section 3.10 presents the most important system performance parameters. Finally, Section 3.11 contains general conclusions.

3.1 Knowledge-driven Information Mining

A classical knowledge-driven information mining system, like the one presented in [171], represents the managed information via a multilevel hierarchical model, as shown in Fig. 3.1. Its initial level is formed by the data sources D , for instance, different EO image data. The second level of the hierarchical model contains the first processed information, θ , extracted from the EO products via statistical or mathematical analysis. For pixel-wise analysis, and being usually θ high dimensional, this stage generally increases the amount of managed data by various orders of magnitude. The third information level is composed of clustered θ features. The resulting cluster identifiers can be interpreted as words, ω . In this stage the data quantity is reduced from the highly dimensional features to an easily manageable single layer of words. At this point, it is possible to represent D by a normalized histogram of word occurrences. Finally, in the last level of the hierarchical model, the user attaches semantic labels, L , to the existing words.

In KIM the user introduces semantic interpretation of a selected image (sub-)scene via AL, by giving positive and negative examples, which are interactively linked with a hierarchical Bayesian network [172] (not shown in Fig. 3.1) to a content-index formed by a combination of different ω_i , where i is an index for the available words. Using the different words, ω_i , for all extracted features θ , one can identify an image uniquely by means of the probabilities $p(\omega_i|D)$, which express the occurrences of the words within the given image. These words will result in forming a BoW [146]. The Bayesian network allows the user to interactively define a link between a specific semantic label L with the existing words ω_i known as the stochastic link $p(\omega_i|L)$. Once the stochastic link is defined, it is possible to calculate the posterior probability $p(L|D)$, which is used to query the database for relevant images, obtaining a probabilistic ranking of the entire image archive as a semantic information representation. A brief description of the theoretical aspects is given in Section 3.5.

3.2 Heterogeneous Data Mining Concept

The HDM conceptual design, as any Bayesian inference system, is composed by two main stages: a data-driven initial stage and the final user-driven stage. In Fig. 3.2 the conceptual stages, their independent modules and connections are shown.

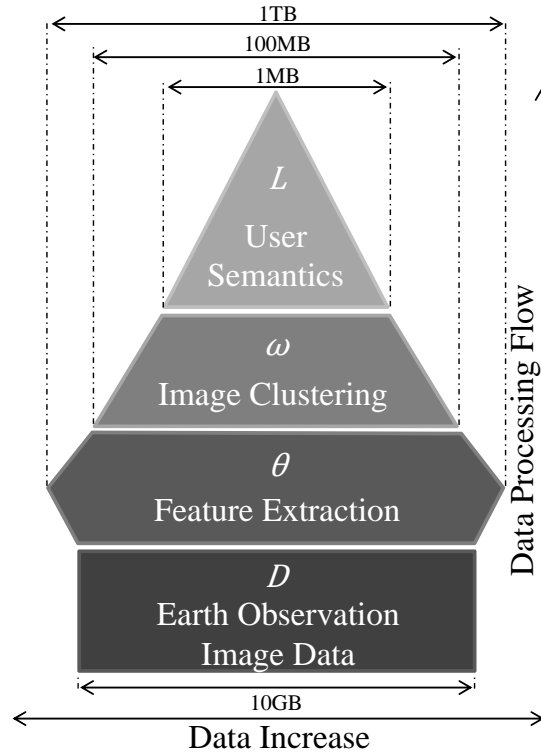


Figure 3.1: System hierarchical levels. The initial level is formed by the data sources D . The second level refers to the information extracted via mathematical analysis, θ . Third information level represents the clustering results ω of the features θ extracted in the previous level. The last level consists of the user's specific concept L , and its relationship with the existent words.

The initial modules of the data-driven stage represent the different heterogeneous EO databases or GIS repositories. From these online linked repositories, HDM gets the data to be analyzed. The analysis stage itself is performed offline. In the first analysis step, different types of features can be extracted at pixel level. In addition, the information available in the GIS repository in form of a vector map is rasterized to the EO image resolution. All the extracted features are clustered automatically using any kind of unsupervised clustering, e.g., k-means. The clustering results, i.e., cluster identifiers, are used for the generation of BoW signatures. At the end of the analysis processes, the word maps, that defines the cluster or word assigned to every pixel in the image, and the calculated BoW signatures, $p(\omega_i|D)$, are stored into a database.

In the second stage, the user interaction enters in scene and guides the processes inside an active learning loop. This loop is composed by three steps: machine learning, probabilistic retrieval and user interaction. This stage is real time from the point of view of the user, who expects a relatively fast response of the system to the required actions. Inside the loop, the user interacts via user interface introducing positive examples of the label, L , and negative examples for $\neg L$, allowing the learning

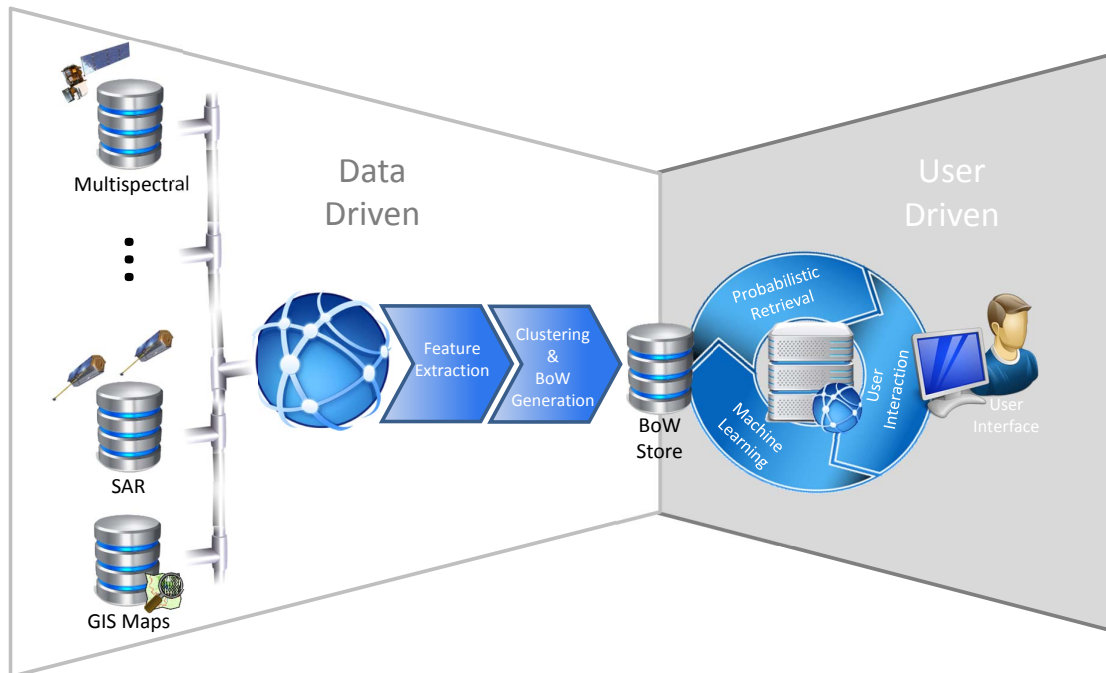


Figure 3.2: The Heterogeneous Data Mining (HDM) conceptual design is composed by two main stages, a data driven initial stage and the final user driven stage. The initial modules of the data driven stage represent the different heterogeneous EO data bases, or GIS repositories, where the data are obtained. The data driven stage extracts the features, clustering them and generating the BoW for each image. The user driven part is composed by the user interaction, the learning and the probabilistic retrieval processes. The user can introduce positive and negative examples about specific semantics that will allow the system learning. After the learning step, the user driven cycle will end with a probabilistic retrieval query to obtain the desired images.

of the system. After the interaction step, the user can perform a probabilistic retrieval. If the results are not satisfactory, the user can refine the learning and retry the probabilistic retrieval. On the other hand, if the results are satisfactory, the interaction ends.

Following, a more detailed HDM prototype review is presented, explaining the main functionalities and algorithms used in each module.

3.3 Feature Extraction

This module extracts at pixel level the image analysis features and map features required by the system to uniquely identify an EO image. For the posterior assumptions of the probabilistic retrieval, the generation of unambiguous stochastic links are required. To achieve this, a full statistical independence of the extracted features θ would be optimal. Thus, a careful selection of features and their coding is necessary to provide sets of descriptors that are as statistically independent as

possible.

In this case the features used are spectral parameters, such as multispectral features or intensity values; and texture features, as the Weber Local Descriptor (WLD) [132]. The system modular design allows to easily add new feature descriptors.

HDM system prototype also relies on the information extracted from existing maps stored in GIS platforms. It is possible to access map information from two different sources. The first one comes from official sources like national governments, government agencies, or supra national organisms. This kind of source is everyday more abundant due to Open Data and Open Government initiatives. The second one is made available by open collaborative projects like OpenStreetMap [7], where the platform users upload geographical information every day and worldwide.

For this prototype implementation, we have used the information available in OpenStreetMap. The layer extraction for this particular initial experiment has been manually done using open source Quantum GIS program [173], but it can be easily automatized in future implementations.

At the end of feature extraction operations we have available different feature sets.

3.4 Clustering and Bag-of-Words

The statistical independence of the features will help obtaining non correlated BoW dictionaries, which in theory, provide more meaningful results when combined. The features obtained from the Feature Extraction module are used as input for the Clustering and Bag-of-Words module. This module produces two different outputs for every image and analyzed feature. The first output is the word map that defines the cluster or word assigned to every pixel in the image. The second output is the probability of every word in the image, which is built using the previously calculated word maps. At the end of the data driven part the word maps and BoW probabilities are linked together with the analyzed heterogeneous EO images defining a query object in a database.

For the EO images, the clustering of the features relies on the unsupervised K-means clustering algorithm. Clustering processes can take from days to weeks depending on the size of the dataset and the feature length. In the case of GIS maps, the selected layers from OpenStreetMap will represent the cluster identifiers of the map model in the system. In this case, we have chosen seven different classes: water bodies, roads, railways, buildings, urban areas, sport areas and green areas. The map information is in vector form and, hence, a vector-to-raster conversion is needed. The map rasterization is done using the GDAL library. In order to do the conversion, the raster image must have the same resolution of the EO images with which it will be fused. The resulting raster image is composed by simple integers losing all the semantic meaning of the OpenStreetMap classes. The integer layer is

now treated like the output of a clustering process required to generate a word map. Using the word map the BoW probabilities are calculated. Finally the word map and the BoW probabilities are stored into the database and linked to the existing query objects created with the EO image analysis.

As stated before, a single feature is in general not enough to generate a meaningful representation of a semantic label. For this reason, even if initially the feature dictionaries are calculated independently, when running the system a combination of the independent feature dictionaries are used for the learning,

$$\omega_i = \omega_{f_1,j} \otimes \omega_{f_2,k} \otimes \dots \otimes \omega_{f_c,z} \quad (3.1)$$

where ω_i is the combined dictionary, the different f values represent each of the independent feature used in the learning and the subscripts associated to them (j, k, \dots, z) represent the length of that specific dictionary. The length of ω_i will be the product of the independent dictionary lengths.

3.5 Machine Learning

In the following, we present a brief summary of the theoretical concepts of the learning stage based on a naive Bayes Classifier [172, 174]. The learning is based on the posterior probabilities of a user-defined semantic label L given an image D expressed as,

$$p(L|D) = \sum_i p(L|\omega_i) \cdot p(\omega_i|D). \quad (3.2)$$

As an alternative, one can apply the Bayesian theorem:

$$p(L|D) = p(L) \cdot \sum_i \frac{p(\omega_i|L) \cdot p(\omega_i|D)}{p(\omega_i)} \quad (3.3)$$

where $p(L)$ is the prior probability of the semantic label L , $p(\omega_i|D)$ are the probabilities of the words in a given image, $p(\omega_i|L)$ denotes the probabilistic links of the words with a label, which can be expressed as the probability of the words updated with the examples defined by the user. Finally, $p(\omega_i)$ is the prior of the words ω_i given by:

$$p(\omega_i) = \sum_L p(\omega_i|L) \cdot p(L) \quad (3.4)$$

where the labels are restricted to L and $\neg L$ and the probabilistic link $p(\omega_i|L)$.

At this point, the computation of $p(\omega_i|L)$ and $p(\omega_i|\neg L)$ should be considered. Once these terms are known, $p(L|D)$ can be calculated using (3.3) and (3.4). In order to calculate these probability links we will make use of the user inputs by

means of training samples. The training samples can be positive, represented by the presence of the label L , or negative referring to the absence of L and defined as $\neg L$. For sake of simplicity, the next equations present the calculation of $p(\omega_i|L)$, but they are applied to $p(\omega_i|\neg L)$ in the same way. Therefore, we define T as a set of user provided positive training data in the form of $T = \{N_1, \dots, N_r\}$ with N_i as the number of occurrences of ω_i and r as an index of the existing words. T presents a multinomial distribution that can be parametrized via $\phi = \{\phi_1, \dots, \phi_r\}$. As introduced in [3], and widely described in, [175] and [176], we can express our desired probability as,

$$p(\omega_i|L) = E[\phi_i] = \int \phi_i p(\phi|T) d\phi_i, \quad (3.5)$$

where $p(\phi|T)$ is modelled as a Dirichlet distribution,

$$p(\phi|T) = Dir(\phi|\alpha), \quad (3.6)$$

and α is a hyperparameter vector which represents the user interaction (i.e., introduction of training examples by the user). It has the same dimension as the used BoW dictionary, and is initialized to one, $\alpha = \{1, 1, 1, \dots, 1\}$. One of the Dirichlet model characteristic is the property to perform the learning incrementally. Moreover, a Dirichlet distribution is the conjugate prior of a multinomial distribution in Bayesian statistics. Thus, new user interactions can update the posterior probability by means of,

$$p(\phi|T) \sim Dir(\alpha_1^k + N_1^{k+1}, \dots, \alpha_r^k + N_r^{k+1}) = Dir(\alpha^{k+1}), \quad (3.7)$$

where k refers to the user interaction, N_i^{k+1} are the new training examples and the updated hyperparameters α_i^{k+1} are defined by the following expression,

$$\alpha_i^{k+1} = \alpha_i^k + N_i^{k+1}. \quad (3.8)$$

Once the hyperparameters and the update procedures are defined, we can rewrite (3.5) as,

$$p(\omega_i|L) = \frac{\alpha_i^{k+1}}{\sum \alpha_i^{k+1}}. \quad (3.9)$$

Since $p(\omega_i)$, expressed in (3.4), is the sum of L and $\neg L$, using the negative samples introduced by the user we define another hyperparameter vector set for the required $p(\omega_i|\neg L)$ calculation. With these hyperparameter sets we are finally able to calculate the posterior probabilities $p(L|D)$, defined in (3.3), and proceed to the probabilistic search.

3.5.1 Assumption of Feature Probability Independence

The original KIM implementation assumes the full statistical independence of the features and the resulting clusters. Thus, the calculation of the stochastic link can be performed by a simple multiplication of probabilities

$$p(\omega_i|L) = p(\omega_{f_1,j}|L) \cdot p(\omega_{f_2,k}|L) \dots \cdot p(\omega_{f_c,z}|L) \quad (3.10)$$

where c is the number of feature models used in the learning, f identifies the feature model and the associated subscript represents its length. The same statistical independence assumption is made for the joint probability of the words in a given image $p(\omega_i|D)$.

This approach was already computationally fast, with the imposed restriction of the use of only two different feature models in the learning stage: spectral and texture features were used to ensure the required statistical independence. The restriction was imposed due to the computational complexity from the calculations of $p(\omega_i|L)$ and $p(\omega_i|D)$ which can be represented as multidimensional matrices where each element refers to the probability of occurrence of a certain word combination. Therefore, every additional feature model increases the dimensionality of these matrices by one. As a consequence, the number of operations are multiplied by the number of words of each feature model. Defining n as the number of operations required for the independent calculation of the posterior probabilities for each feature model, and c being the identifier of the model, we obtain the computational complexity as,

$$O = n_1 \cdot n_2 \cdot \dots \cdot n_c. \quad (3.11)$$

The different sizes of the dictionaries in ω_i are relatively small in comparison with the total amount of operations needed for the calculation of the posterior probabilities. Therefore, the different numbers of operations n_c can be equalized to $n_c = n$. This results in a final polynomial complexity of the algorithm, increasing with c ,

$$O(n^c). \quad (3.12)$$

3.5.2 Assumption of Posterior Probability Independence

The complexity of the KIM algorithm should be reduced since its algorithm, based on the statistical independence of the features (3.10), is not fast enough in a high resolution EO Big Data scenario. Our proposed approach extends the statistical independence assumption from the features, which has been proved valid in [177] and [178], to the posterior probabilities. The proposed approach is derived from the belief that the statistical independence can be inherited if the extracted features from

the original data are independent. In this case, the proposed statistical independence assumption is defined as,

$$p(L|D) = p(L|D)_1 \cdot p(L|D)_2 \cdot p(L|D)_3 \cdot \dots \cdot p(L|D)_c \quad (3.13)$$

where $p(L|D)$ is the product of the individual posterior probabilities of each feature model, and c is the total number of feature models used.

As stated in (3.3) the posterior probability calculation requires the knowledge of $p(\omega_i|L)$ and $p(\omega_i|D)$. Assuming the posterior probability independence, we will treat each feature model dictionary independently. Furthermore, we avoid the calculations of the joint probabilities and the iterations over multidimensional representations of $p(\omega_i|L)$ and $p(\omega_i|D)$. Thus, it is possible to greatly reduce the number of operations required for the calculation of $p(L|D)$. Moreover, the computational complexity of the new algorithm is simplified. With our new statistical independence assumption the complexity can be determined as the addition of the different feature model complexities,

$$O = n_1 + n_2 \dots + n_c. \quad (3.14)$$

Simplifying the different complexities to n , as in the previous case, and assuming c is not meaningful when compared with n , the complexity changes from polynomial to linear as follows,

$$O(n \cdot c) = O(n). \quad (3.15)$$

The complexity reduction due to the new statistical independence assumption results in a huge acceleration of the required computational effort. This acceleration can be used for the inclusion of new features models in the learning stage. Since more feature models mean an extension of the possible combinations of words ω , and in consequence extended discrimination capabilities, this will be useful in more complex user semantics definition processes.

3.6 Probabilistic Retrieval

For the last decades the machine learning community has used multiple feature distances for the classification and retrieval of different multimedia assets [179]. In [20], once the information from images is captured in a feature set, two different ways to endow images with meaning are presented. The first compares the feature set with the elements in a training set, leading to conditional probabilities that sketch an interpretation of the image, but does not determine it completely. This approach is described in Section 3.6.1. The second approach relies exclusively on the feature set to generate visual signatures and compute the similarities. Examples of this approach can be found in [16]. In Section 3.6.2 we propose a modification

of the classical approach, consisting in the calculation of the similarity between the elements contributing to the posterior probability of the query image and the image signatures obtained with the BoW.

3.6.1 Retrieval Based on the posterior Probability Value

This retrieval method is the one originally implemented in KIM. The method proposed a probabilistic retrieval that relies on the $p(L|D)$ value of the images in the database. Thus, the $p(L|D)$ of every image in the database is calculated and ranked by its value. The images with a higher probability of containing the user requested semantic label, L , appear in the initial positions of the ranking.

3.6.2 Retrieval Based on Similarity Metrics

A classical retrieval by similarity relies on the image signatures computed exclusively from features. In our system the visual signatures are represented by the BoW probabilities $p(\omega_i|D)$. Thus, a classical similarity retrieval would include the similarity calculation between the stored $p(\omega_i|D)$.

Our contribution to the probabilistic retrieval modifies the classical approach calculating the similarity distance among the $p(\omega_i|D_{DB})$, BoW signature, of each element in the database, and the elements used for the calculation of $p(L|D_Q)$ in the query example according to (3.2). By doing so, we introduce the user-specific semantics into the similarity computation.

The use of similarities or distances for the retrieval allows us to introduce a new parameter in the retrieval process, namely the distance metrics. We have implemented a set of different metrics to calculate distances, d :

- **Euclidian:**

$$d_E = \sqrt{\sum_i ((p(L|\omega_i) \cdot p(\omega_i|D_Q)) - p(\omega_i|D_{DB}))^2} \quad (3.16)$$

- **Kullback-Leibler:**

$$d_{KL} = \sum_i p(L|\omega_i) \cdot p(\omega_i|D_Q) \cdot \ln \left(\frac{p(L|\omega_i) \cdot p(\omega_i|D_Q)}{p(\omega_i|D_{DB})} \right) \quad (3.17)$$

- **Kullback-Leibler symmetric variant:**

$$\begin{aligned} d_{KLS} = & \sum_i p(L|\omega_i) \cdot p(\omega_i|D_Q) \cdot \ln \left(\frac{p(L|\omega_i) \cdot p(\omega_i|D_Q)}{p(\omega_i|D_{DB})} \right) \\ & + \sum_i p(\omega_i|D_{DB}) \cdot \ln \left(\frac{p(\omega_i|D_{DB})}{p(L|\omega_i) \cdot p(\omega_i|D_Q)} \right) \end{aligned} \quad (3.18)$$

- **Jensen-Shannon Divergence:**

$$d_{JSD} = \frac{1}{2} \cdot \sum_i p(L|\omega_i) \cdot p(\omega_i|D_Q) \cdot \ln \left(\frac{p(L|\omega_i) \cdot p(\omega_i|D_Q)}{M} \right) + \frac{1}{2} \cdot \sum_i p(\omega_i|D_{DB}) \cdot \ln \left(\frac{p(\omega_i|D_{DB})}{M} \right) \quad (3.19)$$

where $M = 1/2 \cdot (p(L|\omega_i) \cdot p(\omega_i|D_Q) + p(\omega_i|D_{DB}))$

- **Manhattan:**

$$d_M = \sum_i |(p(L|\omega_i) \cdot p(\omega_i|D_Q)) - p(\omega_i|D_{DB})| \quad (3.20)$$

- **Chebychev:**

$$d_{Ch} = \max_i \{ |(p(L|\omega_i) \cdot p(\omega_i|D_Q)) - p(\omega_i|D_{DB})| \} \quad (3.21)$$

The availability of different metrics is another resource that the user can exploit in order to improve the image retrieval. As we will present in Section 3.10, the use of a specific distance metric can be useful for certain user concepts.

3.7 User Interface

The User Interface (UI), shown in Fig. 3.3, is presented as a QbE interface where the user can load an image from the repository. The user can select the example image directly navigating the repository or simply by selecting one of the 20 images that randomly are shown in the right part of the UI. These random images can be refreshed at any time just pressing the "Random" button. The first main canvas is used to represent the query example. Depending on the dataset and the features selected for the learning, it is possible that the query element contains more than one analyzed EO image, see Section 3.3. In these cases, the query canvas representation can be switched between these source images. The second main canvas, located in the center of the UI, represents a Posterior Probability Map (PPM) defined as the posterior probability ratio of each pixel in the image,

$$PPM = \frac{p(L|d_n)}{p(L|d_n) + p(\neg L|d_n)} \quad (3.22)$$

where d_n are the individual pixels in D , $p(L|d_n)$ is the posterior probability of the label L given a pixel d_n and $p(\neg L|d_n)$ the posterior probability of $\neg L$ in the pixel.

The PPM is updated with every user input providing an useful interactive tool to check the validity of the learning process. Moreover, the UI also implements different drop lists where the user can select:

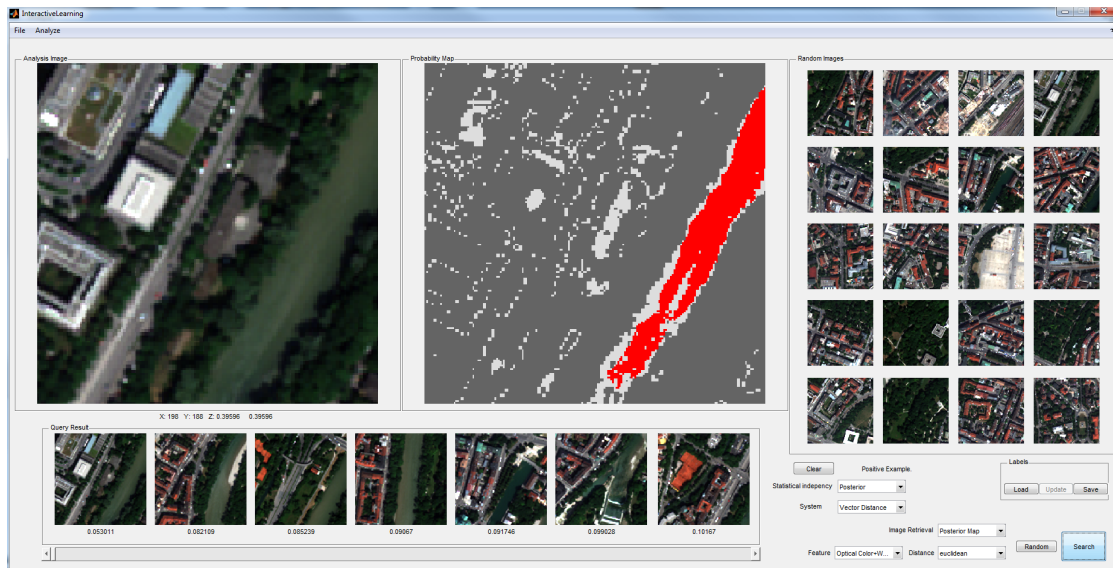


Figure 3.3: System UI. The first main graphical canvas represents the query image example, the second one the posterior probability, $p(L|d_n)$, values of each pixel. On the right, a group of random images from the database are shown. At the bottom under the main canvases the query results are shown. On the right, under the random images, the parameter selection drop lists, label load/save and search buttons are shown.

- **System Algorithm:** The user can select between KIM or HDM algorithm.
- **Retrieval Method:** It is possible to select the probabilistic retrieval method based on posterior probability value or the one based on similarity metrics.
- **Feature Models:** The user can choose different combinations of feature models which will be used in the learning-retrieval processes.
- **Similarity Metric:** In case of the probabilistic retrieval based on similarity metrics is selected, a list for selecting the desired metric is enabled. The available metrics are listed in section 3.6.2.
- **Image Source:** Certain feature model combinations have different EO image source. For this cases the user can switch at any moment the image source shown in the main canvas. This option can be helpful during the learning process in order to improve the quality of the input introduced by the user.

After every query, the retrieved images are ranked under the main canvas. Those ranked images can be clicked so they can be loaded as example image for continuation of the training. Finally there is the option to save, load and update the user defined labels, L and the associated training hyperparameters.

3.8 User Interaction

In a common search procedure, the posterior probability map starts all in grey, representing the unknown state of every pixel in the image due to the lack of positive or negative examples. With every provided example, N_i^{k+1} , the hyperparameters α_i^{k+1} and the stochastic link, $p(\omega_i|L)$, are updated. As a consequence, the posterior probabilities, $p(L|d_i)$, are also updated and with them the PPM, defined in (3.22), that is shown to the user. Black pixels will represent low probability and white ones high probability. Once the probability of a pixel is over 0.9, the pixel is highlighted in red.

At any moment, the user can perform a query to check the retrieved results. By checking the results, it is possible to refine the learning process introducing, for instance, negative examples over an image retrieved in the first positions, but which does not contain the label L . As a direct response of this negative example, the image will be penalized in the next search appearing in a lower ranking position.

During the learning retrieval process, the user can decide to try one of the different approaches implemented for probabilistic retrieval. Using the probabilistic retrieval method based on similarity metrics, it is also possible to try one of the several implemented distance metrics described in Section 3.6. Once the user agrees with the semantic label definition, it can be stored in the database to be reused in further analysis.

3.9 Application Scenarios

In order to test the applicability of the system and the implemented new algorithms, we present three different scenarios. The analysis of these scenarios is only possible due to the speed-up achieved with the new learning algorithm which allows the introduction of a higher number of feature models. In the first scenario, called Multispectral-SAR fusion, the goal is to demonstrate the system speed-up and performance in comparison with the original KIM implementation for urban assessment. The second scenario, Multitemporal-SAR fusion, uses the system with image time series for change detection applications. And, in the final scenario, Multispectral-SAR-Map fusion, we combine the classical image analysis feature models with a model extracted from map information stored in a GIS server.

3.9.1 Case: Multispectral-SAR Fusion

For validating the system we have chosen Munich city multispectral images from WorldView-2 (WV-2) and SAR images from TS-X both with 1.25 meter pixel spacing, covering an area of 24 km^2 , as shown in Fig. 3.4. The size of the total scene is 4890×3202 pixels cut into tiles of 200×200 pixels, with a total number of 500 tiles.

	Search Proc.	1 Model	2 Model	4 Model
Feature Independence (KIM)	Posterior Probability	0.62 s	6.39 s	2354 s
	Sim. Metric <i>JSD</i>	0.045 s	0.15 s	43.5 s
Posterior Independence (HDM)	Posterior Probability	0.155 s	0.196 s	0.31 s
	Sim. Metric <i>JSD</i>	0.042 s	0.126 s	34.61 s

Table 3.1: System query run-time for different statistical assumptions, query ranking types and feature model numbers. The first row is used as a threshold and corresponds to an emulation of the original KIM implementation. The difference among the four models using the new statistical assumption is four orders of magnitude.

The clustering and BoW generation of this dataset resulted in 256 words for the intensity feature and 19 words for the WLD texture feature in the multispectral image, along with 124 words for the intensity and 8 words for the texture features of the SAR image. Summarizing, each database object is composed by one multispectral patch, one SAR patch, four word maps and the associated probabilities.

The evaluation of the system and algorithm performance is done by measuring the time required for the completion of a query and the quality of the query results. We compare the execution time of the original KIM algorithm with the modified HDM algorithm presented in Section 3.5.2.

The first experiment of this scenario calculates the query processing speed of the system for KIM and HDM algorithms using the two available retrieving methods, presented in Section 3.6. The feature model combination used are: multispectral intensity for the unique model case, multispectral intensity and texture when two models are used; and multispectral and SAR intensity and texture in the four model tests. Table 3.1 summarizes the obtained results. In the case of the KIM algorithm, the improvement of speed using similarity metric retrievals amounts to one or two orders of magnitude compared with the posterior probability approach. When comparing KIM and HDM algorithm performances, HDM turns out to be faster. The HDM similarity metric retrieval method performs faster than KIM but in the same order of magnitude. In contrast, using the posterior probability retrieval method, the HDM performs four orders of magnitude faster than the original KIM implementation for the four feature model case.

In a second experiment we define a fixed learning process based on positive and negative examples over the same pixels in the same images. The first stage of this experiment, *A.1* in Table 3.2, shows the initial results and the first error, (i.e., the first misclassification result), of the similarity metric based retrieval versus the posterior



(a) WorldView-2



(b) TerraSAR-X

Figure 3.4: Multispectral-SAR case scenario of Munich, Germany.

3. Knowledge-driven Heterogeneous Data Mining



























Exp.	Label	Query Example	Alg.	Retrieval Method	Posterior Map	Retrieved Patches				Error Position	Error
A.1	River		HDM	Posterior Probability						1	
			HDM	Sim. Metric KL						11	
A.2	Railway		KIM	Posterior Probability						10	
			HDM	Posterior Probability						13	

Table 3.2: System query results. First experiment shows a better performance of the similarity metric retrieval method in initial learning stages. The second experiment shows how the simplification of the calculation processes obtained with the HDM algorithm does not affect the quality of the retrieved results. Moreover, for some cases the query results are even better.

probability ranking methods when looking for the user query *river*. The four-feature models are used in the learning stage. After a few user interactions the query output shows an improvement of the results using the similarity metric retrieval method. For this case, the patches shown in the initial positions match the query image; meanwhile, the posterior probability retrieval method is still insufficiently trained to provide correct results.

In the second part of this experiment we evaluate the final query response using the feature statistical assumption implemented in KIM and the new HDM algorithm. We show the first ranked images and the first misclassified element, with its position in the rank. In the results, shown in A.2 of the Table 3.2, the user searches for *railways*. We can observe how the HDM algorithm provides also good results. Moreover, in some cases, like in the one shown in the experiment results, the first error appears even later, at position 13th.

3.9.2 Case: Multitemporal-SAR Fusion

For the second application scenario a sequence of two TS-X images has been chosen in order to show the system capabilities to detect changes in an image time series. In this case, the query example involves at the same time feature models generated from both SAR images. The images cover a part of the Elbe river course in Germany, see Fig. 3.5. This region suffered a severe flood during 2013. The images correspond to an initial acquisition on June 26, 2008 and a second one during the flooding on June 15, 2013. Both images have the same characteristics, they correspond to a stripmap level 1B product with horizontal polarization and 5.88 meter azimuth and ground range resolutions. The size of the scene is 11453x20528 pixels with a pixel spacing of 2.75 meters and a covered area of 1778 km^2 . Each image has been cut in small tiles of 200x200 pixels, resulting in a total of 5814 patches. Like in the previous case,

the system manages pixel based feature clustering results. The four features used are the intensities, from which we obtain 56 words, and WLDs, 31 words, from both SAR images. For this dataset each database object is composed by two SAR image patches, four word maps and the associated probabilities. In this case we assume the statistical independence based on the large time interval between acquisitions, 5 years. To verify the statistical independence, we computed a similarity map based on the Normalized Compression Distance (NCD) [180], obtaining high dissimilarity values. This experimental approach was employed since the analytical verification of statistical independence is a highly complex problem [181]. In the next three experiments we retrieve all the patches in the database. The ranking is done by using the posterior probability value in order to get first the patches with a higher probability of containing the semantic concept defined by the user.

The first experiment, *B.1* in Table 3.3, aims the search of flooded areas in the scene, where a flooded area is represented as an object in the database containing areas with no water in 2008 and which are covered by water in 2013. After giving just one positive example, the first error (i.e., the one corresponding to an image with no flooding) appears at position 132nd. Continuing with the training, misclassifications start to be common only after position 1300th.

The second experiment of the multitemporal scenario, *B.2*, aims to retrieve non flooded patches. This includes images with no change at all (e.g., permanent course of the river) and images with changes not related with the flooding (e.g., crop changes). The first patch with severe flood appears at position 338th just with a unique positive example. Following the tuning of the learning process with more positive and negative examples, the error proliferation starts at 1771st position and they start to be regularly ranked after position 4600th.

In the third experiment, *B.3*, the user retrieves images with agricultural fields that have changes in the the crop. The initial query, with only one positive example, provides correct results until position 33rd. Continuing the learning, the first misclassification goes backwards to position 364th and their appearances become more regular after position 1200th.

3.9.3 Case: Multispectral-SAR-Map Fusion

In this scenario the system is tested with the same dataset used in Section 3.9.1, i.e., optical bands of a multispectral image (WV-2) and a SAR (TS-X) image (see Fig. 3.4). Additionally, as GIS map feature model, we generate a raster image using information from the OpenStreetMap collaborative project (see Fig. 3.6). From this map feature model we obtain 7 different words corresponding to the extracted classes in Section 3.3. Thus, the database objects are composed by one multispectral patch, one SAR patch, four word maps with the associated probabilities from the EO images and a word map with the associated probabilities from the GIS map.

The first experiment intends to demonstrate the acceleration of the learning

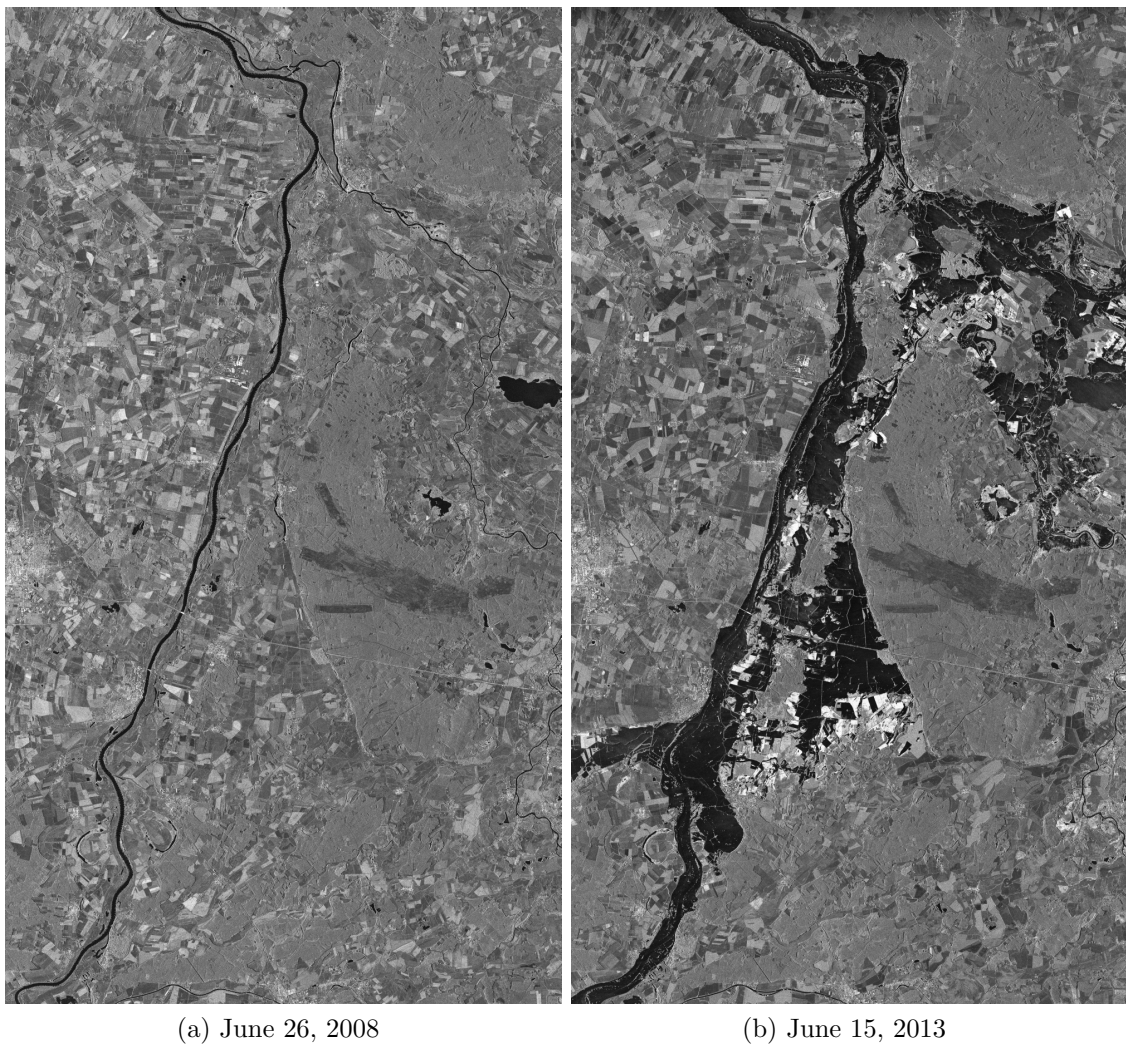


Figure 3.5: TerraSAR-X images of Elbe river years before and during the 2013 flooding. The flooded areas can be seen in black.






















Exp.	Label	Query Example	Year	Retrieval Method	Posterior Map	Retrieved Patches	Error Position	Error
B.1	Flood		2008	Posterior Probability			1300	
			2013					
B.2	No Flood		2008	Posterior Probability			1771	
			2013					
B.3	Crop Change		2008	Posterior Probability			1200	
			2013					

Table 3.3: Multitemporal-SAR fusion case scenario. *B.1* experiment shows the case where the user searches for flooded regions. The experiment *B.2* shows the case in which not flooded images are requested. In the last experiment of this case scenario, *B.3*, images with crop change during the time series are requested.

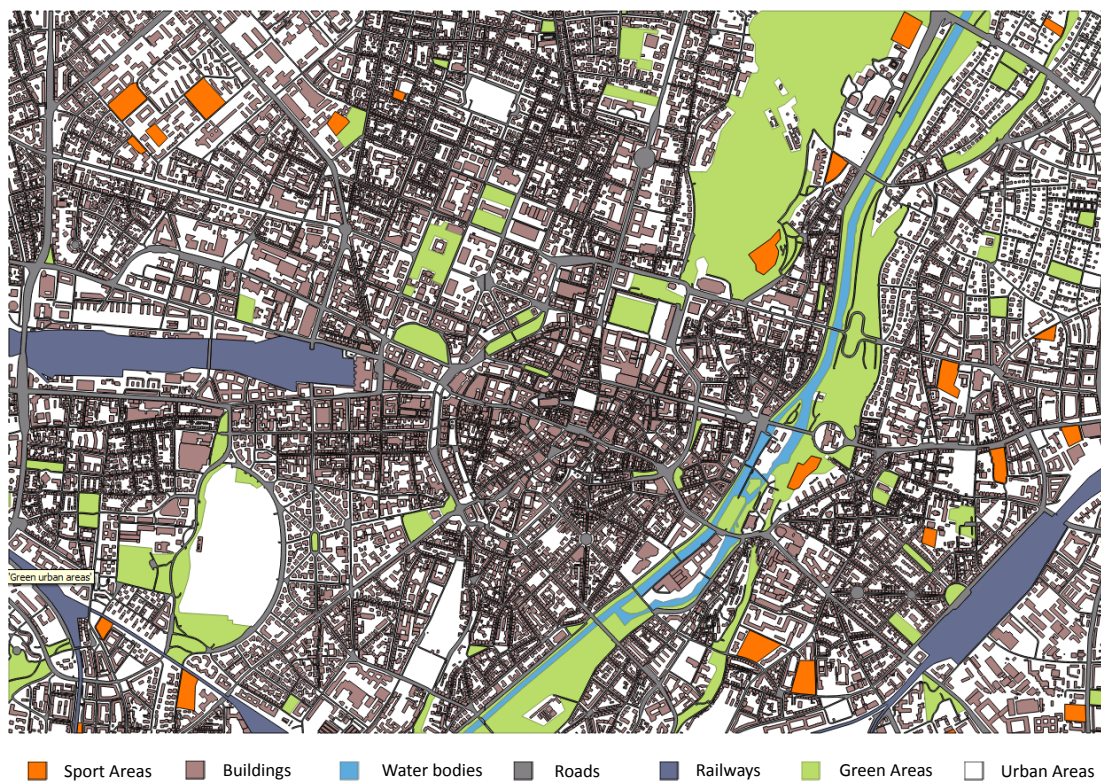


Figure 3.6: Raster map with several classes extracted from OpenStreetMap. Munich city, Germany.

process due to the insertion of a generic map feature model. The features used in this experiment are intensity and WLD features from both images, and the GIS map. The retrieval method is the one based on the total sum of the posterior probabilities. In this experiment, the user looks for tiles with *small paths* in parks or gardens. The first stage, *C.1.1* in Table 3.4, shows the speed-up of the learning process since the first positive example. The first row presents unsatisfactory results obtained without the inclusion of the map feature model. By contrast, including the map, the first ranked images contain the user-defined concept. Here, due to the optical intensity feature model, white pixels are highly pushed up in the ranking. The use of the posterior probability based retrieval method, brings the tiles containing a bigger amount of white pixels (e.g., tiles containing clouds or places under construction) to the top of the ranking. Nevertheless, for the query including a GIS feature model, the inclusion of extra information related with roads, adds an extra discrimination capability. This enables, from the very beginning, the retrieval of positive matches in the top of the rank.

C.1.2, in Table 3.4, shows the first top four positions of the rank after adding one more positive and one more negative examples. The initial results are good in both cases. However, in the case without a map feature model, the first error appears at the 5th position. Meanwhile, in the case with a GIS feature model, it appears at position 17th.

The second experiment shows the system performance for the discovery of more complex classes like *dry river borders* with and without the addition of the feature model from OpenStreetMap. The query method in the two first stages, *C.2.1* and *C.2.2*, is based on the posterior probability. The third stage, *C.2.3*, uses the similarity metric retrieval in order to show the importance of the chosen retrieval method after the same learning process.

As in the previous experiments, *C.2.1* shows the response of the system after one initial positive example. Only the experiment with a GIS map feature model shows initial good results. Continuing the active learning, *C.2.2*, the case without the map feature model is unable to learn the concept, providing only a correct tile in position 16th of the rank. However, for the query including the map feature model, the learning improves. After three positive and two negatives additional examples, the first wrong result moves from the initial 6th position, in the *C.2.1*, to the 22nd.

In the last stage of this experiment, *C.2.3* in Table 3.4, we compare the retrieved results following the same learning process of *C.2.2*, but using now the similarity metric retrieval method. In contrast to *C.2.2*, even not using a GIS feature model, it is possible now for the system to learn the user concept and obtain positive results. The first error appears in this case at the 9th position. In the second query, a map feature model is used, obtaining results with the same quality with a smaller number of interactions.

Exp.	Label	Query Example	GIS	Retrieval Method	Posterior Map	Retrieved Patches					Error Position	Error
C.1.1	Small Path		No	Posterior Probability							1	
			Yes								1	
C.1.2	Small Path		No	Posterior Probability							5	
			Yes								17	
C.2.1	Dry River Border		No	Posterior Probability							1	
			Yes								6	
C.2.2	Dry River Border		No	Posterior Probability							1	
			Yes								22	
C.2.3	Dry River Border		No	Similarity Metric KL							9	
			Yes								22	

Table 3.4: Multispectral-Map fusion case scenario. In the first experiment *C.1*, the user searches for *smallpaths* surrounded by vegetation. *C.1.1* shows the results after the first user example. *C.1.2* shows the results after two positive and one negative example. The last experiment, *C.2*, aims to search for the user defined concept *dryriverborder*. In *C.2.1* and *C.2.2* the system is unable to learn the user concept if the map model is not used. For the case with map layer the retrieved images contains the defined label. *C.2.3* shows that when using matrix distance search, even without the map model, the concept learning is possible and for the case with map model the learning process tends to be faster.

3.10 Performance Evaluation and Discussion

More general system results can be extracted by manually annotating the previously introduced datasets. The annotation process involved the generation of a record with the presence of different semantic concepts over the whole datasets. Due to the highly time consuming nature of the annotation task, it was possible to calculate quantitative statistical parameters only for the concepts presented in this section.

Extending the first experiment of the Multispectral-SAR case scenario, we can see in Fig. 3.7 the system response with different configurations retrieving several dataset sizes. It is possible to detect that for an unique model and for the fusion of two models, the retrieval based on similarity metrics performs faster than the posterior probability retrieval. The reasons are two: first, we avoid the calculation

of the posterior probability on the entire database and compute it just on the query image; and second, the dictionary sizes for the distance calculations are short enough to ensure a fast computation over all the elements of the database. However, this tendency changes when the number of feature models to fuse is more than four. In this case, the number of elements in the dictionary increases to a point where the cost of calculating the posterior probabilities in comparison of the computation cost to calculate the distances is trivial. This is exactly the reason why the computation times for four feature models using similarity metric retrieval are alike for both KIM and HDM.

Summarizing, we can say that the distance metric retrieval method is the most efficient one up to the combination of two models. If more feature models are used, the posterior probability value based retrieval performs faster. This can be explained, first, due to the simplification in the posterior probability calculation over the whole database. And second, because the subsequent ranking of scalar values is less expensive than the one based on similarity metrics.

To validate the introduced probabilistic retrieval based on similarity metrics, different queries were performed using the distance metric as a parameter instead of using a fixed unique metric. Table 3.5 shows a summary with the best query results for two different semantic labels with different combination of feature models. Specifically Table 3.5 provides precision, recall, accuracy and F_1 measures, i.e, the equally weighted harmonic mean of precision and recall. A detailed explanation of the measurements can be found in [182].

In the first test we define the user concept *river* using the four feature model employed for the experiment in Section 3.9.1 in the learning process. The first ranked 25 tiles are retrieved using all the implemented metrics. The results show the Chebychev metric outperforming the rest of the metrics. The precision, 80%, and recall, 83%, are at least a 4% better than the rest of the metrics. The obtained accuracy is 97% with a value of F1 measure of 81%, three percent better than the second best metric. The second experiment presents the query results for a *turquoise roof* user concept using in the learning stage the two feature models from the multispectral image, intensity and WLD texture. The number of images retrieved using the similarity metric method is 100. It is remarkable how Kullback-Leibler and Jensen-Shannon divergence outperform greatly the rest of the metrics. In this case, Kullback-Leibler provides an 84% of precision, 74% of recall.

Finally, we have extended the Multitemporal-SAR case scenario to present more exhaustive quantitative measures of the retrieved results. In Fig. 3.8 we provide precision, recall, accuracy and F_1 measure for different user defined concepts. The graphics show how the different measures vary depending on the number of retrieved tiles from the database. In almost all the cases, the precision of the system for retrieved tiles from 1 to 1000 remains over the 90%. The recall increment varies depending on the amount of images in the database annotated with the concept. The total amount of tiles is 5814, from which 1409 are annotated with the *flooded*

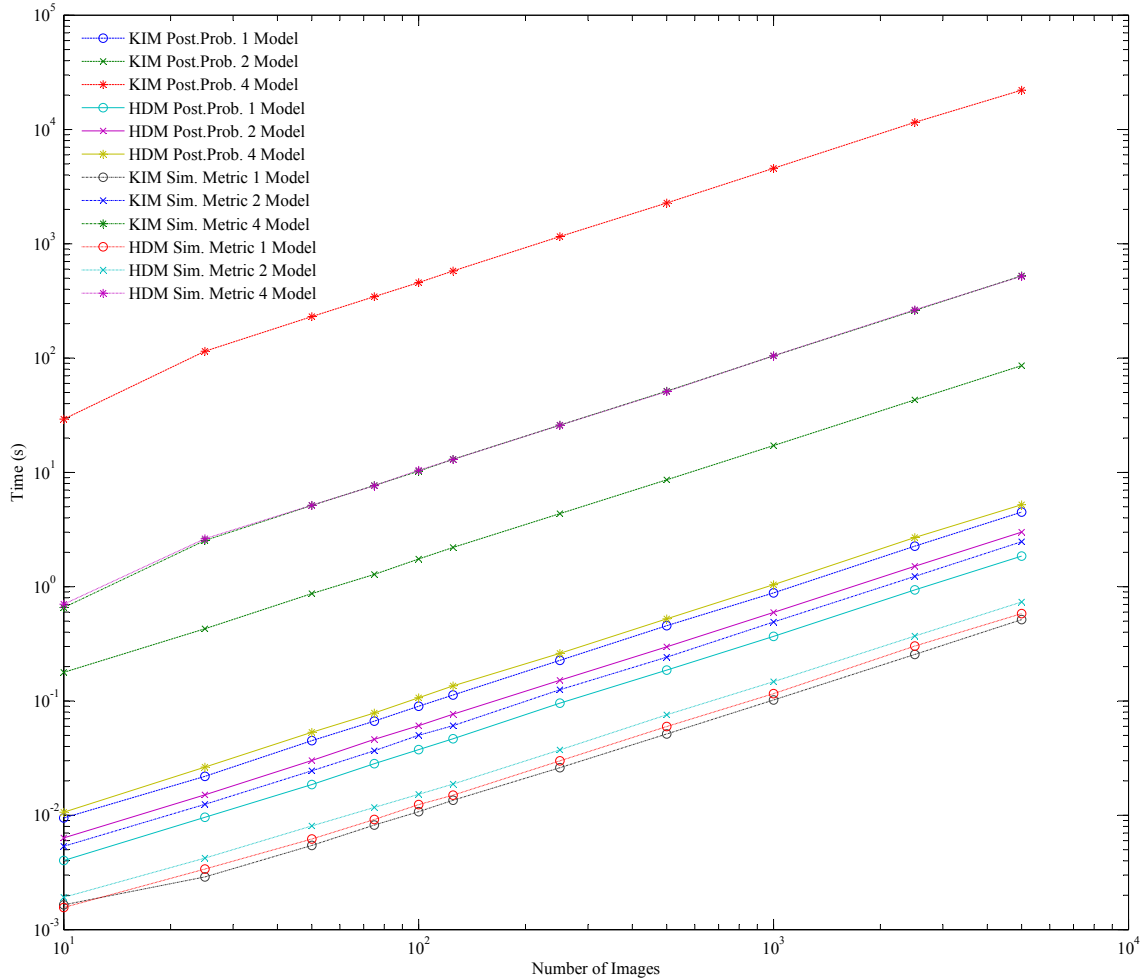


Figure 3.7: Query computation time curves. It is clearly visible the difference in orders of magnitude between retrieval time using KIM and HDM algorithms. The complexity of the KIM algorithm limits in practice the number of feature models in the learning stage. The used of HDM allows us to convert the learning process into a linear complexity problem. This enables a real time response to every user interaction and for the database querying process. The similarity metric retrieval method is the most efficient one up to the combination of 2 models. After that, the posterior probability retrieval method performs better.

3. Knowledge-driven Heterogeneous Data Mining

User Concept	Feature Models	Images Retrieved	Metric	Precision	Recall	Accuracy	F1 measure
River	Multispectral Intensity & WLD, SAR Intensity, & WLD	25	Euclidian	44%	45%	92%	45%
			Kullback-Leibler	72%	75%	96%	73%
			Kullback-Leibler S.	76%	79%	97%	78%
			Jensen-Shannon	72%	75%	96%	78%
			Manhattan	76%	79%	97%	78%
			Chebychev	80%	83%	97%	81%
Turquoise Roof	Multispectral Intensity, WLD	100	Euclidian	45%	40%	66%	42%
			Kullback-Leibler	84%	74%	88%	79%
			Kullback-Leibler S.	53%	47%	70%	50%
			Jensen-Shannon	80%	70%	85%	74%
			Manhattan	50%	44%	69%	47%
			Chebychev	41%	36%	64%	38%

Table 3.5: System query results. First test defines the user concept *river* using a four feature models in the learning process ranking 25 tiles and using all the implemented metrics. The results show the Chebychev metric outperforms the others in overall. The precision, 80%, and recall, 83%, are at least a 4% better than the rest of the metrics. The second experiment presents the first one hundred query results for a *turquoise roof* user concept using in the learning stage the two feature models from the multispectral image, intensity and WLD texture. Kullback-Leibler and Jensen-Shannon divergence outperform greatly the rest of the metrics. In this case, Kullback-Leibler outperforms the rest of the metrics with an 84% of precision, 74% of recall.

label, 4405 tiles with the *non flooded* label, 1930 tiles annotated as *non changed* and 2474 tiles as *crop change*.

3.11 Conclusions

We have presented an HDM prototype inspired by and following the main concept summarized in Section 3.1 and previously implemented in the KIM system in [3]. The HDM enhances the original KIM system overcoming the two-model limitation. HDM introduces a faster active learning algorithm modifying the required statistical independence from the features to the posterior probabilities. The obtained speed-up allows the introduction of new feature models in the learning stage and the definition of more complex user semantics. The acceleration can also open new ways for knowledge-driven information mining systems to Big Data scenarios.

For comparison purposes we re-implemented the original KIM method, and based on it, we introduced new search methods and theoretical probabilistic assumptions which may outperform in speed the previous one by various orders of magnitude. The proposed probabilistic search method based on the distances between the elements used for the calculation of the posterior probabilities and image BoW in the database performs better for weakly defined labels. However, for two different reasons, this search approach cannot replace completely the approach based on total posterior probabilities. First, the posterior probability based retrieval yields the

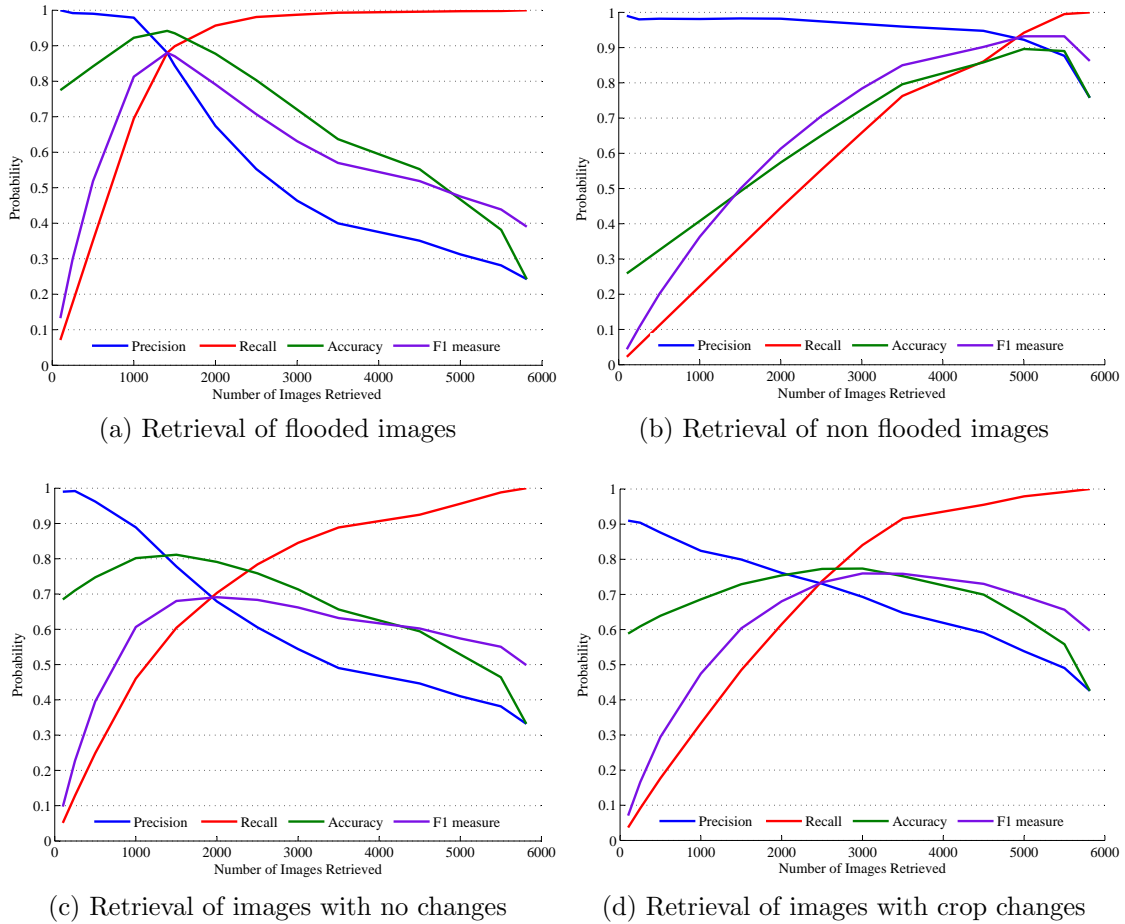


Figure 3.8: System quantitative results for the retrieval of different user-specific semantics.

tiles with a higher probability of containing the required user semantics. Second, due to the simplicity of the scalar ranking of posterior probabilities, this approach performs faster for scenarios with a feature model number greater than two. The latter fact is due to the increase of the computation cost of the similarity distances with each extra feature model.

We experience a considerable speed-up of the learning algorithm by introducing the assumption of posterior probability statistical independence, which does not seem to introduce biases in the learning processes. Moreover, for some cases, it outperforms the original concept when looking at the first misclassification in the ranking.

Furthermore, we have demonstrated the system performance in a time series case scenario. The system did detect successfully different types of image changes, such as flooded areas or even the crop rotation in agricultural fields. The system is also able to retrieve efficiently unchanged patches.

3. Knowledge-driven Heterogeneous Data Mining

Finally, we have implemented a link to an external data infrastructure which allows us to include a feature model in the learning processes based on information independent of the image content (e.g., information extracted from GIS maps). This link provides a new tool for the improvement of the active learning processes and the posterior search and retrieval operations. Our tests have shown promising results, allowing the definition of more complex semantic concepts.

Multilayer Architecture for Heterogeneous Geospatial Data Analytics

The constantly growing process of the Earth Observation (EO) data and their heterogeneity require new systems and tools for effectively querying and understanding the available data archives. In this chapter we present the system architecture of a tool for heterogeneous geospatial data analytics. The system implements different web technologies in a multilayer server-client architecture allowing the user to visually analyze satellite images, maps and in-situ information. Specifically, the information managed is composed of EO multispectral and Synthetic Aperture Radar (SAR) products along with the multitemporal in-situ Land Use/Cover Area frame Survey (LUCAS). The integration of these data provides a very useful information during the EO scene interpretation process. The system also offers interactive tools for the detection of optimal datasets for EO multitemporal image change detection, providing at the same time ground truth points for both, human and machine analysis.

The chapter continues with Section 4.1 where the importance of data in integration in EO is pointed out. Section 4.2 deeply describes the architecture of the system and the multiple layers that compose it. Section 4.3 shows the performance evaluation of the system in multiuser and multidevice environments. Section 4.4 presents different functionalities of the system, such as capabilities for a better understanding of EO images (Section 4.4.1) and tools for optimum dataset selection (Section 4.4.2). Finally, Section 4.5 contains the conclusions summing up the chapter.

The content of this charter will be published in: K. Alonso, D. Espinoza-Molina, and M. Datcu, Multilayer Architecture for Heterogeneous Geospatial Data Analytics: Querying and Understanding EO Archives, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 791-801, March 2017.

4.1 Heterogeneous EO Data Integration

The EO data heterogeneity as described in Section 2.1 has different sources where sensor diversity, which includes a different variety of imagery and related metadata; digital cartography; and in-situ data are very prominent elements. Regarding data integration, Geographical Information Systems (GIS) play a key role. GIS are defined in [183] as "computer-based information systems that enable capture, modeling, storage, retrieval, sharing, manipulation, analysis, and presentation of geographically referenced data". From the initial standalone GIS architectures, the internet development has promoted the intercommunication among GIS specially via Web services [184].

In the same way, plenty of the scientific community work has been focused in the link of information sources by means of integration and fusion of the different information sources. Regarding the information integration, implementations with EO data have been presented for security and hazard decision makers like GEODec [185] or the system introduced in [186] which aims to support the Earthquake research and disaster response. On the same subject, the work in [187] presents a geospatial data processing functionality to support collaborative and more efficient emergency response. This is achieved by integrating distributed in-situ data with very high resolution optical EO images providing: geospatial data queries, on demand image processing, and fast map visualizations. We also find projects like EOLib [161] or TELEIOS [162] where EO image metadata and linked data are used as query parameters in order to improve EO image retrieval results.

The heterogeneous data integration brings new possibilities for data representation and visualization. The value of data integration in geospatial infrastructures was shown in [188] where the correct design of data and information models was mandatory in order to assure the interoperability at metadata, data and semantic levels.

4.2 Architecture

This system is designed to support EO analysts and expert users through analytical processes. Thus, it is meant to provide tools for data visualization, statistical analysis and data management, with the objective to improve the EO image understanding and help in the dataset selection for multitemporal change detection. Another technical requirement is the capability to handle multiple users accessing simultaneously the system from different devices running different Operative Systems (OS). Aiming to offer such an ubiquitous tool, the system is based on web technologies following a server-client philosophy. The architecture of the system has been designed using a multilayer approach, see Fig. 4.1. The main server side is composed by: the data source layer, data ingestion layer, database management

system, and user oriented web functionality layer. The system is also specifically designed to rely all the computational complexity over the server, making the client side lightweight. In this way, the client is only composed by the Graphic User Interface (GUI) layer, which can be accessed from any electronic device capable of running a web browser with HTML-5 compatibility.

4.2.1 Data Sources

One of the system's main feature is the possibility to work with heterogeneous information sources. The heterogeneity on the data offers big possibilities to the researchers, allowing them to analyze specific scenarios from different perspectives. The presented system is able to manage and process information from the LUCAS survey, introduced in Section 3, and different satellite imagery, optical (Section 2.1.1) and SAR (Section 2.1.5).

Despite of the actual implementation, the designed system architecture, based on well-known standards, makes possible to easily integrate more data sources, e.g., hyperspectral images (Section 2.1.4).

4.2.2 Data Ingestion

The Data Ingestion layer performs one time non iterative processes in order to populate the system data repositories. Two main processes can be differentiate: metadata ingestion, and tile generation. The first process extracts the LUCAS survey metadata, in CSV format, to be ingested into the system geographical database. The ingestion is made by parsing the extracted metadata to series of Standard Query Language (SQL) queries that insert and submit the information to the geographical database. The second process of the ingestion module produces tiles of the EO products at different zoom levels in order to make more efficient the visualization of those via Tile Map Services (TMS) [189]. The system base projection is WGS-84. If the data source has a different projection, a reprojection process can be performed before the tile generation. A TMS only provides access to the geographical map representation of the EO data, not to the data. Additionally, thumbnails of the LUCAS images are generated in order to optimize server-client communications.

4.2.3 Database Management System

The Database Management System is composed by the main geographical database and a data repository in the system archive. The geographical database rests on PostgreSQL (object-relational database) technology [190, 191] with the spatial database extension PostGIS. The extension adds support for geographic objects allowing location queries to be run in SQL. PostGIS also enables the creation of a

4. Multilayer Architecture for Heterogeneous Geospatial Data Analytics

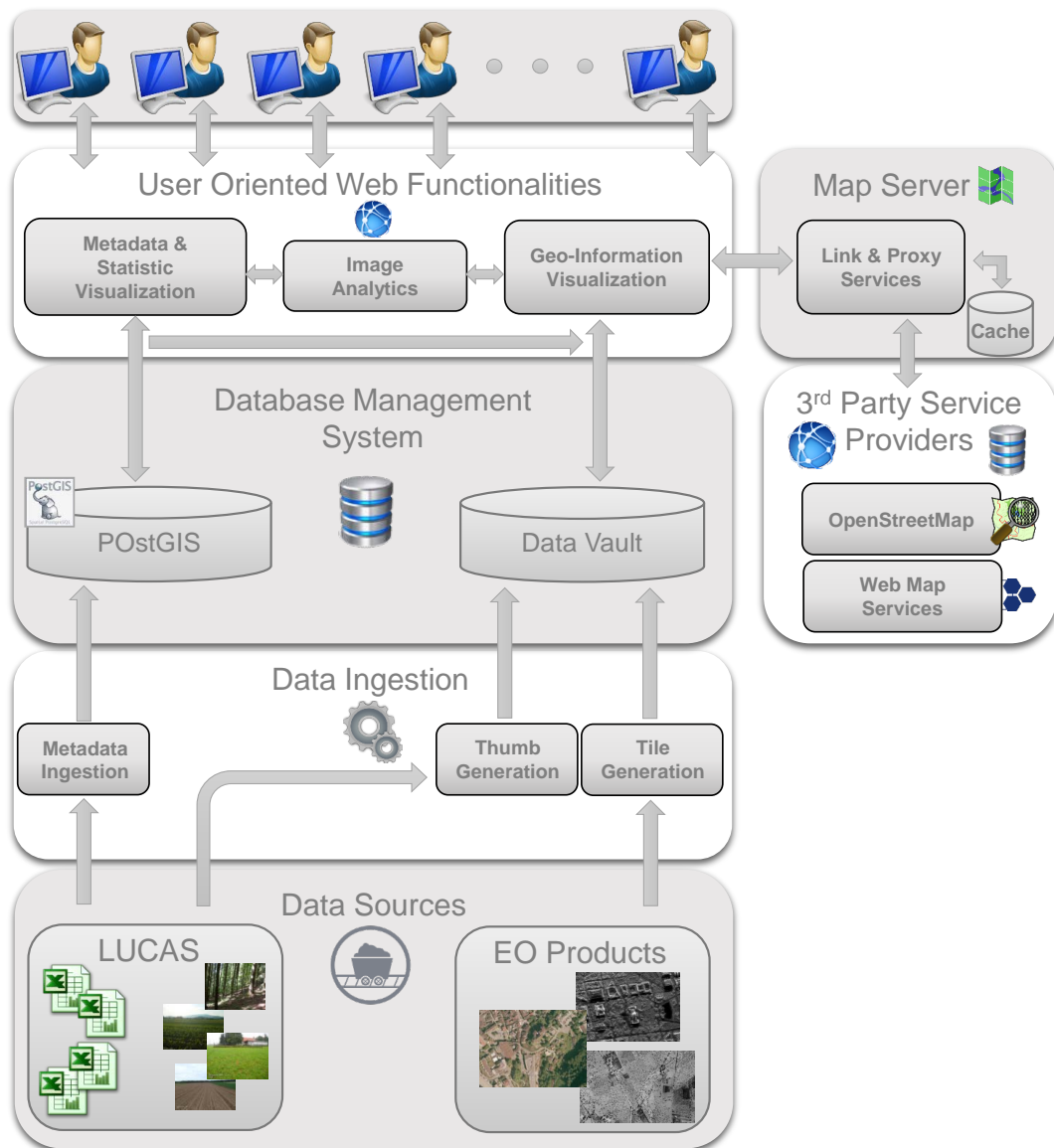


Figure 4.1: System architecture. Following a server-client philosophy, the system is designed to rely all the computational complexity over the server making the client side lightweight.

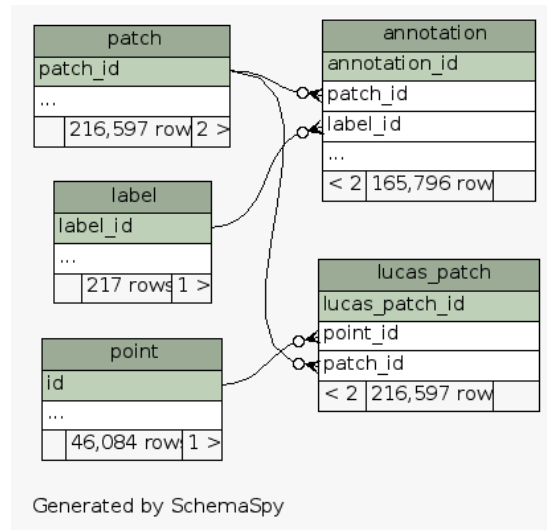


Figure 4.2: Compact relationship of the Lucas database scheme

database schema defining spatially the content of LUCAS survey of different years, and in consequence, allowing spatio-temporal queries.

A compact scheme of the geographical database for the LUCAS data is shown in Fig. 4.2. The *Point* table contains all the geographical locations of the points where the survey was performed. In addition, this table stores information about the date of survey, latitude and longitude coordinates, and geometry. The *Patch* table stores the information of the photos taken on the observed points. This table provides an URL to each picture in the data repository, thus linking the survey metadata with the multimedia images. The table *Lucas-Patch* consists of the relation between the point and path tables. *Label* table comprises the semantic labels, which describe the land use and land cover categories. The labels are stored following a hierarchy, which is specified in the level field of the table. The *Annotation* table stores the annotation of one patch with several semantic labels, that is the relation between patch and semantic labels.

4.2.4 User Oriented Web Functionalities

The User Oriented Web Functionalities module processes every user request. This layer implements all the server logic procedures required. It is divided in three main logic blocks: 1) Geo-Information Visualization, 2) Metadata and Statistic Visualization, and 3) Image analytics.

The first logic block, the Geo-Information Visualization block, performs the communication protocols with third party service providers and/or the Database Management System in order to retrieve the required visual information. Going into

detail in the communication with the third party service providers, the system implements in parallel an instance of MapServer [192]. The system centralizes all the communication with the third party providers through the MapServer, who works as a proxy. In this way the user and the system main logic remain isolated and avoid cross-domain communications. Furthermore, MapServer already implements most of the standardised protocols for communication and publication of spatial data on the web, e.g., Web Map Service (WMS) [193], allowing an easy connection between the main server, acting as a client, and the third party service providers which serve the data. WMS is the most spread geographical map producer standard on the web. WMS providers generate on-live map representations in a pictorial format, e.g., PNG, of the geographic information to every request. In this sense, one last useful functionality of MapServer is the data caching. Mapserver implements tile caching capabilities through the MapCache project which can improve the system performance in an multi-user environment by reducing the data request number to WMS providers. All the obtained visual content can be used in the analytical block or just be directly presented to the user.

The Metadata and Statistic Visualization block is composed by a set of functions and procedures to collect the data from the database required to generate the required data visualization. It is important to point out that the visualization is done in the user machine and this block functionality is limited to the data acquisition, processing and parsing. There are two types of data to handle: 1) raw data, and 2) statistical data. The raw data comprises the information of a single LUCAS point using direct metadata retrieval from the database without any processing step. On the other hand, the statistical data generally involves the retrieval and statistical processing of a group of LUCAS points in a geographical region. Both data are finally parsed to the structure required by the GUI.

The products generated by the previous processes are presented together to the user by the Image Analytic block. This process collects the interaction of the user with the data and sends the required instructions to Geo-Information and statistical visualization processes in case an update is required.

4.2.5 Graphic User Interface

The computational complexity relies on the server making the client lightweight and operational in any electronic device capable of running a web browser with HTML-5. Through the user interface, shown in Fig. 4.3, the user interacts with the system. The main map canvas, upper-center, is developed using WebGLEarth library [194] [195], which takes advantage of the Web Graphic Library (WebGL) [196] technology to render a 3D Earth globe on the browser without any external plug-in requirement. The canvas supports several WMS or TMS layers simultaneously, which can be enable/disable at will. Thus, the user is able to visualize simultaneously or to switch from OpenStreetMap to the EO SAR layer just by clicking a button. The

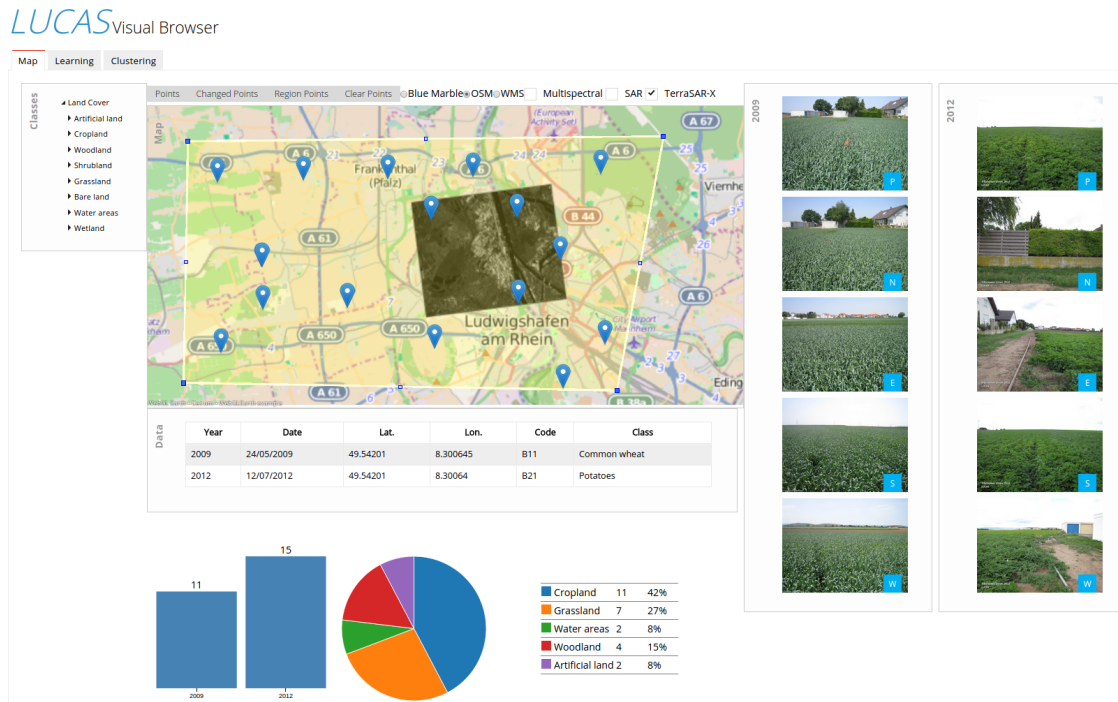


Figure 4.3: Graphic user interface of the system.

communications required to obtain the WMS layers from third party services are transparent to the user, who receives all the information from the Image Analytics module.

Another feature of the map canvas is the capability to use a polygon based region selection tool, which allows to focus the analysis in an specific area of interest. Once a region is selected, it is possible to: 1) query all the points inside the region, 2) get the survey points with land cover changes among the surveys or even 3) ask for the points with a specific land cover. While the first two options are done by pressing the buttons over the map canvas, the specific land cover query is done using the hierarchical land cover tree in the upper-left of the GUI. Besides the preconstructed queries, the users can produce more tailored queries by selecting specific combination of land cover changes, survey dates or even maximum distances between points.

The points are displayed on the map by using markers. Marker generation is a build-in feature of WebGLEarth library, which we customized in order to introduce Scalable Vector Graphic (SVG) markers with the capability to include a variety of color codes to represent different information of the data.

Along with the region points, interactive statistical charts with the point information are presented, lower-center in Fig. 5.2. The charts are generated using D3JS [197] library which provides a wide variety of interactive visualization components. It is also possible to click over each point for checking its specific information. This

information is shown under the map canvas in a table form, in which the user can check specific details about the location, land cover or acquisition times. While the table is generated, the in-situ images of the different surveys are loaded for comparison and analysis in the right side of the GUI.

4.3 Performance Evaluation

The performance of the system is evaluated on a development workstation with an Intel Core i7-2760QM CPU, 8 Gigabyte of RAM memory, Gigabit Ethernet network adapter and Ubuntu 12.04 as operative system. We present two different performance evaluations: 1) multiuser evaluation, and 2) multidevice evaluation.

The multiuser tests have been performed on an Ethernet Local Area Network (LAN) with a total of ten simultaneous users running Ubuntu or Windows7 as OS with Firefox or Chrome as Browser. During the tests the loading of main page of the system turned out to be the mayor bottle neck in the system performance. The initial HTML file links all the required libraries and has a size of 4.15 Megabyte. This initial loading is unique per user session but it is the most data intensive request the server must handle. The rest of the requests consist mainly of tiles and specific data requests which are numerous but small in size. Typical loading time for the main page is 4 seconds, but it goes up to 12 seconds when al the users start a session synchronized. The multiuser tests also show that once this initial data transmission peak is over the system can handle the user interactions seamlessly. The users are able to use the system with a mean latency around 250 ms with spikes of 700 ms when requesting high amount of tiles.

The multidevice test is carried out on a Wireless Local Area Network (WLAN) using the following devices: laptops running Windows 7, Windows 10 and Ubuntu; a smartphone running Android; and two different tablet devices running Android and iOS. As mentioned in Section 4.2 by relying on web technologies the system is independent of the OS and it only requires a web browser supporting HTML-5 and WebGL. Thus the system has been successfully tested in the most used browsers in the market: Chrome, Firefox, Microsoft Edge, Safari and Opera.

Nevertheless, before the system enters into production state, a precise study of the user community and server requirements will be performed.

4.4 Case Studies

This section aims to show the potential of the system. We present three different use cases. First, we show how the system helps the user to understand the image content. Second, we present a case of study where the system is used for optimum dataset recognition and selection. Both cases exploit the integration of LUCAS

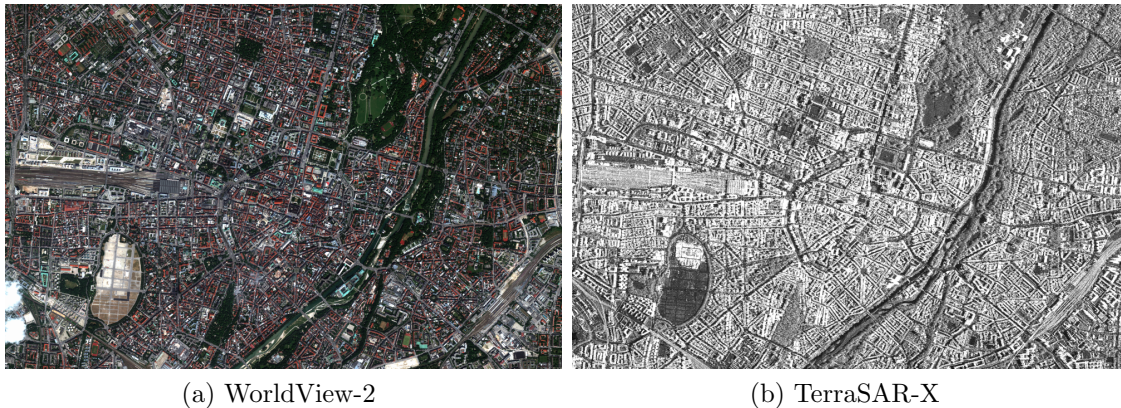


Figure 4.4: Image understanding case scenario of Munich, Germany.

survey data in order to enhance or improve the analysis and work performed with EO data.

4.4.1 Image Understanding

Image understanding refers to the capability of the users to interpret the content of the image they are studying. For the purpose of introducing the image understanding capabilities of the system we present three different scenarios in different cities.

4.4.1.1 Munich

This initial case of study is placed in the city of Munich, Germany. The system is ingested with the LUCAS information of Germany, linked with an OpenStreetMap layer; and two EO products of Munich: a multispectral image from WorldView-2 (WV-2) and a SAR image from TerraSAR-X (TS-X). Both EO images have pixel spacing of 1.25 meter, covering an area of 24 km^2 . The size of the total scene is 4890×3202 pixels. The EO images used are the same used in Section 3.9.1. Due to readability the data set is shown again in Fig. 4.4, but a higher resolution representation is available in Fig. 3.4

In this case we present a scenario where the availability of heterogeneous data sources from a same location allows a better understanding of the EO scene by expert and non-expert users. The data used on the experiment are shown in Fig. 4.5. Analysing just the SAR image, Fig. 4.5a, it is possible for both users to deduce that the main vertical structures of the image correspond to bridges. Moreover, an expert user most probably would interpret, due the intensity of the surrounding pixels, that the bridges are over several lanes of railways.

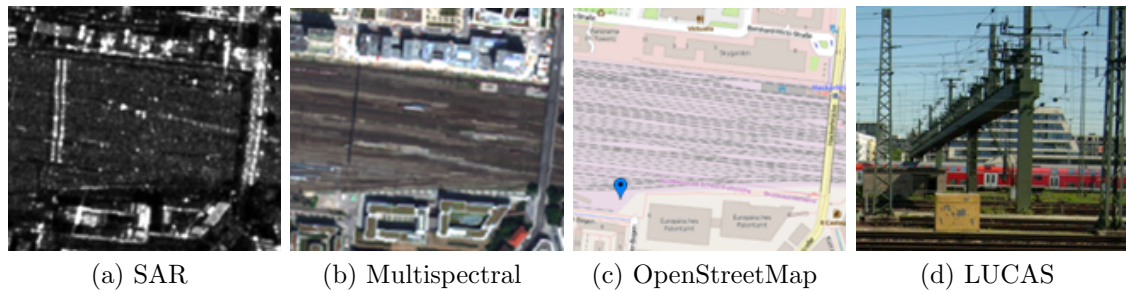


Figure 4.5: Munich EO scene understanding. From the (a) SAR image an user can recognize two different bridge structures. When adding (b) Multispectral image to the scene interpretation it is clear that the left structure, because of the small width, can be at most a gangway for pedestrians but the resolution of the image does not allow a correct identification. Adding the (c) Map layer the user realises that the structure is not appearing what practically discard the gangway. Including the in-situ information from (d) LUCAS surveys the user can finally recognise the unidentified structure as a main overhead line supporting infrastructure for the trains.

For the second step of the experiment, the users have also available a multispectral image, shown in Fig. 4.5b, for the scene interpretation. In this case the railway assumption would be clear. With the multispectral image the initial assumption about the bridges can be modified. The right structure of the image corresponds clearly to a bridge, but a question rises concerning the element on the left. Due to its width the left structure can not be a bridge where the cars can transit, but it could still be a gangway for pedestrians.

In the next step of this experiment we add one more data source to the scene interpretation process, the map layer with the OpenStreetMap information, Fig. 4.5c. The map clearly identifies the railways and the big bridge, providing at the same time more detailed information about the surrounding buildings, street names, etc. On the other hand, it does not help with the interpretation of the unidentified structure clearly visible in the SAR image.

The last step uses the remaining source integrated by our system, the LUCAS surveys. Going back to Fig. 4.5c it is visible a blue marker pointing out the availability of information from the LUCAS survey. Retrieving this information and adding it to the scene interpretation, the users would know that the survey point is classified as *non built-up area* inside the *artificial land* land cover category. Moreover, analysing one of the available photos, see Fig. 4.5d, the users can finally get an interpretation of the unidentified structure. The structure corresponds to a main overhead line supporting infrastructure for the trains.

4.4.1.2 Karlsruhe

The second case of study is located in the city of Karlsruhe, Germany. The users have available the LUCAS data, OpenStreetMap information and a TS-X image.



Figure 4.6: Image understanding case scenario of Karlsruhe, Germany.

The EO image, Fig. 4.6, is a spotlight image with 1.25 meter pixel space and a size of 4343x5741 pixels covering an area of 60 km^2 .

Starting again from the SAR image, Fig. 4.7a, the users are able to identify a slightly brighter striped element in the middle of the street that diagonally crosses the image. Generally streets present a low and homogeneous back-scatter coefficient in SAR images, since the flat surface of such targets favours specular reflection. Moving now to the OpenStreetMap layer, Fig. 4.7b, it is possible to identify clearly two separate roads on the street probably one for each traffic direction. However there is nothing that gives us an explanation to the stripped element in the middle of the street. As in the previous case there is available in-situ information on that street, Fig. 4.7c, where you can clearly see a tram line. The more intense back-scatter then is explained by the track ballast that are used to level and hold the rails in place. The ballast is made of crushed stone with sharp edges which increase the back-scatter coefficient in comparison with the road explaining the increase of brightness and, therefore, the unidentified stripped element.



Figure 4.7: Karlsruhe EO scene understanding. From the (a) SAR image an user can recognize a slightly brighter striped element in the middle of the street that diagonally crosses the image. When adding (b) OpenStreetMap image to the scene interpretation it is possible to identify clearly two separate roads on the street probably one for each traffic direction. Nevertheless the striped element still remains unknown. Adding the in-situ information from (d) LUCAS surveys the user can finally recognise the unidentified structure as the ballast made of sharp stones and used to keep the rails in place.

4.4.1.3 Stuttgart

The third image understanding scenario is based on Stuttgart, Germany. The available data are composed of the LUCAS survey, OpenStreetMap information and a TS-X image. The EO image, Fig. 4.8, is a spotlight image with 1.25 meter pixel space and a size of 6472x3617 pixels covering an area of 56 km^2 .

Focusing our analysis in Fig. 4.9a, the SAR image shows several low back-scattering rectangular elements, some of them surrounded by high back-scattering elements. The low back-scattering, as explained in the previous example, is normally due to a flat surfaces, e.g., streets, water bodies or sport courts. In this case such amount of small water bodies can only be explained due to agri/aqua-cultural exploitation. Aquaculture exploitation can be discarded because these type of installations are usually along river/lake borders. The agricultural use matching this kind of SAR pattern is related with flooded crops like rice. In this case the probability of the image to contain rice fields is very low but nevertheless the use of the map information can help discarding totally this hypothesis. Analysing the map, Fig. 4.9b, we can see that the element corresponds to the legend of *sportpitch* in OpenStreetMap. Finally analysing the LUCAS in-situ information we can know that the small pitches correspond to tennis courts and the surrounding high back-scattering is probably due to the metallic fences and light poles.

4.4.2 Optimum Dataset Selection

In this case of study we intend to show the possibilities of the presented tool for the selection of optimal datasets and ground truth information. This capabilities can



Figure 4.8: Image understanding case scenario of Stuttgart, Germany.

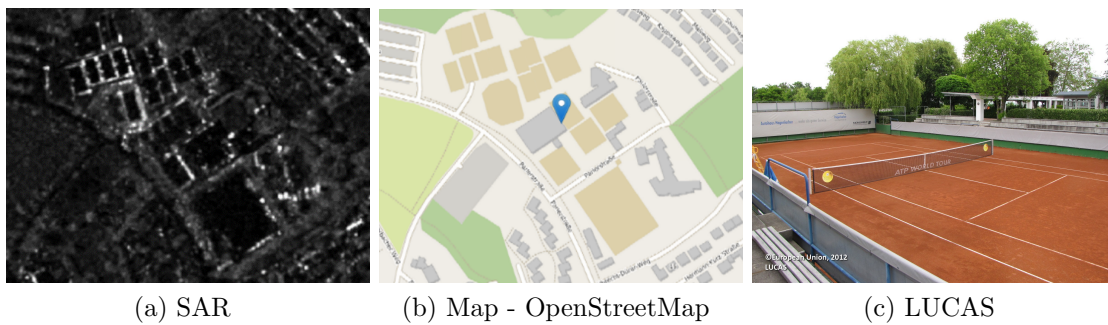


Figure 4.9: Stuttgart EO image understanding. The (a) SAR image shows several low back-scattering rectangular elements, some of them surrounded by high back-scattering elements. By means of (b) OpenStreetMap image it is possible to discard agri/aqua-cultural exploitation and to know the actual use of unidentified elements as sport pitches. Finally, analysing the in-situ information from (d) LUCAS surveys, it is possible to know that at least some of them are tennis courts.



Figure 4.10: Optimum dataset selection example for sunflower fields. If EO analysts would like to look for EO images containing sunflower crops in Germany, the system will show them the location of this type of field. The EO analysts would notice the proliferation of sunflower fields in the north-east of Germany, surrounding the capital, Berlin.

be proved useful for the EO analyst in different contexts such as change detection on EO image time series or during new EO analyst training activities.

With this purpose we define a region that encloses the whole Germany and make use of the in-situ information provided by LUCAS surveys of 2009 and 2012. Moreover, taking advantage of the query capabilities described in Section 4.2.5 we present different hypothetical scenarios where the presented system can be useful.

In the initial scenario EO analysts need to define a location for the acquisition of EO products for sunflower crop analysis and change detection in time series. Operating the system, it is possible to easily retrieve the existing points with *Sunflower* crop. The procedure is as follows: 1) define a region, 2) deploy the tree showing the land cover classes in order to find the *Sunflower* crop, and 3) double-click to perform the request and visualize the results. The obtained results presented in Fig. 4.10 show the distribution of sunflower fields in Germany. Additionally, it is clearly visible a bigger concentration of sunflower fields in the north-east, in the region surrounding the capital, Berlin. Taking into account this results, the EO analysts should delimit their analysis region around Berlin and order EO products of this zone. They could also take advantage of the information from the LUCAS points to classify the fields to which the GPS coordinates correspond to a sunflower field.

In the second scenario two EO analysts need to generate a data set for vineyard

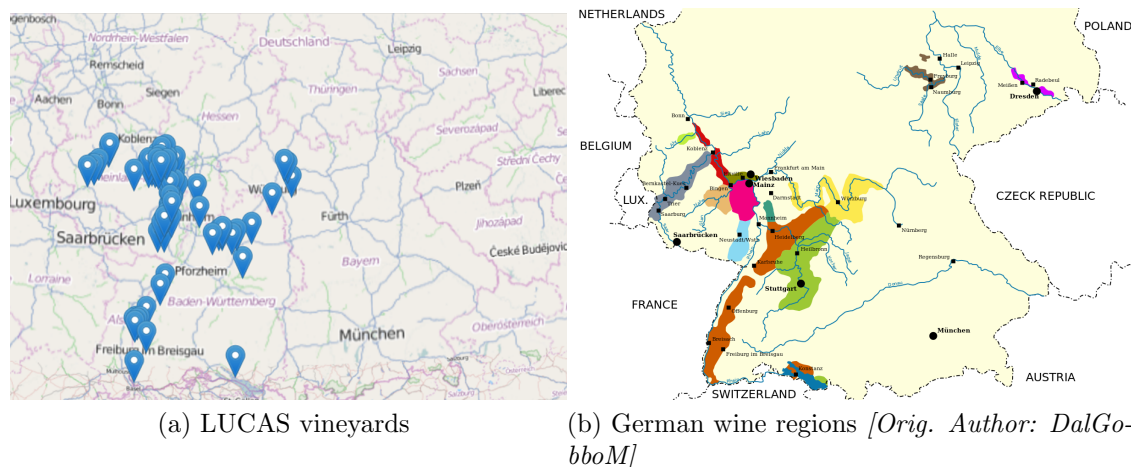


Figure 4.11: Optimum dataset selection, vineyard crop detection. Two EO analyst look for vineyards using the system and in parallel, for comparison purposes, a search engine. The first EO analyst obtains (a) using the system feed with the LUCAS information. The second EO analyst obtains (b), a map from *wikipedia.org* which identifies the main German wine regions. It is visible the resemblance with the result shown in (a) with the main regions shown in (b). The use of the system supports the users with ground truth data and offers results in accordance to the information available on the internet.

crop detection. One of them has available our system and the other will make use of traditional web search tools.

The EO analyst using the system follows a procedure similar to the one explained in the previous example, but now selecting the land cover *Vineyard*. The results, see Fig. 4.11a, show a concentration of vineyards in the west and south-west, with a bigger concentration around the west. The bigger density area corresponds to the Baden-Württemberg and Rhineland-Palatinate region and the fields appear to be along the Rhine river. Hence, the EO analyst should focus the efforts on these regions.

The procedure of the second EO analyst is also very simple. Open a web browser and using any of the available search engines type *German wine regions*. In most of the cases, the three first search results include the Wikipedia entry, where you can get access to the map shown in Fig. 4.11b. This map shows the 13 most important German wine regions, providing the answer to the second EO analyst.

Comparing the obtained results, we can see how both of them are consistent. Our system, making use of the in-situ information provided by LUCAS survey, shows points belonging to the most of the wine regions. It is clearly visible a similarity in the form of the map with the distribution of the LUCAS points. Nevertheless, our system does not provide any result from the smaller regions located in the upper top of the map. This is probably related with the 2 km^2 grid between survey points, which make it difficult to register smaller wine regions.

Ending this case scenario we can sentence that the use of our system is reliable and mostly in accordance with the information obtained from the web. Moreover, as mentioned previously the use of our system also provides with the additional information that can be used as ground truth in posterior analysis.

Finally, we can state that using these query capabilities and the interactive statistical graphic representations it is possible to detect regions with generic land cover changes or even specific changes, e.g., crop changes. Some examples of specific crop changes are crops moving from *barley* to *potato* or from *common wheat* to *rape*. The land cover detectable changes are the combination of the classes registered in LUCAS. After locating a region with the desired change type, or a big change diversity, the user can easily get the region coordinates and acquisition dates. With them the user could contact the data suppliers in order to get the desired EO data and proceed with the change detection analysis using available LUCAS information as ground truth in the validation processes.

4.5 Conclusions

We have presented the architecture and a prototype of a multilayer system for heterogeneous geospatial data analytics. The system implements a server-client architecture, which integrates several web technologies. One of the benefits related to the server-client approach is the simplicity of the client. The server is responsible of the most complex processing tasks making possible to offer lightweight clients for different devices. The presented architecture manages the data from the source. The initial layers read the original data and perform transformations to make viable the data integration. These heterogeneous data are linked and stored in a geographical database or in a system repository. The link among the data allows the User Oriented Web Functionality layer to exploit the database capabilities in order to perform geographical queries over the stored data. This layer also implements all the communication protocols to the linked third party services and the server logic that interacts with the user via the GUI.

The presented case studies show the system capabilities managing heterogeneous EO and in-situ data sources. In the first case of study the system proves its utility helping to get a better understanding of EO images for expert and non-expert users. The second case of study use the presented system as a tool for the selection of optimal datasets. Exploiting the in-situ information of LUCAS survey it is possible to use the surveyed point data as ground truth information for change detection on EO image time series.

Data Mining Methodology for the Assurance of in-situ Recorded Land Cover Changes

This chapter presents a data mining methodology to filter and validate land cover change detections obtained from multitemporal in-situ surveys. As in-situ data we use the measurements from the European Land Use/Cover Area frame Survey (LU-CAS), which provides images with standardized metadata about land cover and land use within the whole territory of the European Union. Multitemporal LU-CAS surveys present an anomaly in the amount of land cover changes that disagree with the estimated by experts. Therefore, our methodology analyses the available data in order to explain the existing irregularities in them. The initial step of our methodology is based on database query refinements. The data mining methodology continues with an image analysis process. This analysis calculates similarity measures of the multitemporal images which are used to identify the potential misclassifications. The final step involves a GIS based on web technologies. By defining different color codes assigned by the similarity measures, the system represents the examined points on a digital Earth globe. There, a user can easily discriminate potentially misclassified points for subsequent detailed analysis or corrections. The final output of the methodology shows remarkable results for detecting misclassified land cover changes.

The rest of the chapter continues with Section 5.1 where the motivation for the presented data mining methodology is described. Section 5.2 presents the architecture of the system. Section 5.3 explain the data mining methodology which is composed by three different steps: Section 5.3.1 describes the data refinement step,

The content of this charter will be published in: K. Alonso, D. Espinoza-Molina, and M. Datcu, Mining Multitemporal in-situ Heterogeneous Monitoring Information for the Assurance of Recorded Land Cover Changes, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 877-887, March 2017.

Section 5.3.2 the image analysis and similarity computation step, and Section 5.3.3 presents the last step composed by data visualization and filtering processes. After the methodology presentation, each step is evaluated independently in Section 5.4, Section 5.5, and Section 5.6. Finalizing with a full methodology evaluation, Section 5.7 and conclusions, Section 5.8.

5.1 LUCAS Anomaly

As described in Section 2.1.7, every three years since 2006 a survey campaign has been carried out to monitor, in a standardized way, the state and change dynamics in land use and cover in the European Union. The amount of information collected until now reaches tens of Terabytes. This volume of information is already big enough to make it impossible for the data to be supervised at small scale. In consequence, the task of collecting and supervising the data relies exclusively on the field surveyors.

Analyses of the acquired multitemporal data have shown a very high variability in the land covers, exceeding the expectations of the experts. Using this peculiarity as motivation we started a deeper analysis of the LUCAS surveys aiming to identify the real land cover changes from the potential inaccuracies introduced during the recording or annotation of the data. Table 5.1 shows two LUCAS survey points. While the first point shows a clear example of land cover change, the other one shows a point with a non-visible land cover change. In reality, both points are marked as land cover changes. In our understanding, the latter point is an example of the aforementioned anomalies, where the recorded land cover change is uncertain.

With this study we aim to provide tools for the quality assurance of the existing and future LUCAS surveys. Furthermore, we aim to improve the impact and integration of the in-situ observations in EO applications. Consequently, the following presents a data mining methodology to filter and validate land cover change detections obtained from multitemporal LUCAS in-situ surveys.

5.2 Data Mining System Architecture

The base of the presented mining methodology is the system for heterogeneous geospatial data analytics described in Chapter 4. The system, as shown in Fig. 3.2, follows a server-client philosophy. The server side is composed of four different layers: (1) the raw data layer, (2) data ingestion layer, (3) database management system, and (4) user oriented web functionality layer. The raw data layer contains the original information sources that are analyzed and processed in the data ingestion layer in order to pass the obtained information to the database management system layer which will store it in a more accessible way, facilitating the querying and visualization operations performed by the user oriented web functionalities layer. In





	2009	2012
Real Land Cover Change		
Uncertain Land Cover Change		

Table 5.1: Multitemporal LUCAS survey points. The first point shows a clear example of land cover change. In contrast the second point presents a non visible land cover change. In the LUCAS survey records both points contain land cover changes. In this way, the second point could be considered an example of the existing irregularities.

this last layer, the Geo-Information Visualization block performs the communication protocols with third party geographical information providers using a parallel instance of MapServer [192]. Mapserver centralizes all the communication with the third party providers working as a proxy for the main system which remains isolated and avoids cross-domain communications. The client is constituted of the Graphic User Interface (GUI) layer accessible from any electronic device capable of running a HTML-5 compatible web browser.

The data mining methodology requires new or modified system modules, represented with a darker color in Fig. 5.1. The Feature Extraction module is newly introduced in the architecture and it is responsible of the in-situ image analysis processes. The geographical database, the Metadata and Statistical Visualization module, along with the Image Analytics module were already part of the system architecture but their functionalities were upgraded considering the requirements of the presented data mining methodology. The PostGIS [198] is a community developed open source spatial database extender which allows the geographical querying of the spatial data and it is based on PostgreSQL technology [190, 191]. The Metadata and Statistic Visualization module collects the data from the database to generate the required data visualization. The Image Analytics module manages the interaction of the user with the data and sends the required instructions to Geo-Information and statistical visualization processes in case an update is required. In Section 5.3 the specific functionalities of each module are described in detail.

5. Data Mining Methodology for the Assurance of in-situ Recorded Land Cover Changes

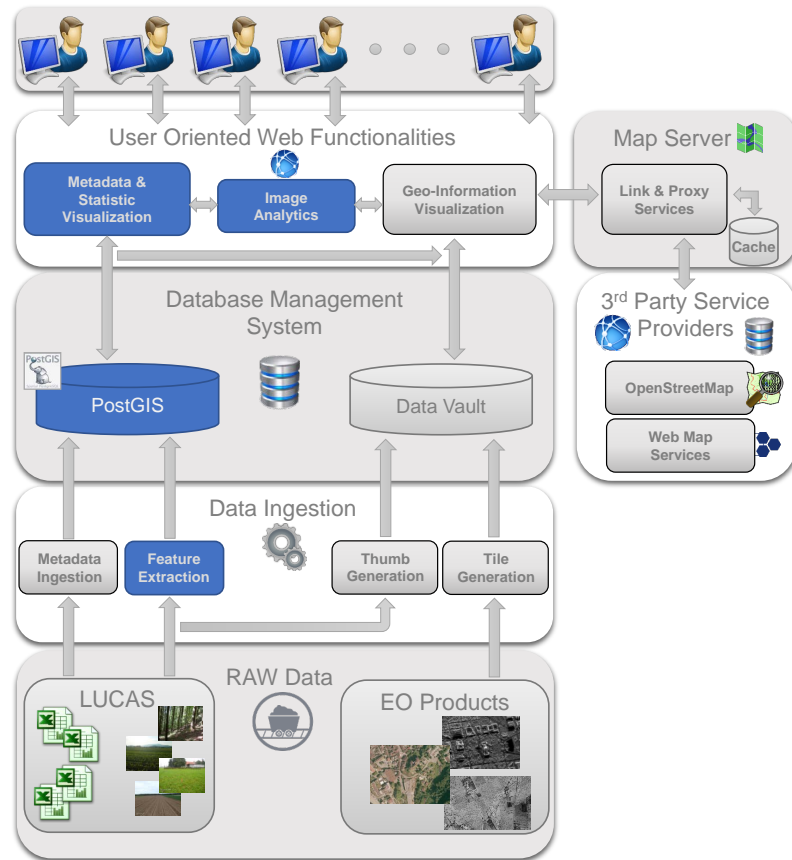


Figure 5.1: System architecture. Following a server-client philosophy, the system is designed to rely all the computational complexity over the server making the client side lightweight.

5.3 Data Mining Methodology

A preliminary study of the survey methodology shows small improvements in the survey protocol with the pass of the years. One remarkable change is the increase number of the surveyed points. Consequently, different amounts of information are available in the database for each surveyed point. Although it is important to be aware of this fact, it does not impact the change detection procedures. A second change, the one that could explain at a certain level the high multitemporal land cover changes, is the update in the land cover hierarchical class structure between the surveys done in 2009 and 2012. The multilevel hierarchy starts with general land cover classes at the lower level and extends to more specific classes with each higher level. The hierarchy changes were limited to the inclusion of third hierarchical level classes inside the first level classes *Woodland* and *Bare Land*. Nonetheless, even this hierarchy structure change solely has produced a non-realistic increase in the detected changes. In order to quantify the non-realistic land cover changes due to the

Common Hierarchy		Hierarchy Update 2012
Woodland	Broadleaved	-
	Coniferous	Spruce Dominated Pine Dominated Other
	Mixed	Spruce Dominated Pine Dominated Other
Bare Land		Rocks and Stones Sand Lichens and Moss Other

Table 5.2: Land cover class hierarchy modifications between 2009 and 2012 LUCAS surveys. The Woodland second level hierarchy members *Coniferous Woodland* and *Mixed Woodland* were extended to a third hierarchical level formed by: *Spruce Dominated*, *Pine Dominated* and *Other*. Furthermore, the first hierarchical level *Bare Land* was extended with a second level hierarchy.

hierarchy modifications and detect other possible misclassification sources we introduce a data mining methodology which comprises three different steps: 1) database query refinement, 2) in-situ image analysis and similarity measure computation, and 3) on map data visualization and filtering.

5.3.1 Database Query Refinement

The data mining methodology for land cover change detection starts with the mapping of the hierarchical structure changes. The objective of the mapping is to exclude the points that only present land cover changes due to the land cover class hierarchy modification. The changes are shown in Table 5.2. The changes were limited to only two of the first hierarchical level classes *Woodland* and *Bare Land*. *Woodland*'s second level hierarchy members *Coniferous Woodland* and *Mixed Woodland* were extended with a third hierarchical level formed by: *Spruce Dominated*, *Pine Dominated* and *Other*. In 2009 *Bare Land* was just defined as a first hierarchical level, but in 2012 it's definition was extended with a second level hierarchy formed by: *Rocks and Stones*, *Sand*, *liches and Moss* and *Other*.

The mapping of these hierarchy changes is implemented at the Metadata and Statistic Visualization module. There, it is possible to refine the database requests in Standard Query Language (SQL) [199]. As a result, the queries to the database requesting points containing land cover changes will exclude the points whose land cover change was among the mapped ones, and hence it will avoid the introduction

of false-positives.

5.3.2 In-situ Image Analysis and Similarity Measure Computation

After the query refinement, the methodology exploits the available point images. Our integration of the Feature Extraction module in the architecture offers valuable new data obtained from image analyses which extend the LUCAS information. In this way, our PostGIS database is extended to link and store the results from two different image analyses. The first one corresponds to the Bag of Words (BoW) [146] generated using a common dictionary of Red-Green-Blue (RGB) colors. The second analysis extracts at image level the texture information using the Weber Local Descriptor (WLD)[132].

Being D a given image for analysis, the first image processing step results in an image color quantization based on a predefined color map. The assigned color map is created by dividing uniformly the color space in 256 elements. Once the D is quantized the BoW is generated defining $p(\omega_{RGB}|D)$ as the probabilities of the words in a given image. Where ω_{RGB} represents the 256 words in the dictionary and the index RGB the identifier of each word.

The second analysis takes again an image D as input for our WLD algorithm that generates as output a WLD histogram that we will use as a second BoW dictionary, $p(\omega_{WLD}|D)$. In this case the words ω_{WLD} represent the different combinations of excitation levels and orientations taken into account in the WLD algorithm. In our implementation we decided to use 18 different excitation levels and 8 orientations.

The described analyses procedures are done over all the available LUCAS images. The obtained results are stored in a database, linking the source images and the corresponding surveyed point. After the information extraction and analysis phase comes the result ranking process. This process is implemented in the Metadata and Stastic Visualization module following a classical approach used by the image analysis community. We use the extracted feature sets as visual signatures to compute similarities among images. Examples of these uses can be found in [16] and [200]. In this specific scenario we use the stored $p(\omega_i|D)$ to compute similarity distances and generate a ranking of the LUCAS points, where i is the index indicating the length of the used dictionaries. The available dictionaries are the previously described $p(\omega_{RGB}|D)$, $p(\omega_{WLD}|D)$ and the joined dictionary obtained from the concatenation of these two $p(\omega_{RGB}|D) \cup p(\omega_{WLD}|D)$.

For the calculation of the similarities of the images from the multitemporal survey we decide to use Kullback-Leibler distance

$$d = \sum_i p(\omega_{i_{2009}}|D) \cdot p(\omega_{i_{2012}}|D) \cdot \ln(p(\omega_{i_{2009}}|D)) \quad (5.1)$$

where $p(\omega_{i_{2009}}|D)$ is the probability vector of a given image from 2009 survey and $p(\omega_{i_{2012}}|D)$ it's equivalent in the 2012 survey.

As mentioned in the Section 2.1.7, each LUCAS point is composed of five different images. The main image shows the exact GPS (p)oint that has been surveyed and the other four photos cover the surroundings of the location by showing the different cardinal directions: (N)orth, (E)ast, (S)outh, and (W)est. Here, we define each LUCAS point as a five element vector P ,

$$P = [d_p, d_N, d_E, d_S, d_W] \quad (5.2)$$

where the distances d_* are computed according to (5.1) for each pair of multitemporal images available in the original LUCAS point.

For the sake of posterior visualization simplicity we aim for an unique similarity value for each LUCAS point. Thus, we calculate a similarity value as a weighted mean of the elements in P . To that end, we define a weight vector

$$W = [w_p, w_N, w_E, w_S, w_W] \quad (5.3)$$

where the different weights w_* are assigned as follows: $w_p = 0.3$ given that it always contains the analyzed land cover; and the remaining weights are set to 0.175 in contemplation of possible changes of the surroundings which also affect the analyzed land cover. The similarity can then be computed as

$$Similarity = \frac{\sum_{n=1}^5 w_n \cdot d_n}{\sum_{n=1}^5 w_n}. \quad (5.4)$$

Finally, by means of the calculated point similarity values it is trivial to generate a ranking, listing the points in order of image similarity between their images among surveys.

5.3.3 On Map Data Visualization and Filtering

The implementation of the on map visualization and filtering procedures requires modifications in the user oriented web functionalities layer. The Metadata and Statics Visualization module capabilities are extended in pursuance of a better and faster visual discrimination of the data differences. The approach followed includes a redefinition of the markers used to represent the survey points. The markers implement a color codification which allows us to represent the results obtained from the similarity rankings which we use as confidence value of the annotation. Additionally, we are able to include another color indicators to visually represent the time span and the distance between the survey acquisitions. The time of the year

when the information of the point was collected in each survey can be meaningful to explain some of the dissimilarities in the images.

Regarding the Image Analytics module, different User Interface (UI) elements have been introduced in order to help the data filtering process and the annotation modification. Furthermore, new server-client and inter-module communication processes have been implemented to support the new functionalities required. Fig. 5.2 shows part of the user interface during the visualization and filtering step. The points shown are the result of the previous two steps. In this specific case the color intensity from brighter to darker indicates the similarity between the images of the multitemporal survey points. The black color indicates high similarity while the brighter tonalities represent lower similarity. The slider-bar over the map can be used to filter the points drawn over the map. It can set different distance thresholds which are used in the querying process as condition that the points must fulfill in order to be retrieved. The slider-bar has three different operation modes. It can retrieve the points: 1) under the threshold, 2) over the threshold, and 3) in the range between to thresholds. The specific point information is visualized individually by showing a table summarizing the most relevant data (under the map frame), and the acquired multitemporal images (right). The multitemporal images are group by survey year. The top image corresponds to the top the exact Global Positioning System (GPS) point surveyed and it is followed by the images pointing to the different cardinal directions, starting with the north and continuing clock-wise. The selection of a specific point can be done in three different ways. First, by selecting the markers over the map. Second, by using the buttons in the last column of the information table. And third, by using the pagination roulette at the bottom of the UI. This roulette indexes all the points represented on the map.

5.4 Query Refinement Evaluation

For the evaluation of the query refinement processes we use the LUCAS data of Germany and Spain from the surveys of 2009 and 2012. A summary of the evaluation is presented in Table 5.3.

5.4.1 Case Study Germany

Germany's LUCAS data sum a total of 46084 surveyed points. The initial points are reduced to the points that share the same geolocation, 37504. The difference in points, as explained in Section 5.3, is due to the increase of the surveyed points in the 2012 survey. Hence, we have a total of 18752 pairs of points in which we can perform the temporal land cover change detection. Querying only by the change on land cover will return a total of 9240 geographical points with land cover changes, the 49.35% of the total. The query refinement, described in 5.3.1, implements the

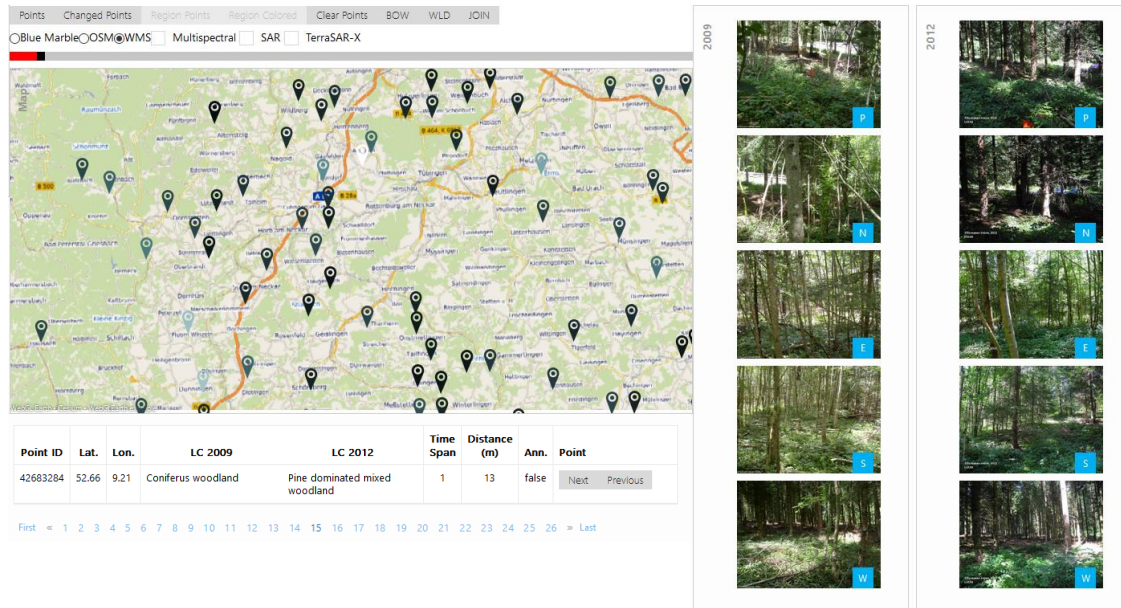


Figure 5.2: User interface used during the step three of the data mining methodology. The tool allows the visualization of the degree of reliability of the detected land cover changes. It also implements capabilities of filtering, selection and correction of annotations.

modifications of the land cover hierarchy and detects 2496 points annotated as land cover changes, the 13.31%, which should not be marked as land cover change. These points can be differentiated by the hierarchy change type. Thus, an 8.67% correspond to *Coniferous Woodland* hierarchy change, a 4.43% to the *Mixed Woodland* change and a 0.21% to the *Bare Land* change. Examples of the points that are discarded are shown in Table 5.4. At this stage, the number of points with possible land cover changes has been reduced to 6744, a 35.96% of the total. Comparing the number of points discarded with the originally annotated as change, we can say that in Germany the 27.01% of the detected land cover changes were related to the modification of the class hierarchy and not real land cover changes.

5.4.2 Case Study Spain

The LUCAS data for Spain contains a total of 65290 surveyed points. The number of multitemporal points available is 25016. The initial query for the retrieval of the land cover changes returns 12172 points, the 48.66% of the total. After the query refinement procedures we can discard 2020 points, the 8.07%. Looking at the hierarchy change type, the 5.43% corresponds to *Coniferous Woodland* hierarchy change, a 1.60% to the *Mixed Woodland* change and a 1.04% to the *Bare Land* change. Therefore, the number of points with possible land cover changes can be reduced to 10152, a 40.59% of the total. In this case, the number of points discarded

5. Data Mining Methodology for the Assurance of in-situ Recorded Land Cover Changes

	Germany		Spain	
	Number of Points	Percentage	Number of Points	Percentage
Multitemporal Points	18752	100	25016	100
Annotated Changes	9240	49.27	12172	48.66
Detected Hierarchy Changes	2496	13.31	2020	8.07
<i>Coniferous Woodland</i>	1625	8.67	1360	5.43
<i>Mixed Woodland</i>	832	4.43	400	1.60
<i>Bare Land</i>	39	0.21	260	1.04
Changes After Refinement	6744	35.96	10152	40.59
Discarded vs. Annotated	-	27.01	-	16.60

Table 5.3: Evaluation of the query refinement step using LUCAS data of Germany and Spain from 2009 and 2012. The annotated land cover changes in both cases are around 50%. Applying the query refinement procedures to filter the class hierarchy modifications between 2009 and 2012 LUCAS surveys the points discarded as land cover change are 13% for Germany and 8% for Spain. The ratio of the discarded land cover changes relative to the annotated changes is of a 27% in Germany and of a 16% in Spain.

versus the initially annotated with land cover change is the 16.60%.

5.5 Image Analysis and Similarity Measure Evaluation

The methodology's second step generates the similarity measures used for ranking and color coding the multitemporal points. The system implements the possibility to generate three different rankings based on the analysis described in Section 5.3.2. The rankings obtained by using the RGB dictionary and the join dictionary present a better results comparing with the WLD dictionary. The ranking of the former two shows a higher visual coherence clearly ranking the most similar and dissimilar point at the extremes of the ranking. The ranking obtained with the WLD dictionary presents less consistent results, interleaving high similarity points with low similarity ones.

Table 5.5 presents the information of different points with high and low similarities using the join dictionary. The first four examples correspond to the high similarity values. High similarity points are marked as low certainty of containing land cover changes. While most of the high similarity points annotated with multitemporal change do not seem to have any land cover change, there are some of the points, e.g. middle-left point in Table 5.5, that even showing a high similarity also contain a land cover change. The last row shows points with low image similarity.













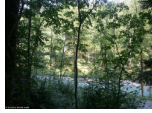





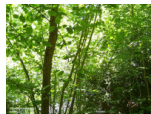











Year	Land Cover	Point Images				
		Point	North	East	South	West
2009	Coniferus Woodland					
2012	Other Coniferous Woodland					
2009	Mixed Woodland					
2012	Spruce Dominated Mixed Woodland					
2009	Bare Land					
2012	Other Bare Soil					

Table 5.4: Examples of the points affected by the land cover hierarchy modifications: *Coniferous Woodland*, *Mixed Woodland*, and *Bare Land*. Similar points are discarded during the query refinement step of the methodology because they do not contain a real land cover change.

At this side of the ranking, the majority of the points correspond to agricultural lands where the change in the crop type is clearly visible. These points with low image similarity are the ones that should be marked with the higher certainty of land cover change.

5.6 Data Visualization and Filtering Evaluation

At this point of the methodology we will focus our evaluation on the usability of the developed tools for the data visualization, filtering and correction. To evaluate the performance of the complete methodology and the usability of the developed tools we focused our analysis in the LUCAS data of Germany. At this stage the user will exploit the outputs of the previous methodology steps. The initial step provided the query refinement, as described in Section 5.4.1, where the number of points containing land cover changes was reduced a 27.01%, from the initial 9240 to

5. Data Mining Methodology for the Assurance of in-situ Recorded Land Cover Changes








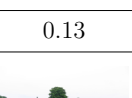
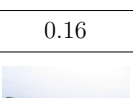








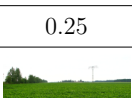



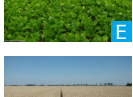
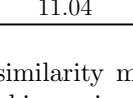
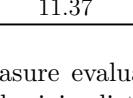
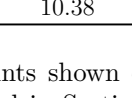
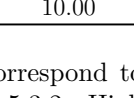
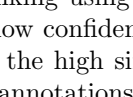
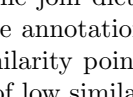
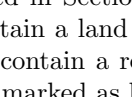
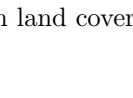
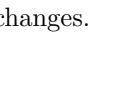


Similarity	Year	Land Cover	Point Images		Land Cover	Point Images	
High	2009	Coniferous Woodland			Non built-up Linear Features		
		Spruce Dominated Woodland				Grassland without Tree Cover	
	2012	Mixed Woodland			Woodland		
	Distance		0.13	0.16	Distance	0.43	0.64
Medium	2009	Barley			Grassland without Tree/Shrub Cover		
		Maize				Spontaneous Re-vegetated Surfaces	
	2012						
	Distance		0.18	0.20	Distance	0.25	0.26
Low	2009	Barley			Sugar Beet		
		Spontaneous Re-vegetated Surfaces				Common Wheat	
	2012						
	Distance		11.04	11.37	Distance	10.38	10.00

Table 5.5: Image analysis and similarity measure evaluation. The points shown correspond to the opposite extremes of the ranking using the join dictionary described in Section 5.3.2. High similarity points are marked as low confidence annotation points to contain a land cover change. We can appreciate that most of the high similarity points don't really contain a real land cover change. On the other hand, the annotations of low similarity points are marked as highly reliable considering most of them contain land cover changes.

6744. The second step generated the similarity ranking of the points that are used here to represent the confidence level of the land cover change annotation.

The developed tools allowed a user to review effectively the remaining points in around one working day. At the end of the data visualization and filtering process all the points were reviewed. The land cover changes of the 72.2% of the points were validated. The multitemporal changes of the other 27.8% were discarded.

5.7 Data Mining Methodology Evaluation

After evaluating independently each of the methodology's steps we proceed to evaluate the performance of the entire methodology. It is clear to us the query refinement

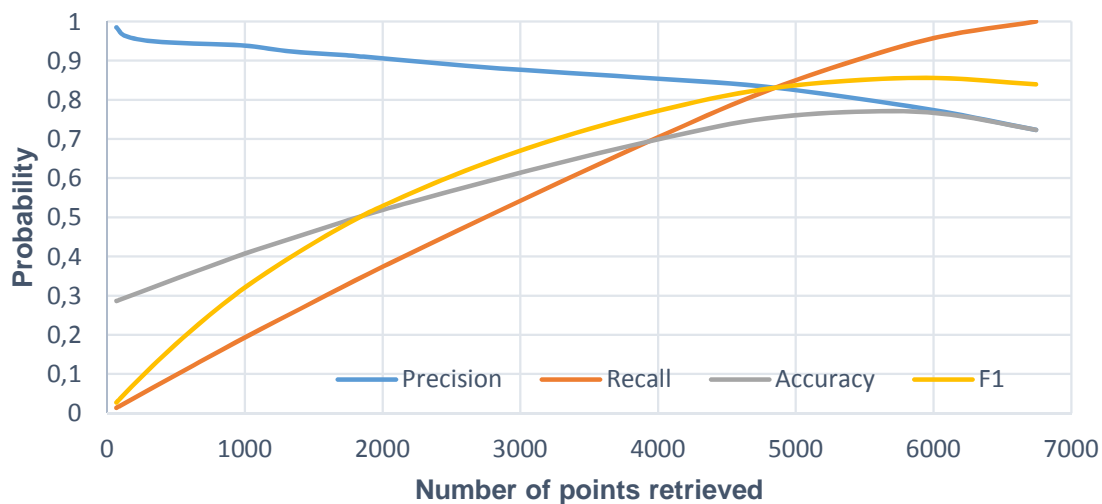
step is a valuable process in order to initially narrow down the number of points to be analyzed. The image analysis and similarity computation step can be re-evaluated by using the results obtained in the evaluation of the third step in Section 5.6. Thus, using the corrected annotation of the land cover changes as a ground truth, a quantitative analysis of the quality of the similarity ranking for land cover change detection can be computed. In other words, we use the validated results of the land cover changes, obtained from the data mining methodology, to measure the performance of similarity rankings for detecting real/non-real land cover changes. The analysis is performed using the ranking generated by the RGB dictionary. For the case of retrieving the points with a real land cover change, the obtained results are presented in Fig. 5.3a. Here, there are represented the precision, recall, accuracy and F_1 measures, i.e, the equally weighted harmonic mean of precision and recall. A detailed explanation of the used measurements can be found in [182]. The results show very high precision values, over 90%, when limiting the ranking up to 2000 points. The accuracy and $F1$ parameters start at lower values but they values increase with the amount of retrieved images and the improvement of the recall. While retrieving around 5000 points the maximum performance is offered obtaining a precision of 83.67%, a recall of 81%, an accuracy rate of 74.8% and a $F1$ value of 82.3%.

On the contrary, if we inverse the approach and analyze the performance of the ranking for retrieving the points with a non-real land cover change, the obtained results are not so promising. The Fig. 5.3b shows how the precision value rapidly decays to 75% when limiting the retrieved points to 500. When trying to retrieve the same amount of points annotated as non-real land cover change, 1875, the precision value is just a 55.25% with a recall of 55.9%, an accuracy of 75.27% and a $F1$ value of 55.57%. These poor results are explained by the points similar to the one shown in Table 5.5, where even having high visual similarity, the land cover changes exist.

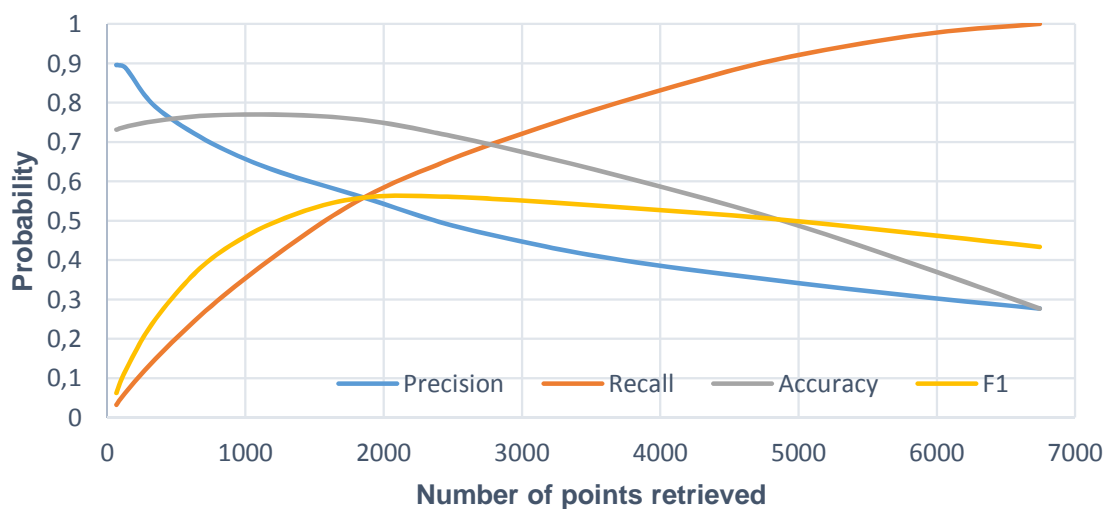
In our opinion the results obtained with the similarity ranking for the case of low similarity points could offer good enough results for some kind of automatization. On the other hand, the results with the high similarity points are not good enough. Hence, we think the data visualization and filtering tools play an important role in the proposed data mining methodology. This last step uses the generated ranking result in order to facilitate the user task of reviewing, filtering and correcting the land cover annotations. The entire process can be performed with an affordable time investment, offering better final results than any possible automatization. Thus, at the end of the three steps of the data mining methodology we reduce the original 49.27% of the points annotated as land cover changes, to only the 25.97%.

Additionally, the data visualization tools have helped to identify and understand the most common land cover misclassifications. Some of the examples of the errors are listed in Table 5.6. One of the most common misclassification mistakes are related to the linear features, i.e., roads. The first case shows a point of a road inside a forest. The landscape did not change but in 2009 the surveyors decided to classify

5. Data Mining Methodology for the Assurance of in-situ Recorded Land Cover Changes



(a)



(b)

Figure 5.3: Quantitative results for the retrieval of points using the similarity metric for retrieving: (a) real land cover changes and (b) non-real land cover changes.

it as *Non Built-up Linear Feature* while in 2012 they preferred to focus more on the surroundings assigning the point the *Broadleaved Woodland* class. Another common mistake is the one shown in the second case where the criteria for the definition of a *Non Built-up Area Feature* or *Non Built-up Linear Feature* is not totally clear. Case 3 and Case 4 show errors in the classification due to small distance differences in the surveyed points. We have also noticed different criteria when classifying grass fields. Specifically troublesome appear to be the land covers *Grassland without Tree/Shrub Cover*, *Grassland with Sparse Tree/Shrub Cover*, *Temporary Grassland*, and *Spontaneous Re-vegetated Surfaces*. Examples of these class misclassifications are shown in the Cases 5 to 7. *Inland Running Water* class also appears to have classification problems, see Case 8. Here, as in previous cases, the error is due to small distance differences between the points. Some less common classification errors include *Apple Tree* and *Cherry fruit* classes. Finally, there are two common misclassifications that appear to happen inside residential areas. In the first one, the surveyors usually change the classification criteria. In the initial survey they decided to annotate a building while in the second survey they decided to annotate the garden of the building. The second classification error in residential areas involves the classes *Buildings with 1 to 3 Floors* and *Buildings with more than 3 Floors*, example of this error is shown in Case 10.

5.8 Conclusions

We have presented a data mining methodology that is able to successfully filter false land cover changes from the real land cover change detections in multitemporal LUCAS in-situ surveys. As base for the data mining methodology we use the heterogeneous geospatial data analytics system presented in Chapter 4. We have described the three methodology steps and evaluated them independently. The database query refinement step maps the changes in the class hierarchy in order to exclude the points that only present land cover changes due to the hierarchy's modification. The evaluation of this step has shown relevant land cover change filtering capabilities. The query refinement was able to discard the 27.01% of the data annotated as land cover changes in Germany and a 16.6% in Spain. The second step, the image analysis and similarity computation step, showed big capabilities generating similarity rankings with the point images. In the third step, the visual evaluation of the ranking is very good. It clearly positions at the extremes the most similar and dissimilar points. Unfortunately, a correct ranking based on similarity does not ensure a good discrimination of the land cover changes. This fact can be seen in the quantitative analysis performed in Section 5.5 where the dissimilar images offer a very good land cover change detection but failed to detect the changes in more similar images. The data visualization step takes advantage of the previous results in order to offer simplicity and efficiency to the users in their data reviewing tasks.

5. Data Mining Methodology for the Assurance of in-situ Recorded Land Cover Changes

The developed tools allow the reviewing task with a small investment in manpower and time. The final data mining results show a clear reduction in the total number of land cover changes which go from the initial 49.27% to only the 25.97%. Additionally, the data mining methodology has improved our knowledge of the data and has helped us to identify common mistakes done during the surveying campaigns. In our understanding the final quality of future surveys could be improved in two different ways. First, the surveyor training could be improved by presenting the detected common mistakes during the training sessions. Second, the developed system can easily be accessible to the surveyors on the field which will provide fast information of the previous surveys reducing the uncertainty and the subjective criteria in the decision making process.

Year	Case	Land Cover	Point Images		Case	Land Cover	Point Images	
2009	1	Non Built-up Linear Features			2	Non Built-up Area Features		
2012		Broadleaved Woodland				Non Built-up Linear Features		
2009	3	Maize			4	Non Built-up Linear Features		
2012		Grassland without Tree/Shrub Cover				Vineyards		
2009	5	Grassland without Tree/Shrub Cover			6	Temporary Grassland		
2012		Grassland with Sparse Tree/Shrub Cover				Grassland without Tree/Shrub Cover		
2009	7	Spontaneous Re-vegetated Surfaces			8	Broadleaved Woodland		
2012		Temporary Grassland				Inland Running Water		
2009	9	Apple Tree			10	Buildings with 1 to 3 Floors		
2012		Cherry Fruit				Buildings with more than 3 Floors		

Table 5.6: Common misclassification patterns encountered after concluding the data mining methodology over the German 2009 and 2012 LUCAS surveys. Some of the most common misclassification mistakes are related to the linear features, i.e., roads (Cases 1-2). Other errors are due to small distance differences between the surveyed points (Cases 3-4 and 8). Also, grass fields and different type of fruit trees are difficult to classify (Cases 5-7 and 9). Finally, the residential areas have shown common misclassifications (Case 10).

6

Visual Analytics for EO Archives and In-Situ Data

In conjunction to image information retrieval systems that provide a wide variety of tools for interactive exploration of big image archives, visual analytic techniques can offer additional tools to interactively analyze big amounts of data. In general, these techniques exploit the results of different automatic analysis processes and generate interactive visualization that facilitate the understanding, reasoning and decision making over large and complex datasets. Due to the nature, extension, and complex composition of Earth Observation (EO) archives, they make perfect candidates for the implementation of visual analytic techniques. Hence, this chapter showcases some of the possibilities available when different data visualizations are implemented.

The chapter firstly introduces the main data categories used for the visual representations of the content of the EO archives, in Section 6.1. In Section 6.2, the different EO archive data models used for the visual analyses are presented. Section 6.3 illustrates the content of the TerraSAR-X (TS-X) archive and presents different visual representations resulting from the querying of the dataset. In Section 6.4 we introduce various visualizations of the Land Use/Cover Area frame Survey (LU-CAS) dataset which use specific data from Germany. The analysis moves toward the integration of databases which is showcased using the previous datasets, and whose results are presented in Section 6.5. The chapter ends with the conclusion statements of Section 6.6.

6.1 Data Categories

Remote sensing archives are usually composed of three main data categories: numerical, semantic and statistical. Examples of numerical data are very diverse in EO archives. Common numerical metadata related to the EO images are: acquisition

time, file size, number of rows/columns, number of layers, ground range azimuth resolution, and row/column pixel spacing among many others. Among the numerical data, geospatial data have a special relevance in EO. In general, geospatial data can be seen as a conglomerate of different data that contain locational information or geographic data. These data go from classical ZIP codes or addresses to geographical coordinates. Focusing in remote sensing, spaceborne and airborne imagery includes geographical coordinates delimiting the extension of the captured scene. Additionally, information detailing the spacecraft/aircraft position and instrument orientation (i.e., pitch, yaw and roll) during the acquisition is enclosed. Concerning the positioning, the spaceborne images enclosed data defining the exact orbital position of the satellite. For the airborne images the position is set via GPS.

Since 1995, when GPS started to be fully operational, the development of civilian applications based on this global positioning system has reach unprecedented levels. The integration of GPS receivers was initially limited to navigation devices but it soon spread to more general use devices, such as cameras or telephones. This proliferation of GPS compliant devices produces huge amount of geolocated data that are enclosed in images, routes or interest points.

The semantic data are composed of categorical data which represent specific characteristics that cannot be expressed simply using numeric values. Examples of these data can be found directly on the EO image metadata, e.g., instrument identifier, polarization, acquisition mode or product level. On the other hand, EO archives also contain semantic data resulting from human annotation or machine learning techniques. The generation of semantic annotations pursue different goals. Some studies use them as the mean to perform detailed analyses of specific EO images. Additionally, broader approaches make use of the semantics in order to facilitate the navigation and study of the EO archives. In any case, the used semantics always implement an unequivocal interconnection method that make possible the link between the different types of semantics. Possible interconnections methods go from simple class hierarchies to more complex ontologies. The ontologies can define in very high detail level the semantic frame of the study, and consequently, they are not only able to provide the explicit information regarding the semantic data instances but are also able to infer implicit information which is not specifically introduced on it.

Finally, the last data category is the result of the statistical analysis of the previous data types.

6.2 Data Models

A key aspect required for the generation of visual representation of heterogeneous information is the definition and posterior integration of the data models. Data models are a key tool for describing the data contained in the managed datasets.

In the following, we introduce the two data models used for the generation of the visual representations. First we present the data model describing the TS-X data which is followed by the data model used for describing the LUCAS dataset.

6.2.1 TerraSAR-X Data Model

The TS-X data model was originally designed for the system developed in the TELEIOS project [201]. The multimodule system was presented in detail in [202, 203]. This system distinguishes three kinds of system users: the ground segment system administrator (i.e., operator), the EO expert, and the end-user. Each user profile will interact with the system in a different way and consequently the access to the system is done using different set of modules. The data model designed for the system, shown in Fig. 6.1, considers all the profiles, their specificities, needs and constrains. EO Products usually are delivered in collections which are described by the table *collection* in this specific data model. During the initial storage, table *ingestion* describes some of the parameters needed for the processing of the images in the collection. The processing starts extracting the metadata from the annotation files and storing the information in *metadata*. Later, the images, which have been ingested in the *image* table, are tiled in several patches and stored in *patch*. Once the patches are created, several feature extraction methods are applied. The feature extraction methods are represented with pink headers in the data model. Among these, we can highlight the different Gabor features which are stored in *features_gafs* and *features_glc*. By combining all the generated information, the system implements different annotation tools that allow the user to generate semantic labels that are stored in *annotation* and *label*, where the predefined taxonomy is stored, see Section 6.3.1. Finally, GeoNames [204] geographical database information is linked as auxiliary data for toponymical purposes, i.e., translation from geographical coordinates to place names.

6.2.2 LUCAS Data Model

The initial LUCAS data model is presented in Section 4.2.3 but the inclusion of the data mining methodology presented in Chapter 5 extended the scope of the managed information. A detailed representation of the extended data model is shown in Fig. 6.2. The table *point* mimics the content in which the European Commission disseminate the LUCAS metadata. In general, elements of this table contain an unequivocal point identifier, several geospatial data in the form of GPS coordinates and codes referencing the administrative boundaries where they are located. The administrative areas are defined table *nuts-area* using the Nomenclature of territorial Units for Statistic (NUTS) [205] geocode standard. The EC defines NUTS levels from level 1 to level 3, but it is common use to define a level 0 that indicates the EU Member State. Thus, NUTS level 1 identify major socio-economic regions, 98 in

6. Visual Analytics for EO Archives and In-Situ Data

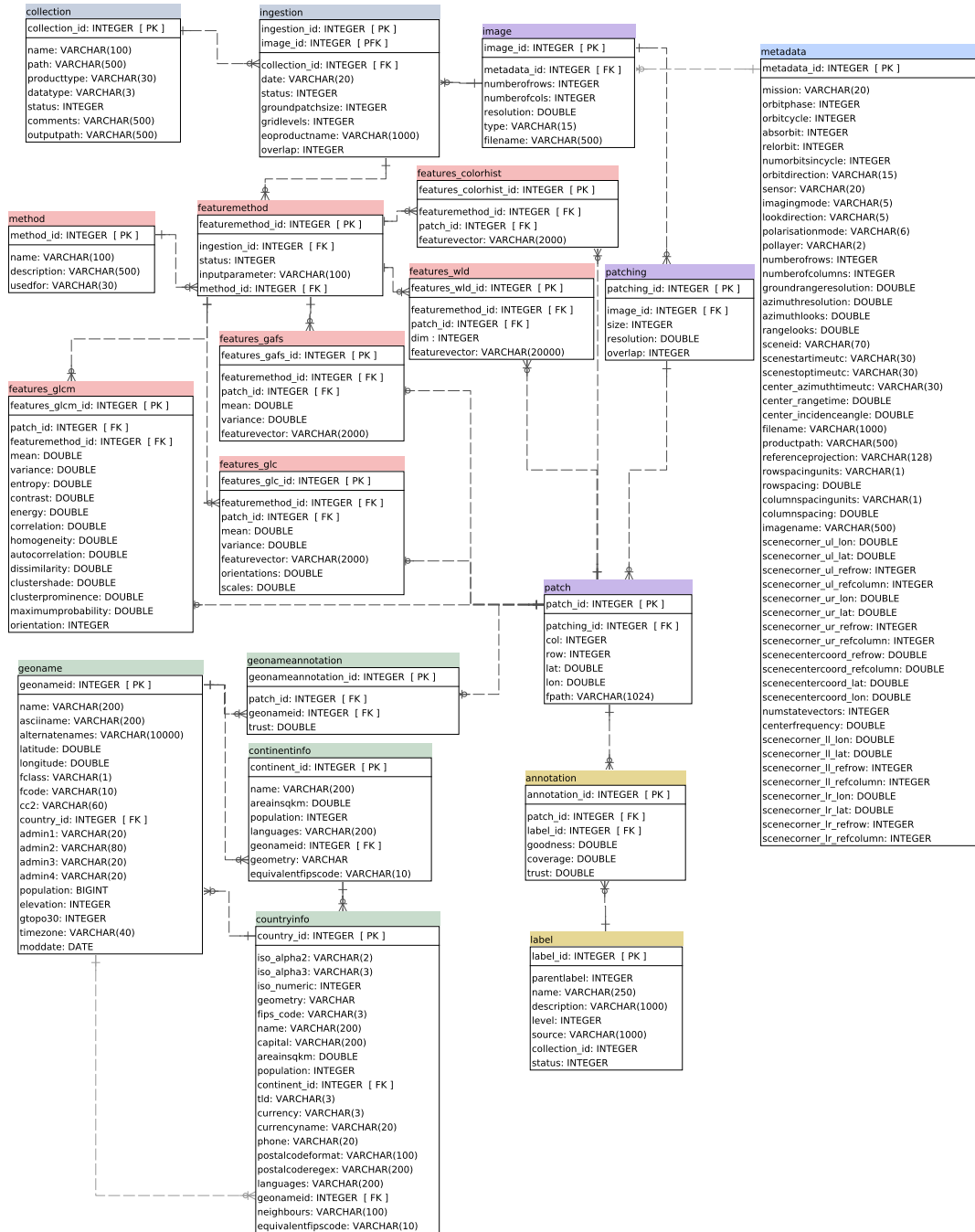


Figure 6.1: Representation of the data model used for the description of the TerraSAR-X data.

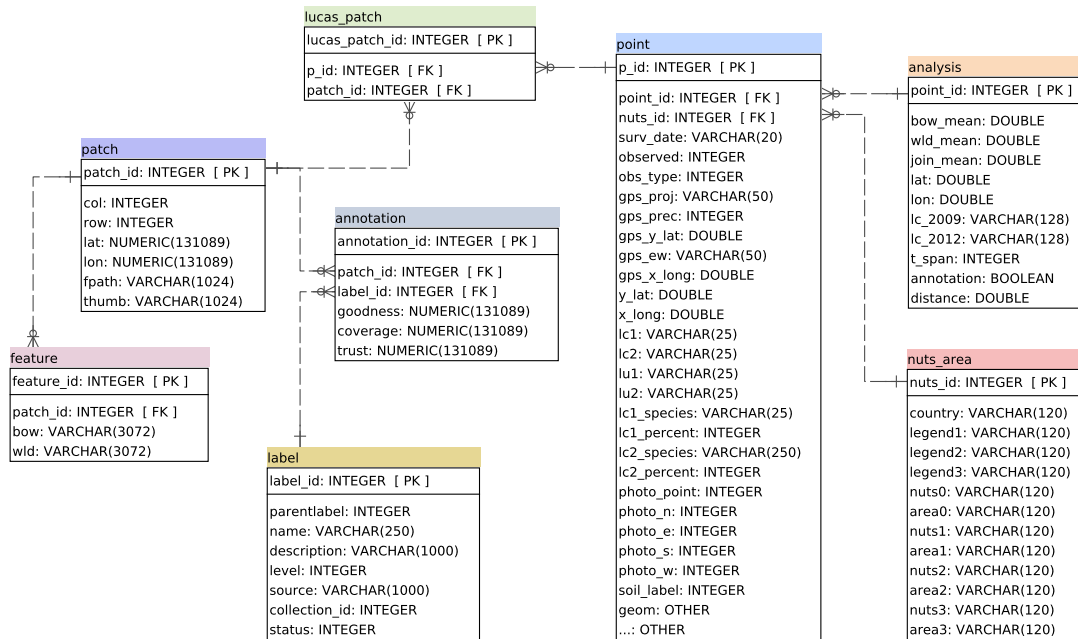


Figure 6.2: Representation of the data model used for the description of the LUCAS dataset.

the EU. Level 2 refers to basic administrative regions (e.g., Government regions in Germany, autonomous communities and cities in Spain) which sum a total of 1,276. Finally, EU recognizes a total of 1,342 small regions at NUTS level 3.

The *patch* table stores the information of every image in the archive. The table contains information about the size of the photo and the URLs pointing to the images in the data repository and the corresponding image thumbnail. The table *lucas-patch* links the survey metadata in table *point* with the multimedia images in table *patch*. The *annotation* table is the link between the images in *patch* table with the semantic labels in *label* table and defines some specific quality parameters regarding the annotation. Regarding the semantic concepts stored in table *label*, they maintain the defined hierarchy by storing the corresponding parent and its level in the hierarchy. Finally, the *analysis* table stores the results of the statistical image analysis applied to the multitemporal LUCAS points along with other pre processed information used in the data mining methodology presented in Chapter 5.

6.3 Analytics of the TerraSAR-X Archive

This section initially presents the content of the TS-X archive followed by visual representations based on different visualization approaches and set of variables.

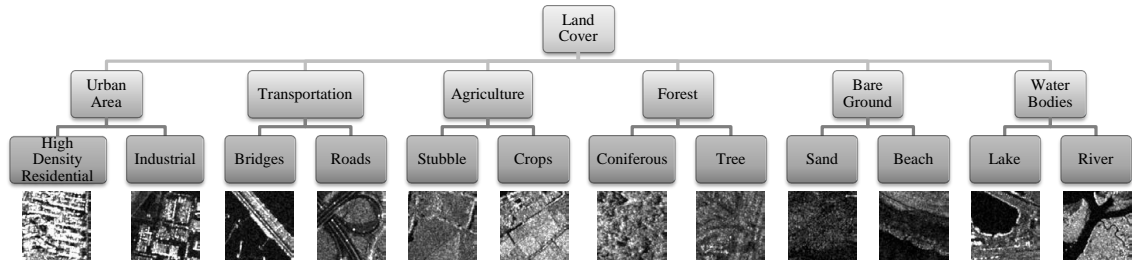


Figure 6.3: Example of the semantic catalogue of TerraSAR-X dataset. From left to right the second hierarchical level land cover are: high density residential area, industrial area, bridge, road, stubble, crop, coniferous, tree, sand, beach, lake, and river.

6.3.1 TerraSAR-X Archive

The TS-X dataset is generated from a seed of 364 L1B products which includes the image information along with 28490 metadata entries. The images are Radiometrically Enhanced (RE) Multilook Ground range Detected (MGD). The MGD offers an optimization with respect to radiometry reducing the speckle (i.e., granular noise) and MGD provides square resolution cells. The imaging mode is spotlight with a pixel spacing of 1.25 meters, and a resolution of 2.9 meters. The products include ascending and descending orbits with single polarization captures, HH and VV. The images are tiled to 160×160 pixel, resulting in a total of 406277 high resolution patches. Before storing the patches, Gabor and Weber primitive features are extracted. This features will be used in the posterior semantic annotation.

The semantic annotation process resulted in a total of 157661 annotations with around 300 semantic labels. The annotation process followed the recommendations described in [206], where the authors described via taxonomies the semantic classes that can be successfully retrieved from TS-X products. A reduced representation of the taxonomy is shown in Fig. 6.3. There, the semantic categories are divided in two hierarchical levels. The top category describes generic land covers such as: *Urban Area*, *Transportation*, *Agriculture*, *Forest*, *Bare Ground*, and *Water Bodies*. The second hierarchical level defines more specific land covers like *High Density Urban Areas*, *Trees*, *Lakes* or *Crops*. The Fig. 6.3 also shows some of the more common low hierarchy classes including patch examples.

6.3.2 Visualizing EO Image Metadata

The EO product metadata are diverse. In this section we present two possible representations of the stored metadata. The first case makes use of the available geospatial information to represent the geographical distribution of the EO scenes in the database. The second example uses various metadata related to the instrument characteristics during the image acquisition and exploits the relationship between them.



Figure 6.4: Geographical distribution of the ingested TerraSAR-X scenes around the world. The green points represent the localization of the scenes.

6.3.2.1 Geographical Distribution by Geospatial Metadata

A very efficient way to represent the content of archives containing geographical information is by using the geospatial information to represent it over a digital map. In Fig. 6.4 it is possible to see the geographical distribution of the ingested EO scenes. The dataset contains scenes spread all over the world, including: Europe, Asia, America and Africa. In this case, the way to query the archive is relatively simple. Query 1 retrieves from the database the central latitude and longitude of the images of all existing *metadata* table entries.

Query 1 Distribution of the image patches in the TerraSAR-X archive.

Input: –

Output: [*scenecentercoord_lat*, *scenecentercoord_lon*]

- 1: **procedure** GET LATITUDE AND LONGITUDE COORDINATES OF THE EO PRODUCTS IN THE DATABASE
 - 2: **SELECT** *scenecentercoord_lat*, *scenecentercoord_lon* **FROM** *metadata*
-

6.3.2.2 Relationship among the Instrument Metadata

A novel way to study the database content is representing it exploiting the relationship between the image resolution with the incidence angle and the number of annotations. The Query 2 shows the procedure to query the TS-X database. The basic parameters to query are the metadata entries referring to the looking direction

of the instrument, i.e., incidence angle; and the resolution. The number of annotations for every specific angle and resolution are counted with the addition on some extra location information to help differentiate the origin location of the patches by continent.

Query 2 Number of patches in the archive with specific incidence angle and resolution combination.

Input: –

Output: [*number of patches, lookdirection, groundrangeresolution, continent*]

- 1: **procedure** COUNT THE NUMBER OF IMAGE PATCHES WITH SPECIFIC INCIDENCE ANGLE AND RESOLUTION IN THE DATABASE
 - 2: **SELECT** Count(*patch_id*), *lookdirection*, *groundrangeresolution*
 - 3: **FROM** *patch*, *metadata*, *continentinfo*
 - 4: **WHERE**(*patch.lat*, *patch.lon*) **IN** *continentinfo.geometry*
-

In Fig. 6.5, the query results are divided by continents and the total number of annotations is represented by the size of the bubbles. Focusing on the results, we can highlight the largest number of annotations for a unique incidence angle and resolution corresponds to south America, with an angle between 22 and 24 degrees and a resolution around 3 m. The scenes corresponding to Europa spread basically around 34 and 50 degrees, while the scenes in Africa can be found at 40 to 42 degrees. Europa is the continent with more annotated scenes, followed by Asia. In general, with this representation we can clearly notice the physical dependence of the EO product resolution with the acquisition angle. The shallower the incidence angle the better the obtained final resolution of the image.

6.3.3 Visualizing Semantics and Geospatial Data

The geospatial data can be a productive information source when generating visual representations. By enhancing the use of geographical information it is possible to generate various visualizations. The following examples showcase visualizations that use geographic areas with different extensions (e.g. continent, Country, city, etc.) in order to visualize the annotated semantic labels in the archive.

6.3.3.1 Main Land Covers by Continent

As pointed out in Section 6.3.1, the TS-X database content is annotated using a two level taxonomy of land cover concepts. These land cover annotations along with the geographical locations can be used to analyzed the content of the archive. By using the Query 3 we can obtain the distribution of the six main land cover categories around the world: *Urban Area*, *Agriculture*, *Water Bodies*, *Transportation*, *Forest* and *Bare Ground*. The obtained results are plotted in Fig. 6.6. The largest

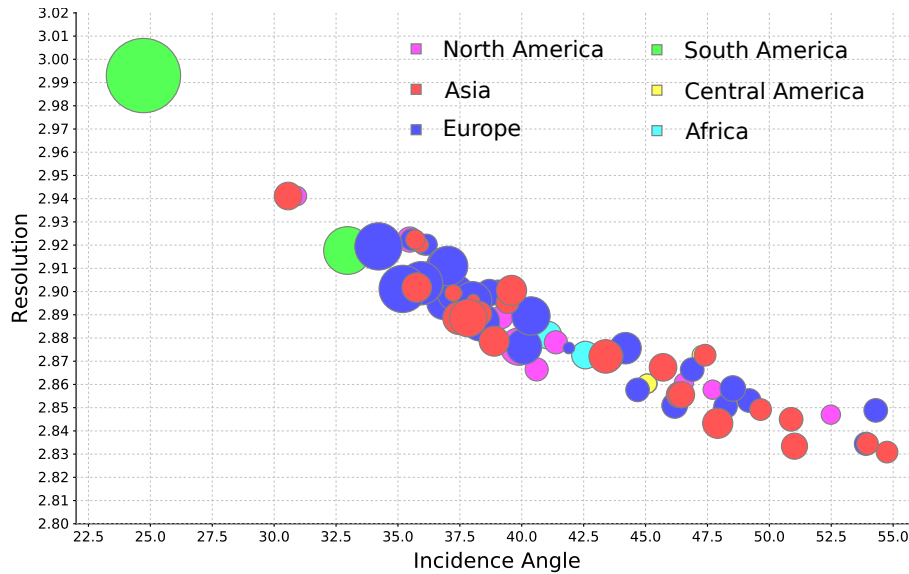


Figure 6.5: Incidence angle versus ground range resolution and total number of annotations per continent used as test database.

land cover in the TS-X dataset corresponds to *Bare Ground* category in South America. Europe stands out with the highest number of annotations of *Forest* and *Agriculture*. North America has a big amount of *Urban Area* annotations. The semantic annotations related to Africa are prolific in *Urban Areas* and *Water Bodies*. Regarding the Asian continent, it is noticeable the high number of annotations corresponding to *Water Bodies* semantic category. Besides, it is remarkable that *Agriculture* is the semantic concept with smaller number of annotations.

Query 3 Number of patches in the database sorted by main land cover and continent.

Input: –

Output: [*number of patches*, *parentlabel*, *continent*]

- 1: **procedure** RETRIEVE THE IMAGE PATCHES SORTING THEM BY THEIR LAND COVER AND GEOLOCATION
 - 2: **SELECT** COUNT(*patch_id*), *parentlabel*, *continent*
 - 3: **FROM** *patch*, *label*, *continentinfo*
 - 4: **WHERE** (*patch.lat*, *patch.lon*) **IN** *continentinfo.geometry*
-

6.3.3.2 Land Cover Distribution Around the World

From continental analysis we can reduce the scope to cities around the world. Thus, just with small changes on Query 3 it is possible to produce Fig. 6.7 which shows a comparison of some relevant land cover categories in different cities. Some remarks

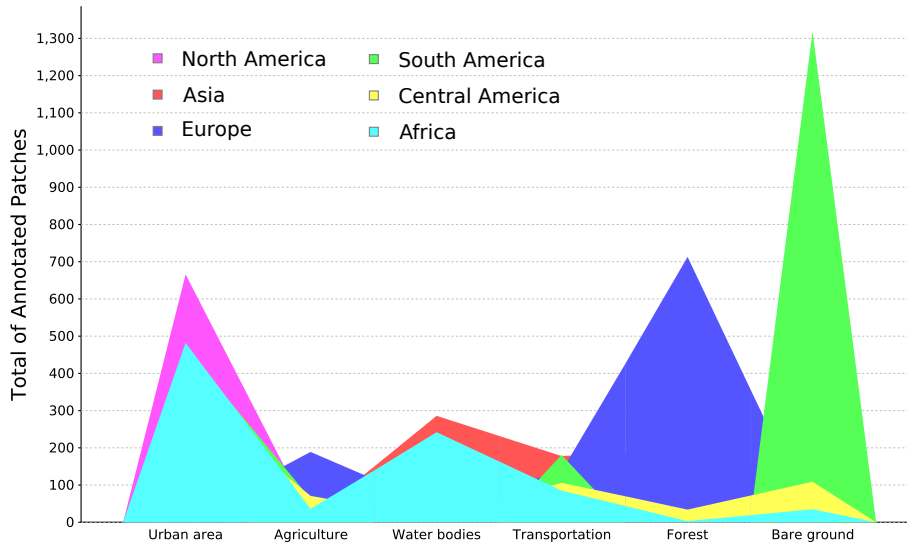


Figure 6.6: Distribution of the main land cover categories around the world.

that can be extracted from the representation include the fact that Madrid in Spain has the highest number of elements annotated with the category *Road*, in contrast to Bogota in Colombia which has the lowest. The semantic category *Sand* appears only in cities of the middle east Countries like Saudi Arabia, Irak, and Iran. Besides, the concept *Forest* appears to be highly annotated in Singapore and other neighboring cities of Malaysia and Thailand. *Skyscraper* category is similarly distributed between Ottawa in Canada, Kuala Lumpur in Malasia, and San Francisco in USA.

Continuing with the analysis of the semantic annotations in the database, it is even possible to focus the study area in a way that we can produce analytics of land cover distribution of some of the most important cities in Germany, see Fig. 6.8. In this specific case the graphical representation is done using the tools provided by Gephi [207]. The figure shows the distribution of the land cover in eleven German cities; Berlin, Stuttgart, Oldenburg, Munich, Mannheim, Lindau, Kiel, Karlsruhe, Cologne, Bremen, and Bonn. Each axis represents a city which extends in a spiral of circles representing different land covers. The size of the circle represents the amount of elements in the database and the color of the circles the highest hierarchical class that the specific land cover belongs. The main land cover categories are listed in the legend. The generated analytic has some interactive possibilities that allow the user to visualize each circle information just by hovering the mouse cursor over specific circles. An analysis of the image can provide some interesting information. For example, Berlin is the city with the highest number of annotations corresponding to *Urban Area*. Lindau possesses the highest number of annotations with *Water Body* but limited just to one type. On the other hand, Kiel has four different *Water Body* classes and Oldenburg has no annotations related to water.

6.3. Analytics of the TerraSAR-X Archive

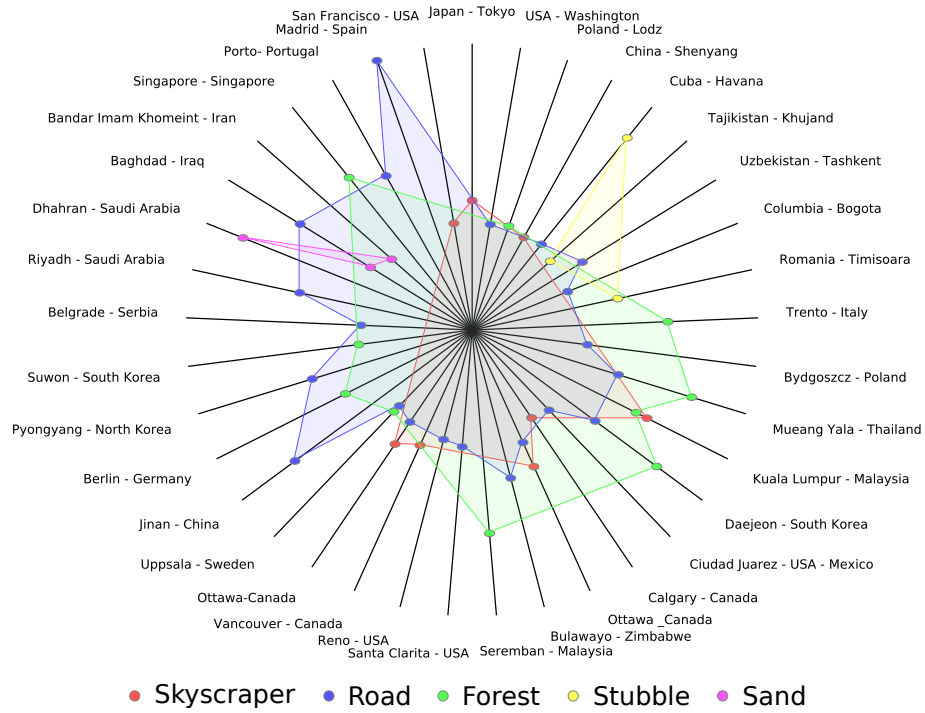


Figure 6.7: Distribution of different land covers over different cities in the world using TerraSAR-X data.

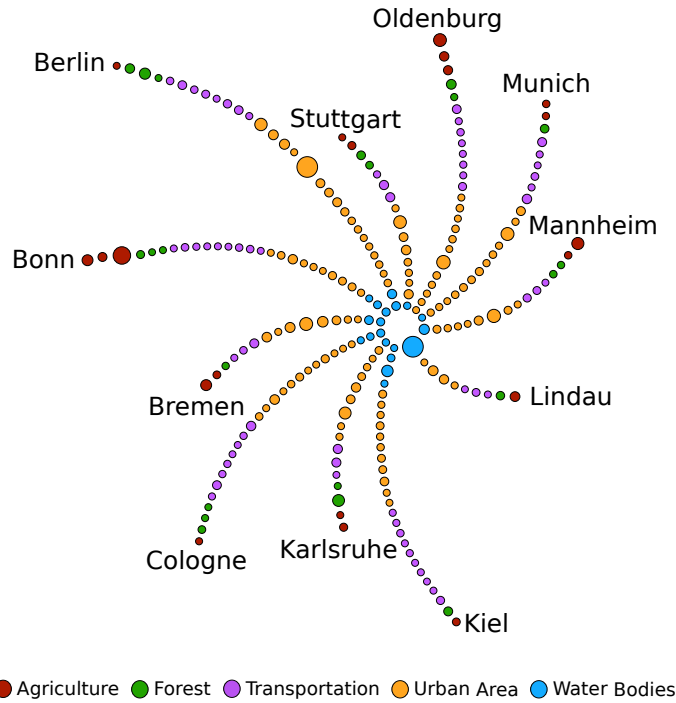


Figure 6.8: Distribution of Land Cover over Germany using TerraSAR-X data

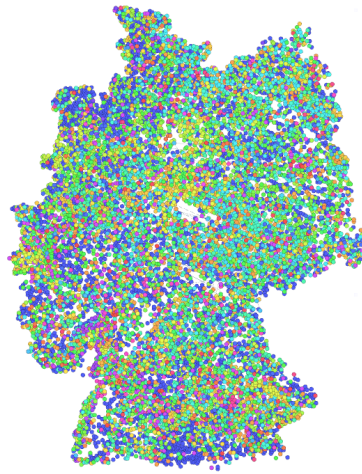


Figure 6.9: Distribution of the different land covers over Germany using LUCAS database

6.4 Analytics of the LUCAS Archive

The following discussion on the LUCAS survey dataset is focused on Germany. At this point we employ some of the visualization capabilities offered by the system presented in Chapter 4 which uses the query capabilities and the interactive statistical graphic representations. By executing the Query 4 it is possible to represent geographically the surveyed points and the annotated classes. This is exactly what it is represented in Fig. 6.9, where we can observe the diversity of the semantic catalog and its huge extension. The figure shows the last hierarchical level of the classes with a total of 83 different land cover classes. The most common land covers are Grassland, 29.66%, Common wheat 13.05% and Maize 9.75%.

Query 4 LUCAS survey points in a Country

Input: *country*

Output: [*label.name, latitude, longitude*]

- 1: **procedure** FILTER THE POINTS BY GEOGRAPHIC LOCATION
 - 2: **SELECT** *label.name, latitude, longitude*
 - 3: **FROM** *point, label*
 - 4: **WHERE** *nuts0 = country*
-

It is also possible to analyze LUCAS database content dividing the study in regions and changing the visualization concept. Hence, using Query 5 it is possible to obtain the distribution per year of the main land covers for any major socio-economic region. In comparison to the queries performed over the TS-X archive,

the LUCAS dataset land cover hierarchy is more complex regarding the number of levels. Thus, it requires recursive queries in order to obtain the parent labels representing the top in the land cover hierarchy. Furthermore, the query is divided in different procedures in order to produce the final answer.

The results of the query are shown in Fig. 6.10, where 16 German states are listed along the percentage of each LUCAS main class using the 2009 data. Analysing the figure, we can observe how *Cropland*, *Woodland* and *Grassland* are the more common land covers in Germany. Exceptions to this are the city states of Berlin, Bremen and Hamburg, where due to the limited extension of land the biggest percentage is assigned to *Artificial land*.

Query 5 Distribution of main land covers in the hierarchy of major socio-economic regions for an specific year

Input: $[interest_region], year$

Output: $Distribution = [interest_region, parent_land_cover, value]$

1: **procedure** FILTER THE POINTS BY YEAR AND LOCATION

2: **SELECT** $land_cover, region$ **FROM** $point$

3: **WHERE** $region$ **IN** $[interest_region]$ **AND** $survey_date$ **IN** $year$

1: **procedure** OBTAIN THE PARENT LAND COVER OF THE ALL RETRIEVED INSTANCES

1: **procedure** GENERATE THE STATISTICS FOR EVERY POSSIBLE COMBINATION OF LAND COVER THE POINTS

6.5 Joint Analytics of LUCAS and TerraSAR-X

In this section we present the available geospatial information from different cities, linking the information of LUCAS with the TS-X archive. For the generation of this visualization, the semantic annotations from both databases are parsed to a common land cover semantic, allowing the data integration and comparison. The required queries are modifications of the ones presented in Query 3, for the Terrasar-X dataset; and Query 5, for the LUCAS dataset. The resulting visualization is presented in Fig. 6.11. There, the inner circle colours correspond to different German cities: Stuttgart, Karlsruhe, Berlin, Bremen, and Cologne. The middle circle colors correspond to the data source of the information, as explained above, LUCAS or TS-X databases. Finally, the outer circle colours represents the general Land cover classes defined for this visualization: Artificial Land, Agriculture, Forest, Bareground, and Water.

Analysing the visualization is clear that the elements belonging to the LUCAS database are more common in general, with a predominance of elements corresponding to *Agriculture*, *Forest* and in a lesser amount to *Artificial Land*. As in

6. Visual Analytics for EO Archives and In-Situ Data

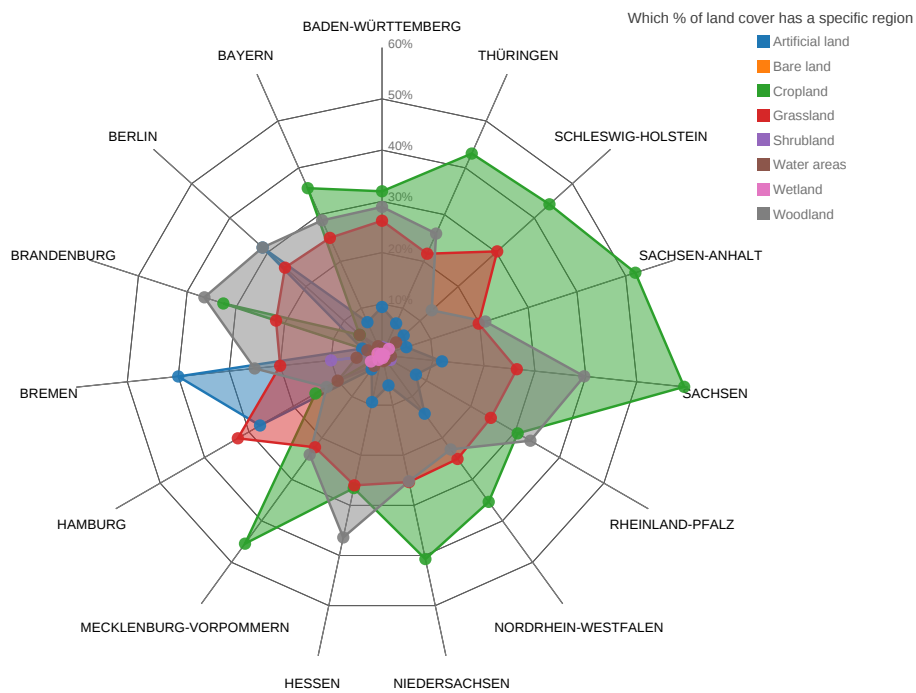


Figure 6.10: LUCAS Data analytics. Each axis corresponds to one of the 16 German states and show the percentage of each LUCAS main class using the 2009 data. We can observe how *Cropland*, *Woodland* and *Grassland* are the more common land covers in Germany. The city states of Berlin, Bremen and Hamburg appear as an exception due to their limited extension of land, where the most common land cover is *Artificial land*.

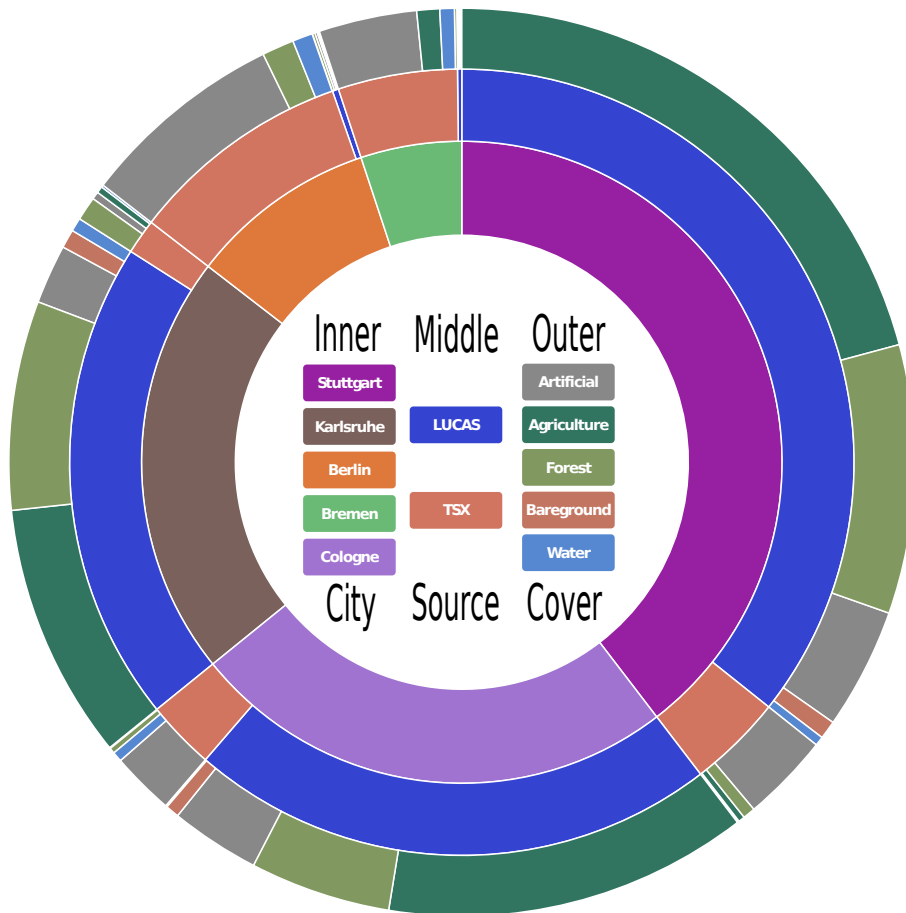


Figure 6.11: LUCAS and TerraSAR-X Data analytics. Available geospatial information of different cities joining the information from LUCAS database and a database containing TerraSAR-X EO product semantic annotations. The inner circle elements correspond to different German cities. The middle circle elements correspond to the different data sources. The outer circle elements represent the different Land cover classes.

the previous example, the differences between the city states and the others can be easily detected. For Berlin and Bremen, the available LUCAS data are limited and most of the TS-X patches belong to *Artificial Land*. For all the others, with a longer extension in terrain, much more LUCAS data are available in comparison to TS-X.

6.6 Conclusions

In this chapter we have presented firstly the main data categories that comprise the remote sensing archives. After that we have introduced the data models used for defining the TS-X dataset and LUCAS dataset. Next, along with a short description of the TS-X archive content, we introduced some visual analytics generated

after performing some semantic queries of TS-X image content. The visual representations showcase some of possible query approaches focusing the studies from big geographical regions to smaller regions, i.e., cities. Besides, we introduced some interactive analytics regarding LUCAS dataset. Using the integration capabilities and interactive visualization tools of the system presented in Chapter 4, we analyzed the presence of the main land cover in each German region. Finally, we introduced a visual representation that joints the previous datasets and compares their content. As a closing remark we can conclude that visual analytics provide very powerful tools that help the user by: 1) summarizing the query results, 2) inferring new information, and 3) improving the general understanding of the content of the image archives.

Summary, Conclusions and Future Works

7.1 Summary and Conclusions

The actual and future space Earth Observation (EO) missions along with other EO initiatives will continue to evolve and to provide great amount of heterogeneous data. The diversity of EO data includes: images (e.g., satellite or airborne), maps, semantic labels, and various metadata among others. For a more efficient exploitation of these data, we have presented different data mining tools, techniques, algorithms and concepts running on systems that promote the data integration. The outcome of this thesis validates the proposed approaches, confirming their capability of quick extracting valuable information for a better understanding of EO data. These approaches, at different degrees, rely mainly on data analytics and system engineering disciplines. More specifically, the research scope covers data integration, fusion, mining and visualization of heterogeneous content.

Our main contributions for fusing heterogeneous EO are introduced with the novel Heterogeneous Data Mining (HDM) concept. HDM enhances the original Knowledge-driven Information Mining (KIM) in two different ways. First, by introducing a faster active learning algorithm which shifts the required statistical independence from the features to the posterior probabilities. Second, by linking an external data infrastructure which allows the inclusion of information independent of the image content in the learning process, e.g., information extracted from maps in Geographical Information Systems (GIS). HDM obtains a remarkable speed-up that allows us to overcome KIM's two-model limitation in the learning, and therefore, enables the introduction of new feature models in the learning stage and the definition of more complex user semantics. The carried out tests quantified the performance of our algorithm obtaining accelerations of various orders of magnitude in comparison to the previous one. In addition, the extension of the assumption of statistical

independence in the learning algorithm from features to posterior probabilities did not introduce biases in the learning process.

Along with the learning algorithm, we introduced a probabilistic search method based on the distances between the elements in the database. This search method computes the distance among the posterior probabilities of the query image and the image Bag of Words (BoW) in the database. The obtained performance is optimum for weakly defined labels, but it cannot totally replace the retrieval method based on total posterior probabilities. In contrast to the presented probabilistic search, the posterior probability retrieval is capable of yielding the tiles with a higher probability of containing the required user semantics. Furthermore, the scalar ranking used by the posterior probability method performs faster for scenarios where several feature models are taken into account in the learning process. On the contrary, the computation cost of the probabilistic retrieval increases with each extra feature model.

The data integration aspect of the work is covered all along the thesis, but it takes special significance with the definition of the multilayer system architecture for heterogeneous geospatial data analytics. Under the scope of this system we successfully integrate information from EO products, cartographic products and in-situ data. The implemented server-client prototype integrates several web technologies that balance the complex and heavy processes toward the server. In this way, it is possible to run the lightweight client in a great variety devices with HTML-5 compatibility, e.g., laptops, tablets or phones. In the multilayer architecture the initial layers read the original data and perform the required transformations to make the data integration viable. Once transformed, the data are linked and stored in a geographical database or in a system repository. The link among the data allows the top layer to exploit the database capabilities in order to perform geographical queries over the stored data. At the same time, the top layer also implements all the communication protocols to the linked third party services and the server logic that interacts with the user via graphical interfaces. The data integration capabilities offered by the proposed system have proven their utility by helping to get a better understanding of EO images for expert and non-expert users. Furthermore, exploiting the in-situ information, the system can define optimal datasets, as well as ground truth information for change detection on EO image time series.

The integration of in-situ Land Use/Cover Area frame Survey (LUCAS) data gave rise to its analysis via data mining procedures. Analyses of the acquired multitemporal LUCAS data had shown a very high variability in the land covers, which exceeded experts' expectations. Thus, we performed a deeper analysis of the LUCAS surveys aiming to differentiate the real land cover changes from the potential inaccuracies introduced during the acquisition of the data. Our analysis proposes a data mining methodology that takes advantage of the new heterogeneous geospatial data analytics system introduced in this thesis. Divided in three different steps, the methodology is able to successfully filter false land cover changes. The first step

refines the database query processes mapping the changes in the land cover class hierarchy, excluding the points that only present land cover changes due to the hierarchy's modification. The second step analyses the in-situ images, computes the similarity among the multitemporal images of each survey point, and then, generates a ranking of the points based on the computed similarity values. The last step requires the user interaction and includes on map data visualization and filtering tools. With the help of these tools and a small investment in manpower and time, the filtering reviewing task is performed. The final data mining procedure results in a clear reduction in the total number of land cover changes, validating the presented methodology and tools for the assurance of the in-situ recorded land cover changes.

Finally, we focused our efforts in the visualization of heterogeneous EO data. In order to generate visualizations that facilitate the data understanding of the EO information, it is necessary to combine the results obtained from automatic analysis methodologies with interactive visualization tools. These tools allow the analysis of massive amounts of information in real-time, and in consequence, they provide the means to navigate, understand and exploit the data more efficiently.

We presented data models of two different EO systems that are queried to generate visual analytics. First, we performed queries to the TerraSAR-X content combining different image metadata. Some of those metadata are used directly on the visualization (e.g., incidence angle) but others like the location (i.e. latitude and longitude) need to be processed in order to infer their corresponding geographical regions (e.g., continent, country or city). Secondly, we developed some interactive analytics that query the LUCAS dataset visualizing the division of the main land covers in a country. In line with the rest of the thesis, and using the integration capabilities and interactive visualization tools of the proposed geospatial analytic system, we introduced a novel visual representation that joints the previous datasets. In general, the presented visual analytics help the user summarizing the query results, inferring new information, and improving the general understanding of the content of the image archives.

7.2 Future Works

A clear further step is the addition of more data mining and machine learning tools based on in-situ and EO imagery fusion. This step will integrate and extend the HDM concept within the presented geospatial analytic system. In this way the system will be able to integrate and fuse the information obtained from in-situ sources with the maps and the EO products for machine learning purposes.

The algorithm acceleration obtained with the proposed HDM concept opens new ways for knowledge-driven information mining systems to Big Data scenarios. In this regard, the introduction of additional models based on heterogeneous sources should be followed by intensive validation tests over larger datasets which would

7. Summary, Conclusions and Future Works

include a wide variety of images from different sensors, scenarios and third party data sources.

Future developments may address LUCAS information and the presented data mining methodology for the quality control of the recorded data. Works in this specific topic could target the enhancement of the methodology via the inclusion of interactive graphical visualizations allowing a better comprehension of the mistakes done on every specific land cover change.

A

Related Publications

A.1 Journals

- K. Alonso and M. Datcu, "Accelerated Probabilistic Learning Concept for Mining Heterogeneous Earth Observation Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 7, pp. 3356-3371, July 2015.
- K. Alonso, D. Espinoza-Molina, and M. Datcu, Multilayer Architecture for Heterogeneous Geospatial Data Analytics: Querying and Understanding EO Archives, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 791-801, March 2017.
- K. Alonso, D. Espinoza-Molina, and M. Datcu, Mining Multitemporal in-situ Heterogeneous Monitoring Information for the Assurance of Recorded Land Cover Changes, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 877-887, March 2017.

A.2 Conferences

- K. Alonso and M. Datcu, "Accelerated Knowledge-Driven Image Mining System for Data Fusion in Big Data," in *Proc. ESA-EUSC-JRC 9th Conference on Image Information Mining. The Sentinels Era*, Bucharest, pp. 97-100, 2014.
- K. Alonso and M. Datcu, "Image Information Mining: an Accelerated Bayesian Algorithm for Data Fusion of SAR Big Data," in *Proc. 10th European Conference on Synthetic Aperture Radar (EUSAR)*, Berlin, pp. 1-4, 2014.
- K. Alonso and M. Datcu, "Knowledge-driven image mining system for Big Earth Observation data fusion: GIS maps inclusion in active learning stage," in

A. Related Publications

- Proc. IEEE Geoscience and Remote Sensing Symposium (IGARSS)*, Quebec City, QC, pp. 3538-3541, 2014.
- K. Alonso, D. Espinoza-Molina and M. Datcu, "LUCAS Visual Browser: A tool for land cover visual analytics," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Milan, pp. 1484-1487, 2015.
 - D. Espinoza-Molina, K. Alonso, G. Schwarz and M. Datcu, "Big Data Analytics for Detailed Urban Mapping," in *Proc. Mapping Urban Areas from Space (MUAS)*, Frascati, 2015.
 - D. Espinoza-Molina, K. Alonso and M. Datcu, "Visual Analytics for Semantic Queries of TerraSAR-X Image Content," in *Proc. SPIE Remote Sensing Conference*, Toulouse, pp. 1-10, 2015.
 - D. Espinoza-Molina, K. Alonso and M. Datcu, "Semantic Indexing of TerraSAR-X and in situ data for urban analytics," in *Proc. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS)*, Kish Island, pp. 185-188, 2015.
 - D. Espinoza-Molina, K. Alonso and M. Datcu, "Visual Data Mining for Exploration of the Feature Space Using In-situ Data," in *Proc. IEEE Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, pp. 3538-3541, 2016.

Acronyms

AL	Active Learning.
ARVI	Atmospherically Resistant Vegetation Indexes.
B	Blue spectral band.
BoW	Bag of Words.
CBIR	Content-Based Image Retrieval.
DoG	Diference of Gaussians.
EC	European Commission.
EO	Earth Observation.
EU	European Union.
EVI	Enhanced Vegetation Index.
FrFT	Fractional Fourier Transform.
G	Green spectral band.
GARI	Green Atmospherically Resilient Index.
GEMI	Global Environmental Monitoring Index.
GeoIRIS	Geospatial Information Retrieval and Indexing System.

GFD	Gabor Feature Descriptor.
GFD	Generic Fourier Descriptors.
GIS	Geographical Information Systems.
GLCM	Gray Level Co-occurrence Matrix.
GLDH	Gray Level Difference Histograms.
GMRF	Gaussian Markov Random Fields.
GPP	Gross Primary Production.
GPS	Global Positioning System.
GRF	Gibbs Random Fields.
GUI	Graphic User Interface.
HDM	Heterogeneous Data Mining.
HF	Haralick Features.
HSV	Hue-Saturation-Lightness.
HTML	Hyper Text Markup Language.
I ³ KR	Intelligent Interactive Knowledge Retrieval.
IIM	Image Information Mining.
IIM-TS	Image Information Mining in Time Series.
InSAR	Interferometric SAR.
IR	InfraRed.
IRS	Information Retrieval Systems.
KIM	Knowledge-driven Information Mining.
LAN	Local Area Network.
LBP	Local Binary Pattern.
LDA	Latent Dirichlet Allocation.
LSI	Latent Semantic Indexing.
LUCAS	Land Use/Cover Area frame Survey.
MARS	Multimedia Analysis and Retrieval System.
MEA	Multi-sensor Evolution Analysis.
MFT	Matched Fourier Transform.
MGD	Multilook Ground range Detected.
MIRS	Multimedia Information Retrieval Systems.
MODIS	Moderate Resolution Imaging Spectroradiometer.

NCD	Normalized Compression Distance.
NIR	Near-InfraRed.
NUTS	Nomenclature of territorial Units for Statistic.
NVDI	Normalized Difference Vegetation Index.
OS	Operative Systems.
pLSA	probabilistic Latent Semantic Analysis.
PNG	Portable Network Graphics.
PolSAR	Polarimetric SAR.
PPM	Posterior Probability Map.
QbE	Query by Example.
QBIC	Query By Image Content.
QbK	Query by Keyword.
QMF	Quadrature Mirror Filter.
R	Red spectral band.
RE	Radiometrically Enhanced.
RF	Relevance Feedback.
RGB	Red-Green-Blue.
RVI	Ratio Vegetation Index.
SAR	Synthetic Aperture Radar.
SAVI	Soil Adjusted Vegetation Index.
SI	Semantic Index.
SIFT	Scale-Invariant Feature Transform.
SOM	Self-Organizing Maps.
SQL	Standard Query Language.
SR	Simple Ratio.
STFT	Short Time Fourier Transform.
SVG	Scalable Vector Graphic.
SWIR	Short-Wavelength Infrared.
TMS	Tile Map Services.
TS-X	TerraSAR-X.

UI	User Interface.
URL	Uniform Resource Locator.
VARIgreen	Visible Atmospherically Resistant Index Green.
VIg	Vegetation Index green.
VQ	Vector Quantization.
WebGL	Web Graphic Library.
WLAN	Wireless Local Area Network.
WLD	Weber Local Descriptor.
WMS	Web Map Service.
WV-2	WorldView-2.
WWW	World-Wide Web.

List of Symbols

C_{aero}	Aerosol resistance coefficients.
D	Data source.
G	Gain factor.
L	User defined semantic label.
L_{cba}	Canopy background adjustment factor.
L_{soil}	Soil-adjustment factor.
M	Segments of the histogram.
N_{atm}	Atmospherically corrected surface reflectances.
N_i	Number of occurrences of ω_i .
O	Complexity.
R	SIFT region size.
T	Set of user provided positive training data.
Φ	Dominant orientations.
α	Hyperparameter vector representing the user interaction.
$\neg L$	Not user defined semantic label.
ω	Words.
ω_{RGB}	Dictionary obtained from RGB features.
ω_{WLD}	Dictionary obtained from WLD features.
ω_i	Combined dictionary.
ϕ	Parametrized multinomial distribution of T .
θ	Extracted features.

θ_{SIFT}	SIFT orientations.
ζ	Differential excitations.
c	Model identifier.
d	Distance.
d_{Ch}	Chebyshev distance.
d_E	Euclidian distance.
d_{JSD}	Jensen-Shannon Divergence.
d_{KLS}	Kullback-Leibler symmetric variant distance.
d_{KL}	Kullback-Leibler distance.
d_M	Manhattan distance.
d_n	Individual pixels in D .
k	Training iteration.
n	Number of operations.
$p(L D)$	Posterior probability of a label in the image data.
$p(L d_n)$	Posterior probability of a label in image data pixel.
$p(\omega_{RGB} D)$	Probability of ω_{RGB} within the given the image data.
$p(\omega_{WLD} D)$	Probability of ω_{WLD} within the given image data.
$p(\omega_i)$	Prior probability of the words.
$p(\omega_i D)$	Occurrences of the words within the given image data.
$p(\omega_i L)$	Probabilistic link of the words with a label.
$p(\omega_i \neg L)$	Probability of the words outside the user define label.

Bibliography

- [1] K. C. Clarke, “Advances in Geographic Information Systems,” *Computers, Environment and Urban Systems*, vol. 10, no. 3-4, pp. 175–184, 1986.
- [2] R. L. Church, “Geographical Information Systems and Location Science,” *Computers & Operations Research*, vol. 29, pp. 541–562, 2002.
- [3] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P. Marchetti, and S. D’Elia, “Information Mining in Remote Sensing Image Archives: System Concepts,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, pp. 2923–2936, Dec. 2003.
- [4] J. Bieniarz, *Sparse Methods for Hyperspectral Unmixing and Image Fusion*. PhD thesis, University of Osnabrück, Dec. 2015.
- [5] L. Graham, “Synthetic interferometer radar for topographic mapping,” *Proceedings of the IEEE*, vol. 62, pp. 763–768, June 1974.
- [6] European Commission, “INSPIRE Directive,” 2007. online: <http://www.eea.europa.eu/about-us/what/seis-initiatives/inspire-directive>.
- [7] M. Haklay and P. Weber, “OpenStreetMap: User-generated Street Maps,” *IEEE Pervasive Computing*, pp. 12–18, 2008.
- [8] C. J. Date, *An Introduction to Database Systems: Vol. I (4th Ed.)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1986.
- [9] G. Salton, “A simple blueprint for automatic boolean query processing,” *Information Processing and Management*, vol. 24, pp. 269–280, May 1988.
- [10] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

- [11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," Technical Report 1999-66, Stanford InfoLab, Nov. 1999. Previous number = SIDL-WP-1999-0120.
- [12] J. C. Nordbotten, *Multimedia Information Retrieval Systems*. University of Bergen, Norway, 2008. Web-book: http://nordbotten.com/ADM/ADM_book/MIRS-frame.htm.
- [13] G. Lu, *Multimedia database management systems*. Boston, USA: Artech House Computing Library, 1999.
- [14] N. Sebe, M. S. Lew, X. Zhou, T. S. Huang, and E. M. Bakker, "The state of the art in image and video retrieval," in *Proc. Second International Conference on Image and Video Retrieval (CIVR 2003)*, pp. 1–8, 2003.
- [15] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, 2006.
- [16] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, vol. 40, pp. 1–60, Apr. 2008.
- [17] T. Kato, T. Kurita, H. Shimogaki, T. Mizutori, and K. Fujimura, "Cognitive view mechanism for multimedia database system," in *Proc. First International Workshop on Interoperability in Multidatabase Systems (IMS 1991)*, pp. 179–186, Apr. 1991.
- [18] T. Kato, "Database architecture for content-based image retrieval," in *Proc. SPIE Conference on Image Storage and Retrieval Systems*, vol. 1662, pp. 112–123, 1992.
- [19] J. Z. Wang, N. Boujemaa, A. D. Bimbo, D. Geman, A. G. Hauptmann, and J. Te, "Diversity in Multimedia Information Retrieval Research," *Multimedia Information Retrieval*, vol. 1, pp. 5–12, 2006.
- [20] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions in Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.
- [21] Y. Rui, T. S. Huang, and S.-F. Chang, "Image Retrieval : Past , Present , and Future," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 1–23, 1997.

-
- [22] P. Aigrain, H. Zhang, and D. Petkovic, "Content-based Representation and Retrieval of Visual Media: A State-of-the-art Review," *Multimedia Tools and Applications*, vol. 3, no. 3, pp. 179–202, 1996.
- [23] V. N. Gudivada, "Spatial Similarity Measures for Multimedia Applications.," in *Proc. Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 363–372, 1995.
- [24] C. S. McCamy, H. Marcw, and J. G. Davidson, "A Color Rendition Chart," *Journal of Applied Photographic Engineering*, vol. 2, no. 3, pp. 95–99, 1976.
- [25] M. Miyahara and Y. Yoshida, "Mathematical Transform Of (R, G, B) Color Data To Munsell (H, V, C) Color Data," in *Proc. SPIE*, vol. 1001, pp. 650–657, 1988.
- [26] J. Wang, W.-J. Yang, and R. Acharya, "Color clustering techniques for color-content-based image retrieval from image databases," in *Proc. IEEE International Conference on Multimedia Computing and Systems*, pp. 442–449, June 1997.
- [27] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.
- [28] A. H. Munsell, *A Color Notation*. G. H. Ellis Company, 1905.
- [29] R. S. Hunter, "Photoelectric Color Difference Meter," *Journal of the Optical Society of America*, vol. 48, pp. 985–995, Dec. 1958.
- [30] G. H. Joblove and D. Greenberg, "Color Spaces for Computer Graphics," *SIGGRAPH Comput. Graph.*, vol. 12, pp. 20–25, Aug. 1978.
- [31] M. Ioka, *A Method of Defining the Similarity of Images on the Basis of Color Information*. Research report RT, IBM Research, Tokyo Research Laboratory, 1989.
- [32] M. Stricker and M. Orengo, "Similarity of Color Images," in *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1995.
- [33] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1582–1596, Sept. 2010.
- [34] G. Stockman and L. G. Shapiro, *Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1st ed., 2001.

- [35] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, pp. 610–621, Nov. 1973.
- [36] C. C. Gotlieb and H. E. Kreyszig, "Texture Descriptors Based on Co-occurrence Matrices," *Journal Computer Vision, Graphics, and Image Processing*, vol. 51, pp. 70–86, July 1990.
- [37] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, 2005.
- [38] F. Liu and R. W. Picard, "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 722–733, July 1996.
- [39] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 971–987, July 2002.
- [40] J. L. Chen and A. Kundu, "Rotation and Gray Scale Transform Invariant Texture Identification Using Wavelet Decomposition and Hidden Markov Model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, pp. 208–214, Feb. 1994.
- [41] J. R. Smith and S.-F. Chang, "Transform features for texture classification and discrimination in large image databases," in *Proc. IEEE International Conference on Image Processing (ICIP 1994)*, vol. 3, pp. 407–411 vol.3, Nov. 1994.
- [42] A. Kundu and J.-L. Chen, "Texture classification using QMF bank-based subband decomposition," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 5, pp. 369–384, 1992.
- [43] K. S. Thyagarajan, T. Nguyen, and C. E. Persons, "A maximum likelihood approach to texture classification using wavelet transform," in *Proc. IEEE International Conference on Image Processing (ICIP 1994)*, vol. 2, pp. 640–644 vol.2, Nov. 1994.
- [44] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 55–73, Jan. 1990.
- [45] J. R. Smith and S.-F. Chang, "Automated binary texture feature sets for image retrieval," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996)*, vol. 4, pp. 2239–2242, May 1996.

-
- [46] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 837–842, Aug. 1996.
- [47] T. Randen and J. H. Husoy, "Filtering for texture classification: a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 291–310, Apr. 1999.
- [48] Y. Mingqiang, K. K. Idiyo, and R. Joseph, "A Survey of Shape Feature Extraction Techniques," *Pattern Recognition*, pp. 43–90, Nov. 2008.
- [49] I. T. Young, J. E. Walker, and J. E. Bowie, "An analysis technique for biological shape. I," *Information and Control*, vol. 25, no. 4, pp. 357–370, 1974.
- [50] M. Peura and J. Iivarinen, "Efficiency of Simple Shape Descriptors," in *Proc. International Workshop on Visual Form*, (Capri, Italy), 1997.
- [51] D. Chetverikov and Y. Khenokh, "Matching for Shape Defect Detection," in *Proc. 8th International Conference on Computer Analysis of Images and Patterns (CAIP 1999)*, (Berlin, Heidelberg), pp. 367–374, 1999.
- [52] R. Chellappa and R. Bagdazian, "Fourier Coding of Image Boundaries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, pp. 102–105, Jan. 1984.
- [53] Q. M. Tieng and W. W. Boles, "Recognition of 2d object contours using the wavelet transform zero-crossing representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 910–916, Aug. 1997.
- [54] A. D. Bimbo and P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 121–132, Feb. 1997.
- [55] H. Freeman, "On the encoding of arbitrary geometric configurations," *IRE Transactions on Electronic Computers*, vol. EC-10, pp. 260–268, June 1961.
- [56] J. A. Saghri and H. Freeman, "Analysis of the precision of generalized chain codes for the representation of planar curves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, pp. 533–539, Sept. 1981.
- [57] W. I. Grosky and R. Mehrotra, "Index-based object recognition in pictorial data management," *Comput. Vision Graph. Image Process.*, vol. 52, pp. 416–436, Oct. 1990.
- [58] R. Mehrotra and J. E. Gary, "Similar-shape retrieval in shape data management," *Computer*, vol. 28, pp. 57–62, Sept. 1995.

- [59] M. Kliot and E. Rivlin, “Invariant-based shape retrieval in pictorial databases,” in *Proc. 5th European Conference on Computer Vision (ECCV 1998)*, vol. 1, pp. 491–507, 1998.
- [60] A. Goshtasby, “Description and discrimination of planar shapes using shape matrices,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-7, pp. 738–743, Nov. 1985.
- [61] K. Chakrabarti, M. Ortega-Binderberger, K. Porkaew, and S. Mehrotra, “Similar shape retrieval in mars,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME 2000)*, vol. 2, pp. 709–712, 2000.
- [62] C. H. Teh and R. T. Chin, “On image analysis by the methods of moments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, pp. 496–513, July 1988.
- [63] W.-Y. Kim and Y.-S. Kim, “A region-based shape descriptor using Zernike moments,” *Signal Processing: Image Communication*, vol. 16, no. 12, pp. 95–102, 2000.
- [64] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*. New York, NY, USA: John Wiley & Sons, Inc., 2002.
- [65] D. Zhang and G. Lu, “Generic fourier descriptor for shape-based image retrieval,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME 2002)*, vol. 1, pp. 425–428 vol.1, 2002.
- [66] H. Blum, “A Transformation for Extracting New Descriptors of Shape,” in *Models for the Perception of Speech and Visual Form* (W. Wathen-Dunn, ed.), pp. 362–380, Cambridge: MIT Press, 1967.
- [67] B. S. Morse, *Computation of Object Cores from Grey-level Images*. PhD thesis, Chapel Hill, NC, USA, 1995.
- [68] D. Zhang and G. Lu, “Review of shape representation and description techniques,” *Pattern Recognition*, vol. 37, no. 1, pp. 1–19, 2004.
- [69] T. S. Chua, K.-L. Tan, and B. C. Ooi, “Fast signature-based color-spatial image retrieval,” in *Proc. IEEE International Conference on Multimedia Computing and Systems*, pp. 362–369, June 1997.
- [70] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, “Efficient and effective Querying by Image Content,” *Journal of Intelligent Information Systems*, vol. 3, no. 3, pp. 231–262, 1994.

-
- [71] H. Lu, B.-C. Ooi, and K.-L. Tan, "Efficient Image Retrieval By Color Contents," in *Proc. First International Conference on Applications of Databases (ADB 1994)*, pp. 95–108, 1994.
- [72] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 762–768, June 1997.
- [73] K. S. Fu and J. K. Mui, "A survey on image segmentation," *Pattern Recognition*, vol. 13, no. 1, pp. 3–16, 1981.
- [74] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260–280, 2008.
- [75] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The qbic system," *Computer*, vol. 28, pp. 23–32, Sept. 1995.
- [76] W. Niblack, "The QBIC Project: querying Images By Content Using Color, Texture and Shape," in *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pp. 173–187, 1993.
- [77] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu, "Virage image search engine: An open framework for image management.," in *Proc. Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 76–87, 1996.
- [78] A. Gupta and R. Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, no. 5, pp. 70–79, 1997.
- [79] J. R. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-based Image Query System," in *Proc. Fourth ACM International Conference on Multimedia (MULTIMEDIA 1996)*, pp. 87–98, ACM, 1996.
- [80] J. R. Smith and S.-F. Chang, "Visually searching the Web for content," *IEEE MultiMedia*, vol. 4, pp. 12–20, July 1997.
- [81] J. R. Smith and S.-F. Chang, "Image and Video Search Engine for the World Wide Web.," in *Proc. Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 84–95, 1997.
- [82] C. Frankel, M. J. Swain, and V. Athitsos, "WebSeer: An Image Search Engine for the World Wide Web," tech. rep., Chicago, IL, USA, 1996.

- [83] T. Gevers and A. W. M. Smeulders, “Pictoseek: combining color and shape invariant features for image retrieval,” *IEEE Transactions on Image Processing*, vol. 9, pp. 102–119, Jan. 2000.
- [84] T. S. Huang, S. Mehrotra, and K. Ramchandran, “Multimedia analysis and retrieval system (mars) project.,” in *Proc. 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval*, pp. 101–117, 1996.
- [85] E. Hörster, R. Lienhart, and M. Slaney, “Image retrieval on large-scale image databases,” in *Proc. 6th ACM International Conference on Image and Video Retrieval (CIVR 2007)*, pp. 17–24, ACM, 2007.
- [86] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [87] R. Zhao and W. I. Grosky, “Negotiating the semantic gap: from feature maps to semantic landscapes,” *Pattern Recognition*, vol. 35, pp. 593–600, 2001.
- [88] T. Hofmann, “Probabilistic Latent Semantic Indexing,” in *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pp. 50–57, ACM, 1999.
- [89] Y. Takama and S. Hattori, “Mining Association Rules for Adaptive Search Engine Based on RDF Technology,” *IEEE Transactions on Industrial Electronics*, vol. 54, pp. 790–796, Apr. 2007.
- [90] H. Ma, J. Zhu, M. R. T. Lyu, and I. King, “Bridging the semantic gap between image contents and tags,” *IEEE Transactions on Multimedia*, vol. 12, pp. 462–473, Aug. 2010.
- [91] J. Tang, Z. J. Zha, D. Tao, and T. S. Chua, “Semantic-Gap-Oriented Active Learning for Multilabel Image Annotation,” *IEEE Transactions on Image Processing*, vol. 21, pp. 2354–2360, Apr. 2012.
- [92] L. Ballerini, X. Li, R. B. Fisher, and J. Rees, “A Query-by-Example Content-Based Image Retrieval System of Non-melanoma Skin Lesions,” in *Proc. First MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support (MCBR-CDS 2009)*, pp. 31–38, 2010.
- [93] N. Boujemaa and J. Fauqueur, “Ikona: Interactive Specific and Generic Image Retrieval,” *International Workshop on Multimedia Content-Based Indexing and Retrieval*, pp. 2–5, 2001.
- [94] B. Settles, “Active Learning Literature Survey,” *Computer Sciences Technical Report*, vol. 1648, pp. 1–67, 2010.

-
- [95] A. Baraldi, V. Puzzolo, P. Blonda, L. Bruzzone, and C. Tarantino, "Automatic Spectral Rule-Based Preliminary Mapping of Calibrated Landsat TM and ETM+ Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, pp. 2563–2586, Sept. 2006.
- [96] J. S. Tyo, A. Konsolakis, D. I. Diersen, and R. C. Olsen, "Principal-components-based display strategy for spectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, pp. 708–718, Mar. 2003.
- [97] D. Bratasanu, I. Nedelcu, and M. Datcu, "Interactive Spectral Band Discovery for Exploratory Visual Analysis of Satellite Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, pp. 207–224, Feb. 2012.
- [98] G. S. Birth and G. R. McVey, "Measuring the colour of growing turf with a reectance spectrophotometer," *Agronomy Journal*, vol. 60, pp. 640 – 643, 1968.
- [99] C. F. Jordan, "Derivation of Leaf-Area Index from Quality of Light on the Forest Floor," *Ecology*, vol. 50, no. 4, pp. 663–666, 1969.
- [100] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring Vegetation Systems in the Great Plains with ERTS," in *Proc. Third Earth Resources Technology Satellite Symposium*, vol. 1, pp. 309–317, 1973.
- [101] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sensing of Environment*, vol. 8, no. 2, pp. 127–150, 1979.
- [102] P. J. Sellers, "Canopy reflectance, photosynthesis and transpiration," *International Journal of Remote Sensing*, vol. 6, no. 8, pp. 1335–1372, 1985.
- [103] R. B. Myneni, F. G. Hall, P. J. Sellers, and A. Marshak, "The interpretation of spectral vegetation indexes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, pp. 481–486, Mar. 1995.
- [104] J. R. Jensen, *Remote Sensing of the Environment: An Earth Resource Perspective*. Prentice Hall Series in Seographic Information Science, Pearson Prentice Hall, 2007.
- [105] A. Huete, "A soil-adjusted vegetation index (SAVI)," *Remote Sensing of Environment*, vol. 25, no. 3, pp. 295–309, 1988.
- [106] H. Q. Liu and A. Huete, "A feedback based modification of the NDVI to minimize canopy background and atmospheric noise," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 2, pp. 457–465, 1995.

- [107] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, “Overview of the radiometric and biophysical performance of the MODIS vegetation indices,” *Remote Sensing of Environment*, vol. 83, no. 12, pp. 195–213, 2002.
- [108] B. D. Wardlow, S. L. Egbert, and J. H. Kastens, “Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains,” *Remote Sensing of Environment*, vol. 108, no. 3, pp. 290–310, 2007.
- [109] D. A. Sims, A. F. Rahman, V. D. Cordova, B. Z. El-Masri, D. D. Baldocchi, P. V. Bolstad, L. B. Flanagan, A. H. Goldstein, D. Y. Hollinger, L. Misson, R. K. Monson, W. C. Oechel, H. P. Schmid, S. C. Wofsy, and L. Xu, “A new model of gross primary productivity for North American ecosystems based solely on the enhanced vegetation index and land surface temperature from MODIS,” *Remote Sensing of Environment*, vol. 112, no. 4, pp. 1633–1646, 2008.
- [110] Y. J. Kaufman and D. Tanre, “Atmospherically resistant vegetation index (arvi) for eos-modis,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, pp. 261–270, Mar. 1992.
- [111] A. A. Gitelson, Y. J. Kaufman, R. Stark, and D. Rundquist, “Novel algorithms for remote estimation of vegetation fraction,” *Remote Sensing of Environment*, vol. 80, no. 1, pp. 76–87, 2002.
- [112] B. Pinty and M. M. Verstraete, “GEMI: a non-linear index to monitor global vegetation from satellites,” *Vegetatio*, vol. 101, no. 1, pp. 15–20, 1992.
- [113] R. P. Sripada, D. C. Farrer, R. Weisz, R. W. Heiniger, and J. G. White, “Aerial color infrared photography to optimize in-season nitrogen fertilizer recommendations in winter wheat,” *Agronomy Journal*, vol. 99, no. 6, pp. 1424–1435, 2007.
- [114] “The IDB Project. A database for remote sensing indices.,” Dec. 2011. online: <http://www.indexdatabase.de>.
- [115] L. B. Almeida, “The fractional Fourier transform and time-frequency representations,” *IEEE Transactions on Signal Processing*, vol. 42, pp. 3084–3091, Nov. 1994.
- [116] J. Singh and M. Datcu, “SAR Image Categorization With Log Cumulants of the Fractional Fourier Transform Coefficients,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 5273–5282, Dec. 2013.

-
- [117] A. Popescu, C. Patrascu, J. Singh, I. Gavat, and M. Datcu, "Spotlight TerraSAR-X Data Modeling using Spectral Space-Variant Measures, for scene Targets and Structure Indexing," in *Proc. 8th European Conference on Synthetic Aperture Radar (EUSAR 2010)*, pp. 1–4, June 2010.
- [118] C. O. Dumitru and M. Datcu, "Information Content of Very High Resolution SAR Images: Study of Feature Extraction and Imaging Parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 4591–4610, Aug. 2013.
- [119] D. Espinoza-Molina, D. Gleich, and M. Datcu, "Evaluation of Bayesian Despeckling and Texture Extraction Methods Based on GaussMarkov and Auto-Binomial Gibbs Random Fields: Application to TerraSAR-X Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, pp. 2001–2025, May 2012.
- [120] E. Miyamoto and T. Merryman, "Fast Calculation of Haralick Texture Features," tech. rep., 2005.
- [121] R. Cossu, "Segmentation by means of textural analysis," *Pixel*, vol. 1, no. 2, pp. 21–24, 1988.
- [122] A. Baraldi and F. Parmiggiani, "An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 2, pp. 293–304, 1995.
- [123] L. K. Soh and C. Tsatsoulis, "Texture analysis of sar sea ice imagery using gray level co-occurrence matrices," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 780–795, Mar. 1999.
- [124] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Canadian Journal of remote sensing*, vol. 28, no. 1, pp. 45–62, 2002.
- [125] T. M. Kuplich, P. J. Curran, and P. M. Atkinson, "Relating SAR image texture to the biomass of regenerating tropical forests," *International Journal of Remote Sensing*, vol. 26, no. 21, pp. 4829–4854, 2005.
- [126] I. Champion, C. Germain, J. P. D. Costa, A. Alborini, and P. Dubois-Fernandez, "Retrieval of Forest Stand Age From SAR Image Texture for Varying Distance and Orientation Values of the Gray Level Co-Occurrence Matrix," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 5–9, Jan. 2014.

- [127] A. A. Popescu, I. Gavvat, and M. Datcu, "Contextual Descriptors for Scene Classes in Very High Resolution SAR Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, pp. 80–84, Jan. 2012.
- [128] O. Rajadell, P. García-Sevilla, and F. Pla, "Spectral-Spatial Pixel Characterization Using Gabor Filters for Hyperspectral Image Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 860–864, July 2013.
- [129] H. C. Li, T. Celik, N. Longbotham, and W. J. Emery, "Gabor Feature Based Unsupervised Change Detection of Multitemporal SAR Images Based on Two-Level Clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 2458–2462, Dec. 2015.
- [130] J. Singh and M. Datcu, "SAR Image Categorization With Log Cumulants of the Fractional Fourier Transform Coefficients," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 5273–5282, Dec. 2013.
- [131] J. Singh and M. Datcu, "Use of the second-kind statistics for VHR SAR image retrieval," in *Proc. 9th International Conference on Communications (COMM 2012)*, pp. 367–370, June 2012.
- [132] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A Robust Local Image Descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705–1720, 2010.
- [133] J. Chen, S. Shan, G. Zhao, and X. Chen, "A Robust Descriptor Based on Weber's Law," in *Proc. Computer Vision and Pattern Recognition*, 2008.
- [134] "Weber's law," *Encyclopædia Britannica*, 2016. online: <http://www.britannica.com/science/Webers-law>.
- [135] S. Cui, C. O. Dumitru, and M. Datcu, "Ratio-Detector-Based Feature Extraction for Very High Resolution SAR Image Patch Indexing," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 1175–1179, Sept. 2013.
- [136] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal on Computer Vision*, vol. 60, pp. 91–110, Nov. 2004.
- [137] M. A. Michael W. Davidson, "Difference of Gaussians Edge Enhancement," *Molecular Expressions Microscopy Primer: Digital Image Processing*, Olympus America Inc., and Florida State University, 2016. online: <http://micro.magnet.fsu.edu/primer/java/digitalimaging/processing/diffgaussians/index.html>.

-
- [138] B. Sirmacek and C. Unsalan, "Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, pp. 1156–1167, Apr. 2009.
- [139] Q. Li, G. Wang, J. Liu, and S. Chen, "Robust Scale-Invariant Feature Matching for Remote Sensing Image Registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, pp. 287–291, Apr. 2009.
- [140] H. Goncalves, L. Corte-Real, and J. A. Goncalves, "Automatic Image Registration Through Image Segmentation and SIFT," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 2589–2600, July 2011.
- [141] B. Fan, C. Huo, C. Pan, and Q. Kong, "Registration of Optical and SAR Satellite Images by Exploring the Spatial Relationship of the Improved SIFT," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 657–661, July 2013.
- [142] S. Wang, H. You, and K. Fu, "BFSIFT: A Novel Method to Find Feature Matches for SAR Image Registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, pp. 649–653, July 2012.
- [143] Z. Harris, "Distributional Structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.
- [144] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 2, pp. 524–531 vol. 2, June 2005.
- [145] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.
- [146] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature Coding in Image Classification: A Comprehensive Study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 493–506, 2014.
- [147] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A Theoretical Analysis of Feature Pooling in Visual Recognition," in *Proc. ICML*, pp. 111–118, Omnipress, 2010.
- [148] Y. Yang and S. Newsam, "Bag-of-visual-words and Spatial Extensions for Land-use Classification," in *Proc. 18th International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2010)*, pp. 270–279, ACM, 2010.

- [149] L. J. Zhao, P. Tang, and L. Z. Huo, "Land-Use Scene Classification Using a Concentric Circle-Structured Multiscale Bag-of-Visual-Words Model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, pp. 4620–4631, Dec. 2014.
- [150] M. Lienou, H. Maitre, and M. Datcu, "Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, pp. 28–32, Jan. 2010.
- [151] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, pp. 193–204, Mar. 2011.
- [152] M. C. Burl, C. Fowlkes, and J. Roden, "Mining for Image Content," in *Proc. Systemics, Cybernetics, and Informatics / Information Systems: Analysis and Synthesis*, (Orlando, USA), 1999.
- [153] M. Datcu, S. D'Elia, R. King, and L. Bruzzone, "Introduction to the Special Section on Image Information Mining for Earth Observation Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 795–798, 2007.
- [154] W. Hsu, M. L. Lee, and J. Zhang, "Image Mining: Trends and Developments," *Journal of Intelligent Information Systems*, vol. 19, no. 1, pp. 7–23, 2002.
- [155] K. W. Tobin, B. L. Bhaduri, E. A. Bright, A. Cheriyyadat, T. P. Karnowski, P. J. Palathingal, T. E. Potok, and J. R. Price, "Automated Feature Generation in Large-Scale Geospatial Libraries for Content-Based Indexing," *Photogrammetric Engineering & Remote Sensing*, vol. 72, pp. 531–540, May 2006.
- [156] S. Natali, A. Beccati, S. D'Elia, M. Veratelli, P. Campalani, M. Folegani, and S. Mantovani, "Multitemporal Data Management and Exploitation," in *Proc. 6th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multi-Temp 2011)*, no. C, (Trento), pp. 217–220, 2011.
- [157] C.-R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, and K. Palaniappan, "GeoIRIS: Geospatial Information Retrieval and Indexing System-Content Mining, Semantics Modeling, and Complex Queries.," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 839–852, Apr. 2007.
- [158] S. Durbha and R. King, "Semantics-enabled Framework for Knowledge Discovery from Earth Observation Data Archives," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 11, pp. 2563–2572, 2005.

-
- [159] F. Bovolo and L. Bruzzone, “Image Information Mining in Time Series: Algorithms and Methods for Prototyping,” tech. rep., IIMTS-TN-ACS-0102 Issue 1.0, ESA, 2007.
- [160] M. Molinier, J. Laaksonen, and T. Hame, “Detecting Man-made Structures and Changes in Satellite Imagery with a Content-based Information Retrieval System Built on Self-Organizing Maps,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 861–874, 2007.
- [161] “EOlib. Earth Observation Librarian.,” Dec. 2011. online: <http://deepenandlearn.esa.int/tiki-index.php?page=EOLIB+Project>.
- [162] D. Espinoza-Molina and M. Datcu, “Earth-Observation Image Retrieval Based on Content, Semantics, and Metadata,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 5145–5159, Nov. 2013.
- [163] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data - the story so far,” *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.
- [164] M. Quartulli and I. G. Olaizola, “A Review of EO Image Information Mining,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 75, pp. 11–28, Jan. 2013.
- [165] J. Kohlhammer, D. Keim, M. Pohl, G. Santucci, and G. Andrienko, “Solving Problems with Visual Analytics,” in *Proc. 2nd European Future Technologies Conference and Exhibition (FET 11)*, vol. 7, pp. 117–120, 2011.
- [166] D. A. Keim, F. Mansmann, and J. Thomas, “Visual Analytics: How Much Visualization and How Much Analytics?,” *SIGKDD Explor. Newsl.*, vol. 11, pp. 5–8, May 2010.
- [167] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, “Fusing concept detection and geo context for visual search,” in *Proc. 2nd ACM International Conference on Multimedia Retrieval (ICMR 2012)*, pp. 4:1—4:8, ACM, 2012.
- [168] D. A. Keim, C. Panse, M. Sips, and S. C. North, “Visual Data Mining in Large Geospatial Point Sets,” *Computer Graphics and Applications, IEEE*, vol. 24, no. 5, pp. 36–44, 2004.
- [169] T. F. Stepinski, P. Netzel, and J. Jasiewicz, “LandEx - A geoweb tool for query and retrieval of spatial patterns in land cover datasets,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 1, pp. 257–266, 2014.

- [170] M. Babae, G. Rigoll, and M. Datcu, “Immersive interactive information mining with application to earth observation data retrieval,” in *Proc. International Cross Domain Conference and Workshop (CD-ARES 2013)*, vol. 8127 of *Lecture Notes in Computer Science*, pp. 376–386, 2013.
- [171] M. Schröder, H. Rehrauer, K. Seidel, and M. Datcu, “Interactive Learning and Probabilistic Retrieval in Remote Sensing Image Archives,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 5, pp. 2288–2298, 2000.
- [172] D. Heckerman, “A Tutorial on Learning with Bayesian Networks,” tech. rep., Microsoft Research Advanced Technology Division, Nov. 1996.
- [173] QGIS Development Team, *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2014.
- [174] D. Lewis, “Naive (bayes) at forty: The independence assumption in information retrieval,” in *Proc. 10th European Conference on Machine Learning (ECML 1998)*, pp. 4–15, 1998.
- [175] M. Schröder-Brzosniowsky, *Stochastic Modeling of Image Content in Remote Sensing Image Archives*. PhD thesis, University of Ulm, 2000.
- [176] I. M. G. Muñoz, *Concepts Elaboration and System Architectures for Mining Very Large Image Archives*. PhD thesis, University of Siegen, 2009.
- [177] H. Daschiel and M. Datcu, “Information Mining in Remote Sensing Image Archives: System Evaluation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 1, pp. 188–199, 2005.
- [178] H. Daschiel and M. Datcu, “Image Information Mining System Evaluation Using Information-Theoretic Measures,” *EURASIP Journal on Applied Signal Processing*, no. 14, pp. 2153–2163, 2005.
- [179] M. Rehman, M. Iqbal, M. Sharif, and M. Raza, “Content Based Image Retrieval : Survey,” *World Applied Sciences Journal*, vol. 19, no. 3, pp. 404–412, 2012.
- [180] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, “The Similarity Metric,” *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.
- [181] P. Grunwald and P. Vitanyi, “Shannon Information and Kolmogorov Complexity,” *CoRR*, vol. cs.IT/0410, p. 54, 2004.
- [182] T. Fawcett, “An Introduction to ROC Analysis,” *Pattern Recognition Letters*, vol. 27, pp. 861–874, June 2006.

-
- [183] M. Worboys and M. Duckham, *GIS: A Computing Perspective, 2nd Edition*. Boca Raton, FL, USA: CRC Press, Inc., 2004.
- [184] N. Alameh, "Chaining Geographic Information Web Services," *IEEE Internet Computing*, vol. 7, pp. 22–29, Oct. 2003.
- [185] C. Shahabi, F. Banaei-Kashani, A. Khoshgozaran, L. Nocera, and S. Xing, "GeoDec: A Framework to Effectively Visualize and Query Geospatial Data for Decision-Making," *IEEE Multimedia*, vol. 17, no. 3, pp. 14–23, 2010.
- [186] J. Wang, M. Pierce, Y. Ma, G. Fox, A. Donnellan, J. Parker, and M. Glasscoe, "Using service-based gis to support earthquake research and disaster response," *Computing in Science Engineering*, vol. 14, pp. 21–30, Sept. 2012.
- [187] D. Brunner, S. Member, G. Lemoine, S. Member, F. Thoorens, and L. Bruzzone, "Distributed Geospatial Data Processing Functionality to Support Collaborative and Rapid Emergency Response," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 2, no. 1, pp. 33–46, 2009.
- [188] T. Heinen, S. Kiemle, and B. Buckl, "The Geospatial Service Infrastructure for DLR's National Remote Sensing Data Library," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 2, no. 4, pp. 260–269, 2009.
- [189] OSGeo, "Tile Map Service." online: http://wiki.osgeo.org/wiki/Tile_Map_Service_Specification.
- [190] M. Stonebraker and L. A. Rowe, "The Design of POSTGRES," in *Proc. ACM SIGMOD International Conference on Management of Data*, vol. 15, (Washington, D.C., USA), pp. 340–355, ACM, 1986.
- [191] M. Stonebraker, L. A. Rowe, and M. Hirohama, "The Implementation of POSTGRES," *IEEE Transactions on Knowledge and Data Engineering*, vol. 2, no. 1, pp. 125–142, 1990.
- [192] R. Vatsavai, S. Shekhar, T. E. Burk, and S. Lime, "UMN-MapServer: A High-performance, Interoperable, and Open Source Web Mapping and Geospatial Analysis System," in *Proc. 4th International Conference on Geographic Information Science (GIScience 2006)*, pp. 400–417, 2006.
- [193] OGC, "Web Map Service," Mar. 2006. online: <http://www.opengeospatial.org/standards/wms>.
- [194] P. Sloup, "WebGL Earth," bachelor thesis, Masaryk University, Faculty of Informatics, Spring 2011.

- [195] K. T. GmbH, “WebGL Earth,” 2011. online: <http://www.webglearth.org>.
- [196] M. Foundation, “Web Graphics Library,” Mar. 2011. online: <https://www.khronos.org/registry/webgl/specs/1.0/>.
- [197] M. Bostock, “D3 Data-Driven Documents,” Feb. 2011. [Online: <https://d3js.org/>].
- [198] “PostGIS. Spatial and Geographic Objects for PostgreSQL,,” 2016. [Online: <http://postgis.net/>].
- [199] J. Groff and P. Weinberg, *SQL The Complete Reference, 3rd Edition*. New York, NY, USA: McGraw-Hill, Inc., 3 ed., 2010.
- [200] K. Alonso and M. Datcu, “Accelerated Probabilistic Learning Concept for Mining Heterogeneous Earth Observation Images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 7, pp. 3356–3371, 2015.
- [201] “TELEIOS. Virtual Observatory Infrastructure for Earth Observation Data,” Sept. 2010. online: <http://www.earthobservatory.eu/>.
- [202] D. Espinoza Molina and M. Datcu, “Data Mining and Knowledge Discovery tools for exploiting big Earth-Observation data,” *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-7/W3, pp. 627–633, 2015.
- [203] D. Espinoza-Molina, C. O. Dumitru, M. Datcu, and C. members, “KDD Prototype - phase I,” tech. rep., German Aerospace Agency, Issue: 3.2.1.
- [204] M. Wick, “GeoNames.” online: www.geonames.org.
- [205] “NUTS - Nomenclature of Territorial Units for Statistics,” Apr. 2016. online: <http://ec.europa.eu/eurostat/web/nuts/overview>.
- [206] C. O. Dumitru, G. S. Shiyong Cui, and M. Datcu, “A Taxonomy for High Resolution SAR Images,” in *Proc. Image Information Mining: Geospatial Intelligence from Earth Observation Conference (ESA-EUSC-JRC 2014)*, pp. 89–92, 2014.
- [207] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: An Open Source Software for Exploring and Manipulating Networks,” 2009.