Interactive Models for Latent Information Discovery in Satellite Images

DISSERTATION

zur Erlangung des Grades eines Doktors
der Ingenieurwissenschaften

vorgelegt von
Dipl.-Eng. Dragos Bratasanu

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät
der Universität Siegen
Siegen 2014

Koordinator: Prof. Dr. Otmar Loffeld
Koordinator: Prof. Dr. Mihai Datcu

Datum der Mündlichen Prüfung: 30 Mai 2014

To my parents

Abstract

The recent increase in Earth Observation (EO) missions has resulted in unprecedented volumes of multi-modal data to be processed, understood, used and stored in archives. The advanced capabilities of satellite sensors become useful only when translated into accurate, focused information, ready to be used by decision makers from various fields. Two key problems emerge when trying to bridge the gap between research, science and multi-user platforms: (1) The current systems for data access permit only queries by geographic location, time of acquisition, type of sensor, but this information is often less important than the latent, conceptual content of the scenes; (2) simultaneously, many new applications relying on EO data require the knowledge of complex image processing and computer vision methods for understanding and extracting information from the data.

This dissertation designs two important concept modules of a theoretical image information mining (IIM) system for EO: semantic knowledge discovery in large databases and data visualization techniques. These modules allow users to discover and extract relevant conceptual information directly from satellite images and generate an optimum visualization for this information.

The first contribution of this dissertation brings a theoretical solution that bridges the gap and discovers the semantic rules between the output of state-of-the-art classification algorithms and the semantic, human-defined, manually-applied terminology of cartographic data. The set of rules explain in latent, linguistic concepts the contents of satellite images and link the low-level machine language to the high-level human understanding.

The second contribution of this dissertation is an adaptive visualization methodology used to assist the image analyst in understanding the satellite image through optimum representations and to offer cognitive support in discovering relevant information in the scenes. It is an interactive technique applied to discover the optimum combination of three spectral features of a multi-band satellite image that enhance visualization of learned targets and phenomena of interest. The visual mining module is essential for an IIM system because all EO-based applications involve several steps of visual inspection and the final decision about the information derived from satellite data is always made by a human operator. To ensure maximum correlation between the requirements of the analyst and the possibilities of the computer, the visualization tool models the human visual system and secures that a change in the image space is equivalent to a change in the perception space of the operator. This thesis presents novel concepts and methods that help users access and discover latent information in archives and visualize satellite scenes in an interactive, human-centered and information-driven workflow.

Table of Contents

1

Introduction

Our planet is going through unprecedented climate and environmental changes, affecting our society in previously unknown ways. To monitor and predict the effects of these diverse environmental challenges, new generations of space borne imaging sensors with advanced recording capabilities have been recently launched or are scheduled for the next few years (e.g. ESA Sentinels, SPOT 6, 7, Pleiades). This increase in Earth Observation (EO) missions will result in large volumes of data requiring to be processed, understood, used and stored in archives. The need for timely delivery of accurate, focused information for decision making and intervention is constantly growing and the continuous increase in archives' size and EO sensors' variety require new methodologies for information mining and management, supported by shared knowledge.

The analysis and understanding of only a few very high resolution multi-spectral or synthetic aperture radar satellite images has become a highly complex and challenging task in the current operational scenarios of Earth Observation. In addition, major applications relying on remote sensing data (e.g. global monitoring, disaster management support, agriculture and food security) and large programmes and initiatives (e.g. GMES, GEO, GEOSS) support the international trend of launching new, more powerful satellites into orbit to measure phenomena about the Earth. While the technical capabilities of satellites have increased manifold, studies reveal than less than 5% of the data are actually used in applications.What is the reason these numbers are so low?

Classical data access systems in EO archives allow only queries by metadata, i.e. geographical location, time of acquisition, type of sensor and cloud cover. Data analysis and information retrieval are usually performed by remote sensing experts through time consuming and expensive procedures of visual inspection and semi-automated investigations. The manual processes performed by experts to mine information from images are currently too complex and expensive to be applied systematically on even a small subset of the acquired scenes. These limitations might become even more challenging in the future since more missions and constellations are being planned, with broader sensor variety, higher data transmission rates and increasing complexity.

The exponential increase in volume, details, diversity and complexity, complemented by users' demands for simultaneous access to multi-domain data require new methodologies for image information mining, multi-domain information management, knowledge management and sharing. In today's Earth Observation scenario, the research fields of Image Information Mining (IIM) and Content-based Image Retrieval (CBIR) are providing new solutions for querying large remote sensing archives directly by image content and latent information discovered in the scenes. IIM is an interdisciplinary approach for automatic remote sensing analysis that draws on knowledge from signal processing, image analysis, pattern recognition, artificial intelligence, machine learning, information theory, databases, semantics, ontology and knowledge management. CBIR systems aim to model the human behaviour when querying an image archive. They rely on searching a database by image content using (1) techniques from computer vision to interpret and understand the scene and (2) techniques from information retrieval and database management to rapidly locate images suiting a specific query.

1.1 Motivation

The classical procedure in CBIR supports an interactive query-by-example retrieval, allowing users to take active part in the mining process. Searches based on natural language terms rapidly run into issues of intractability due to the very limited progress on the problem of language processing, the need for vast common-sense knowledge about the world and the need to process many queries at once [90]. Most CBIR systems rely on diverse user interaction methods, such as search by association, search by example, and search by sketching. All these methods involve the use of images to search for other related images. Often this search process is iterative: at each stage, the user clicks an image “more like” their target image, refining the set of candidates [90]. A few systems have opened new directions for research in IIM and CBIR for EO: KIM [154, 155] with its following versions KES and KEO, GeoIRIS [153], Rapid Image Information Mining RIIM [150].

A classical CBIR system does not always successfully retrieve the target images because of the semantic gap - the missing link between the image signal represented by low-level machine features and the semantic concepts represented by words defined by a human operator. The semantic gap can be regarded as the key to translate the machine vocabulary into human language and vice-versa.

The spectral gap is another challenge that needs to be addressed in a CBIR system for Earth Observation. The spectral gap is defined as the gap between the information available in a multi-band satellite image (e.g. 8 spectral bands) and the limited amount of information that

can be displayed on the computer screen (e.g. 3 spectral channels). How can this loss of information be reduced to minimum? This dissertation bridges the spectral gap and answers the question: "What features of the satellite image contain the highest amount of visual information in rapport with the class / object of interest?". The answer to this question holds the key to minimizing the loss of information due to the spectral gap – the gap between the number of visual and non-visual spectral bands and the three channels of the display.

1.2 Positioning this Dissertation

The objective of this dissertation is to introduce and apply theoretical models for bridging the semantic gap in Earth Observation data understanding applications. The semantic gap was defined as the most important problem in the field of CBIR. Because an IIM system is based on several complex theoretic fields, this thesis integrates knowledge from parameter estimation, information theory, Bayesian inference, machine learning, color science, color vision, models of the human visual system and image processing. This dissertation brings two theoretical contributions, one from the perspective of latent information discovery in satellite images (briding the semantic gap) and one from the perspective of visual data mining (bridging the spectral gap).

1.3 Contributions

This thesis addresses two important modules an IIM system for EO should implement: (1) semantic information discovery and (2) advanced human-centered data visualization. An IIM system is composed of chain processes, from the raw image files, to visualization, data processing, feature selection, classification, indexing and ingestion. A graphical interface allows users to provide training samples for the mining algorithms to query the archive for similar images. This section presents the contributions of this dissertation and the two theoretical modules.

1.3.1 Latent Information Discovery in Satellite Images

The first contribution of this dissertation brings a solution for bridging the gap between the output of the state-of-the-art automatic classification algorithms and the high-level semantic terminology of cartographic data. We provide a hybrid method to automatically understand and describe the semantic rules that link the outputs of unsupervised information mining methods to cartographic vector data with different specifications. By discovering the set of rules that explain semantic classes in cartographic systems, we introduce the theoretical model of an interactive learning loop that uses the concept of direct semantics applied on satellite images. Thus, we provide a solution for an important problem that emerges while generating cartographic information layers directly for the raw files of the satellite image: semantic annotation of objects and classes in the scenes.

Figure 1.1 shows an example of how the semantic gap between satellite data and the ontology of CORINE LAND COVER (CLC) can be bridged. CLC is the European standard for cartography of land cover and land use, aiming at providing an inventory of Earth's surface features for managing the environment. The experiment was performed on a Landsat ETM 7+

image from Romania aiming to infer the semantic rules that link the image data with the CLC map. Initially, the information is reduced to a spectral map with 27 clusters with intermediate level semantic labels attached to each pixel value and then a number of latent topics are estimated from this map. Each pixel in the spectral map is assigned to the latent topic of maximum probability. Figure 1.2 shows the distribution of visual words (clusters) within each of the five estimated latent topics and how the terminology of CLC can be described by these latent topics. Table 1.1 shows the distribution of latent topics over the classes in CLC and the semantic rules that bridge the intermediate-level semantics to the high-level information classes. In the ideal case, each topic exclusively generates a single CLC class.

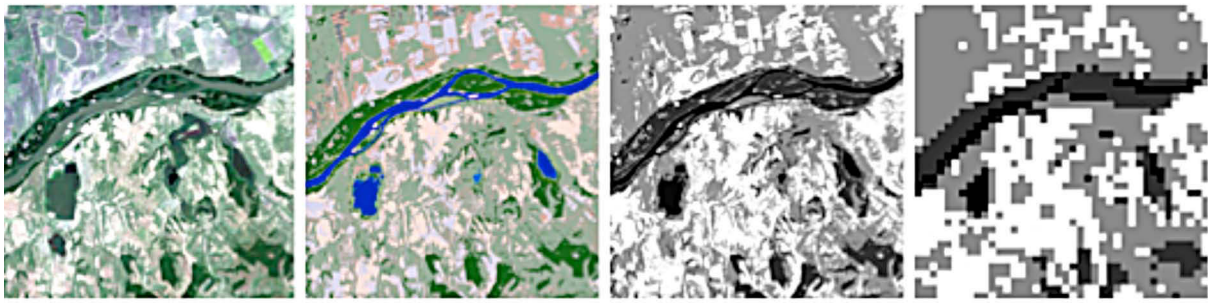


Figure 1.1a, 1.1b, 1.1c, 1.1d – fig. 1.1a shows the Landsat image 600 X 600 pixels, fig. 1.1b shows the index map with 27 classes (i.e. visual words), fig. 1.1c shows how each pixel is classified to one of the latent topics (i.e. CLC classes); fig. 1.1d depicts how each image tile (15x15 pixels) is classified into one of the latent topics (i.e. CLC classes)

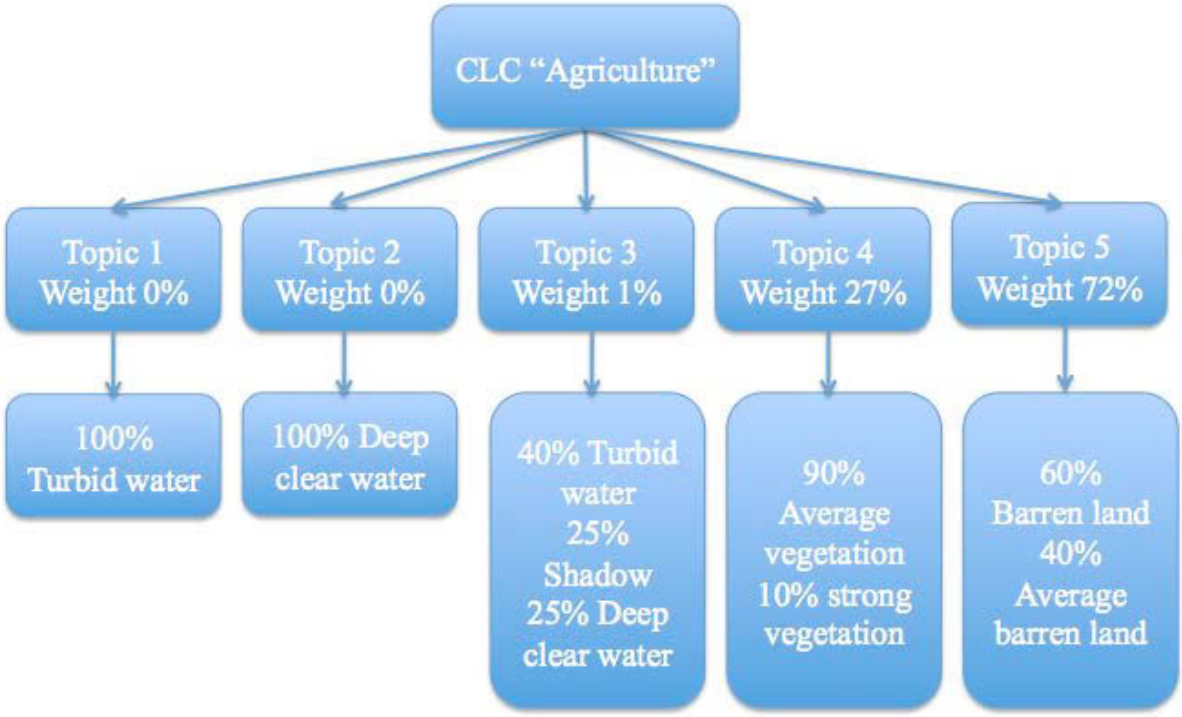


Figure 1.2 – The five latent topics estimated from the satellite image are distributions over clusters and the CLC classes are described by distributions over these latent topics.

	Water bodies	Agricultural areas	Artificial Land	Forest and natural areas	Wetlands
Topic 1	73%	0%	0%	0%	0%
Topic 2	13%	0%	0%	0%	0%
Topic 3	0%	2%	0%	54%	15%
Topic 4	14%	53%	32%	36%	80%
Topic 5	0%	45%	68%	10%	5%

Table 1.1: Semantic Rules Discovery

Topic 1: 100% turbid water
Topic 2: 90% deep clear water, 10% turbid water
Topic 3: 100% peat bogs
Topic 4: 100% average vegetation
Topic 5: 70% average shrub land, 20% bright barren land, 20% strong barren land, 10% average vegetation.

1.3.2 Visual Data Mining

The second contribution of this dissertation introduces an adaptive visualization methodology used for enhancing visual mining of objects and classes in multi-band satellite images. The visual mining module is essential for an IIM system because all remote sensing applications involve several steps of visual inspection – e.g. data quality assessment, operation-oriented area/object search and analysis, algorithm learning, information mining evaluation, etc. Multiple EO-related domains require highly accurate analysis of satellite images and the demands of users working in these areas are so challenging that automated procedures have yet to reach the required quality standards. For this reason, data analysis is still performed through extensive trials of visual interpretation. Because the final decision about the information derived from satellite images is always made by an operator, the visualization system models the automatic response of the human visual system to external stimuli (i.e. the image) and optimizes the combination of the spectral bands mapped to the channels of the display. The algorithm ranks the features of a multi-band satellite image using measures from information theory and automatically feeds the top three bands to the channels of the display. Figure 1.3 shows an experiment performed on WorldView-2 8-band satellite data and the target class selected by the human operator for training. The spectral bands are ranked using the minimum-redundancy-maximum-relevance (mRMR) criterion and the top three features are automatically mapped to the R, G, B channels of the display. Figure 1.4 shows the natural colour display and two enhanced versions of the target class using the method described in this dissertation. Initially, the top spectral band is displayed in the R channel. In the second case, the top spectral band is displayed in the G channel to take advantage of the increased sensitivity of the eye in the green region of the electromagnetic spectrum.

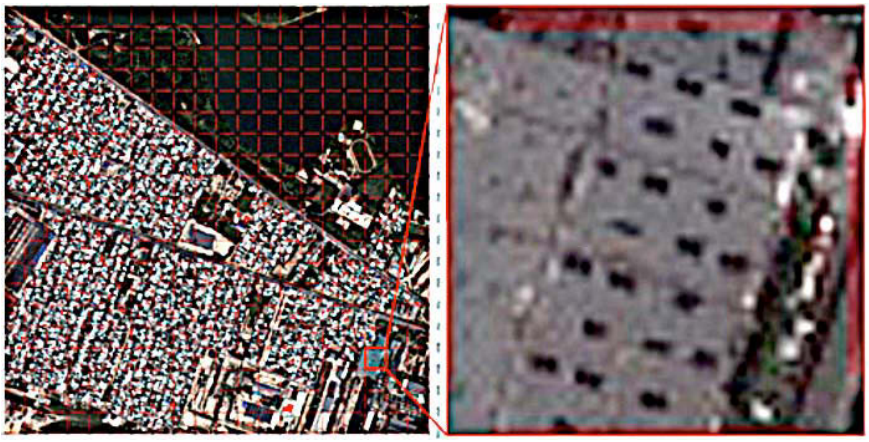


Figure 1.3 – Satellite image, target class

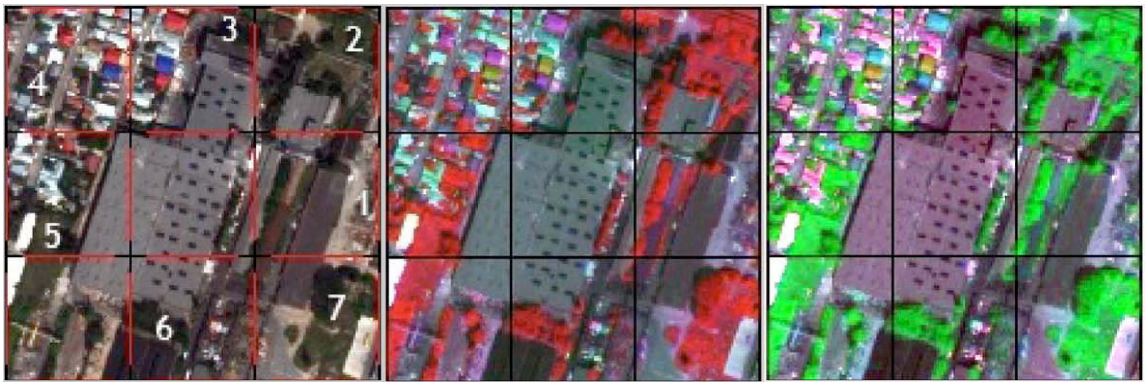


Figure 1.4 – Natural colour display, bands NIR-2, Coastal and Red in the R,G,B channels and bands Coastal, NIR-2 and Red in the R,G,B channels

1.4 Outline of the Dissertation

The thesis presents novel theoretical concepts and methods that help users access and discover latent information in large image archives, visualize, analyze and interpret satellite scenes in an interactive workflow. Chapter 1 has given a short introduction to the new fields of CBIR and IIM in the context of EO, with emphasis on the motivation for developing a new system. The two main contributions of this dissertation are defined and exemplified. Chapter 2 presents an extended review of the state-of-the-art of CBIR in multimedia, medical, forensics and remote sensing. An overview of existent query systems is provided. Chapter 3 gives an outline of current remote sensing sensors and technologies available, explains the spectral gap and presents the future trends in remote sensing for EO. Chapter 4 studies the theoretical background in stochastic modelling, Bayesian statistics, information theory and parameter estimation. Chapter 5 describes the module for latent information discovery in satellite images and provides a solution for bridging the semantic gap, which is regarded as one of the most important problems in the field of CBIR. Chapter 6 presents the module for enhancing visualization of objects and classes of interest in multispectral satellite images. Because the final decision about the information derived from satellite images is always made by an operator, the visualization system models the automatic response of the human visual system to external stimuli (i.e. the image) and optimizes the combination of the spectral bands mapped to the three channels of the display.

2

Information Spaces in Earth Observation Data

Our planet is going through unprecedented environmental changes and our society will have to face ever increasing natural and man-made hazards. To forecast and address the effects of these diverse challenges, new generations of space borne sensors are being designed to monitor and measure phenomena about the Earth. State-of-the-art imaging sensors with advanced recording capabilities have been recently launched or are scheduled for the next few years (e.g. ESA Sentinels, SPOT 6,7, Pleiades). These new data give birth to new applications, new questions and new solutions that need to converge in a common space of user understanding: information described by semantic concepts.

This chapter provides an overview of the current remote sensing technologies for Earth Observation (EO) operating in the visible and infrared part of the electromagnetic spectrum, with emphasis on methods used for transforming data into knowledge. The chapter concludes by emphasizing the need of user communities for standardized, up-to-date, semantic-based systems. The advanced capabilities of satellite sensors become useful only when data are translated into accurate, focused information, ready to be used by decision makers. Several approaches used to generate semantic dimensions in remote sensing applications are analyzed. While the spectrum of requirements has expanded, the technical capacity to keep up with these requirements has yet to reach the standards of quality desired by users from various fields.

2.1 Introduction

Remote sensing is the art and science of collecting information about an object or geographic area from a distant vantage point using remote sensing instruments. Data collection about the Earth was initially performed using cameras mounted on suborbital aircrafts. The first comprehensive definition was adopted in 1997 in [169] and stated that 'photogrammetry and remote sensing are the art, science and technology of obtaining reliable information about physical objects and the environment, through the process of recording, measuring and interpreting imagery and digital representations of energy patterns derived from non-contact

sensor systems'. A modern definition of remote sensing should include also the management and dissemination of information derived from data in an user-centered, easy to understand way. Remote sensing data become useful information only by addressing the needs of users on the ground. For this reason, remote sensing functions in harmony with other complementary related sciences including cartography, surveying and geographic information systems (GIS). Dahlberg and Jensen [170] and Fisher and Lindenberg [171] suggested a model of interaction between remote sensing, cartography, surveying and GIS. All are recognized as having unique yet overlapping areas of knowledge as shown in figure 2.1. It is important to observe that, when many scientific domains partially overlap, the only way to connect them is by using a language that is universally understood: i.e. semantic concepts. Although different sciences use different models to describe different phenomena, the sharing of knowledge is performed only via semantics, i.e. natural language. It is the only common ground that scientists and decision makers share.

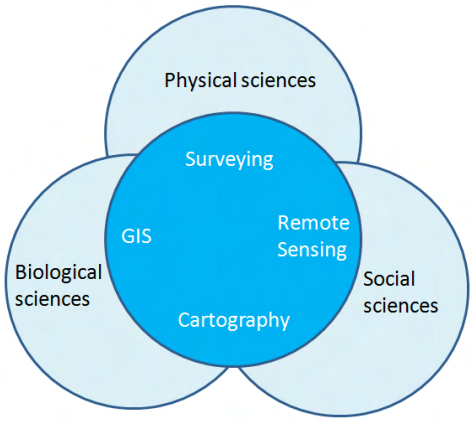


Figure 2.1 - Interaction model depicting the relationship between remote sensing, GIS, surveying, cartography as they relate to biological, social and physical sciences (adapted from [179])

The process of analysis and interpretation of remote sensing data employs not only the scientific knowledge but all the background of the analyst, all information obtained throughout his or her lifetime. The synergism of combining scientific knowledge with real-world experience allows the person to develop heuristic rules to extract information from the images. The information can be disseminated among interested parties only via visual representations and semantic concepts that describe in linguistic terms the findings and the results.

From Photons to Electrons to Neurons

Airborne or space borne sensors record very specific information about an object (e.g. size of the crown of a tree) or the geographic extent of a phenomenon (e.g. forest area). The electromagnetic energy emitted or reflected from a target is transformed into an electrical signal (photons to electrons) by the sensor. This recording is used to discover and analyze the actual property under investigation. The semantic gap requires a method to translate these discoveries from instrument-language (i.e. numbers) into human language (i.e. semantic concepts) and thus turn data into knowledge (electrons to neurons). The most reliable approach to infer information from the data is thorugh visual analysis or manual

investigations. However, remote sensing data most of the times are both visual and non-visual, covering a wide range of the electromagnetic spectrum, including infrared, shortwave infrared and even microwave. The spectral gap represents the loss of information between the number of spectral bands in which data are recorded and the only three channels of the display that can accommodate only three features at any given time. The spectral gap requires a method to select from the large number of attributes only the ones containing the highest amount of relevant information with respect to the application and display them using only the three channels of the computer screen.

Remote sensing technology is unobtrusive if the sensor records the energy emitted or reflected by a source on the ground. Passive remote sensing does not interfere and does not disturb the phenomenon of interest. Nowadays, remote sensing has become critical to the successful modelling of numerous natural and man-driven processes [172].

2.1.1 Data collection & analysis procedures

The data collection and analysis procedures used for remote sensing applications involve the following steps [179]:

1. The hypothesis to be tested is defined using a specific type logic - inductive or deductive - and an appropriate processing model (e.g. deterministic, stochastic)

2. In-situ and complementary data necessary to calibrate the remote sensing data and to evaluate its radiometric, geometric and thematic characteristics are collected.

3. Data are collected passively or actively using remote sensing sensors, ideally at the same time with the ground reference data.

4. Remote sensing data are processed either using image processing techniques, modelling and n-dimensional visualization

5. Metadata, processing workflow and accuracy of information are provided and results are communicated using GIS, maps, statistical tables, semantic labelling, etc.

2.1.2 Data collection

Remotely sensed data are collected using passive or active systems. Passive sensors record electromagnetic radiation that is reflected or emitted from the terrain. Active systems (e.g. RADAR, LIDAR, SONAR) are the sources of electromagnetic energy and the recorders of the amount of radiant flux scattered back from the terrain.

The amount of electromagnetic radiance L (watts per meter squared per steradian), recorded within the instantaneous field of view (IFOV) of a passive, optical remote sensing system is defined as:

$$L = f(\lambda, s_{x,y,z}, t, \theta, P, \Omega) \tag{2.1}$$

- λ = wavelength – the spectral response measured in various frequency channels
- $s_{x,y,z}$ = x, y, z location of the pixel and its size (x, y)
- t = temporal information, date and time of acquisition
- θ = the set of angles that describe the geometric relationships between the radiation source, the terrain target and the remote sensing system
- P = the polarization of back-scattered energy recorded by the sensor
- Ω = the radiometric resolution at which the data (reflected, emitted, back scattered radiation) are recorded by the remote sensing system.

2.2 Resolution

2.2.1 Spectral Resolution

Remote sensing investigations are based on developing a deterministic model between the amount of electromagnetic energy reflected, emitted or back-scattered in specific bands and the physical characteristics of the phenomenon under investigation. Spectral resolution represents the number and width of specific wavelength intervals (bands) in the part of the electromagnetic spectrum in which the sensor is sensitive.

Multispectral remote sensing systems record energy in multiple bands or channels. Figure 2.2 shows the wavelength intervals (nominal spectral resolution) for the spectral channels of the Landsat ETM+ sensor. In practice it is difficult to create a detector that has extremely sharp band-pass boundaries. A more precise method of stating bandwidth is to look at the typical Gaussian shape of the detector sensitivity - figure 2.3.

Index	Spectral Band	Spectral Range (μm)	Spatial Resolution (m)
1	Blue	0.45-0.52	30
2	Green	0.52-0.60	30
3	Red	0.63-0.69	30
4	Near Infrared NIR	0.76-0.90	30
5	Mid-Infrared-1 MIR-1	1.55-1.75	30
6	Thermal Infrared TIR	10.4-12.5	120
7	Mid-Infrared-2 MIR-2	2.08-2.35	30

Figure 2.2 – Band intervals for Landsat ETM+ sensor

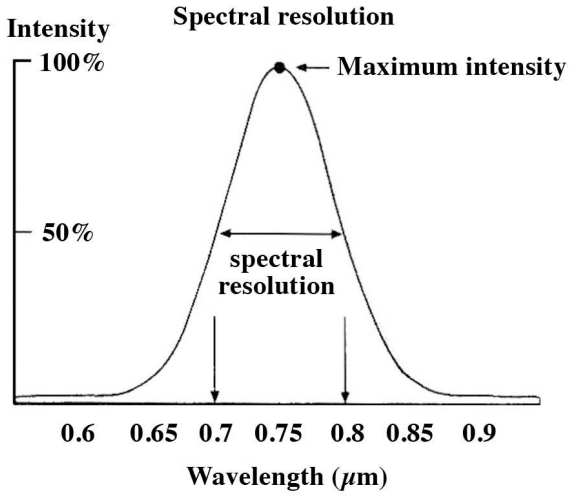


Figure 2.3 – Detector sensitivity

2.2.2 Spatial Resolution

There is a general relationship between the size of an object or an area to be identified and the spatial resolution of the system. The spatial resolution is a measure of the smallest angular or linear separation between two objects that can be resolved by the remote sensing system. A sensor's nominal spatial resolution is defined as the dimension (in meters) of the ground-projected field of view, where the diameter of the circle on the ground is a function of the instantaneous-field-of-view and the altitude of the sensor vehicle. The smaller the nominal spatial resolution, the greater the spatial resolving power of the system. Figure 2.4 depicts an aerial photograph at various spatial resolutions: 10 cm / pixel, 50 cm / pixel, 1 m / pixel, 10 m / pixel.

Figure 2.4 - aerial photograph at various spatial resolutions: 10 cm / pixel, 50 cm / pixel, 1 m / pixel, 10 m / pixel.

2.2.3 Temporal Resolution

Remote sensing systems can record data about the same landscape through time. The information derived is then used for change analysis and prediction. The temporal resolution of a satellite refers to the time period between acquisitions of the same phenomenon of interest. Trade-offs are usually made in association with the various resolutions. Generally, the higher the temporal resolution requirement, the lower the spatial resolution requirement (e.g. weather satellites) and the higher the spatial resolution, the lower the temporal resolution.

2.2.4 Radiometric Resolution

Radiometric resolution is defined as the sensitivity of a remote sensing detector to the differences in signal strength as it records the radiant flux reflected, emitted or back-scattered from the terrain. It defines the number of just discriminable signal levels. Radiometric resolution is sometimes referred to as levels of quantization: 256 levels for 8-bit resolution, 2048 levels for 11-bit resolution and 4096 for 12-bit resolution.

Remote sensing data are usually stored as a matrix of numbers. Each digital value is located at a specific row and column in the matrix. Each pixel is defined as a 2-dimensional picture element that is the smallest non-divisible element of a digital image. Each pixel has an original raw brightness value expressed as a digital number (DN). All bands in the satellite image are geometrically registered to each other.

2.2.5 Angular Information

Remote sensing instruments record very specific angular characteristics associated with each pixel. The angular characteristics are a function of:

- The location in the 3-dimensional sphere of the illumination source (e.g. Sun)
- The orientation of the terrain or phenomenon under investigation
- The location of the suborbital or orbital remote sensing system and its associated azimuth and zenith angles.

There are two angles involved in determining the angular information: the angle of incidence associated with the incoming energy that illuminates the terrain and the angle of exitance from the terrain to the sensor system. The bidirectional nature of data collection influences the spectral and polarization characteristics of the at-sensor radiance L. Angular information is the basis of photogrammetric applications. Stereoscopic image analysis is based on the concept that the same object is sensed from two angles. Viewing the terrain from two different points introduces stereoscopic parallax, which is the foundation of photogrammetry and radargrammetry.

2.3 Data Analysis Methods

A basic remote sensing model has three components: (1) a scene model that specifies the form and nature of the energy and matter within the scene and their spatial and temporal order, (2) an atmospheric model that describes the interaction between the atmosphere and the energy entering and being emitted from the scene and (3) a sensor model that describes the behaviour of the sensor in response to incident energy fluxes.

Significant advances have been made in digital image processing with focus on scientific visualization, modelling and hypothesis testing [172-174]. The reader can refer to [175-178] for extended summaries and reviews of the image processing methods available. Data processing techniques include image pre-processing (radiometric and geometric corrections), image enhancement, pattern recognition, photogrammetric image processing of stereoscopic

imagery, decision-tree and neural network image analysis, multispectral and hyperspectral data analysis, data fusion and change detection.

Parametric classification methods applied on remote sensing data can be either *hard*, with with discrete, mutually exclusive classes or *fuzzy*, case in which the belonging of a pixel to a specific class is defined in terms of probabilities.

Non-parametric clustering algorithms are dependent on how the seed training data are extracted. It is often difficult to label the clusters and create information maps described by semantic concepts. Artificial neural networks have been introduced to solve the problem of labelling. Their drawback is that sometimes it is difficult to understand how the method came up with a certain conclusion because the information is locked within the weights of the hidden layers.

The Multi-Concept

The multi-concept was standardized in the 1960s by Colwell. The most useful and accurate method of scientific image interpretation includes the following types of analysis: multispectral, multidisciplinary, multiscale, multipolarization, multiresolution and multitemporal. Measurements made in multiple discrete wavelength regions of the electromagnetic spectrum (multispectral) are usually more valuable than acquiring a single panchromatic image. Multiscale and multiresolution images taken of an area are very useful for analysis and interpretation. Smaller-scale imagery is useful for placing intermediate scale imagery in its proper regional context. The very large-scale imagery can be used to provide detailed information about local phenomena. Ground reference is the largest scale and these data are very important to calibrate and to verify remote sensing-derived information.

Image analysts should work together with multidisciplinary experts when focusing on a remote sensing analysis or information extraction problem. This approach often yields synergistic and interesting results as multidisciplinary scientists share their expertise. While single date remote sensing investigations can generate valuable information, usually they can't give information on the changes and processes at work. A multitemporal investigation obtains more than one image of an object and allows the understanding of processes and the development of predictive models.

A new instrument that can be added to the multi-concept analysis of satellite imagery is multi-label. When experts analyze the same image data looking for different phenomena, they will label the information according to their own analysis. Although attached on the same image, these labels will vary with respect to the user and phenomena of interest.

2.4 Remote sensing satellite systems

Remote sensing systems record reflected or emitted energy from an object or area of interest on the ground. Multispectral systems record energy in multiple bands, hyperspectral sensors record energy in hundreds of bands and ultraspectral systems in thousands of bands.

Figure 2.5 gives an overview on how digital remote sensing data are turned into information. Initially, the remote sensing sensor detects electromagnetic energy exiting from the phenomena of interest on the ground and passes through the atmosphere. The energy is recorded as an analogue electrical signal, which is converted to a digital value through an A/D converter. If an aircraft platform is used, the data are simply returned to Earth. If a spacecraft platform is used, the digital data are telemetered to Earth receiving stations directly or indirectly via tracking and data relay satellites. Radiometric and geometric corrections are usually necessary to enhance visualization and interpretability of the data. Biophysical and land cover information extracted using visual or computer assisted processing is distributed to end users. This information should be labeled and explained in semantic concepts to ensure a wide availability and understanding throughout various user communities.

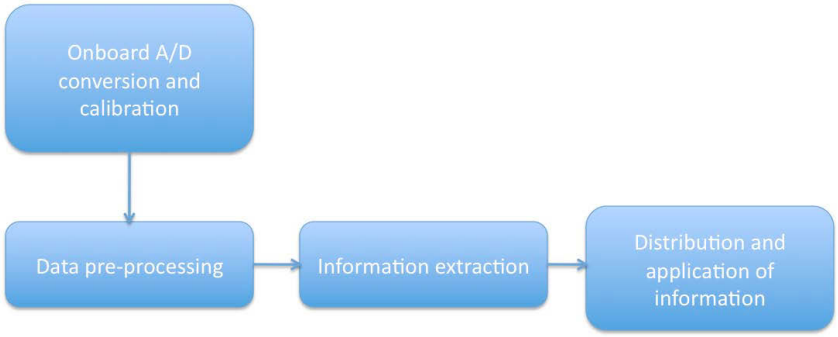


Figure 2.5 - Workflow for transforming remote sensing data into information. The data recorded by the sensor are converted from an analogue electrical signal to digital values and then calibrated. Ground pre-processing removes geometric and radiometric distortions. This may involve the use of ephemeris or ancillary data such as map coordinates, digital elevation model, etc. Future sensors will pre-process the information onboard.

There are six types of remote sensing systems used for multispectral and hyperspectral data collection: (1) traditional aerial photography, (2) multispectral imaging using a scanning mirror and discrete detectors, (3) multispectral imaging with linear arrays, (4) imaging with a scanning mirror and linear arrays, (5) imaging spectrometry using linear and area arrays, (6) digital frame camera aerial photography based on area arrays [179]. There are a lot of multispectral remote sensing systems available and it is beyond the purpose of this dissertation to provide details about them.

2.5 Spectral dimensions in satellite images

The first remote sensing mission, Landsat-1 was launched in July 1972 with the clear objective to study and monitor the land cover of our planet. Its main instrument was the MSS multispectral scanner and it recorded data in four spectral bands - green, red and two infrared bands, with a spatial resolution of 60 meters / pixel. The energy reflected by all objects within a 60 X 60 meters area was integrated into a single cell of the sensor, making the satellite useful only for large-scale observations. Figure 2.6 shows one of the first satellite images in history - Landsat-1 image of the Amazon forest in Brazil in 1972.

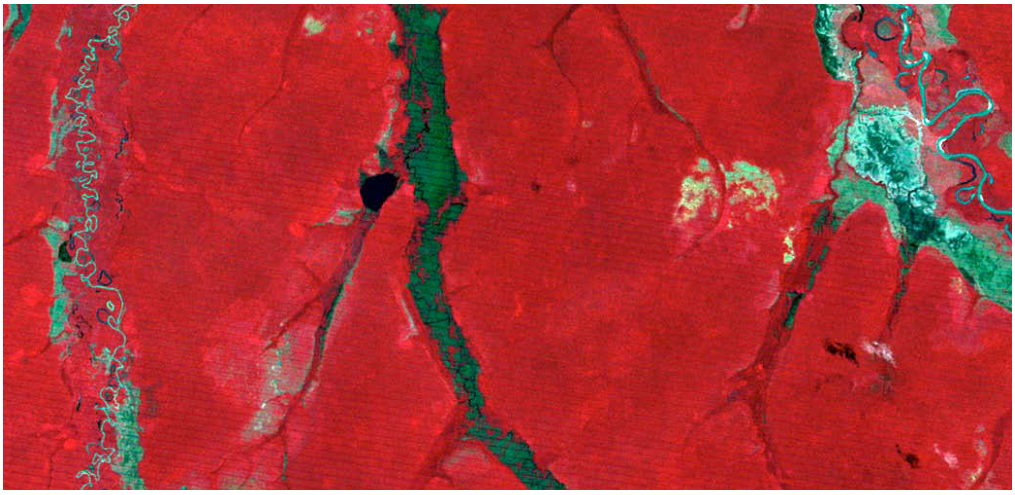


Figure 2.6 - Landsat-1 MSS image from 1972 depicts in false colours (spectral bands 4-2-1) the Amazon forest in Brazil. With a spatial resolution of 60 meters / pixel, land monitoring is possible only at global and regional scales (Credits: United States Geological Survey -USGS)

Landsat 4 satellite carried a sensor with higher spectral and spatial resolutions than its previous versions. The space vehicle flew the thematic mapper TM sensor that recorded information about the Earth in seven spectral bands: blue, green, red, near-infrared, two middle infrared at a spatial resolution of 30 meters/pixel and a thermal band at 120 meters/pixel. Landsat-7 operates the Enhanced Thematic Mapper ETM+ sensor, with eight spectral channels and spectral resolution ranging from 15 meters/pixel (panchromatic) to 30 meters/pixel (multispectral) and 60 meters/ pixel (thermal). The Landsat TM bands were selected after years of analysis for their value in water penetration, discrimination of vegetation type and vigour, plant and soil moisture, differentiation of clouds, snow and ice and identification of hydrothermal alteration in certain rock types. Landsat TM is twice as effective as the Landsat MSS based on its ability to provide twice as many separable classes over a given area [180]. Figure 2.7 shows the spectral bands of a Landsat ETM+ image taken over Bucharest, Romania.

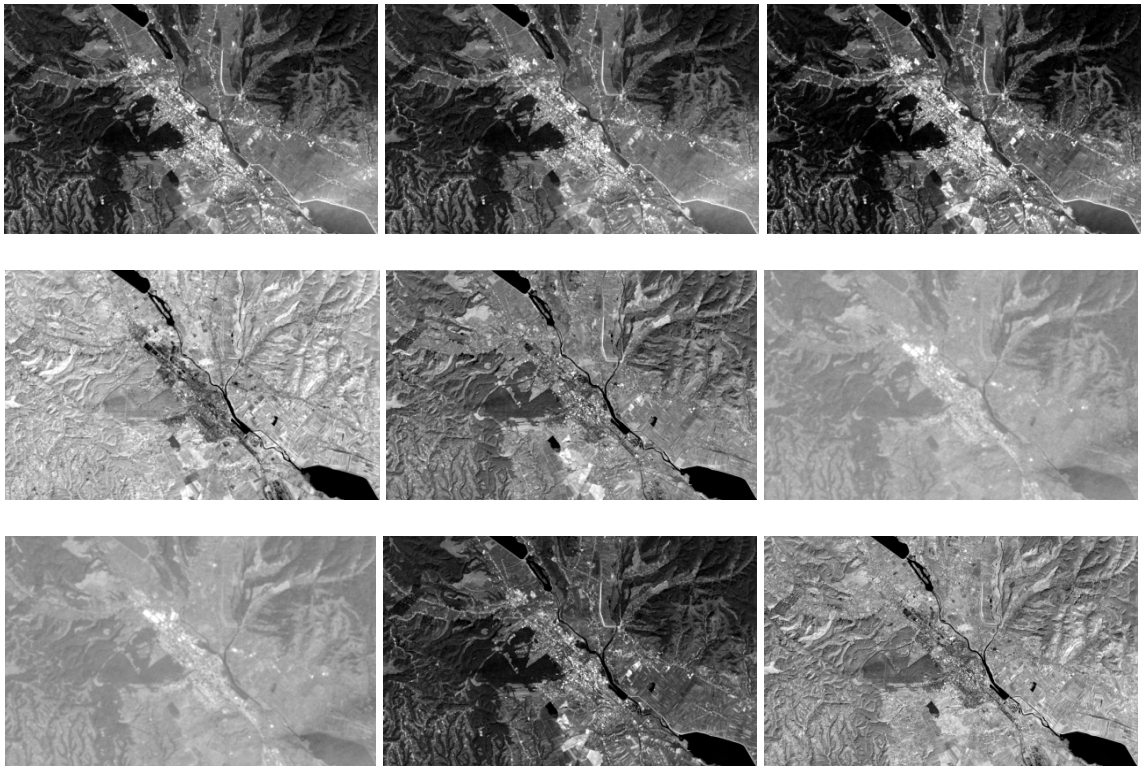


Figure 2.7 - Landsat image of Bucharest, Romania (Image credits: USGS)
From left to right, top to bottom: spectral bands 1, 2, 3, 4, 5, 6-1, 6-2, 7, 8

The beginning of remote sensing missions for Earth Observation (1970s) shows an intense focus on multi-band sensors with higher spectral resolution and lower spatial resolution (e.g. Landsat, 8 spectral bands). Several years later (1990s) the spectral resolution of sensors decreased (e.g. Quickbird, IKONOS, 4 spectral bands), and the focus shifted on building sensors with higher spatial resolution. Today, new satellite sensors integrate both high spatial and spectral resolutions (e.g. WorldView-2, 8 spectral bands). European Space Agency's (ESA) Sentinel-2 satellites [185] will routinely deliver high-resolution optical images globally, providing enhanced continuity of SPOT- and Landsat-type data. Sentinel-2 will carry an optical payload with visible, near infrared and shortwave infrared sensors comprising of 13 spectral bands (figure 2.8) - 4 bands at 10m, 6 bands at 20m and 3 bands at 60m spatial resolution. The additional red channels improve monitoring of vegetation and estimation of related parameters.

Besides the standard spectral bands and the generic land use land cover maps, these sensors will also generate geophysical variables such as leaf coverage, leaf chlorophyll content, leaf water content and fractional vegetation cover. Figure 2.9 shows some examples of ESA Sentinel-2 simulated Level-2 geophysical products derived from simulated Level-1 orthorectified products.

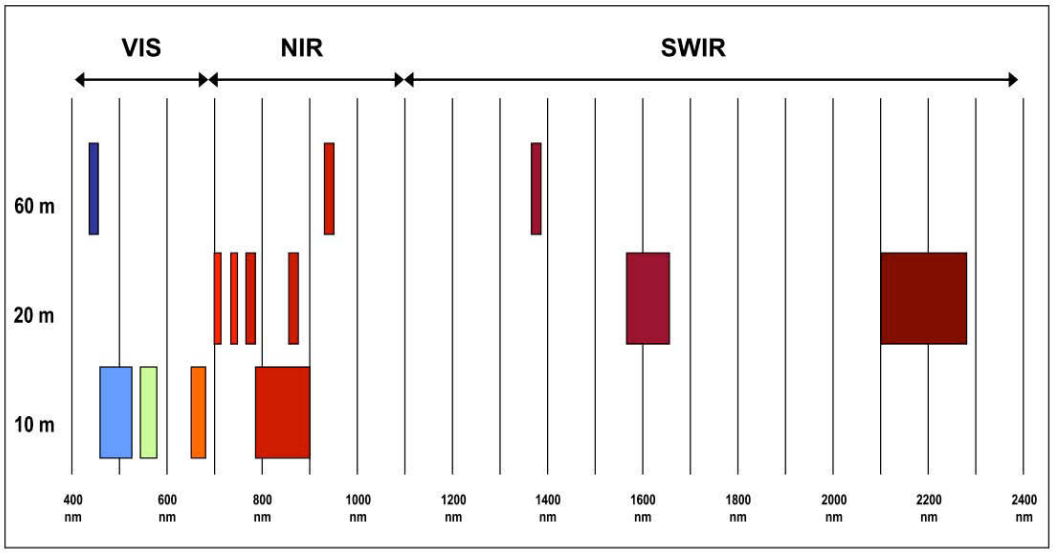


Figure 2.8 - Sentinel-2 sensor works in 13 spectral intervals. The wavelengths are represented on the horizontal and the spatial resolution displayed on the left
(Image Credits: European Space agency - ESA)

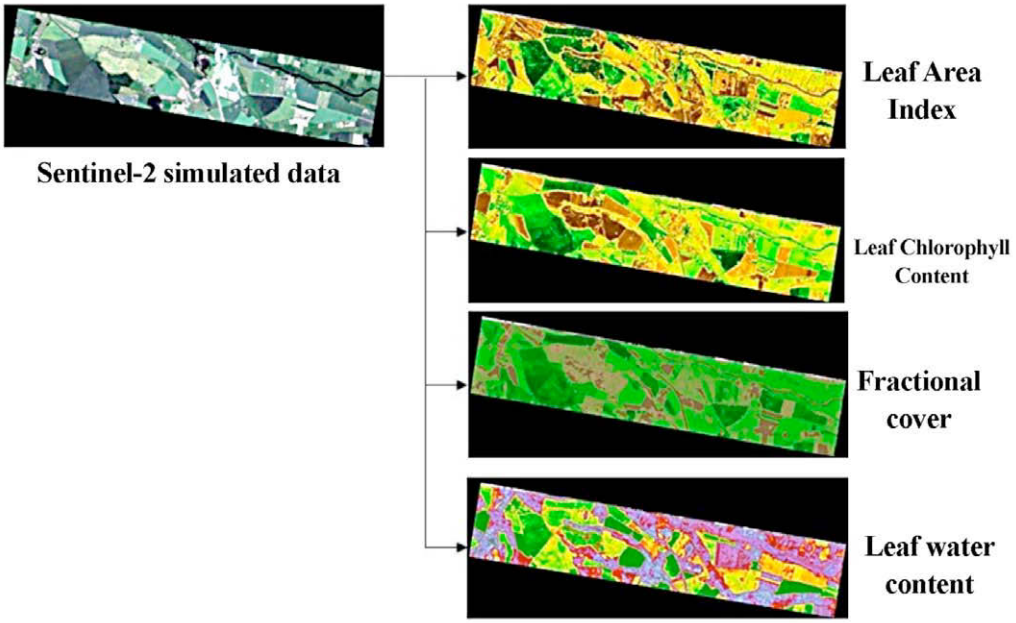


Figure 2.9 - Sentinel-2 simulated level-1 data and level-2 geophysical parameters
(Image Credits: European Space agency - ESA)

The spectral gap, i.e. the gap between the information recorded by a satellite sensor in multiple spectral bands and the information displayed by a computer screen, i.e. always three channels, is different for every sensor. Landsat required many years of studies and experiments to determine what spectral features optimize visualization for specific applications. With the launch of new satellites (e.g. European Space Agency ESA Sentinel-2, WorldView-3), this question will become more difficult to address due to increased spectral resolutions (13, 16 spectral features) and increased number of applications. The spectral gap will remain an unresolved problem, always challenging users to reduce the information available to a limited number of channels for visual display.

2.6 Semantic dimensions in satellite image-based cartography

Earth Observation is a modern science that studies the Earth's changing environment, using remote sensing instruments, space borne and airborne sensors. The study of the Earth's physical components, such as the atmosphere, oceans and land aims at understanding the natural processes as well as the effects of man's actions on the environment. This understanding can be put into practice only by creating optimum communication channels between the multiple parties involved in the decision making processes. Communication between remote sensing experts and users has to be done using a language all parties understand - semantic labels. The need to extract semantic information directly from satellite data has been addressed in several projects, as described in the following paragraphs.

2.6.1 Corine Land Cover

Corine Land Cover (CLC) is the European standard for cartography of land cover and land use, aiming at providing an inventory of Earth's surface features for managing the environment. If the environment and its natural resources are to be properly managed, decision and policy makers have to have an overview of the existing information which is as complete and up-to-date as possible. The CORINE project (Coordination of information on the environment) [187] aimed at providing a comprehensive database with relevant data useful for understanding different features of the environment (e.g. the state of individual environments, the geographical distribution and state of natural areas, of wild fauna and flora, the quality of water resources, land cover structures. As expressed in [187], for environmental purposes the land cover information has to meet special requirements: it must be cartographic as well as statistical and it must reproduce the information at different scales in order to be useful at multiple levels of decision making.

In any land cover cartographic inventory, four elements are linked - the scale (the surface area of the smallest unit to be mapped), the nature of the basic information used (EO satellite data), the structure of the nomenclature and the number of items it contains. On the basis of the first three elements and the provisional nomenclature used for the feasibility study, the land cover team has formulated the definitive nomenclature for the project. Figure 2.10 shows the schematic construction of the land cover nomenclature in CORINE. A key point of the nomenclature is the fact that the heading terminology must be unambiguous and vague terms are avoided.

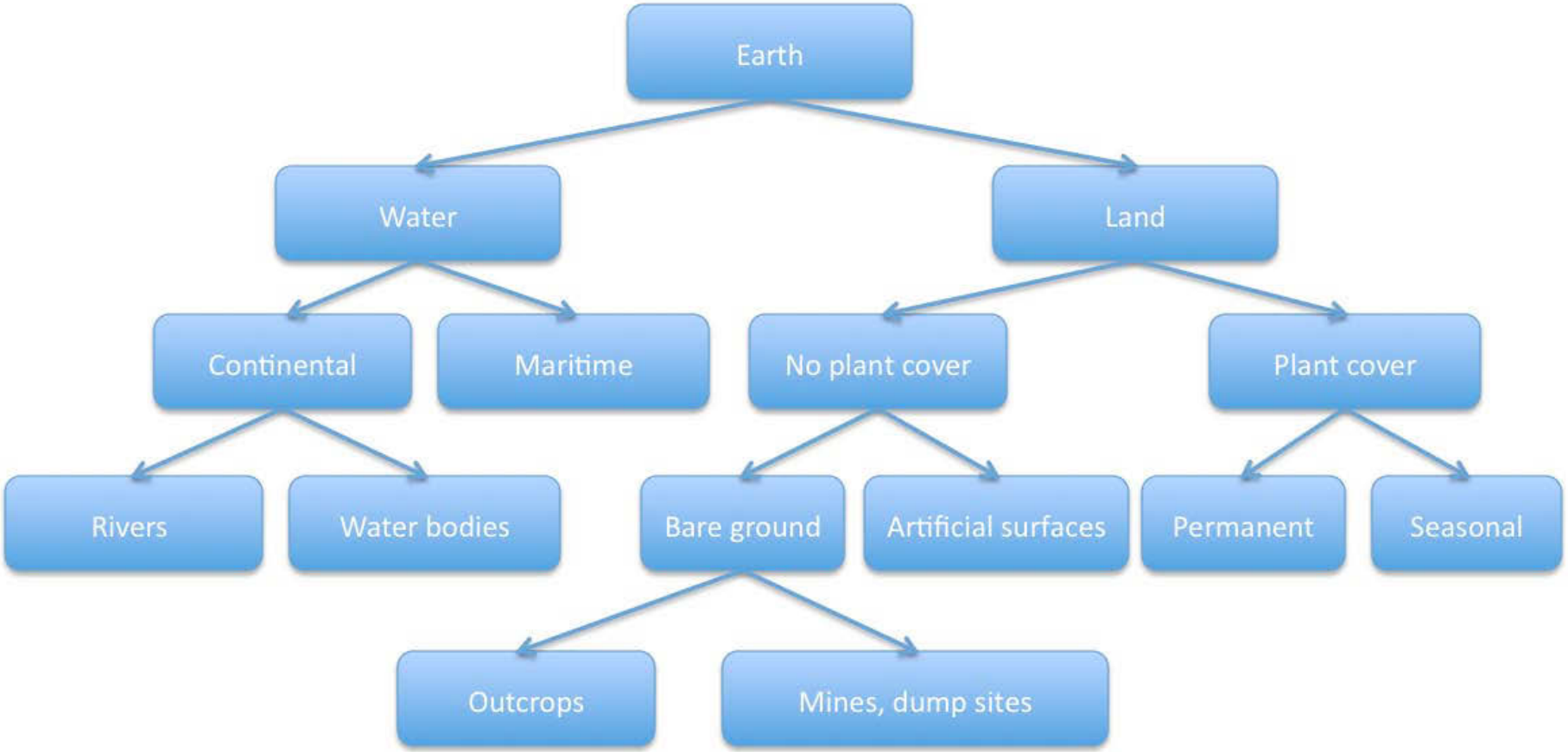


Figure 2.10 - Theoretical schematic construction of a land cover nomenclature

Level 1	Level 2	Level 3
1. Artificial surfaces	1.1. Urban fabric	1.1.1. Continuous urban fabric 1.1.2. Discontinuous urban fabric
	1.2. Industrial, commercial and transport units	1.2.1. Industrial or commercial units 1.2.2. Road and rail networks and associated land 1.2.3. Port areas 1.2.4. Airports
	1.3. Mine, dump and construction sites	1.3.1. Mineral extraction sites 1.3.2. Dump sites 1.3.3. Construction sites
	1.4. Artificial non-agricultural vegetated areas	1.4.1. Green urban areas 1.4.2. Sport and leisure facilities
2. Agricultural areas	2.1.Arable land	2.1.1. Non-irrigated arable land 2.1.2. Permanently irrigated land 2.1.3. Rice fields
	2.2. Permanent crops	2.2.1. Vineyards 2.2.2. Fruit trees and berry plantations 2.2.3. Olive groves
	2.3. Pastures	2.3.1. Pastures
	2.4. Heterogeneous agricultural areas	2.4.1. Annual crops associated with permanent crops 2.4.2. Complex cultivation 2.4.3. Land principally occupied by agriculture, with significant areas of natural vegetation 2.4.4. Agro-forestry areas
3. Forests and semi-natural areas	3.1. Forests	3.1.1. Broad-leaved forest 3.1.2. Coniferous forest 3.1.3. Mixed forest
	3.2. Shrub and/or herbaceous vegetation association	3.2.1. Natural grassland 3.2.2. Moors and heathland 3.2.3. Sclerophyllous vegetation 3.2.4. Transitional woodland shrub
	3.3. Open spaces with little or no vegetation	3.3.1. Beaches, dunes, and sand plains 3.3.2. Bare rock 3.3.3. Sparsely vegetated areas 3.3.4. Burnt areas 3.3.5. Glaciers and perpetual snow
4. Wetlands	4.1. inland wetlands	4.1.1. Inland marshes 4.1.2.Peatbogs
	4.2. Coastal wetlands	4.2.1. Salt marshes 4.2.2. Salines 4.2.3. Intertidal flats

Figure 2.11 - CORINE LAND COVER Nomenclature
(Credits: European Environment Agency - EEA)

2.6.2 URBAN ATLAS

The understanding of urban dynamics is the basis for planning the sustainable development of urban areas and the conservation of Earth's resources. However, monitoring and understanding the urban dynamics are the most complex tasks city planners have to deal with. The complexity and variety of the different urban components and functions as well as of the interactions between them are even more pronounced when available mapping is outdated, of low quality and where standard information is not available.

The Murbandy (Monitoring Urban Dynamics) project was initially launched with the purpose of monitoring and measuring the extent of urban areas and their progress towards sustainability through developing land use databases for various cities. Murbandy has been extended to Moland (Monitoring Land Use Changes) [186], a comprehensive study that uses different layers of information, combined with land use changes for urban areas. The methodology was based on developing an accurate land use database for urban areas using data derived from satellite images and aerial photography. The database provides the starting point in combining environmental, economic and social data to understand the dynamics and characteristics of urban growth and related parameters.

The methodology implemented to understand and map the dynamics of urban areas in Europe consisted of three interrelated parts: (1) Change detection - measuring changes in the spatial extent of urban areas and in urban structures over a period of 40-50 years; (2) Understanding, identifying and testing a number of indicators to be used to assess the sustainability of urban areas; (3) Forecast - developing urban growth scenarios for the areas under surveillance using state-of-the-art urban dynamics models [186].

The accuracy of the coverage refers to a map scale of 1:25.000, with the minimum-mapping-unit of 1 hectare for the artificial surfaces and 3 hectares for non-artificial surfaces. In order to follow the standard European land cover classification standards, the nomenclature for land use land cover structures is based on an extended version of the CLC 2000 legend, described in the previous section.

The CLC nomenclature is not detailed enough for the goal of Urban Atlas. The scale chosen for the CLC project is 1:100, with the minimum mapping unit 25 ha, while the scale for Urban Atlas is 1:25, with a minimum unit of 1 ha. A feature having a certain attribute in CLC might have a different one in Urban Atlas because the unit is smaller and the scale more detailed. An example of Urban Atlas mapping of Copenhagen and the related level of semantic abstraction is presented in figure 2.12.

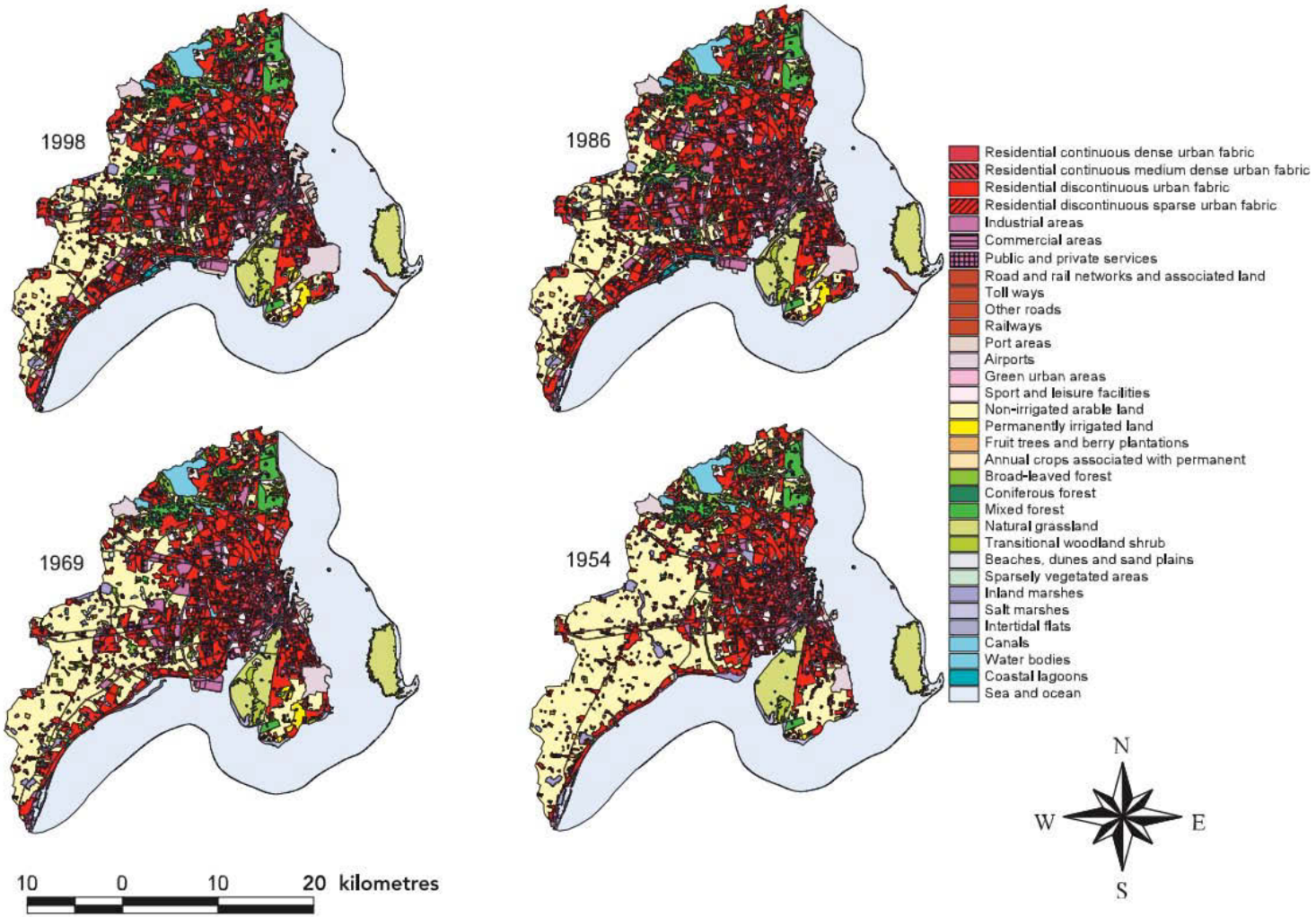


Figure 2.12 - Urban Atlas mapping of the city of Copenhagen, Denmark showing the land use evolution of the study area during four different periods. (Credits: European Environment Agency - EEA)

All datasets in Urban Atlas are used to analyze and understand different European environments at different levels of complexity and decision-making. The semantic concepts attached to the urban maps link map producers, scientists and decision makers through a joint vocabulary. The common understanding between the parties that use the data is ensured via the semantic labels attached to the information layers.

2.6.3 Rapid Mapping Applications

Every year, fires, floods, earthquakes and volcanic eruptions, landslides and other humanitarian crises claim the lives of thousands of citizens in Europe and around the world. In the framework of the GMES program [188], the GMES Emergency Response Service [189] reinforces the European capacity to respond to emergency situations. This service provides a reactive cartographic service to users involved in the management of humanitarian crises, natural disasters and man-made emergency situations. The service aims at developing and publishing timely and high-quality products derived from EO data, showing the extent and impact of the event. The post-disaster satellite data are used to assess and monitor the ongoing crisis situation, i.e. delineate the affected areas and estimate the damages caused by the disaster.

After the satellite data has been downlinked and received, the image is geocoded, rectified, and fused with other data sets. Subsequently, various algorithms and processing chains tailored to the type of event are employed to extract the requested information. The results are integrated into map products or other formats. Additional interpretation texts, legends and overview maps are generated and incorporated into the final map product to enhance understanding among users. Figure 2.13 shows an example of a flood map in Romania, 2010.

While old maps used graphical symbols to represent land cover categories, the legends of new maps created for emergency situations are represented using image patches directly clipped from the satellite image (figure 2.14). Each image patch has a semantic concept attached by the analyst to describe the phenomenon depicted and to bridge understanding with and between users.

2.6.4 EO-based applications

Several applications based on remote sensing data rely on detailed analyses and visual interpretation of satellite images. Similar to previous mapping scenarios, the maps derived from satellite images through visual investigations contain detailed explanations using semantic concepts directly attached to the phenomena of interest. Figure 2.15 depicts a chemical spill in Hungary that occurred in 2010 and severely impacted the environment in neighbouring countries. The map depicts the satellite image as a background, annotated with labels, explanations and semantic concepts directly overlaid on the data. The semantic concepts used by the image analysts to label the data are the only way to ensure understanding and usability of this product among multiple users and decision makers.

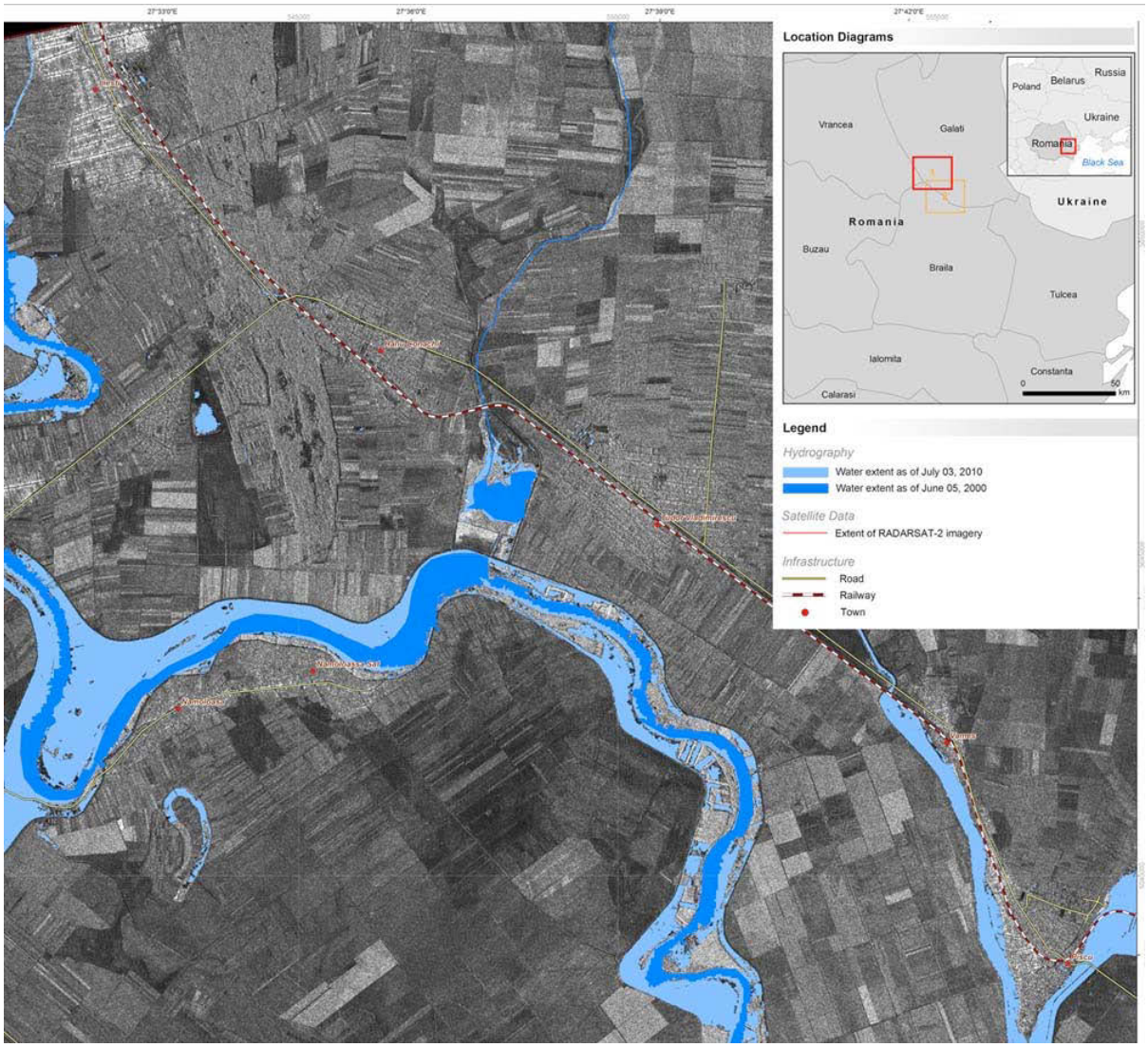


Figure 2.13 - Flood map from Romania (Credits: German Aerospace Agency DLR)

Figure 2.14 - Legend using image patches and semantic concepts

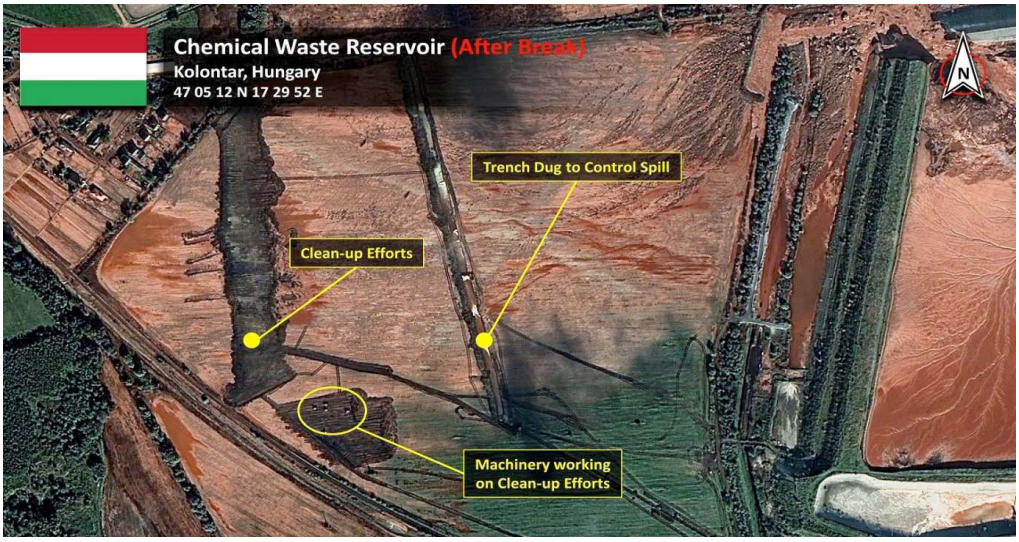


Figure 2.15 - Chemical spill in Hungary depicted in a Quickbird satellite image with semantic explanations directly attached to the phenomenon of interest (Image credits: DigitalGlobe)

Conclusions

The sensor capabilities for recording and processing are constantly improving and new instruments provide more data than users can process and understand in an operational, timely manner. According to the European Space Agency, less than 5% of the data available are actually applied in real-world scenarios.

Although several automatic tools for data processing (e.g. classification, segmentation algorithms) have been developed, these methods still lack the human dimension - the semantic connection with the user. While researchers focus on the inputs (i.e. algorithms), the user communities rely solely on the outputs (i.e. knowledge ready to be applied). For this reason, systems have been developed to bridge the gap between the satellite data and the users' needs and requirements (e.g. CLC, Urban Atlas). These approaches emphasize the need of EO communities for standardized, updated systems able to describe the satellite data using user-friendly latent semantic concepts. Although there are several methods that provide an inventory of the environment, all of them rely on extended, expensive manual-analyses that are subject to developer-bias and can become obsolete before they reach the final users because of the amount of time required to be produced. While requirements have increased, the technical capacity to keep up with these requirements has yet to reach the desired standards of quality.

Next chapter gives an in-depth review of the current state-of-the-art systems developed to link the gap of understanding between machine and human languages. The chapter presents a brief history of the domains of Image Information Mining and Content-based Image Retrieval and concludes with the proposal for a new system concept that addresses two key problems that have been identified to be missing solutions.

3

Image Information Mining: State-of-the-Art

This chapter explains the concept of Image Information Mining, as it evolved since the beginning of the 21st century when Content-Based Image Retrieval was integrated with Data Mining and Knowledge Database Discovery (KDD). The architecture of a classical IIM system containing several processing modules is detailed - feature extraction, indexing, a communication environment for learning and output evaluation - and upgraded with a module for data visualization. The pre-processing module and its functionality are explained and a detailed state-of-the-art on the current techniques for feature extraction, feature selection and dimensionality reduction is provided. The chapter continues with a state-of-the-art review on CBIR methods with emphasis on the current challenges and opportunities in the field of multimedia signal processing. The review is centered around the methods available to bridge the semantic gap between machine features and human-centered latent concepts. The user seeks similarity in semantics while the database can only provide similarity in image processing results. For this reason, bridging the semantic gap is maybe the most challenging puzzle that researchers are trying to solve.

The chapter concludes with a review on the CBIR systems available for EO applications and with the proposal for a new system that addresses the current challenges: bridging the semantic gap, feature ranking and scientific data visualization.

3.1 Introduction

Due to the large volume of data available, the analysis and extraction of information from satellite images has become a complex and challenging task. In addition, the emerging and increasing requirements of major EO-based applications (e.g. mapping, global monitoring of natural resources, disaster management) and large programmes and initiatives (e.g. GMES, GEO, GEOSS) require new methodologies and tools for information mining and management supported by shared knowledge.

The process of analysis and interpretation performed manually by experts to derive information from satellite images is currently too expensive to be applied systematically on even a small subset of the acquired scenes. This limits the full exploitation of the terabytes of archived and new data. Imaging satellite sensors acquire large volumes of data and statistics show that less than 10% of the available non-commercial (i.e. free) images are being downloaded and less than 5% are processed and used [1]. The issue might become even more challenging in the future since more missions - including constellations - are being planned, with broader sensor variety, higher data rates and increased complexity (e.g. ESA Sentinels, or ESA's third party missions). These problems are common across other fields of interest, such as multimedia, medicine, astronomy and planetary remote sensing.

The exponential increase in volume, details, diversity and complexity as well as the users' demand for simultaneous access to multi-domain data created a requirement for new approaches for image information mining, multi-domain information management, knowledge management and sharing. A few examples of EO archives include the following systems: DLR EOWEB [2], Alexandria Digital Library [3], USGS GLOVIS [4] and Earth Explorer [5]. The DLR EOWEB receives hundreds of gigabytes of data per day and users can retrieve images using meta-information such as acquisition time and date, geographical location and sensor. The Alexandria Digital Library (ADL) allows access to remote sensing imagery through its meta information, providing a distributed searching mechanism for retrieving geospatial referenced data collections. ADL has the capability of searching different types of databases placed at different locations and enables the implementation of web clients as Globetrotter [8] and Gazetteer [9].

The idea to integrate database exploration with image processing techniques emerged at the end of the 1970s and evolved into a new field known as Content Based Image Retrieval (CBIR). Also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR), CBIR is the application of computer vision and image processing techniques to give a solution to the problem of searching for images in large databases. "Content-based" implies that algorithms look for the actual contents of the images rather than the metadata such as keywords, tags and descriptions associated with the image. The term “content” refers to image features, colors, shapes, textures, or any other information that can be derived from the image data. Because meta-information is not always useful for retrieving an image depicting a phenomenon of interest and because keyword-based search may sometimes be inconsistent among users, a system that can filter images based on their content provides better indexing and returns more accurate results [6]. The field of Image Information Mining (IIM) emerged at the beginning of the 21st century when CBIR was integrated with Data Mining and Knowledge Database Discovery (KDD).

Data Mining is the process of discovering new patterns from large data sets involving methods from statistics, artificial intelligence and database management. In contrast to machine learning, the emphasis lies on the discovery of previously unknown patterns as opposed to generalizing known patterns to new data. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data that extracts previously unknown patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining).

At abstract level, the KDD field is concerned with the development of methods and techniques for understanding the data. The basic problem addressed by the KDD process is

mapping low-level data, which are typically too voluminous to process, into more compact forms (e.g. a report), more abstract (e.g. a descriptive approximation or model of the process that generated the data) and more useful (e.g. a predictive model for estimating the values of future cases) [7]. The application of specific data mining methods for pattern discovery and extraction is at the core of the process.

IIM, similar to other forms of data mining such as text-based information mining or semantic web technologies, aims at making semantic image content accessible. The new domain of IIM combines expertise from image processing, database organization, pattern recognition, content-based image retrieval and data mining. Several key points need to be addressed:

- Image processing indicates the understanding of patterns from a single image
- Content-based retrieval discovers images based on their semantic and visual contents
- Spatial data mining denotes the extraction of spatial relationships and patterns from remotely sensed images without any link to a common database.

An IIM system allows users the option to operate large collections of images and access the databases to extract information about patterns hidden in the images. The set of relevant images retrieved by the system is dynamic, subjective and unknown. An IIM system can enable the communication between the operating low-level machine language and users that understand only the high level of semantic abstraction. An IIM system usually has two fundamental modules: a component where image processing and classification algorithms are executed and an interactive component where queries are introduced by the user. Figure 3.1 describes the flow of data in an classical IIM system: the image data are imported into the system and the feature extraction module computes the main image characteristics. These features are indexed in the database. In the second module, the archive is queried by the user using similarity measures between the available features derived from the image and the features in the database.

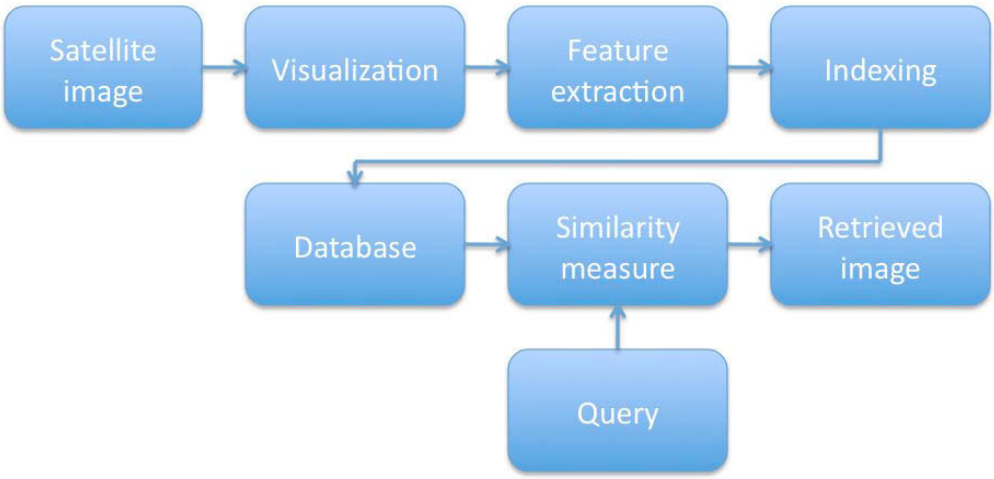


Figure 3.1 - Architecture of a classical Image Information Mining System

3.2 Image Information Mining System Design

The classical design of an IIM system contains several processing modules: visualization, feature extraction, indexing, communication environment for learning and evaluation. In the following paragraphs a short review of these modules is given.

3.2.1 Feature Extraction

Feature extraction involves simplifying the amount of resources required to accurately describe a large dataset. When performing analysis on complex data, one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods to construct combinations of variables to get around these problems while still describing the data with sufficient accuracy. Best results are achieved when an expert constructs a set of application-dependent features. Nevertheless, if no such expert knowledge is available, general dimensionality reduction techniques may help. These include principal components analysis, semi-definite embedding, multifactor dimensionality reduction, multilinear subspace learning, nonlinear dimensional reduction, independent components analysis, etc. In pattern recognition and image processing, feature extraction is a special form of dimensionality reduction. When the input data are too large to be processed and may even be redundant (i.e. much data, but not much information), the dataset will be transformed into a reduced representation of the set of features.

3.2.2 Feature Selection

An image is described by the raw spectral information (e.g. color, radiance, reflectance) and by spatially-derived features at various scales (e.g. texture, shape, morphological features). Each image in the database is represented as a multi-dimensional feature vector - signature.

Color features - Spectral information is maybe the most important feature in image processing and computer vision. Digital imaging systems represent color images using various color spaces and models but the visual information is always displayed using the R, G, B system. This property will become very important in the last part of the dissertation.

Textural features are used to characterize the spatial structure of an image using parametric and non-parametric methods. Textures are complex visual patterns composed of entities with characteristic brightness, color, slope, size, etc and can be regarded as a similarity grouping within local areas of an image [11], [12]. The local sub-pattern properties give rise to the perceived lightness, uniformity, density, roughness, regularity, linearity, frequency, phase, directionality, coarseness, randomness, fineness, smoothness, granulation of the texture as a whole [13]. Materka and Strzelecki [14] give an extended review on the methods of texture analysis and discover four major steps in texture processing:

- Feature extraction: to compute a feature of an image able to numerically describe its texture properties
- Texture discrimination: to partition a textured image into regions, each corresponding to a perceptually homogeneous texture
- Texture classification: to determine to which of a finite number of physically defined classes a homogeneous texture region belongs
- Shape from texture: to reconstruct a 3-D surface geometry from textural information

Texture analysis methods are classified into four major groups [14]:

- structural methods
- statistical methods
- model-based approaches
- transform-based approaches.

In structural approaches, textures are represented by local spatial primitives (micro texture) and a hierarchy of spatial arrangements of the primitives (macro texture) [13,15]. The advantage of structural approaches is that they provide a good symbolic description of the image for synthesis tasks. The drawback is that the abstract descriptions can be ill defined on natural textures because of the variability of lighting and local patterns. Mathematical morphology is another structural approach that has been extensively studied with positive results [16, 17].

Statistical methods describe the texture indirectly by the properties of distributions and relationships between the grey levels of the image. Methods based on second-order statistics have a higher discrimination power than structural and transform-based methods [18]. Textures in grey-level images are discriminated by humans only if their second order moments are different. Niemann [19] showed that because equal first and second order moments but different third order moments require high cognitive effort to discriminate textures, statistics up to second order may be the most important in automatic approaches [20,21]. Methods using multi-dimensional co-occurrence matrices yield better discrimination accuracy than wavelet techniques [22].

Model based analysis methods use stochastic and generative models to interpret image textures [23-28]. The primary drawback of stochastic model based approaches is the computational complexity arising in the estimation of model parameters. Fractal-based models have also been studied for texture interpretation but they lack orientation selectivity and are not suitable for describing local structures [24, 29-31].

Transformation-based methods of texture analysis represent an image in a new space, whose coordinate system is interpreted close to the textural characteristics of the image, e.g. frequency, size. These approaches include the Fourier transform [45], Gabor [46,47] and wavelet transform [48-50]. Fourier transform-based methods show poor results because of the lack of spatial localization. Gabor filters offer better spatial localization but there is usually no single filter resolution at which a spatial structure can be localized. The wavelet transform presents several advantages over Gabor approach [14]: (1) wavelets have varying spatial resolutions that allow representation of textures at different scales; (2) there is a wide range of wavelet functions that can be used for texture analysis.

Geometric features - are usually effective for object detection and analysis. Shape must be invariant to translation, rotation and scale of the image and is characterized either by the boundaries (i.e. the outer contour of the objects) or by the region (i.e. the entire shape of the objects is considered). Several methods for deriving shape information are given in [51,52]. Like any features based on human perception, the major problem in using shapes in CBIR systems is how to describe the shape of an object. Shape representation and description are difficult because one dimension is lost when 3-D real world objects are recorded unto a 2-D image plane. As a result, the shape extracted from the image only partially represents the projected object. Another important problem that emerges is the fact that shape is often corrupted with noise, defects, distortion, occlusion and other artefacts. Searching for images using shape features has raised several questions and brought multiple solutions. Marr and Nishihara [53], Brady [54] debated the methods of representation and sets of criteria for evaluation of shape. Soffer and Samet [55] introduced a pictorial query specification technique that enables the formulation of complex queries, with the possibility of defining spatial constraints represented by shape features (e.g. moment, circularity, eccentricity, rectangularity) and contextual constraints (e.g. how many objects in the target image). Because boundaries can't represent the inside shape of an object, shapes were defined using the lower coefficients of the Fourier expansion of shape tangent angle and arc length [56-60]. Few papers have been published on evaluating shape similarity because a good, robust and image-independent representation is still needed to describe objects in the scenes [61].

Topological features - derived from an image (e.g. the number of connected / disconnected components) are invariant to rotation, scaling, translation, stretching and deformation. The Euler number [62], i.e. the difference between the number of connected components and the number of holes in a binary image, is an example of topological feature used to characterize images. The Euler vector is an extension of the Euler number method defined only on greyscale images.

Spectral indices - Image features, also known as spectral data primitives can be used to create derived features capable of capturing image-independent properties of the spectral signatures of land cover classes. These features are computed as either linear combinations of elementary spectral bands or as ratios between spectral data primitives acquired in different parts of the electromagnetic spectrum. A few examples of extracted spectral indices used for remote sensing applications are explained in [63]:

1) Brightness is defined as the perceived luminance [64]. Initially defined for Landsat ETM+ data, brightness is calculated as:

$$\text{Brightness} = (1/8) \times (TM1 + TM2 + 2 \times TM3 + 2 \times TM4 + TM5 + TM7) \quad (3.1)$$

2) Visible reflectance is the estimated reflectance in the visible portion of the electromagnetic spectrum. It linearly combines bands TM1, TM2 and TM3 that are individually unfeasible for land cover discrimination due to their high correlation.

$$\text{Visible} = (1/3) \times (TM1 + TM2 + TM3) \quad (3.2)$$

3) Cloud detection - Clouds are colder, with a temperature below 300 K and show higher reflectance at 1700 nm wavelength, equivalent to band TM5 of Landsat. A composite developed to enhance this property for cloud detection is [65]:

$$MIRTIR = (1 - MIR) \times TIR \quad (3.3)$$

4) Vegetation indices are calculated to exploit the differences in reflectance patterns of green vegetation from the spectral signatures of other objects:

$$NDVI = (NIR - RED) / (NIR + RED) \quad (3.4)$$

NDVI (Normalized Difference Vegetation Index) has several important properties: (1) within the same leaf-on / off season, NDVI is only rarely affected by the time of acquisition; (2) Similar to the ratio vegetation index defined in [66], NDVI is useful for extracting vegetation areas from shadow; (3) NDVI is unable to highlight subtle differences in canopy density [67] and condition NDVI > 0.35 is necessary but not sufficient to indicate the presence of vegetation areas [63].

5) Normalized difference bare soil index NDBSI enhances bare soil areas, fallow lands and vegetation with marked background response [63]. This index is useful for predicting and assessing bare soil characteristics such as roughness, moisture content, amount of organic matter and relative percentages of clay, silt and sand [67].

$$NDBSI = (TM5 - TM4) / (TM5 + TM4 + 0.001) \quad (3.5)$$

The NDBSI is an adaptation of the original normalized difference bare soil index defined as:

$$BIO = [(TM5+TM3)-(TM4+TM1)] / [(TM5+TM3)+(TM4+TM1)] \quad (3.6)$$

6) Normalized difference snow index NDSI is defined to enhance the difference between typical spectral signature of snow / ice from other objects with similar spectral signature in the visible portion of the electromagnetic spectrum. The mathematical expression of NDSI exploits the particular property of snow being brighter than the vegetation and bare soil in the visible portion of the electromagnetic spectrum and much darker than the clouds at 1700 nm wavelength.

$$NDSI = (TM2 - TM5) / (TM2 + TM5+0.001) \quad (3.7)$$

Another version of NDSI introduced in [63] makes use of the Visible index defined above:

$$NDSI = (Visible - TM5) / (Visible + TM5 + 0.001) \quad (3.8)$$

7) Built-up and Barren Land Area Index is suitable for detecting built-up areas because of their high value of blue-band reflectance.

$$NDBBBI = (TM1 - TM5) / (TM1 + TM5 + 0.001) \quad (3.9)$$

Other examples include the vegetation index (VI), difference vegetation index (DVI), perpendicular vegetation index (PVI), ratio vegetation index (RVI), soil adjusted ratio vegetation index (SARVI), soil adjusted vegetation index (SAVI), transformed soil adjusted

vegetation index (TSAVI) and many others. The interpretation of spectral vegetation indices is explained in [68] and several insights from a user's perspective on their accuracy and uncertainty are given in [69].

Video features - New features and similarity measures based on color, texture and shape have been defined in the video processing domain. The MPEG-7 standard introduced a new set of features [86,91]. The new color features [70,71] (e.g. NF, R-G-B and M colour space) show benefits in areas such as lighting invariance, intuitiveness and perceptual uniformity. By combining relatively simple histograms, Ojala et al [72] increased the accuracy of video image retrieval. A new texture feature based on the Radon transform was introduced in [73] with the advantage of being rotationally invariant. Sebe et al [75] debate how to derive an optimal similarity measure given a training set and conclude that the sum of squared distances gives the worst results. Non-metric distances are evaluated in [76].

3.2.3. Feature Selection Review

Feature selection is a process commonly used in information mining, wherein a subset of the features available from the data are selected to be used as input to a mining algorithm. Feature selection methods reduce the dimensionality of data by selecting only a relevant subset of the recorded features, with the aim of enhancing a validation measure.

The dimensionality problem of the datasets has two common solutions: feature selection (e.g. subset selection) or feature extraction (e.g. Principal Component Analysis, Independent Component Analysis, Singular Value Decomposition, manifold learning, factor analysis). Feature selection is usually preferable to feature transformation (extraction) when the original units and physical meaning of features are important. Feature selection becomes the primary means of dimensionality reduction when categorical features are present and numerical transformations inappropriate. Physical models back the data recorded by imaging satellite sensors and the pixels' values represent physical measurements (e.g. reflectance, radiance) of the natural scene. For this reason, in geospatial applications feature selection is preferred over feature transformation in order to maintain the physical values of pixels, thus welcoming interdisciplinary collaborations - the selected features retain the original meaning domain experts have knowledge of. Another drawback of feature transformation methods is that the new features (components) resulting in linear transformation may not coincide with the discriminatory information required by a classifier.

Feature selection is a process through which a subset is chosen from the original features. According to their modus operandi, feature selection methods can be categorized into filter, wrapper and hybrid models [160], [224]. The filter models use solely the general characteristics of the data to evaluate and select feature subsets without employing any mining algorithm or classifier. They offer more general results by exploiting the data characteristics as their evaluation criteria. Because their searching processes do not require a classification procedure, they can effectively reduce time and processing complexities [161]. The wrapper models integrate a mining algorithm to evaluate the performance of feature selection. The selected attributes are chosen to optimize the respective algorithm but the method requires more computation power than the filter model and its results may not be suitable for other mining algorithms [162]. The hybrid models take advantage of both filter and wrapper models by exploiting their evaluation criteria in different search stages [163].

Several valuable reviews for feature selection methods have been published and the reader is advised to refer to [164] for an extensive study. Other methods are presented in [234-241] and [228-230].

The general procedure of feature selection methods consists of four basis steps – subset generation, subset evaluation, stopping criterion and result evaluation. Subset generation is a search procedure that organizes feature subsets for evaluation based on a search strategy. Each subset is evaluated and compared to the others according to a certain evaluation criterion. This process of generation and evaluation is repeated until a stopping criterion is satisfied. The last step is an extended evaluation of results using ground truth data or prior knowledge. These four steps are as follows:

1. Subset Generation – choosing the candidate subset for evaluation. Two basic issues determine this process: (a) the search starting point and (b) the search strategy. Search may start with an empty set and successively add features (i.e. forward selection) or start with a full set and successively remove features (i.e. backward selection). Another possibility is to start with both ends and add or remove features simultaneously (i.e. bidirectional search) or to start randomly. For a data set with N features, there are 2^N candidate subsets and in order to evaluate this search space, several strategies have been developed: complete search, sequential search and random search. The selection criteria usually implies the maximization of a specific accuracy indicator and by removing the most irrelevant and redundant features from the data, the performance of learning models is improved. In the human-centered approach, feature selection provides a good framework for discovering and understanding knowledge about the data, what attributes are important for what application and how the available attributes relate to each other.

2. Subset Evaluation – every subset must be evaluated using an evaluation criterion. An optimal subset selected using one criterion may not be optimal according to another criterion. The evaluation criterion can be dependent or independent to the mining algorithm applied on the selected feature subset.

- Dependent Criterion used in wrapper models requires a predetermined mining algorithm applied on the selected subset to determine and evaluate which features are selected. It finds features better suited to the learning model but it also tends to be more computationally expensive and the results are not valid for other mining algorithms.

- Independent Criterion evaluates the value of feature subset related to a specific task by exploiting only the intrinsic characteristics of the training data without any learning algorithm. The analysis of the available data attributes is performed by distance measures, information measures, dependency measures and consistency measures [165]. This paper is focused on measures of mutual information and the remainder of this paragraph will be directed to present the trends of this approach.

Shannon's mutual information measure [166] is a good indicator to estimate the dependency, relevance or similarity between two random variables. Mutual information is the gain of information by which the uncertainty about one variable is decreased by the given knowledge of a second variable. Using the properties of this measure, researchers implemented the Mutual Information Feature Selector MIFS [167], which maximizes the relevance of the

input candidate feature to the output class and simultaneously takes into account the redundancy between the candidate features and the already selected features. Because the calculus of the co-occurrence matrix is computationally expensive and time consuming on one hand and the new multispectral satellite sensors offer extended spectral resolutions on the other hand, the work in this dissertation limits the search space only to the available spectral bands. The MIFS method doesn't provide a direct measure to judge whether to add additional new features or not. Another drawback of MIFS is the fact that it operates a linear transformation over three spectral bands, thus loosing the physical model behind pixel values and limiting the method to only three features. The success of an information-based feature selection algorithm depends critically on how much information about the target class is contained in the selected features.

Computation of mutual information between continuous random variables is very difficult because it involves probability density functions and the integration of these functions. To assess these difficulties, two study directions have been followed: (a) derived histograms to approximate the probability density functions, with the risk of degrading the performance as a result of large errors in estimating mutual information and (b) using a Parzen window method to estimate the input distribution instead of dividing the input space into several partitions [168]. Although the latter has better accuracy than MIFS, it is not suitable to be applied with remote sensing data because image histograms are not approximations of probability density functions but discrete distributions of physical measurements.

3. **Stopping Criteria** – determines when the feature selection process stops. This can be accomplished in multiple ways: the search completes, some given bound is specified (e.g. minimum number of features, maximum number of iterations), adding/removing features does not improve the results or a threshold for evaluating results has been reached (e.g. classification error rate is lower than the allowable error rate for a given operation).

4. **Result Validation** – directly measuring the results using prior knowledge about the data.

3.2.4 Multidimensional Indexing

An index is a structure that provides access to a database in terms of record organization. A *N*-dimensional index is a structure computed from the image data that improves the speed of data retrieval operations from a database. Images are assigned to a suitable content-based descriptor extracted from the data and the descriptors are then organized into a data structure for mining and retrieval. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power. For example, while an index of 10,000 documents can be queried within milliseconds, a sequential scan of every word in 10,000 large documents could take hours. The additional computer storage required to store the index, as well as the considerable increase in the time required for an update to take place, are traded off for the time saved during information retrieval.

In multidimensional indexing several components have to be addressed:

1. Dimensionality reduction at the indexing step is important because images and the derived spectral and spatial features contain a high amount of information. Computation is very expensive at this point and complicates vector processing.

2. Clustering - the elements in the feature vectors having similar content can be grouped together using an unsupervised algorithm. Clustering approaches are either model-based [77] (e.g. algorithms based on an apriori specified model such as Gaussian mixture model or Markov chains) or distance-based (e.g. Euclidean, Mahalanobis, Minimum distance). The K-means algorithm is the most popular algorithm used in IIM. It aims at partitioning *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data as well as in the iterative refinement approach employed by both algorithms.

Given a set of observations $(x_1,...,x_n)$ where each observation is a *d*-dimensional vector, k-means clustering aims at partitioning the *n* observations into *k* sets ($k \le n$) $S = \{S_1,...,S_K\}$ so as to minimize the within-cluster sum of squares:

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \tag{3.10}$$

where μ_i is the mean of points in S_i.

Regarding the computational complexity, k-means clustering is NP-hard in a general Euclidean space for 2 clusters [78,79], NP-hard for a general number of clusters *k* even in the same plane [80]. Usually the method uses an iterative refinement technique. It is also called Lloyd's algorithm. Given an initial set of *k* means $m_1^{(1)},...,m_k^{(1)}$, the algorithm proceeds by alternating between two steps.

1. Assignment step: assign each observation to the cluster with the closest mean:

$$S_i^{(t)} = \left\{ x_j : \left\| x_j - m_i^{(t)} \right\| \le \left\| x_j - m_{i*}^{(t)} \right\| \right\} \tag{3.11}$$

for all i* = 1, …, *k*

2. Update step: calculate the new means to be the centroid of the observations:

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{3.12}$$

Commonly used initializations methods are Forgy and Random partition [80]. The Forgy method randomly chooses *k* observations from the data set and uses these as the initial means. The Random partition method first randomly assigns a cluster to each observation and then proceeds to the Update step, by computing the initial means to be the centroid of the cluster's

randomly chosen points. The Forgy method tends to spread the initial means out while Random partition places all of them close to the center of the data set. The Random partition is generally preferable. As it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum and the result may depend on initial clusters. The algorithm is usually very fast, it is common to run multiple times with different starting conditions.

3. Data structure for content-based retrieval - after the data has been clustered using an unsupervised method, a data structure for indexing descriptors to semantic content must be selected. The commonly used methods are tree-based indexing (e.g. multidimensional binary search trees, R-trees) and hashing-based indexing.

3.3 Content-based Information Retrieval: State-of-the-Art in Multimedia

Content-based information retrieval (CBIR) is now extending beyond the boundaries of art, science and culture by providing new paradigms and methods for searching through the variety of media files available. Information retrieval in multimedia aims at searching and extracting knowledge from a database. The great challenge of CBIR is to "make capturing, storing, finding and using digital media an everyday occurrence in our computing environment" [82]. Content-based methods are necessary to bridge the gap between human and machine languages and are especially mandatory when text annotations are incomplete or missing. They can also improve retrieval accuracy even when text annotations are available by providing additional insight into collections. Several domains (e.g. artificial intelligence, image processing, optimization theory, computer vision, pattern recognition, face recognition, robotic guidance) contributed significantly to the underlying foundations of CBIR. Psychology and related fields have given insights for developing interactions with the user.

The beginning of CBIR intensely focused on computed vision applications [83-85] and the algorithms were based on similarity searches of features for images and video. In only a few years, the concept of similarity search was adapted by internet image search engines, e.g. Webseek [86] and Webseer [87]. Extended effort was invested into the direct integration of the feature based similarity search into enterprise databases: IBM DB2 Extenders and Oracle Cartridges [88, 89]. Feature-based similarity search engines became useful in a variety of contexts [90], e.g. searching trademark databases [91], similar video content query [92], image queries.

The development of modern imaging sensor technologies has led to an exponential increase in the volume of visual documents available in databases. With this development comes the need for efficient organization and retrieval of contents. In the early days of content-based information retrieval (CBIR), the leading paradigm for querying multimedia databases was based on keyword search but the method showed many difficulties with practical applications. The manual annotation is very expensive and incomplete; the relation between words and concepts is often complex due to semantic phenomena as synonymy and homonymy and there is no standard criteria to link semantics coming from different users. The semantic gap is the divergence between “information” that comes with the data and the “knowledge” specific for each user and application.

Research activities in the field of CBIR are trying to solve the semantic gap problem by finding ways to connect the discrepancy between low-level features that can be extracted

directly from the image data and the high-level linguistic descriptions applied to the image by the human operator. Keywords and visual features are complementary descriptors and using both of them to query the image databases may offer more meaningful results. Keywords offer a high-level description of the image content and context while the low-level features are parameters providing information about the physical world, very difficult to express in words [90].

Keywords annotating an image can be either linguistic terms corresponding to identifiable items characterizing the visual context of the scene (e.g. words attached to objects or classes) or words that describe the context without any link to the visual information in the image [10]. An example of an annotation that contains a direct link between the image features and the text is presented in figure 3.2. The annotations in figure 3.2 (e.g. Aircraft Airbus 330 and 320) are directly linked to the features of the airplanes in the image. Figure 3.3 presents the results of a Google search using the term "Airbus". In this case, the annotations are embedded in the image metadata, but they have no correlation to the visual features.

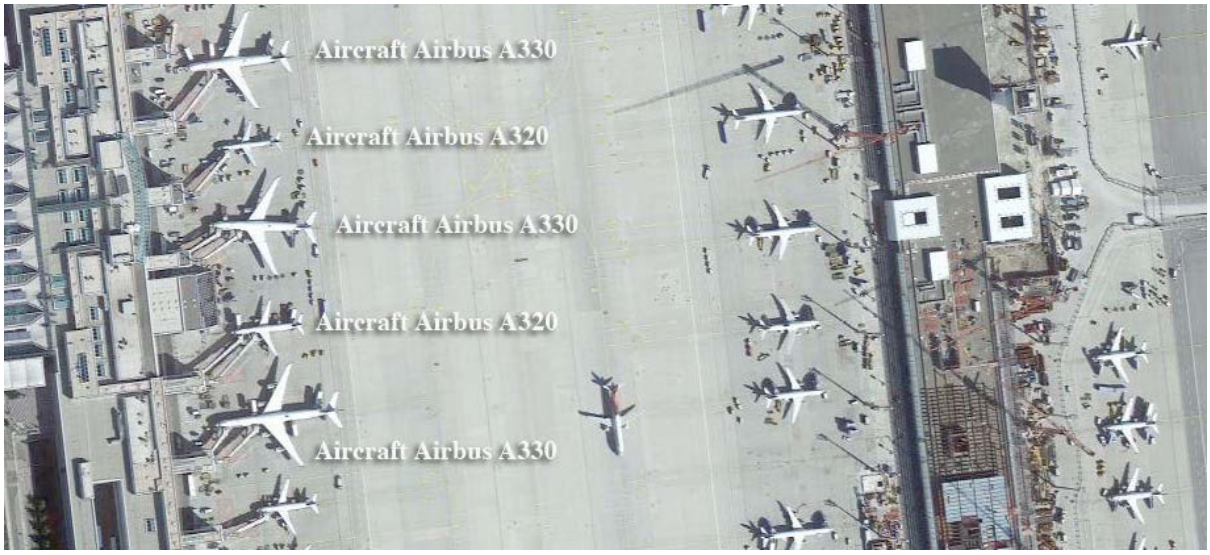


Figure 3.2 - The semantic annotations Aircraft Airbus are directly linked to the objects they describe in the image (Image Credits: DigitalGlobe)

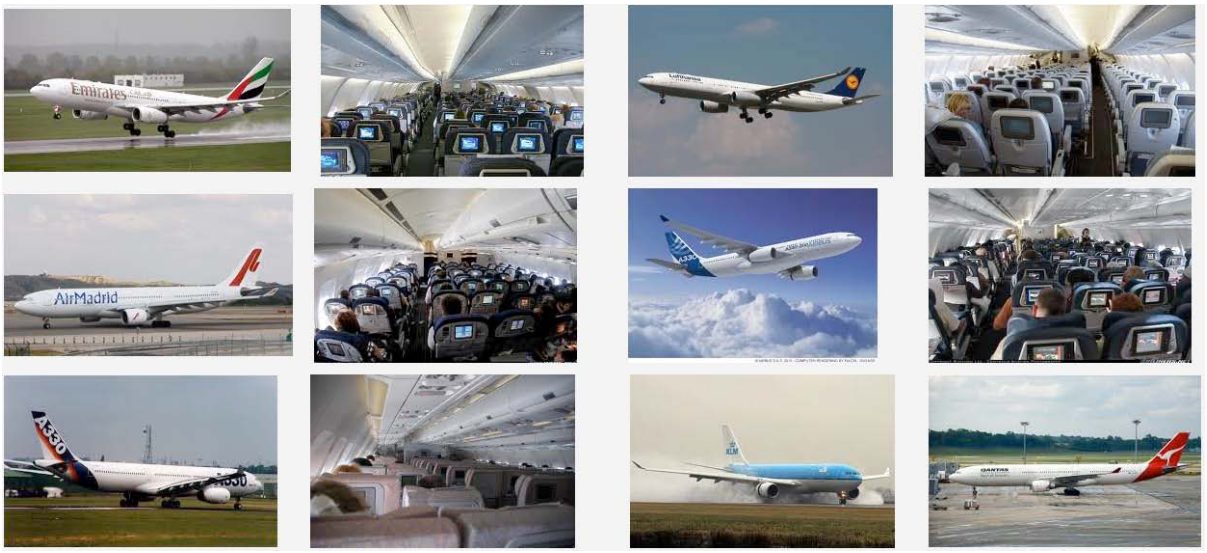


Figure 3.3 - Results of a Google search using the term "Airbus". The semantic annotations are embedded in the image metadata, but they have no correlation to the visual features.

Indexing and retrieval approaches relying on keywords and visual features together are presently being developed to try and solve the semantic gap challenge. Existent search engines make use of image descriptors, a set of data-driven features extracted from the image that may not always be directly connected to the objects the user queries for. The user seeks similarity in semantics while the database can only provide similarity in image processing results. Bridging this gap is maybe the most challenging puzzle that CBIR scientists are trying to address. This means translating the low-level content-based media features into high-level concepts employed by users [93,94]. One of the first pictorial content-based systems addressing the semantic gap in a query was the ImageScape engine [95]. Users were able to perform direct queries for multiple visual objects by using image icons depicting the concepts of interest and the system employed information theory measures to determine the best features for minimizing uncertainty in the classification.

There are two fundamental requirements for a multimedia retrieval system [96]: (1) searching for a particular media item and (2) browsing and summarizing a media collection. The current systems have significant limitations in searching a specific media item because of the inability to understand a wider user vocabulary and the satisfaction level of the operator. Human-centered computing aims at creating for the user the possibility to make queries using his/her own terminology. Experiential computing supports the user explore and gain insights into the media collection. Learning algorithms create a powerful connection between the human and the machine because they allow the computer to understand the media collection at semantic level.

3.3.1 Human-Centered Systems for Multimedia Information Retrieval

The concept of "human-centred" system was defined in [91] as the system that considers the behaviour and needs of the operator, focuses on his/her requirements and integrates the feedback received to improve the results. A very interesting study was performed in 2001 [97] to evaluate if feature-based similarity helped improve image browsing. Results have shown that users prefer to use visual similarity rather than text caption similarity view.

Among the many user requirements [98,99] time is another important aspect to be addressed in designing a CBIR system [100]. To ensure the satisfaction of the operator, the time response of the system needs to be close to near-real time. A concise analysis of the methodologies for interactive retrieval of color images that includes guidelines for selecting methods based on the domains and the types of goals is described in [101]. Researchers also evaluated how users apply the steps of indexing, filtering, browsing and ranking in video retrieval [102].

Another important goal emerging now in media information mining is the extraction of hidden information in user behaviour using large databases [103]. To achieve this goal, researchers undertook multiple directions of study, including experiential computing [104-106] and affective computing [107-112].

3.3.2 Semantic Learning for CBIR

Images can be also represented, described and interpreted using human-centred knowledge (e.g. association between information and semantic concepts attached to the signal features). To perform a query in a database using semantic concepts a comparison between the target image and other images in the archive must be performed using similarity measures at various levels: (1) pixel level - Euclidean or angular distances; (2) multidimensional space - color, texture, shape; (3) object / segment level - semantic labelling of classes and objects is the highest level of abstraction.

Publications have shown the great potential of algorithm learning in multimedia retrieval for bridging the semantic gap [85, 113,114]. The idea behind learning semantics is to discover and create associations between low-level features and semantic descriptors. An initial direction examined the data mining problem by discovering the hidden associations during image indexing. This approach used a visual vocabulary that clusters similar colors and textures [115,116]. Other studies examined fuzzy graph matching algorithms [117] and clustering on space-time regions in the feature space [118] to develop a learning approach. Aksoy et al [119] describe a Bayesian framework for bridging the semantic gap by using a visual grammar that builds a hierarchical semantic model from pixel level to region and scene levels. The pixel-level characteristics are generated via an automatic clustering of primitive features; the region-level signatures through a segmentation algorithm and the scene-level represents the spatial relationships among segments. Two learning steps are needed to generate the visual grammar: a probabilistic link between features and semantic labels and a fuzzy model to link regions and scenes. The image classification process searches representative region groups to describe the scene using semantic labels. The labelled regions are modelled with a Dirichlet distribution based on a number of training examples that depict a certain region and the best matching class is assigned to the image via the maximum a posteriori rule.

A color image segmentation algorithm based on mean shift for estimating density gradients is described in [52]. The users identify the segmented regions by directly labelling the features. A Bayesian hierarchical model for learning and retrieving natural scene categories through intermediate "themes" is introduced in [120]: an image is modelled as a collection of local tiles and each tile is represented by a codeword from a large vocabulary derived from training

examples. A Bayesian hierarchical model is learned for each word. To generate semantic meaning to a previously unseen image, the words are extracted and compared with the predefined models. The drawback of the method is the fact that categories are fixed and the user is not able to assign labels to other classes.

The semantic pathfinder for multimedia indexing described in [121] uses a predefined lexicon of semantic concepts. Given a pattern X, the goal is to discover a semantic concept W from an image I using conditional probability $f_{W|X}$. Each step in the semantic pathfinder analysis extracts the pattern X from data and learns $f_{X|W}$ for all concepts W.

Hudelot et al. [122] describe a learning method based on Support Vector Machine using positive and negative examples in the training phase. The link between semantic concepts and sensor data are also stored in a symbol and modelled as a fuzzy linguistic variable that enables representation of imprecision.

A difficult challenge in semantic understanding of multimedia information is the detection of visual concepts in the presence of complex backgrounds. While the initial goal was to classify the entire image, the clutter and granularity of objects is too coarse for operational applications. The challenge becomes the detection and labelling of the majority of semantic concepts within an image and studies now focus on real world images instead of laboratory-recorded images. In 1996 Lew and Huijsmans [94] used Shannon's theory of information to minimize the uncertainty in face detection problems in greyscale images with complex backgrounds. This method was updated through ulterior studies [93]. The reader may refer to [123] for a comprehensive review on the topic of face recognition.

Classification of multiple features in an interactive way has proven to be fairly difficult. The design of a multilayer neural network model used to merge the results of basic queries on individual features is described in [124]. The first level of semantics is generated by dividing the image in semantic clusters. Then each cluster is divided in feature subclusters (color, texture, shape) and the semantic label is assigned to each subcluster. Fan et al [125] created a multilevel system for annotation of natural scenes with complex backgrounds using dominant image components and semantic concepts. Li and Wang [126] followed a statistical modelling approach for converting images into keywords.

3.3.3 Relevance Feedback

Relevance feedback is the integration of continuous feedback from the user towards learning more about the query. The idea behind relevance feedback is to take the results that are initially returned by a given query and to use information about whether or not those results are relevant to perform a new query. Relevance feedback is the interactive process of communication between human and machine while performing interactive learning for CBIR. The fundamental idea in the interaction loop is to show the user a set of candidate images, allow the user to give negative and positive examples on what images are relevant for a specific query and then let the computer modify the parameter space, semantic space, feature or clustering space to extract relevant examples. The responses are labelled as relevance feedback.

The first relevance feedback method was introduced in [127] and the idea was to move the query point toward the relevant examples and away from the irrelevant ones. Every supervised algorithm can incorporate a relevance feedback loop into the workflow. It has been found that the positive examples are more important than the negative examples for maximizing the accuracy of mining results [128]. Rui and Huang [129] discovered that the optimization-based parameter update achieves higher accuracy than heuristic methods. Li et al [130] propose a computationally optimized composite relevance feedback approach.

A one-class SVM method for updating the feedback space described in [131] showed positive results. He et al. [132] introduced the short term and long term perspectives for inferring a semantic space from the feedback provided by the user. The short term perspective implied the marking of the top three incorrect examples as irrelevant and the marking of the top three correct images as relevant examples. The long term perspective was discovered by updating the semantic space from the results of the short term perspective. Combining multiple relevance feedback strategies improves the results as compared to a single strategy [133]. The reader may refer to [134] for a detailed review on relevance feedback methods.

The QBIC system (Query by Image Content) [135] was developed by IBM as a commercial tool for allowing queries on large image and video archives. QBIC is the first system developed for content-based image retrieval and contains two main components: (1) database population and (2) database query. The population is responsible with the processes related to image processing and image/video database development while the queries are based on learned color and texture patterns, on image examples and user drawings. QBIC also offers an interface for graphical queries and matching the input query to the database. Initially, images are tiled and annotated with text information and an automatic unsupervised segmentation technique is used to generalize the semantic information. The flood fill approach is also employed to automatically identify and annotate objects in the scenes. The algorithm starts from a single pixel and incrementally adds neighbouring pixels with values under a certain threshold.

Photobook [136] was developed by MIT as a content-based image retrieval system based on the main concept of compressing images for a quick query-time performance, preserving essential image similarities. A method derived from the Karhunen–Loève transform is applied on the spectral features to characterize object classes while preserving their geometrical properties. The textural features are analyzed with a method based on the Wold decomposition that separates structured and random texture components. The data are linked with classes through a method based on color difference that provides an efficient way to discriminate between foreground objects and image background. After that, shape, appearance, motion and texture of foreground objects are analyzed and ingested in the archive together with a description. Multiple human-machine interactions are performed to assign semantic labels and through a relevance-feedback loop the system learns the relations between image regions and the semantic content.

PicHunter [137] is a prototype system that represents a simple instance of a general Bayesian framework used to direct a search. With an explicit model of what users would do given a specific target class, PicHunter uses Bayes' rule to predict what is their next target image given their actions. Thus, the retrieval problem turns into the problem of predicting users. This is done via a probability distribution over possible image targets rather than by defining a specific query. Searches can be categorized into three main categories: (1) target specific /

target search; (2) category search and (3) open-ended search-browsing. The Bayesian network can be adapted to all these three search strategies. An entropy-minimizing display algorithm attempts to maximize the information obtained from the user at each iteration of the search. The predictive model can be simulated to estimate how effective a particular kind of interaction will be and design an optimal interaction scheme. PicHunter also makes use of hidden annotations rather than a possibly inaccurate or inconsistent annotation structure that the user must learn and make queries in. The system introduced two experimental paradigms to quantitatively evaluate the performance of the system supported by psychophysical evidence. PicHunter is the first system to introduce latent semantic concepts within images.

Motion content-based video collections are rapidly growing in both the professional and consumer environment and are characterized by increasing capacity and content variety. Our world is adapting to visual communication via video-based applications. To-date, most commercial video search engines (e.g. YouTube, Vimeo) use queries based on text annotations, file name, surrounding text, captions or speech transcript. This query approach works well when specific conditions are met but fails when the visual content is not mentioned or the captions are in different languages. The ideal interactive video approach depends on many factors including type of query, the browsing interface, the interaction scheme and the user's level of expertise.

Content-based video indexing and retrieval systems [207] should assist users to retrieve sequences of video within a large database. Video retrieval is a natural extension of CBIR systems [206] both focusing on accessing image and video by content, spatial (image) and spatio-temporal (video) information. Video indexing adds several orders of complexity to the retrieval problem resulting from indexing, analysis and browsing over the inherently temporal aspect of video.

To overcome some of the recent challenges, researchers developed MediaMill 2007 [184], a semantic video search engine using a 572 concept lexicon and an updated version - TRECVID 2011 [44]. Authors conclude that monitoring retrieval behaviour of users, together with real time active learning helps the human operator improve efficiency in finding results. Another learning point is that, as the number of sources of information increases, so does the retrieval performance.

Other approaches for mining video archives include [43-37] but a detailed review of the methods is beyond the scope of this dissertation because they include additional features that are not applicable to static image retrieval.

3.3.4 Content-based Image Retrieval - Forensics

Forensic investigators often face the challenge to manually analyze a large number of digital images to query for potential evidence [220]. Two typical tasks are the identification of fake images and the identification of case-specific images. Similar problems have also emerged in traditional forensics (e.g. fingerprint identification). Current forensic tools are inadequate in facilitating this process and are confined to generate pages of thumbnail images [138].

3.3.5 Content-based Image Retrieval - Medical Imaging

In most biomedical disciplines, digital image data are rapidly expanding in both quantity and heterogeneity and there is an increasing trend towards the formation of archives adequate to support diagnosis and preventive medicine. Exploration, query and consolidation of the immense collections lead to new tools used to access structurally different data for research, diagnosis and teaching. Overviews of the current state of research and applications in medical imaging are given in [139-141] and [221].

3.3.6 Concluding remarks

The revolutionary imaging technologies have created the need to have a system to organize the abundantly available digital images for easy categorization and retrieval. The requirement for a versatile CBIR system operating on very large databases has attracted the focus and interest of many researchers and led to promising results. The methods and techniques developed encompass diverse areas of interest, e.g. visual image segmentation, feature extraction, feature representation, bridging the semantic gap, mapping features and semantics, storage and indexing, image similarity distance measurement and retrieval, etc. In the past decade, many CBIR systems have been developed. A few examples include the IBM QBIC System [135], MIT Photobook System [136], Berkeley Chabot System [142], Blobworld System [143], Virage System, [144], Visual SEEK and WebSEEK [86], the PicHunter [137], UCSB NeTra System [145], UIUC MARS System [146], PicToSeek system [147], WBIIS from Stanford [148], SIMPLIcity [149].

The major problem for CBIR systems is to incorporate versatile techniques to address a wide variety of query tasks. An important difficulty in CBIR is bridging the semantic gap, the lack of connection between low-level features and the semantic dimension of a given image. The dimensionality of the problem increases with the number of users operating a database as a result of subjectivity in the visually perceived concepts. The image retrieval system contains multiple interdependent tasks with the final goal of translating the subjective phenomena of human perception and understanding into machine language and vice-versa. Feature selection, complex space compression and parameter tuning are mandatory for optimum results.

3.4 Image Information Mining Systems for Earth Observation - A Brief Review

There are only a few CBIR systems operating on large archives of EO data. Most remote sensing retrieval systems allow only simple queries based on sensor, location, time and date but current research shows positive results in the development of advanced query methods in satellite imagery databases.

The Rapid Image Information Mining (RIIM) prototype [150] is a system designed for coastal monitoring and disaster management with an interface for exploration of EO data based on scene content. RIIM offers a possible solution to reduce the feature space by computing only the relevant features that describe a particular concept. The ingestion chain begins with a generation of tiles and an unsupervised segmentation algorithm. A two-step feature extraction algorithm is applied: a first module with a genetic algorithm for the selection of a particular set of features that improves identification of a specific semantic

class and a second module that generates feature models through genetic algorithms. When the user provides a query having a specific label, the feature extraction module will be performed only with the optimal features for prediction thus speeding up the algorithm. The last step is a classification using SVMs. The system automatically computes the confidence value of a selected region and allows the retrieval of regions with a confidence score above a specified threshold.

The VisiMine system [74] is an interactive mining system operating EO data. Because of its ability to distinguish between multiple levels of features (pixel, region and tile), VisiMine uses several feature extraction algorithms for each level. Pixel level features contain spectral and textural information, segments are characterized by their boundaries, shape and size, tile and scene level features contain the spectral and textural information of the whole image scene. Texture features are computed using the Gabor wavelets and Haralick's co-occurrence matrix. Image moments define the geometrical properties of objects. Features are clustered using k-medoid and k-means approaches that perform a partitioning of the set of objects into clusters. K-means clusters objects to their nearest mean (i.e. the centroid of clusters) and k-medoid clusters objects to their nearest average distance (i.e. a medoid is the object whose average distance to all the objects in the cluster is minimal). The center of each cluster in k-medoid method is a member of the data set while the centroid of each cluster in k-means method may not belong to the set. General statistics measures (e.g. histograms, minimum, maximum, mean and standard deviation of pixel characteristics) are computed for regions and tiles. In the training phase, naive Bayesian classifiers and decision trees are used. VisiMine is connected directly to S-PLUS, an interactive environment for graphics, data analysis, statistics and mathematical computing containing over 3000 statistical functions for scientific data analysis. It also includes generic image processing tools, such as histogram processing, spectral balancing, false colors, multiband spectral mixing and data mining tools.

GeoIRIS [153] is a content-based high-resolution satellite imagery retrieval system. GeoIRIS currently supports Query-by-Example using either image regions (tiles) or anthropogenic objects, geospatial queries and geospatial enabled QBE queries. Query content can be chosen from pre-selected semantic examples: region content or objects. The system includes automatic feature extraction at tile level, such as spectral, textural and shape characteristics and object level as high-dimensional database indexing and visual content mining. It also offers the possibility to query the archive by image example, by object and the relationships between object and semantics. The key point of the system is its capability to merge information from heterogeneous sources to dynamically create maps. The image database currently contains 45 GB of 1m and 0.6m pan-sharpened multispectral satellite imagery. GeoIRIS Enterprise Architecture is a Service-Oriented Architecture (SOA) consisting of web services, geospatial query services, content-based retrieval services, map server, distributed middleware and relational database. GeoIRIS Client is a graphical user interface delivered via web browsers and linkage to geoweb visualization software.

KIM [154,155] and the following versions of KES and KEO [151,152, 156] are currently the most advanced IIM systems for EO data. The first prototype of KIM built a theoretical framework of collaborative methods for the extraction and exploration of the image content in large EO data archives and established the link between user knowledge and information content of images. It also provided a solution for communicating at high level of semantic abstraction between users from different domains and heterogeneous sources of information. The semantic interpretation of the image content is linked through Bayesian networks to a

unsupervised content index and, based on this stochastic link the user can query the archive for relevant images. The output is a probabilistic classification of the entire image archive as an intuitive representation of information. The concept was developed and extended into a prototype with high level of semantic concepts [158, 159].

The hierarchy of information in KIM is classified into two components: (1) a resource-expensive, offline, unsupervised computational module responsible for extracting information from data by grouping features in classes and indexing classes in the database and (2) a supervised, online component used by the analyst to define semantic labels. The information representation hierarchy of KIM is based on a five-level Bayesian learning model. In the first phase of the workflow a dyadic k-means clustering is employed to generate a vocabulary of indexed classes. The semantic gap problem is solved by KIM using Bayesian networks, learning the posterior probabilities among classes and user defined semantic labels. The output product (i.e. thematic map) is automatically generated according to predefined cover types.

OMAR [36] is an open source software that allows remote access to imagery and video. The software is developed and maintained in an online distributed environment and although it is available only to classified networks of users it can provide a good example for a wide range of retrieval applications. Key features of OMAR include (1) remote discovery, viewing and manipulation of imagery, (2) on-the-fly orthorectification, precision terrain correction and sensor model projection. The remote user can interactively search for data using a reference map. Results can be filtered only using metadata parameters including date, time, sensor, target ID and combinations of these attributes but do not include any semantic-based queries.

Conclusions and Proposal For a New System

This chapter presented the current status in the fields of CBIR and IIM, with emphasis on the systems available for EO applications. This thesis proposes a new system concept that complements the ones available and addresses the topics that were identified as mandatory by the user communities and haven't been implemented in previous methods. Our system consists of two modules required in an IIM system for EO: (1) semantic rules discovery in large databases; (2) advanced data visualization. These modules secure an optimum workflow and provide the capability to discover and extract relevant conceptual information directly from satellite images and generate an enhanced visualization for the data.

The first module bridges the semantic gap and discovers the latent rules between the low-level output of the state-of-the-art classification algorithms and the semantic, human-defined, manually-applied terminologies of cartographic data. The set of rules explain the content of satellite images using linguistic concepts and link the low-level machine language to the high-level human understanding.

The second module implements an adaptive visualization methodology that can be used in several steps of the workflow. The algorithm assists the image analyst in understanding the satellite image through optimum representations and offers cognitive support in discovering relevant information in the scenes. It is an interactive technique applied to discover the optimum combination of three spectral features of a multi-band satellite image that enhance visualization of learned targets and phenomena of interest. The visual mining module is

essential for an IIM system because all EO-based applications involve several steps of visual inspection and the final decision about the information derived from satellite data is always made by a human operator. To ensure maximum correlation between the requirements of the analyst and the possibilities of the computer, the visualization tool models the human visual system and secures that a change in the image space is equivalent to a change in the eye-brain system of the operator. The next chapter presents the theoretical concepts that have been used in the implementation of this system.

4

Basics of Inference and Stochastic Image Analysis

The CBIR and IIM systems presented in previous chapter operate on different levels of abstraction of the image content, modelled using hierarchical Bayesian networks. These levels that are either image features, spectral bands and indices, clusters or semantic labels, are considered as realizations of stochastic models. Information theory concepts are applied to achieve robust results or to evaluate the steps and connections in the hierarchies.

This chapter focuses on the theoretical aspects used to design the new system concept introduced in this dissertation. Concepts of stochastic image analysis, stochastic processes, Bayesian inference are described with emphasis on generative probabilistic models. These models have been applied to develop the methods presented in the contribution sections of this dissertation and are a key element in discovering latent semantic information in satellite images. The chapter presents the theoretical aspects of Gaussian Mixture Models, Latent Semantic Analysis, Probabilistic Latent Semantic Indexing and Latent Dirichlet Allocation.

Every statistical analysis must be built upon a mathematical model linking the observable reality with the mechanism generating the observations. This model should be a mathematical description of nature: its functional form should be simple and the number of its parameters and components should be minimum. The model should be parameterized in such a way that each parameter can be interpreted easily and identified with some aspects of reality. The functional form should be sufficiently tractable to permit the sort of mathematical manipulations required for the estimation of its parameters and other inferences about nature.

4.1 Stochastic Image Analysis

Two lines of thought have been followed in statistics: descriptive and inferential statistics. The first one analyzes, summarizes and interprets the sample without inferring any property of the population from which the sample was extracted. Inferential statistics introduce the concepts of probability and hypothesis and infers properties of the population from the analysis of the sample. The inferential statistics have followed two distinct trends: the frequentist approach that assumes constant values of unknown parameters of the population and the Bayesian approach that assumes the unknown parameters are continually revised in the light of new data by using weights assigned to previous assumptions. The unknown parameters are regarded as random variables with an associated probability distribution called prior.

4.1.1 Probability

According to the inferential statistics theory, probability can be defined in two ways:

- Definition 1: The probability of one event is the ratio of the number of cases favourable to it to the number of all cases possible when nothings leads to the expectation that every one of these cases should occur more than any other, which renders them as equally possible. This is the frequentist approach probability as defined by Laplace in 1812.

- Definition 2: The probability is a representation of degrees of plausibility by real numbers. This is the Bayesian definition of probability

- Definition 3: In Kolmogorov's probability theory, the probability P of some event E, denoted $P(E)$ is defined in such a way that P satisfies the Kolmogorov axioms, named after the famous Russian mathematician Andrey Kolmogorov, as follows:

Let (Ω,F,P) be a measure space with $P(\Omega)=1$. Then $(\Omega,\Im,P)$ is a probability space, with sample space Ω, event space $\Im$ and probability measure P. The probability of an event E, $P(E)$ must satisfy the Kolmogorov's axioms:

1. Positivity – the probability of an event is a non-negative real number

$$P(E)\geq 0 \tag{4.1}$$

2. Certain event - This is the assumption of unit measure, the probability that some elementary event in the entire sample space will occur is 1. There are no elementary events outside the sample space.

$$P(\Omega)=1 \tag{4.2}$$

3. Sum

$$P\left(\bigcup_{i=1}^{N} E_i\right)=\sum_{i=1}^{N} P\left(E_i\right) \text{ if } E_i\cap E_j=\varnothing, i\neq j \tag{4.3}$$

The product axiom does not assume the exclusivity of events. As a remainder, the notation $P(E_1 \mid E_2)$ refers to the probability of event E_1 conditioned by E_2. In consequence, $P(\varnothing) = 0$, the impossible event has zero probability. However, it does not follow that an event of zero probability is impossible.

Properties:

1. Independence - Two events are statistically independent if:

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2) \tag{4.4}$$

$$P(E_2 \mid E_1) = P(E_2) \tag{4.5}$$

$$P(E_1 \mid E_2) = P(E_1) \tag{4.6}$$

The probability of an event is not influenced by the fact that another event takes place.

2. Bayes Formula

Consider a mutually exclusive and complete set of events $\{H_1,....,H_n\}$ that is not independent of an event *E* in a certain experiment. The events *H* are called hypotheses and are interpreted as hypothetical causes of the event *E*. The following decomposition can be written:

$$P(E) = \sum_{i=1}^{n} P(E \mid H_i) \times P(H_i) \tag{4.7}$$

and the Bayes formula:

$$P(H_i \mid E) = \frac{P(E \mid H_i) \times P(H_i)}{P(E)} \tag{4.8}$$

The probability $P(H_i \mid E)$ is the probability satisfying the hypothesis *H* knowing that the event *E* was produced. This is called the a posteriori probability of *H* and $P(H_i)$ is called the priori probability. The Bayes formula can be understood as a formula for inverting conditional probabilities, i.e. compute $P(H_i \mid E)$ given $P(E \mid H_i)$ and $P(H_i)$.

4.1.2 Random variables

Consider an experiment characterized by its elementary, mutually independent and exclusive events. A particular event consists of the union of several elementary events. A random variable (R.V.) is defined by the biunivoque correspondence with an ensemble of elementary events and is characterized by the probability distribution of these events.

Ω is the sample space composed of the elements ω that are the elementary outcomes of an experiment. Subserts $A \subset \Omega$ are called events and specifically these sets are from a class $\Im (A \in \Im)$. For any of $A \in \Im$ we evaluate the set function $P(\bullet)$ - a mapping from $\Im$ into [0,1] - to generate probabilities $P(A)$. The triplet $(\Omega, \Im, P)$ is the probability space.

The point function $X(\bullet)$ is called a random variable, a mapping from $\Omega \rightarrow R^n$ evaluated for each $\omega \in \Omega$ to yield realizations $X(\omega)$. The probabilities $P(A)$ and the realizations $X(\omega)$ of the random variable X are related by the probability distribution function $F_X(\bullet)$ - a mapping from $R^n \rightarrow [0,1]$, that yields $F_X(\xi)$ as the probability of the set of $\omega \in \Omega$ such that $X(\omega) \leq \xi$.

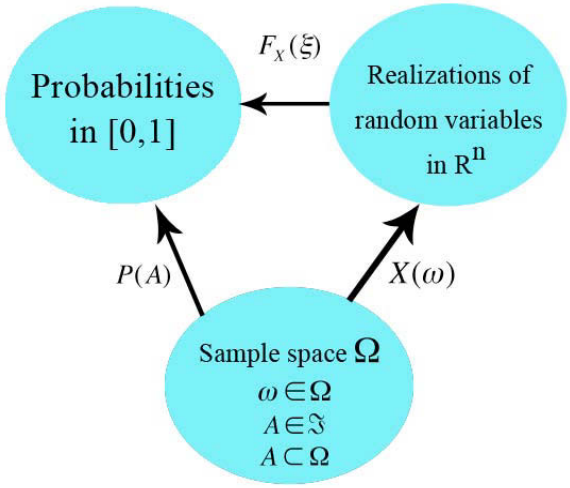


Figure 4.1 – Probability and random variable
(Credits – "Stochastic models estimation and control", Maybeck P.S., Academic Press, 1979)

4.1.2.1 Distribution function and probability density function

The distribution function $F_X : R \rightarrow [0,1]$ of a continuous random variable $X : \Omega \rightarrow R$ - with $\Omega = \{\omega_1, ..., \omega_n\}$ - is the probability that the realizations $X(\omega)$ of the random variable X are less or equal to ξ :

$$F_X(\xi) = P\{\omega : X(\omega) \leq \xi\} \tag{4.9}$$

The distribution function has the following properties:

$$\begin{aligned} &F_X(-\infty) = 0 \\ &F_X(\infty) = 1 \\ &F_X(\xi_2) - F_X(\xi_1) = P(\xi_1 < X < \xi_2) \end{aligned} \tag{4.10}$$

If a scalar-valued function $f_X(\bullet)$ exists such that $F_X(\xi) = \int_{-\infty}^{\xi} f_X(\rho)d\rho$ holds for all values of $\underline{\xi}$, the probability density function (p.d.f.) is defined as:

$$f_X(\xi) = \frac{dF_X(\xi)}{d\xi} \tag{4.11}$$

and is interpreted as $P\{\omega : X(\omega) \in (\xi, \xi + \Delta\xi]\} = f_X(\xi)d\xi$

Unlike the probability distribution function, we are not always assured of the existence of the probability density function. If F_X is absolutely continuous, then the density function does exist. The p.d.f has the following properties:

1. $f_X(\xi) \geq 0$, $-\infty < \xi < \infty$ (4.12)

2. $\int_{-\infty}^{\infty} f_X(\xi)d\xi = 1$ (4.13)

3. $P\{\omega : X(\omega) \in (a,b]\} = F(a < X \leq b) = \int_a^b f_X(\xi)d\xi$ (4.14)

The distribution function and the probability density function can be defined for a multidimensional random variable $X = (X_1,...,X_n)^T$ as:

$$F_{X_1X_2...X_n}(\xi_1,...,\xi_n) = P\{\omega : X_1(\omega) \leq \xi_1, X_2(\omega) \leq \xi_2, ..., X_n(\omega) \leq \xi_n\}$$

$$F_{\underline{X}}(\underline{\xi}) = P\{\omega : \underline{X}(\omega) \leq \underline{\xi}\} \tag{4.15}$$

$$f_{X_1...X_n}(\xi_1,...,\xi_n) = \frac{\partial^n F_X(\xi_1,..\xi_n)}{\partial\xi_1..\partial\xi_n}$$

$$f_{\underline{X}}(\underline{\xi}) = \frac{d}{d\underline{\xi}} F_{\underline{X}}(\underline{\xi}) = \frac{d^n}{d\xi_1..d\xi_n} F_{X_1...X_n}(\xi_1..\xi_n) \tag{4.16}$$

$$P\{\omega : \underline{X}(\omega) \in (\underline{\xi}, \underline{\xi} + d\underline{\xi}]\} = f_{\underline{X}}(\underline{\xi})d\underline{\xi} = f_{X_1...X_n}(\xi_1..\xi_n)d\xi_1..d\xi_n \tag{4.17}$$

Given a n-dimensional random variable from which only $k < n$ components are of interest, a marginal probability density function is defined as:

$$f_{X_1,...,X_k}(\xi_1..\xi_k) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1...X_k, X_{k+1}...X_n}(\xi_1..\xi_n)d\xi_{k+1}d\xi_{k+2}..d\xi_n \tag{4.18}$$

Conditional probability and densities - An n-dimensional joint probability density function is called conditional relative to $(n - k)$ variables if these $(n - k)$ variables have predefined values:

$$\begin{aligned} &f_{X_1,\ldots,X_k|X_{k+1},\ldots,X_n}\left(\xi_1 \ldots \xi_k \mid \xi_{k+1} \ldots \xi_n\right) d\xi_1 d\xi_2 \ldots d\xi_n \\ &= P\left\{\omega : X_1(\omega) \in \left(\xi_1, \xi_1 + d\xi_1\right], \ldots, X_k(\omega) \in \left(\xi_k, \xi_k + d\xi_k\right] \mid X_{k+1}(\omega) = \xi_{k+1}, \ldots X_n(\omega) = \xi_n\right\} \end{aligned} \tag{4.19}$$

Using the Bayes formula, the conditional probability density function is:

$$f_{X_1,\ldots,X_k|X_{k+1},\ldots,X_n}\left(\xi_1 \ldots \xi_k \mid \xi_{k+1} \ldots \xi_n\right) = \frac{f_{X_1 \ldots X_n}\left(\xi_1 \ldots \xi_n\right)}{f_{X_{k+1},\ldots,X_n}\left(\xi_{k+1} \ldots \xi_n\right)} \tag{4.20}$$

Using vector notation, $\underline{X} = [X_1 \ldots X_n]^T$, $\underline{Y} = \left[Y_1, \ldots, Y_m\right]^T$
with $\underline{X}(\omega) \in R^n$ and $\underline{Y}(\omega) \in R^m$:

$$f_{\underline{X}\,\underline{Y}}\left(\underline{\xi}, \underline{\rho}\right) = f_{\underline{X}|\underline{Y}}\left(\underline{\xi} \mid \underline{\rho}\right) \bullet f_{\underline{Y}}\left(\underline{\rho}\right) = f_{\underline{Y}|\underline{X}}\left(\underline{\rho} \mid \underline{\xi}\right) \bullet f_{\underline{X}}\left(\underline{\xi}\right) \tag{4.21}$$

$$f_{\underline{X}}\left(\underline{\xi}\right) = \int_{-\infty}^{\infty} f_{\underline{X}\,\underline{Y}}\left(\underline{\xi}, \underline{\rho}\right) d\underline{\rho} = \int_{-\infty}^{\infty} f_{\underline{X}|\underline{Y}}\left(\underline{\xi} \mid \underline{\rho}\right) \bullet f_{\underline{Y}}(\underline{\rho}) d\underline{\rho} \tag{4.22}$$

$$f_{\underline{Y}}(\rho) = \int_{-\infty}^{\infty} f_{\underline{X}\,\underline{Y}}\left(\underline{\xi}, \underline{\rho}\right) d\underline{\xi} = \int_{-\infty}^{\infty} f_{\underline{Y}|\underline{X}}\left(\underline{\rho} \mid \underline{\xi}\right) \bullet f_{\underline{X}}\left(\underline{\xi}\right) d\underline{\xi} \tag{4.23}$$

Condtional probability density is defined as:

$$f_{\underline{X}|\underline{Y}}\left(\underline{\xi} \mid \underline{\rho}\right) = f_{\underline{X}}\left(\underline{\xi}\right) \tag{4.24}$$

$$f_{\underline{X}\,\underline{Y}}\left(\underline{\xi}, \underline{\rho}\right) = f_{\underline{X}}(\underline{\xi}) f_{\underline{Y}}(\underline{\rho}) \tag{4.25}$$

Through conditional probabilities and densities we are specifying interrelationships among random variables. The two extremes of such relationships are independence and functional dependence. Considering two random variables X and Y (similar for the vector case), they are independent if:

$$P\left(\{\omega : X(\omega) \in A \,\&\, Y(\omega) \in B\}\right) = P\left(\{\omega : X(\omega) \in A\}\right) P\left(\{\omega : Y(\omega) \in B\}\right) \tag{4.26}$$

$$F_{X,Y}\left(\xi, \rho\right) = F_X(\xi) F_Y(\rho), \text{ for all } \xi \text{ and } \rho \tag{4.27}$$

$$f_{X,Y}(\xi,\rho) = f_X(\xi) f_Y(\rho), \text{ for all } \xi \text{ and } \rho \tag{4.28}$$

If X and Y are independent, applying the Bayes' rules gives:

$$f_{X|Y}(\xi,\rho) = \frac{f_{X,Y}(\xi,\rho)}{f_Y(\rho)} = \frac{f_X(\xi) f_Y(\rho)}{f_Y(\rho)} = f_X(\xi) \quad \text{for all } \xi \text{ and } \rho \tag{4.29}$$

4.1.2.2 Statistical Moments

The distribution and density functions for a random variable are fundamental for Bayesian estimation, containing all the information known about the variable. Using these functions, an optimal estimate can be defined using a chosen criterion or it can be used to calculate the expected value of a function of the random variable, where this expected value is the average value obtained over the ensemble of outcomes of an experiment. The expected value of particular functions will generate the moments of a random variable – parameters that characterize the distribution or density function [263].

If $\underline{X}$ is a n-dimensional random variable vector described through the density function $f_{\underline{X}}(\underline{\xi})$ and let $\underline{Y}$ be an m-dimensional vector function of $\underline{X}$: $\underline{Y}(\bullet) = g[\underline{X}(\bullet)]$, where $g(\bullet)$ is continuous. The moments of a random variable $\underline{X}$ are the expected values of certain functions of it. Given the function $g(\underline{\xi})$ of the random variable $\underline{X} = \{X_1,...,X_n\}$, the expectation operator $E[\bullet]$ is introduced and the expectation of $\underline{Y}$ is:

$$E[\underline{Y}] = E\{\underline{g}(\underline{X})\} = \int_{-\infty}^{\infty} \underline{g}(\underline{\xi}) \bullet f_{\underline{X}}(\underline{\xi}) d\underline{\xi} = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \underline{g}(\xi_1 ... \xi_n) \bullet f_{X_1 ... X_n}(\xi_1 ... \xi_n) d\xi_1 .. d\xi_n \tag{4.30}$$

The expectation operator allows the definition of moments of a random variable. Let us consider some specific functions $g(\bullet)$. First, for $g(\underline{X}) = \underline{X}$, we generate the first moment of $\underline{X}$, i.e. the mean of $\underline{X}$.

$$\underline{m}_{\underline{X}} = E\{\underline{X}\} = \int_{-\infty}^{\infty} \underline{\xi} \bullet f_{\underline{X}}(\underline{\xi}) d\underline{\xi} = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \begin{bmatrix} \xi_1 \\ ... \\ \xi_n \end{bmatrix} \bullet f_{X_1 ... X_n}(\xi_1 ... \xi_n) d\xi_1 .. d\xi_n \tag{4.31}$$

$$\text{where} \quad E\{\underline{X}\} = E\left\{\begin{bmatrix} X_1 \\ ... \\ X_n \end{bmatrix}\right\} = \begin{bmatrix} E[X_1] \\ ... \\ E[X_n] \end{bmatrix} = \begin{bmatrix} m_1 \\ ... \\ m_n \end{bmatrix} \tag{4.32}$$

The i –order moment is:

$$m_i = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \xi_i \bullet f_{X_1 ... X_n}(\xi_1,..,\xi_n) d\xi_1 .. d\xi_n$$
$$= \int_{-\infty}^{\infty} \xi_i \left[\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1 ... X_n}(\xi_1,....,\xi_n) d\xi_1 .. d\xi_{i-1} d\xi_{i+1} .. d\xi_n \right] d\xi_i \quad (4.33)$$
$$= \int_{-\infty}^{\infty} \xi_i f_{X_i}(\xi_{X_i}) d\xi_i$$

$$m_{\underline{X}}^{(k)} = \int_{-\infty}^{\infty} \xi^k f_{\underline{X}}(\underline{\xi}) d\underline{\xi} \quad (4.34)$$

with $k = 1,\ldots n$

$\overline{X} = m_1 = \int_{-\infty}^{\infty} \underline{\xi} \cdot f_{\underline{X}}(\underline{\xi}) d\underline{\xi}$ - the mean of the random variable

$\overline{X^2} = m_2 = \int_{-\infty}^{\infty} \underline{\xi}^2 f_{\underline{X}}(\underline{\xi}) d\underline{\xi}$ - the square mean of the random variable

Now, consider $g(\underline{X}) = \underline{X}\underline{X}^T = \begin{bmatrix} X_1^2 & ... & X_1 X_n \\ ... & ... & ... \\ X_n X_1 & ... & X_n^2 \end{bmatrix}$ and we define the matrix Ψ as the n-by-n matrix whose $(i-j)$ component is the *correlation* of X_i and X_j.
The diagonal terms of this matrix are autocorrelations:

$$\Psi_{ij} = E\left[X_i X_j\right] = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \xi_i \xi_j f_{\underline{X}}(\underline{\xi}) d\xi_1 ... \xi_n \quad (4.35)$$

The matrix Ψ is the second noncentral moment of X or the autocorrelation if X.

$$\Psi = E[\underline{X}\underline{X}^T] = \int_{-\infty}^{\infty} \underline{\xi}\underline{\xi}^T f_{\underline{X}}(\underline{\xi}) d\underline{\xi} \quad (4.36)$$

By assuming another function $g(\underline{X}) = [(\underline{X} - m)(\underline{X} - m)^T]$, we define the n-by-n matrix P whose $(i-j)$ component is the *covariance* matrix of $\underline{X}$ and can be written as:

$$P = E[(\underline{X} - m)(\underline{X} - m)^T] = \int_{-\infty}^{\infty} (\underline{\xi} - m)(\underline{\xi} - m)^T f_{\underline{X}}(\underline{\xi}) d\underline{\xi} \quad (4.37)$$

The covariance matrix P is symmetric, the variances of the separate components of $\underline{X}$ are along the first diagonal $P_{ii} = E[(X_i - m_i)^2]^T$. The square root of the variance P_{ii} is called the standard deviation of X_i, denoted as σ_i. The diagonal terms can be expressed as $P_{ii} = \sigma_i^2$

The *correlation coefficient* is X_i and X_j is defined with:

$$r_{ij} = \frac{E[(X_i - m_i)(X_j - m_j)]}{\sqrt{E[(X_i - m_i)^2]E[(X_j - m_j)^2]}} = \frac{P_{ij}}{\sigma_i \sigma_j} \tag{4.38}$$

If the correlation coefficient is zero, then the components X_i and X_j are uncorrelated. Another expression for the covariance matrix is the following:

$$P[(\underline{X} - m)(\underline{X} - m)^T] = E[\underline{X}\underline{X}^T] - mm^T \tag{4.39}$$

which reduces to $P = E[X^2] - \left(E[X]\right)^2$ for the scalar case.

Two random vectors X and Y are uncorellated if their corellation matrix is equal to the outer product of their first order moments:

$$E[XY^T] = E[X]E[Y^T] = m_x m_y^T \tag{4.40}$$

Or, for any i and j:

$$E[X_i Y_j] = E[X_i]E[Y_j] \tag{4.41}$$

The root square of the variance is called the standard deviation of the variable:

$$\sigma_i = \sqrt{P_{ii}} \tag{4.42}$$

If the random variables are independent, their covariance is zero. However, if the covariance of two random variables is zero, one cannot conclude that the random variables are independent. When a feature increases and the other one also increases, the covariance value is positive. Contrary, if the covariance value is negative, when a feature increases the other one decreases [263].

4.2 Transformation-Based Analysis

4.2.1 Principal Component Analysis

Principal Components Analysis (PCA) [32] is a classical statistical method used in data analysis, feature space reduction and compression. PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated random variables (i.e. spectral bands of an image) into a set of values of uncorrelated variables called principal components. The number of principal components chosen for analysis is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the highest possible variance, i.e. accounts for as much of the variability in the data as possible, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The principal components are independent only if the data set is *jointly normally distributed*. PCA is sensitive to the relative scaling of the original variables. The transformation is also known as the discrete Karhunen–Loève transform (KLT).

Let the image X be an $n \times m$ array of pixels with each pixel being a vector of n_c numbers, one for each spectral band of the sensor. PCA transforms X to a vector Y with the first component having the highest variability and the last component the least variability, i.e. the elements of Y are uncorrelated. Another way of saying this is that S_y - the covariance matrix of Y is a diagonal matrix. We are looking for the matrix G such that $Y = G \times X$ and allows the transformation of X into Y. The formula for the covariance matrix S_y is:

$$S_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T \tag{4.43}$$

Considering that:

$$X^T G^T = (GX)^T \tag{4.44}$$

$$S_y = G\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T\right)G^T \tag{4.45}$$

$$S_y = G \times S_x \times G^T \tag{4.46}$$

with the requirement that S_y is a diagonal matrix. The square symmetric matrix A may be decomposed in $A = U \times \Lambda \times U^T$ where the matrix U has the normalized eigenvectors of A as its columns and Λ is a diagonal matrix with the eigenvalues of A along its diagonal. Since U is an orthogonal matrix $UU^T = I$ we can also write $\Lambda = U^T \times A \times U$. This leads to the conclusion that $G = U^T$. To rotate the image to a coordinate system where channels are uncorrelated, we need to multiply the data with U^T, where the columns of U are the eigenvectors of the covariance matrix. The capabilities of PCA to create optimum visual representations of satellite images will be tested in the last contribution of this dissertation. The detailed mathematics of PCA are described in the Appenix.

4.2.2 Independent Components Analysis

Independent component analysis (ICA) is a computational method for separating a multivariate signal into additive subcomponents assuming the mutual statistical independence of the non-Gaussian source signals.

ICA [191] finds the independent components by maximizing the statistical independence of the estimated components. Typical algorithms for ICA use centering, whitening, with eigenvalue decomposition and dimensionality reduction as pre-processing steps in order to simplify and reduce the complexity of the problem for the actual iterative algorithm. Whitening and dimension reduction can be achieved with principal component analysis or singular value decomposition. Whitening ensures that all dimensions are treated equally apriori before the algorithm is run.

In general, ICA cannot identify the actual number of source signals, a uniquely correct ordering of the source signals, nor the proper scaling (including sign) of the source signals. If the data are represented by the random vector $X = (X_1,...,X_m)$ and the components as the random vector $S = (S_1,...,S_n)$ the task is to transform the observed data X using a linear transformation W as $S = W \times X$ into maximally independent components measured by a function of independence $F(S_1,...,S_n)$. The last chapter of the contributions provides a detailed comparison between the capabilities of PCA and ICA to create optimum visual representations of satellite data. The details of ICA and the main differences between PCA and ICA are described in the Appenix.

4.3 Bayesian Inference

Highly complex statistical models made up by multiple, possibly interdependent variables can be addressed by considering conditional independence assumptions. This allows efficient inference to be carried out even for models involving a large number of variables. Bayesian inference considers the unknown parameter θ as a representation of a random variable that is described by a certain probability distribution called prior $f_\theta(\upsilon)$. The prior represents the probabilistic behaviour of the parameter before X was observed. A random sample X brings new information about the prior – if the assumptions were right or wrong - and the prior has to be modified or not. The modified probability distribution taking into consideration the sample knowledge is called a posteriori distribution and is represented by $f_{\theta/X}(\upsilon \mid \xi)$. By applying Bayes' theorem, the corresponding a posteriori probabilities can be determined for each value of θ as:

$$f_{\theta|X}(\upsilon \mid \xi) = \frac{f_{X|\theta}(\xi \mid \upsilon) f_\theta(\upsilon)}{f_X(\xi)} \tag{4.47}$$

where $f_{X|\theta}(\xi \mid \upsilon)$ is the probability distribution of the observed data X for a certain θ. $f_x(\xi)$ is the marginal distribution or "model evidence" of X defined in the parameter space Θ as $f_X(\xi) = \int f_{X|\theta}(\xi \mid \upsilon) \bullet f_\theta(\upsilon) d\upsilon = \int f_{X,\theta}(\xi,\upsilon) d\upsilon$ and is the same for all possible hypotheses being considered [263].

In signal processing and CBIR systems, Bayes' theorem can be applied to extract features from original data or to incorporate information in the mining processes as prior knowledge due to the stochastic nature of the signal models. The image signal can be represented at various levels using the Bayesian hierarchical model. The basic idea in a hierarchical model is that, looking at the likelihood function $f_{X|\theta}$, it may be appropriate to use priors that depend on other parameters not mentioned in the likelihood. These parameters themselves will require priors that depend on new parameters and so on. The process finishes when no more new parameters are introduced.

A non-hierarchical Bayesian model is represented by $\{f_\theta, f_{X|\theta}\}$ with f_θ being the prior and $f_{X|\theta}$ being the likelihood function. Simply stated, knowledge of X leads to an update of θ.
A hierarchical Bayesian model is described by:

$$\left\{f_{\theta_n}, f_{\theta_1|\theta_2}, f_{\theta_2|\theta_3}, \ldots, f_{X|\theta_1}\right\} \tag{4.48}$$

$\left\{f_{\theta_n}, f_{\theta_1|\theta_2}, f_{\theta_2|\theta_3}, \ldots, f_{X|\theta_1}\right\}$ is a hierarchical model because of the way in which the distribution of the parameters in each level of the hierarchy depends on the parameters in the previous levels. The distribution of parameters at any level of the hierarchy depends on the parameters at the next lower level and, conditional on those parameters, is independent of parameters at all levels below that. Figure 4.2 shows an example of how Bayesian networks are applied to represent information at different levels.

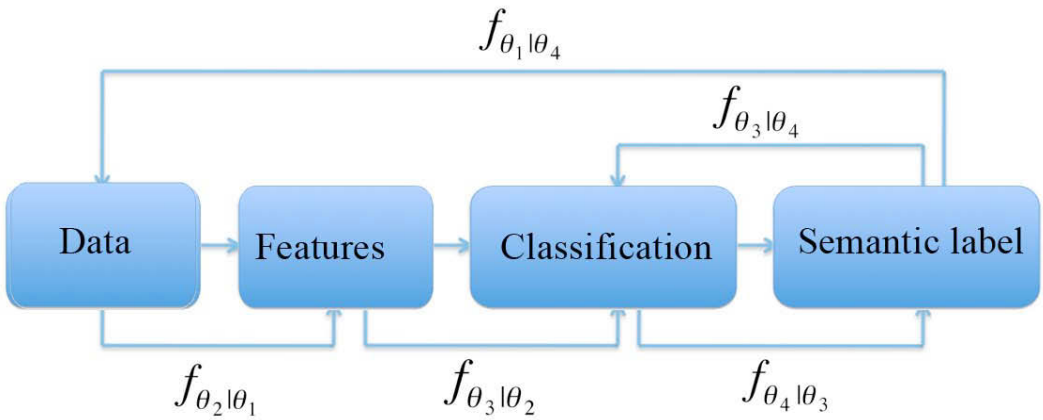


Figure 4.2 – Hierarchical Bayesian Model. $f_{\theta_2|\theta_1}$ represents the stochastic link between the features and the data. $f_{\theta_3|\theta_2}$ learns the unsupervised signal classes (e.g. k-means) knowing the features. $f_{\theta_4|\theta_3}$ infers the semantic content from the spectral map. $f_{\theta_3|\theta_4}$ deduces the cluster of a known semantic label. $f_{\theta_1|\theta_4}$ obtains the image data from its semantic content. The next section explains how Bayesian inference can be applied to image understanding tasks.

4.3.1 Parameter Estimation

Parameter estimation and decision theory consider the problem of selecting a representative value for a given probabilistic distribution. While full probabilistic inference produces as output full posterior distributions, in a number of situations scalar values are required to compare theoretical results with the actual measured quantities. Obviously there is a loss in the quality of the descriptor when moving from a full probabilistic description of a phenomenon to a single value.

The parameter estimation theory is part of the statistical decision theory. The goal of parameter estimation is to evaluate a parameter generated by a source of information in noisy conditions. There are five fundamental components of an estimation problem:

- The variables to be estimated
- The measurements or observations available
- The mathematical model describing how measurements are related to the variables The mathematical model of the uncertainties present
- The performance evaluation criterion to judge which estimation algorithms are best

Let image $\underline{X} = [x_1,...,x_N]$ be the source of information observed in the presence of noise $\underline{n} = [n_1,...,n_N]$. $\underline{Y} = [y_1,...,y_N]$ are the measured pixel intensities.

$$\underline{Y} = \underline{X} + \underline{n} \tag{4.49}$$

Image $\underline{X}$ is a random signal, a realization of a stochastic process. The noise $\underline{n}$ is a 2-dimensional random signal. The problem statement is: given the observations $\underline{Y}$ and possibly some knowledge about $\underline{X}$ and $\underline{n}$, find a guess of $\underline{X}$. We consider the linearly ordered, observed pixel intensities of a random image $\underline{Y} = [y_1,...,y_N]$ characterized by the conditional probability density function

$$f_{y_1...y_n|\underline{X}} = f_{\underline{Y}|\underline{X}} \tag{4.50}$$

Further we consider an estimate $\hat{x}$ of the unknown random gray-level x and the estimation error ε_x. The error ε_x is only hypothetically defined, since the true value of x is unknown. The loss of information that emerges when going from a full probabilistic descrition to a single parameter is expressed by a cost function, leading to the definition of an optimal estimator - the classifcal paramter estimation is formulated as the minimization of the Bayes risk defined over the signal space. The estimation of $\hat{x}$ has to be made by minimizing a cost function C defined as a distance between the actual y and the desired but unknown value of x.

$$\begin{aligned} C &= C(\varepsilon_x) \\ \varepsilon_x &= x - \hat{x} \end{aligned} \tag{4.51}$$

The estimated $\hat{x}$ is a function of the N values of the observed signal and C is a function of (N+1) variables, i.e. the N samples of the signal and the parameter to be estimated. The mean value of the cost $\overline{C}$ is calculated using:

$$\overline{C} = \int_{N+1} C(x - \hat{x}) f_{\underline{Y}\,\underline{X}}(y,x)\,dy\,dx \tag{4.52}$$

The estimated $\hat{x}$ has to be computed so that the mean cost of the estimation $\overline{C}$ is minimum.

$$f_{\underline{Y}\,\underline{X}}(y,x) = f_{\underline{X}|(Y=y)}(x) \bullet f_{\underline{Y}}(y) = f_{\underline{Y}|(X=x)}(y) \bullet f_{\underline{X}}(x) \tag{4.53}$$

$$\overline{C} = \int_{N} f_{\underline{Y}}(y) \left(\int_{-\infty}^{\infty} C(x-\hat{x}) f_{\underline{X}|Y=y}(x) dx \right) dy \tag{4.54}$$

Using the notation $I(y,\hat{x}) = \int_{-\infty}^{\infty} C(x-\hat{x}) f_{\underline{X}|Y=y}(x) dx$, the cost function $\overline{C}$ is minimum if for every observed y we choose $\hat{x}$ so that $I(y,\hat{x})$ is minimum [263].

4.3.2 Case studies

Case 1 – Minimum mean square estimator (MMSE)

In this case we use the cost function $C(\varepsilon_x) = \varepsilon_x^2$, to express the fact that the bigger the estimation error is, the more important the estimation cost becomes. In this case,

$$I(y,\hat{x}) = \int_{-\infty}^{\infty} (x-\hat{x})^2 f_{\underline{X}|Y=y}(x) dx \tag{4.55}$$

The value of $\hat{x}$ that minimizes $I(y,\hat{x})$ is obtained from $\frac{\partial I(y,\hat{x})}{d\hat{x}} = 0$, leading to:

$$\frac{\partial I(y,\hat{x})}{\partial \hat{x}} = -2\int_{-\infty}^{\infty} (x-\hat{x}) f_{\underline{X}|Y=y}(x) dx = 2\hat{x}\int_{-\infty}^{\infty} f_{\underline{X}|Y=y}(x) dx - 2\int_{-\infty}^{\infty} x f_{\underline{X}|Y=y}(x) dx \tag{4.56}$$

Considering $\frac{\partial I(y,\hat{x})}{d\hat{x}} = 0$ and $\int_{-\infty}^{\infty} f_{\underline{X}|\underline{Y}=y}(x) dx = 1$ the estimator becomes:

$$\hat{x} = \int_{-\infty}^{\infty} x \cdot f_{\underline{X}|Y=y}(x) dx \tag{4.57}$$

Case 2 - Maximum a posteriori estimator (MAP)

In this case, the uniform cost function is being employed. The cost function is defined as:

$$C(\varepsilon_x) = \begin{cases} 0 \rightarrow |\varepsilon_x| \leq E/2 \\ 1 \rightarrow |\varepsilon_x| > E/2 \end{cases} \tag{4.58}$$

In this case we are only interested in the estimation error between specific limits E. If the estimation error is between those limits, the cost is null. The previous relationship can be written as:

$$C(\varepsilon_x) = 1 - \begin{cases} 1 \rightarrow (\hat{x} - E/2) \le x \le (\hat{x} + E/2) \\ 0 \rightarrow (\hat{x} - E/2) > x > (\hat{x} + E/2) \end{cases} \tag{4.59}$$

$$I(y,\hat{x}) = \int_{-\infty}^{\infty} f_{\underline{X}|Y=y}(x)dx - \int_{\hat{x}-\frac{E}{2}}^{\hat{x}+\frac{E}{2}} f_{\underline{X}|Y=y}(x)dx \tag{4.60}$$

$$\int_{-\infty}^{\infty} f_{\underline{X}|Y=y}(x)dx = 1 \Rightarrow \text{minimizing } I \text{ implies maximizing } I' = \int_{\hat{x}-\frac{E}{2}}^{\hat{x}+\frac{E}{2}} f_{\underline{X}|Y=y}(x)dx$$

Minimzing I' can be accomplished only by assuming that the interval E for the error estimate is very small. Therefore, we can write:

$$f_{\underline{X}|Y=y}(x) \approx \text{constant} = f_{\underline{X}|Y=y}(\hat{x}),\ \forall x \in [\hat{x} - E/2; \hat{x} + E/2] \tag{4.61}$$

$$I?(y,\hat{x}) \approx f_{\underline{X}|Y=y}(\hat{x}) \int_{\hat{x}-E/2}^{\hat{x}+E/2} dx = E \cdot f_{\underline{X}|Y=y}(\hat{x}) \tag{4.62}$$

Thus, the maximization of I' implies the maximization of the probability density function $f_{\underline{X}|Y=y}(x)$. The value of x that maximizes the aposteriori probability density function is called maximum aposteriori estimator $f_{\underline{X}|Y=y}(x)_{|x=\hat{x}_{MAP}} = \max$.

Initially, we express the probability density function in logarithmic form:

$$\ln f_{\underline{X}|Y=y}(x)\,|_{x=\hat{x}_{MAP}} = \max \tag{4.63}$$

$$\frac{\partial}{\partial x} \ln f_{\underline{X}|Y=y}(x)\,|_{x=\hat{x}_{MAP}} = 0 \tag{4.64}$$

$$f_{\underline{X}|Y=y}(x) = \frac{f_{\underline{Y}|X=x}(y) f_{\underline{X}}(x)}{f_{\underline{Y}}(y)} \tag{4.65}$$

$$\frac{\partial}{\partial x} \ln \left(\frac{f_{\underline{Y}|X=x}(y) f_{\underline{X}}(x)}{f_{\underline{Y}}(y)} \right)_{|x=\hat{x}_{MAP}} = 0 \tag{4.66}$$

$$\left(\frac{\partial}{\partial x} \ln f_{\underline{Y}|X=x}(y) + \frac{\partial}{\partial x} \ln f_{\underline{X}}(x) - \ln f_{\underline{Y}}(y) \right)_{|x=\hat{x}_{MAP}} = 0 \tag{4.67}$$

$$\left(\frac{\partial}{\partial x} \ln f_{\underline{Y}|X=x}(y) + \frac{\partial}{\partial x} \ln f_{\underline{X}}(x) \right)_{|x=\hat{x}_{MAP}} = 0 \tag{4.68}$$

Both MMSE and MAP estimators use the posterior probability density function but they extract different information and do not provide the same solution. The MMSE is the center of mass and the MAP is the mode of the probability density function. The expression for the posterior density includes the deterministic prior knowledge represented by the forward model. The knowledge about the observation noise and the apriori information about the desired paramter are also included. MAP is a complete framework for model-based approaches in information extraction.

Maximum Likelihood Estimator (ML)

If there is no information about the apriori statistical model of the unknown parameter x, we can declare that in the absence of any measurement or observation, x can have any value, with equal probabilites. Simply stated,

$$f_{\underline{X}}(x) = \text{constant}, \ \forall x \in R \tag{4.69}$$

$\hat{x}_{ML}$ is that value of x that maximizes the apriori probability density function $f_{\underline{Y}|X=x}(y)$,

$$\frac{\partial}{\partial \theta} \ln f_{\underline{Y}|X=x}(y)_{|x=\hat{x}_{ML}} \tag{4.70}$$

Maximum Likelihood estimator is not an alternative for MMSE or MAP, ML is the solution in a particular case when there is no apriori information about the unknown parameter.
The details of the Maximum Likelihood estimation are presented in the appendix.

$$\hat{x}_{ML} = \text{argmax}\, f_{\underline{Y}X=x}(y) \tag{4.71}$$

As the prior distribution becomes much wider, less informative, than the posterior distribution, the MAP estimate approaches the ML estimate. The MAP estimate is the mode of the posterior distribution. For unimodal, symmetric posterior distributions, the MAP estimate equals the conditional mean MMSE estimate. The MAP estimator requires a model for the prior distribution of the parameters. Since it uses the information in this apriori model, it is more accurate than the ML estimate.

4.3.3 Generative Probabilistic Models

Many content-based multimedia retrieval tasks can be seen as decision theory problems. In classification cases, a system has to decide whether an image belongs to one class or another. Even the ad hoc retrieval tasks, where the goal is to find relevant documents given a description of an information need, can be seen as decision theory problems: documents can be classified into relevant and non-relevant classes, or each document in a collection can be treated as a separate class and classify a query as belonging to one of these.

The generative probabilistic approach to information retrieval – finding the generating source of a piece of information – has proved successful in media specific tasks, like language modelling for text retrieval [192-194] and Gaussian mixture modelling for image retrieval [195-198].

Because in information retrieval, the goal is to find the best document given a query, one could model the probability of a document given a document directly. This way of modelling the problem is known as discriminative classification. When there are many different possibilities, direct mapping becomes hard to learn and in such cases it is more useful to apply Bayesian inversion and estimates for each possibility. This approach is known as generative classification. Many possible sources for a query exist in information retrieval applications: each document in a collection can be a source. In the generative approach, a separate distribution is estimated for each of the documents in the collection.
The generative image models are probability distributions over a high dimensional (continuous) feature space. The number of different samples that can be drawn is infinite. The models describe the location in the feature space where the likelihood to observe samples is higher and the type of variance that can be expected. If an image is represented by a set of samples $X = \{x_1,...,x_S\}$ each described by an n-dimensional feature vector $V = (v_1,...,v_n)$, the nature of samples is independent of the models.

4.3.3.1 Gaussian Mixture Models

This dissertation is centered around image processing techniques, therefore the following paragraphs will explain various probabilistic models from the perspective of image modeling applications. The Gaussian image models are appropriate models to describe an ideal point in a feature space where all observations are assumed to be versions of this ideal feature vector that are randomly corrupted by many independent minimal influences [199]. In the text domain, it is easy to imagine an ideal concept that has several synonyms that have more or less the same meaning as the concept. In the image domain, this is equivalent to having one point in the feature space ideally representing the class of interest. All observations from this class can be seen as "synonyms", versions of the ideal point that have been corrupted by independent causes (e.g. lighting, angle, haze). Because images are usually representations of complex scenes, instead of using a Gaussian distribution, mixtures of Gaussians are defined to model the images with multiple colors and textures [197]. In general, a finite mixture density is a weighted sum of a finite number C of density functions:

$$f_X(\xi) = \sum_{i=1}^{C} P_C(c_i) f_{X|C}(\xi \mid c_i) \tag{4.72}$$

The mixing weights $P_C(c_i)$ are the prior probabilities of the components c_i in the mixture. The density functions $f_{X|C}(\xi \mid c_i)$ are Gaussian and describe part of the total density. In this case, P_C is the probability mass function and f is the density function.

The usage of mixture models has been classified in [200] in two distinct classes: direct applications and indirect applications. Direct applications refer to situations in which it is believed that a number C of underlying categories or sources exist such that all observed samples belong to one of these categories. Indirect applications refer to situations in which

the link between probability distributions and categories is less clear and a mixture model is used only as a mathematical way of obtaining a tractable way to analyze the data. Modelling images using finite mixture models is usually between the direct and indirect applications. An image can contain only a finite number of classes and only one of the mixture components generates a class (e.g. one component describes the grass, another component describes the sky and another describes the water) and at the same time, the mixture model describes image samples without explicitly separating the components.

Gaussian Mixture Models for Image Representation – The Gaussian mixture model can describe an image by capturing the main characteristics of that image. The samples in an image are assumed to be generated by a mixture of Gaussian sources, where the number of Gaussian components C is fixed for all images in the collection. A Gaussian mixture model is described by a set of parameters $\theta = \{\theta_1 .. \theta_C\}$, each defining a single component. Each component c_i is described by its prior probability $P_{C|\theta}(c_i \mid \theta_i)$, the mean μ_i and the variance Σ_i. The process of generating an image is the following:

1. Take the Gaussian mixture model θ for the image
2. For each sample ξ:

(a) Pick a random component c_i from Gaussian mixture model θ according to the prior over components $P_{C|\theta}(c_i \mid \theta_i)$

(b) Draw a random sample from c_i according to the Gaussian distribution $N(\mu_i, \Sigma_i)$.

In this case θ_i is the observed variable. The mixture model from which the samples for a given image are drawn is known. However, which component c_i generated a given sample is unknown, meaning that the components c_i are unknown variables. The probability of drawing a single sample ξ from a Gaussian mixture model is defined as the marginalization over all possible components:

$$f_{X|\theta}(\xi \mid \theta) = \sum_{i=1}^{C} P_{C|\theta}(c_i \mid \theta_i) f_{X|C,\theta}(\xi \mid c_i, \theta) \tag{4.73}$$

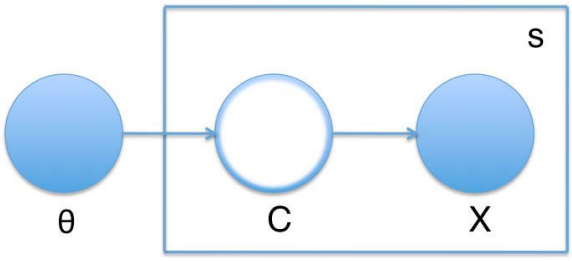


Figure 4.3 - Graphical representation of a Gaussian mixture model

Any distribution can be approximated closely by a mixture of Gaussians. The higher the number of components in the mixture the better the approximation can be. Keeping in mind that these models will be used for retrieval, a perfect description of the image is not the ultimate goal. The purpose is to find images that are similar to a query image. A perfect model would only be able to find exact matches but usually an operator seeks images showing similar latent concepts or classes and not identical feature signatures. Images can represent the same class at conceptual level but at signal level they can be very different. It is important also to avoid over-fitting. Experiments have shown that around eight components are usually enough to capture the most important aspects of an image.

In general, the parameters of a specific document model are unknown and the only available information is the representation of documents – i.e. the feature vectors [201]. A common way to use the data is to assume that the feature vectors are observations from the models and use them as training samples to estimate the unknown model parameters.

4.3.3.2 Latent Semantic Analysis (LSA)

One typical scenario of human-machine interaction for information retrieval involves queries based on human natural language. The user formulates a request and expects the system to return a list of ranked relevant documents. Latent semantic analysis (LSA) [202] is an indexing and retrieval approach that maps documents and words in a collection to a new representation called latent semantic space. LSA takes the high dimensional representation of documents based on word frequencies [203] and applies a linear projection to reduce the dimensionality of the data set: singular value decomposition (SVD). The rationale behind this approach is that similarities between documents or between documents and queries can be more reliably estimated in the representation of a reduced latent space than in the feature vector representation.

4.3.3.3 Probabilistic Latent Semantic Indexing (pLSI)

The concept behind pLSI is a statistical model called the aspect model [204,205] – a latent variable model representing the general co-occurrence data. The model associates an unobserved class variable $Z = \{z_1,...,z_K\}$ with each observation – with each occurrence of a word $w \in W = \{w_1,...,w_M\}$ in a document $d \in D = \{d_1,...,d_N\}$.

pLSI is a generative probabilistic model that is used usually to describe large but finite collections of text or image data and for this reason we use ***discrete*** probability functions to explain it. The model can be defined with the following steps:

- Select a document d_i with probability $p_D(d_i)$
- Pick a latent class z_i with probability $p_{Z|D}(z_i \mid d_i)$
- Generate a word w_i with probability $p_{W|Z}(w_i \mid z_i)$

As a result one obtains an observed pair (d,w) while the latent class variable Z is discarded. Translating the process into a joint probability model results in the expression:

$$p_{D,W}(d_i,w_i) = p_D(d_i) \cdot p_{W|D}(w_i \mid d_i)$$
$$p_{W|D}(w_i \mid d_i) = \sum_z p_{W|Z}(w_i \mid z_i) p_{Z|D}(z_i \mid d_i) \tag{4.74}$$

To derive $p_{W|D}$ one has to sum over the possible choices of *Z* which could have generated the observation. The aspect model is a statistcal mixture model based on two assumptions of independence: (1) the pair of observations (d_i,w_i) are assumed to be generated independently and (2) conditioned on latent class z_i, words w_i are generated independently of the specific document identity. The number of topics z_i is smaller than the number of documents and thus *Z* acts as a bottleneck variable in predicting *W* conditioned on *D*. Document-specific densities $p_{W|D}$ are obtained by a combination of aspects $p_{W|Z}$. Documents are not assigned to clusters, they are characterized by a specific mixture of topics with weights $p_{Z|D}$. Following the likelihood principle, p_D, $p_{Z|D}$ and $p_{W|Z}$ are determined by maximizing the log-likelihood function:

$$L = \sum_{d_i \in D} \sum_{w_i \in W} n(d_i,w_i) \log p_{D,W}(d_i,w_i) \tag{4.75}$$

where $n(d_i,w_i)$ is the term-frequency matrix, i.e. the number of times word w_i occurred in document d_i. The same probabilistic model can be described using the Bayes rule:

$$p_{D,W}(d_i,w_i) = \sum_{z_i \in Z} p_Z(z_i) p_{W|Z}(w_i \mid z_i) p_{D|Z}(d_i \mid z_i) \tag{4.76}$$

The standard procedure for maximum likelihood estimation in latent variable models is the Expectation-Maximization (EM) algorithm. EM alternates two steps: (1) an expectation step where posterior probabilities are computed from the latent variables *Z* based on current estimates of the parameters and (2) a maximization step where parameters are updated for given posterior probabilities computed in the previous E-step. For the aspect model in the symmetric parametrization, the E-step derived with Bayes' rule is:

$$p_{Z|D,W}(z_i \mid d_i,w_i) = \frac{p_Z(z_i) p_{D|Z}(d_i \mid z_i) p_{W|Z}(w_i \mid z_i)}{\sum_{z'} p_{Z'}(z_i') p_{D|Z'}(d_i \mid z_i') p_{W|Z'}(w_i \mid z_i')} \tag{4.77}$$

which is the probability that a word *w* in a particular document or context is explained by the factor corresponding to *Z*. The M-step re-estimation is:

$$p_{W|Z}(w_i \mid z_i) = \frac{\sum_{d_i} n(d_i,w_i) p_{Z|D,W}(z_i \mid d_i,w_i)}{\sum_{d_i,w_i'} n(d_i,w_i') p_{Z|D,W'}(z_i \mid d_i,w_i')} \tag{4.78}$$

$$p_{D|Z}(d_i \mid z_i) = \frac{\sum_{w_i} n(d_i, w_i) p_{Z|D,W}(z_i \mid d_i, w_i)}{\sum_{d_i', w_i} p_{D'|W}(d_i', w_i) p_{Z|D',W}(z_i \mid d_i', w_i)} \tag{4.79}$$

$$f_Z(z_i) = \frac{1}{R} \sum_{d_i, w_i} n(d_i, w_i) f_{Z|D,W}(z_i \mid d_i, w_i) \tag{4.80}$$

with $R = \sum_{d,w} n(d_i, w_i)$

4.3.3.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [216] is a generative model that allows the sets of observations to be described by unobserved (i.e. latent) variables that explain why some parts of the data are similar. Before presenting the mathematical model of LDA, prior explanations are necessary.

- **A generative model** randomly generates observable data according to the latent variables and it specifies a joint probability distribution over observations and label sequences. Generative models are used for modeling data directly (i.e. modeling observations drawn from a probability density function), or as an intermediate step to forming a conditional probability density function. For example, if the observations are words grouped into text documents, LDA regards each document as a mixture of a small number of hidden topics and considers each word as drawn from a topic.

- Latent Dirichlet Allocation is a **topic model** – i.e. a statistical model for discovering the hidden or abstract concepts that occur in a collection of documents. Intuitively, if a text document is about a particular topic, the expectation is that particular words will appear more frequently than others, e.g. words such as “elections”, “government”, “laws” will appear more often in documents about the topic “politics”; words such as “live”, “concert”, “band” will appear more often in documents about the topic “music”. Connecting words such as “the”, “and”, “is” will appear equally in both topics. A document usually contains multiple topics in various proportions, e.g. a document that is 80% about music and 20% politics will intuitively contain more “music” words than “politics” words. In the image domain, a large satellite image taken over the sea will usually contain more “water” pixels than “island” or “land” pixels. A topic can be regarded as the semantic context of a document and is not strongly defined but it is identified through supervised learning and manual labeling. A word may occur in several topics with various probabilities but with very different semantic contexts. For example the word “heart” can belong to the topic “medicine” together with other contextual words such as “cardiology”, “EKG”, “doctor”, “chest”, “physical” and the same word “heart” can belong to the topic “relationships” together with other words such as “feelings”, “St. Valentine”, “love”, etc. The topic is the semantic context describing an idea or a concept. A topic model describes this intuitive understanding using mathematical frameworks and allows the examination

of various sets of documents and the discovery of the latent topics and of the topics in each document.

- LDA is a **Bayesian model** in which each document is represented as a mixture of particular hidden topics and each topic is a discrete probability distribution that explains how common each word is in each topic. LDA understands documents are collections of weighted topics from which words are generated. The topic distribution is assumed to have a Dirichlet prior. For an uniform prior distribution, the pLSA model is similar to the LDA model.

- The **Dirichlet distribution** is over multinomials that are implemented using arrays of floating point values that sum to 1. The learning algorithm is designed to simultaneously infer the Dirichlet hyperparameters that generate both the topic distribution for each document and the word distribution for each topic. A hyperparameter is a parameter of a probability distribution rather than of the model itself. The only real parameter is the chosen number of topics.

- A **mixture model** is a probabilistic model that describes the presence of subpopulations within an overall population, without requiring that an observed dataset should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. Mixture models are used to make statistical inferences about the properties of the sub-populations given only observations on the sampled population, without sub-population identity information.

LDA treats data as observations that arise from a generative model composed of latent variables. The latent variables reflect the semantic structure of the document. Inference aims at answering the question “what are the topics that summarize the document network?”. New data are predicted by the estimated topic model, thus explaining how the new data fit into the estimated topic structures. The goal is to infer the underlying topic structure from observations.

Variable	Type	Definition
K	Integer	The number of topics
V	Integer	The number of words in the vocabulary
M	Integer	The number of documents in the corpus
$N_{d=1...M}$	Integer	The number of words in document d
N	Integer	Total number of words in all documents, $N = \sum_{d=1}^{M} N_d$
$\alpha_{k=1...K}$	Positive real	Prior weight of topic k in a document, usually the same for all topics, normally a number less than 1.
$\underline{\boldsymbol{\alpha}}$	K-dimension vector of positive reals	Collection of all α_K values, is single vector
$\beta_{w=1...V}$	Positive real	Prior weight of word w in a topic, usually the same for all words; normally a number much less than 1.
$\underline{\boldsymbol{\beta}}$	V-dimension vector of positive reals	Collection of all β_w values, is a single vector
$\phi_{k=1...K,w=1...V}$	Probability, [0,1]	Probability of word w occuring in topic k
$\underline{\phi}_{k=1...K}$	V-dimension vector of probabilities, sum to 1	Distribution of words in topic k
$\theta_{d=1...M,k=1...K}$	Probability, [0,1]	Probability of topic k occuring in document d for a given word
$\underline{\theta}_{d=1...M}$	K-dimension vector of probabilties, sum to 1	Distribution of topics in document d
$z_{d=1...M,w=1...N_d}$	Integer between 1 and K	Mixture indicator that chooses the topic for the word w in document d
$\underline{\boldsymbol{Z}}$	N-dimension vector of integers between 1 and K	Identity of topic of all words in all documents
$w_{d=1...M,w=1...N_d}$	Integer between 1 and V	Identity of word w in document d
$\underline{\boldsymbol{W}}$	N-dimension vector of integers between 1 and V	Identity of all words in all documents

Table 1 – Definition of variables in the LDA model

The random variables can be described with the following:

$$\begin{aligned}
&\underline{\phi}_{k=1...K} \sim Dirichlet_V\left(\underline{\beta}\right) \\
&\underline{\theta}_{d=1...M} \sim Dirichlet_K(\underline{\alpha}) \\
&z_{d=1...M,w=1...N_d} \sim Categorical_K\left(\underline{\theta}_d\right) \\
&w_{d,w} \sim Categorical_V\left(\underline{\phi}_{z_{dw}}\right)
\end{aligned} \tag{4.81}$$

Note – these notations will be used again in chapter 5

$\underline{\varphi} = \left\{\underline{\varphi}_k\right\}_{k=1}^{K}$ - is a $K \times V$ matrix

$\underline{\theta} = \left\{\underline{\theta}_d\right\}_{d=1}^{M}$ - is a $M \times K$ matrix

Figure 4.4 describes the dependencies among the variables of the LDA model. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M is the number of documents, N is the number of words in a document.

- α is the parameter of the Dirichlet prior on the per-document topic distributions
- β is the parameter of the Dirichlet prior on the per-topic word distribution
- θ_i is the topic distribution for document i
- ϕ_k is the word distribution for topic k
- z_{ij} is the topic for the j-*th* word in document i
- w_{ij} is the specific word
- K – the number of topics considered in the model
- ϕ is a $K \times V$ matrix, with each row representing the word distribution of a topic.

NOTE - w_{ij} are the only observable variables, while all the other are latent variables.

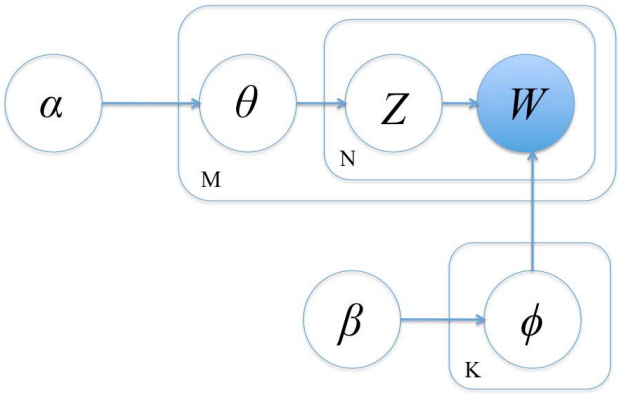


Figure 4.4 – Graphical representation of the LDA model. The boxes represent replicates. The outer rectangle represents documents while the inner rectangle represents the repeated choice of topics and words within a document.

The model is using the following three levels of data descriptors: corpus (the collection the documents), documents and words in the vocabulary, defined as:

- Word = basic unit w defined to be an item from a vocabulary
- Document = a sequence of N words denoted by $W = \{w_1,...w_N\}$
- Corpus = collection of M documents denoted by $D = \{d_1,...,d_M\}$

The generative process implies that documents are represented as random mixtures over latent topics, where each topic is described by a distribution over words. Considering a corpus *D,* with *M* documents, each of length *m* the generative process for each document is:

1. Choose $\theta_i \sim Dir(\alpha)$, $i \in \{1,...,M\}$ and Dir (α) is the Dirichlet distribution for parameter α
2. Choose $\phi_k \sim Dir(\beta)$, $k \in \{1,...,K\}$
3. For each of the word positions i, j where $j \in \{1,...,N_i\}$ and $i \in \{1,...,M\}$

   (a) Choose a topic $z_{i,j} \sim Multinomial(\theta_i)$

   (b) Choose a word $\omega_{i,j} \sim Multinomial(\phi_{zi,j})$

The Multinomial distribution refers to multinomial with only one trial – equivalent with the Categorial distribution. The categorical, the multinomial and the Dirichlet distributions are explained in detail in the Appendix. The Dirichlet random variable has the probability density function on the Euclidean space R^{K-1} defined as:

$$f_{\theta;\alpha}(\theta;\alpha) = \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\theta_1^{\alpha_i-1}..\theta_K^{\alpha_K-1}$$

Exchangeability - The core assumption in LDA is 'bag-of-words'. The words are generated by topics – i.e. fixed conditional distribution functions - and these topics are infinitely exchangeable within a document. The theory of exchangeability states that a finite set of random variables is exchangeable if the joint distribution is invariant to permutation. An infinite sequence of random variables is infinitely exchangeable if every finite subsequence is exchangeable.

The total probability of the Latent Dirichlet Allocation model is:

$$P(\underline{W},\underline{Z},\underline{\theta},\underline{\phi};\alpha,\beta) = \prod_{i=1}^{K}P(\phi_i;\beta)\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)P(W_{j,t}\mid\phi_{Z_{j,t}}) \quad (4.82)$$

The key inferential problem that requires solution is computing the posterior distribution of the latent variables. Approximation methods for the posterior distribution include variational Bayes, Gibbs sampling (described in the Appendix) and expectation propagation.

LDA and other topic models are new, interesting developments in machine learning that promise to increase the reliability and performance of data mining processes. Recently, the LDA model has been applied for data mining in different domains:

- Mining business topics in source codes for open source and commercial systems
- Extraction of 400 topics such as 'September 11 attacks', 'Harry Potter', 'basketball' from a corpus of 330.000 New York Time news articles
- Classification of topics from abstracts published in scientific proceedings
- Extraction of topics from social networks and applications on a collection of emails
- Analysis of the 160.000 abstracts from the 'Cite seer' computer science collection
- Clustering of various biological concepts from a protein related corpus

- Annotating large collections of digital photographs in multimedia databases
- Classifying and annotating satellite images [10, 217]

4.4 Information Theory

This section briefly describes the concept of information and the essential properties in estimation theory. The best-known information measures are Shannon's entropy [166] and Kullback-Leibler divergence [33, 34]. In this disseration, the following measures have been applied in the design of image processing and data mining algorithms. For this reason, we present only the definitions for ***discrete*** random variables.

4.4.1 Shannon's measure of information

For the discrete random variable X with n outcomes $\{\xi_1,...,\xi_n\}$, the probability that the outcome will be ξ_i is $p_X(\xi_i)$. The *information* contained in a message about the outcome of X is $-\log p_X(\xi_i)$. The base of the algorithm is 2 and the unit of information is the bit [263].

The average information or entropy of a message about the outcome of X is:

$$H_X = -\sum_{i=1}^{n} p_X(\xi_i)\log p_X(\xi_i) \tag{4.83}$$

For the discrete random variable X with n outcomes $\{\xi_1,...,\xi_n\}$ and the discrete random variable Y with m outcomes $\{\rho_1,...,\rho_m\}$, the probability that the outcome of X is ξ_i is $p_X(\xi_i)$ **and** the outcome of Y is ρ_j is $p_{XY}(\xi_i,\rho_j)$. The amount of information contained in a message about the outcome of X and Y is $-\log p_{XY}(\xi_i,\rho_j)$. The average information or joint entropy of a message about the outcome X and Y is:

$$H_{XY} = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(\xi_i,\rho_j)\log p_{XY}(\xi_i,\rho_j) \tag{4.84}$$

Properties of Shannon's Measure of Information

- H_X is continuous in the $p_X(\xi_i)$
- H_X is symmetric, that is $H_X = H_Y$ when $p_Y(\xi_1) = p_X(\xi_2)$ and $p_Y(\xi_2) = p_X(\xi_1)$. More generally, H_X is invariant under permutation of the distribution function p_X
- H_X is additive – when X and Y are independent random variables, $H_{XY} = H_X + H_Y$
- H_X is maximum when all the $p_X(\xi_i)$'s are equal
- H_X is minimum when one of the $p_X(\xi_i) = 1$

Theorem - X is a random variable with n outcomes $\{\xi_1,...,\xi_n\}$ and the probability that the outcome will be ξ_i is $p_X(\xi_i)$. Then:

- $H_X \le \log n$, with $H_X = \log n$ if and only if for all *i* it is true that $p_X(\xi_i) = 1/n$
- $H_X \ge 0$, with $H_X = 0$ if and only if there exists a *k* such that $p_X(\xi_k) = 1$

Maximum Entropy - *X* is a random variable with *n* outcomes $\{\xi_1,...,\xi_n\}$. These outcomes occur with probability $p_X(\xi_i) = 1/n$ for all *i.* The average amount of information contained in a message about the outcome of *X:*

$$H_X = -\sum_{i=1}^{n} p_X(\xi_i)\log p_X(\xi_i) = -\sum_{i=1}^{n}\frac{1}{n}\log\frac{1}{n} = -\left(\frac{1}{n}\log\frac{1}{n}\right)\sum_{i=1}^{n}1 = -\left(\frac{1}{n}\log\frac{1}{n}\right)n = -\log\frac{1}{n} = \log n \tag{4.86}$$

Conditional Entropy – The conditional entropy of a random variable *X given another random variable Y* is the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable:

$$H_{Y|X} = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}(\xi_i,\rho_j)\log p_{Y|X}\left(\rho_j \mid x_i\right) \tag{4.87}$$

4.4.2 Mutual information

The entropy of a random variable is a measure of the uncertainity of the random variable, it is a measure of the amount of information required on average to describe the random variable. The relative entropy is a measure of the distance between two distributions. Mutual information is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainity of one random variable due to the knowledge of the other. Consider two random variables *X* and *Y* with a joint probability mass function $p_{XY}\left(\xi_i,\rho_j\right)$ and marginal probability mass functions $p_X(\xi_i)$ and $p_Y(\rho_j)$. The mutual information $I(X,Y)$ is the relative entropy between the joint distribution and the product distribution $p_X(\xi_i)p_Y(\rho_i)$:

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{XY}\left(\xi_i,\rho_j\right)\log\frac{p_{XY}\left(\xi_i,\rho_j\right)}{p_X(\xi_i)p_Y(\rho_j)} \tag{4.88}$$

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X\mid Y) = H(Y) - H(Y\mid X) \tag{4.89}$$

Mutual information is a measure of the information in the random variable *Y* about the random variable *X*. If no information about *X* is recognized in *Y*, then $I(X,Y) = 0$. The main properties of mutual information are presented in the Appendix [263].

4.4.3 Kullback-Leibler divergence

The relative entropy is a measure of the distance between two distributions. In statistics it arises as the log of the likelihood function. The relative entropy $KL(p \| q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p.

$$KL(p \| q) = \sum_{i=1}^{n} p_X(\xi_i) \log \frac{p_X(\xi_i)}{q_X(\xi_i)} \tag{4.90}$$

The divergence is interpreted as a measure describing the amount of information that a measurement gives about the truth of a given model as compared to a second model. If $q_X(\xi)$ is uniform, the divergence is nothing else but the Shannon entropy for the random variable X. The Shannon entropy can be interpreted as the amount of information in a model $p_X(\xi)$ of X compared to the maximum incertitude model – the uniform distribution. The uniform distribution has the maximum entropy. The main properties of the relative entropy and the relationships to mutual information and entropy are presented in the Appendix.

Conclusion

This chapter focused on the theoretical concepts of this dissertation that are used to build the basis of the contributions. The concepts of stochastic image analysis, stochastic processes, Bayesian inference with emphasis on generative probabilistic models were defined and explained. Greater emphasis was put on Latent Dirichlet Allocation and information theory because these topics are used to design the algorithms in the contributions sections. In the next section of this dissertation these concepts are applied to bridge the semantic gap between machine and human languages and to design a concept for advanced visualization of satellite images.

5

Bridging the Gap for Semantic Annotation of Satellite Images

Although state-of-the-art image processing tools are broadly available, all mapping operations are still performed manually or in a hybrid manner because there are still important difficulties in creating thematic maps of the land cover using automatic classification methods. Output products derived from classification algorithms do not contain exclusive mapping units but mixtures of heterogeneous land cover classes (i.e. clusters), as opposed to the cartographic data that use a homogeneous array of information classes (i.e. conceptual areas).

In Earth Observation as well as in other applications of computer vision and image processing, the end-users require a homogenous, conceptual object and not a pixel-based map. The first part of the contributions describes a method to group together similar pixels belonging to the same information classes (i.e. concept) with high-level semantics (e.g. user defined taxonomy in GIS layers) and discover the semantic rules that bridge the three processing layers, from (1) primitive features with no semantic meaning, to (2) intermediate-level semantics indexing a spectral map and to (3) high-level human-centerd semantics revealed in cartographic products.

This chapter introduces a method to semantically annotate satellite images and to map the low-level features to the high-level human concepts, thus bridging the semantic gap between the human and computer languages. (e.g. CORINE LAND COVER CLC 2000 [187]).

5.1 Introduction

The books in a library, the journals in an online database are usually organized by the domains to which they belong, the newspapers on the stands are displayed by topics of interest that easily guide the reader, the articles in a scientific library are grouped together by their relevance to specific fields of study. If you take any daily newspaper, you will observe that it is divided into sections (politics, science, economics, world news). Take another newspaper and there is a high chance that its sections will be very similar to the previous paper, maybe with slightly different names (politics, science & tech, economy, world). If you found a newspaper that has no sections but only articles written and laid out arbitrarily, you would have no problem to classify every piece of news to the section it belongs to. The librarian doesn't have to read all the books in order to know to which domain they belong to, the person at the news stand doesn't read every paper to know where to display it and we don't have to go through all the newspapers to categorize articles into specific sections. It is the same with the chapter you are reading right now, you only have to skim through the text to know what kind of subject it covers. This is a dilemma that has plagued philosophy and science since Plato 2400 years ago - people have much more knowledge about an observation than appears to be present in the information to which they have been exposed. How do people acquire as much knowledge as they do on the basis of as little information as they receive and how does a small set of events lead to generalizations that are usually correct is a persistent mystery of human cognition that has been observed in various fields of science, philosophy, psychology, linguistics, computer science, artificial intelligence, etc. Centuries ago, Plato's solution to this puzzle was that people already possess their knowledge and need only hints and contemplation to retrieve it from the observations. However, time has brought up new views upon this matter and scientists explained it through simple mechanisms of induction [209] that were later translated into computer language and implemented for information retrieval [203].

Some domains of knowledge (e.g. text) contain vast numbers of interrelations between the observations (words) of a specific event (e.g. article, book) that can greatly amplify learning by a process of inference. Learning leads directly to the discovery of latent information in the events (documents, text corpus) that in turn may be generalized to previously unseen ones. In order to create and implement computer models that discover latent knowledge in various sets of data, scientists needed to understand how the human cognition systems reveal information from these datasets.

First, some definitions are required. A concept is a category of observations linked through the similarity of features [214]. For example, all the houses in the world are clustered around the concept house that defines a large set of objects with similar features, properties and functions. This is the first level of understanding – the observation level. The natural inference of fuzzy probabilistic concepts relies on some understanding, some mechanism by which experience with examples can lead to treating new instances more or less equivalently [209]. While the basic features (words) create the first level of understanding concepts, these features also combine into a common latent human-centered feature identity of the event they belong to (text document) which is called *topic* and they create the second level of understanding – the latent level. For example, a group of first level concepts (house, garden, street) combined with different weights may be grouped in second level topics under another name (Residential Areas). By analyzing and associating first-level concepts (words) into

measurable and comparable groups (topics), it is possible to categorize events (documents). As a simple example, the articles in the newspaper (events) can be classified into topics of interest by the words they contain. A political item will contain different words than the science reviews. The problem of synonymy appears both in text and image domains. While in text applications the concept of synonymy is clear, in imaging tasks, an ideal point exists in the feature space and all observations are assumed to be versions of this ideal feature vector that are randomly corrupted by many independent small influences (e.g. optical, electrical) [215]. Spectral mapping and clustering techniques are created to solve this challenge and reduce the size of the vocabulary (visual vocabulary in the case of images) from 2^8, 2^{11} visual words to only a few ideal clusters, or stable points that conceptually represent the same class. (figure 5.1)

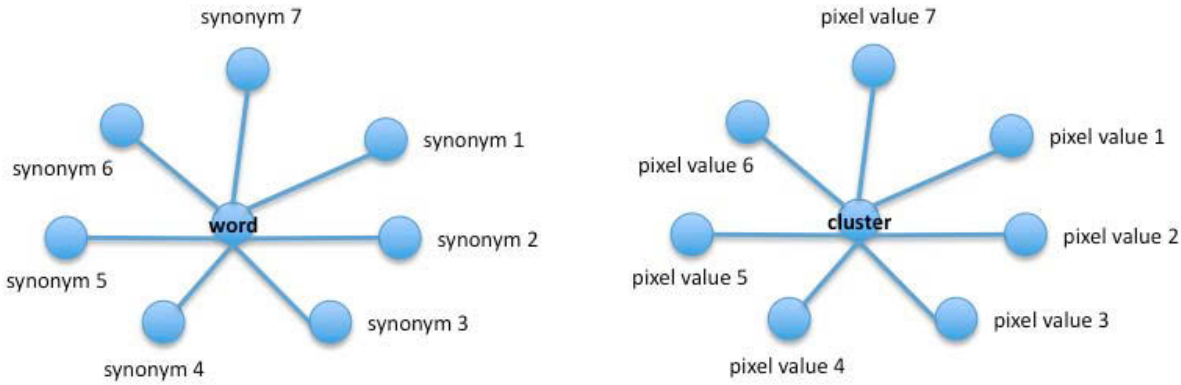


Figure 5.1 – Synonymy in text and image domains

The semantic rules that link the primitive features with the user-defined high-level semantics were discovered by using the LDA model presented in the previous chapter. Image clustering techniques provide as output maps with heterogeneous spectral classes having low-level semantic meaning and are not able to exclusively generate existent homogeneous land use land cover maps. Figure 5.2 shows the workflow of our method for discovering the semantic rules. Each step of the processing chain will be explained in detail in the following sections.

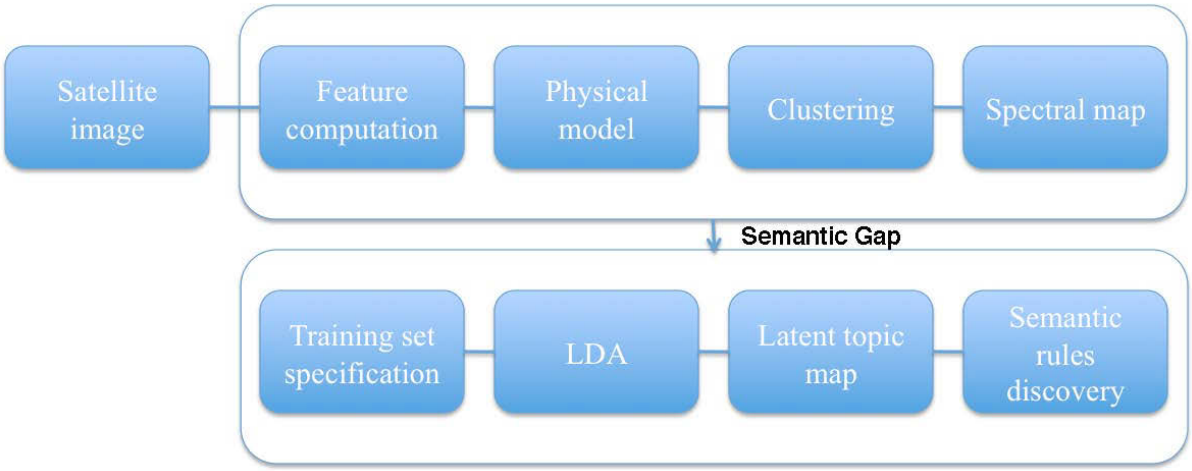


Figure 5.2: Workflow for discovering the semantic rules between the raw satellite data, the intermediate-level spectral map with limited semantic labels and the high-level taxonomy of cartographic products.

5.2 Spectral Signatures And Semantic Content Extraction

Geospatial users require information or information-related services that are focused, concise and reliable, with as low as possible time and money expenses, provided in forms compatible with the user's own activities. In the current Earth Observation (EO) scenario, the archiving centers offer mainly data, images and other 'low-level' products. Usually these products require expert-users to extract specific information from those data or images. With all these considered, during the past few years several research projects developed state-of-the-art tools that add value to the satellite products trying to reach the users' final needs.

To discover the semantic rules that link primitive features to the high-level semantics ontologies of cartographic product, we begin by generating a spectral map, i.e. a cluster map. To achieve this goal we employ the automatic classifier described in [63]. Soilmapper is a purely spectral, per-pixel rule-based classifier, based solely on the spectral domain and prior knowledge retrieved from the remote sensing literature. It requires no training and performs a fully unsupervised preliminary classification over multiple sensors' images calibrated into planetary reflectance. The degree of user supervision required to detect spectral rule-based categories is the same as unsupervised data clustering and far inferior to reference sample selection required by supervised classifiers. The symbolic meaning (i.e. level of abstraction) of the spectral categories (e.g. strong vegetation, deep water) is intermediate between those (low) of clusters (e.g. n-th cluster) and segments (e.g. m-th segment) and those (high) of land cover (information) classes (e.g. forest, water bodies, agriculture field). The classifier is based on prior spectral knowledge. The following paragraph describes the implementation characteristics [63].

Pattern recognition is based exclusively on known spectral signatures of the target classes taken from the remote sensing literature and adapted as fuzzy data templates. This implies that the classification system is pixel-based (context-insensitive) and purely spectral. It uses a set of spectral rules and the mapping system employs no supervised data learning mechanism to dynamically generate new rules. The system maps each pixel data vector into a finite set of discrete, mutually exclusive and exhaustive spectral categories (labels) adapted for Landsat TM and Landsat ETM+ imagery calibrated into planetary reflectance (albedo) and at-satellite temperature. The labels are adaptable to other space borne imaging satellite sensors sensitive to multi-spectral and panchromatic portions of the electromagnetic spectrum (e.g. Advanced Spaceborne Thermal Emission and Reflection Radiometer ASTER and System pour l'Observation de la Terre SPOT-4 and SPOT-5). A property that is very important is the consistency of the system in terms of one-to-one or many-to-one relationships with the set of information classes in other high-level semantic systems. Figure 5.3 shows the possible relations between the spectral categories and other classification schemes [63].

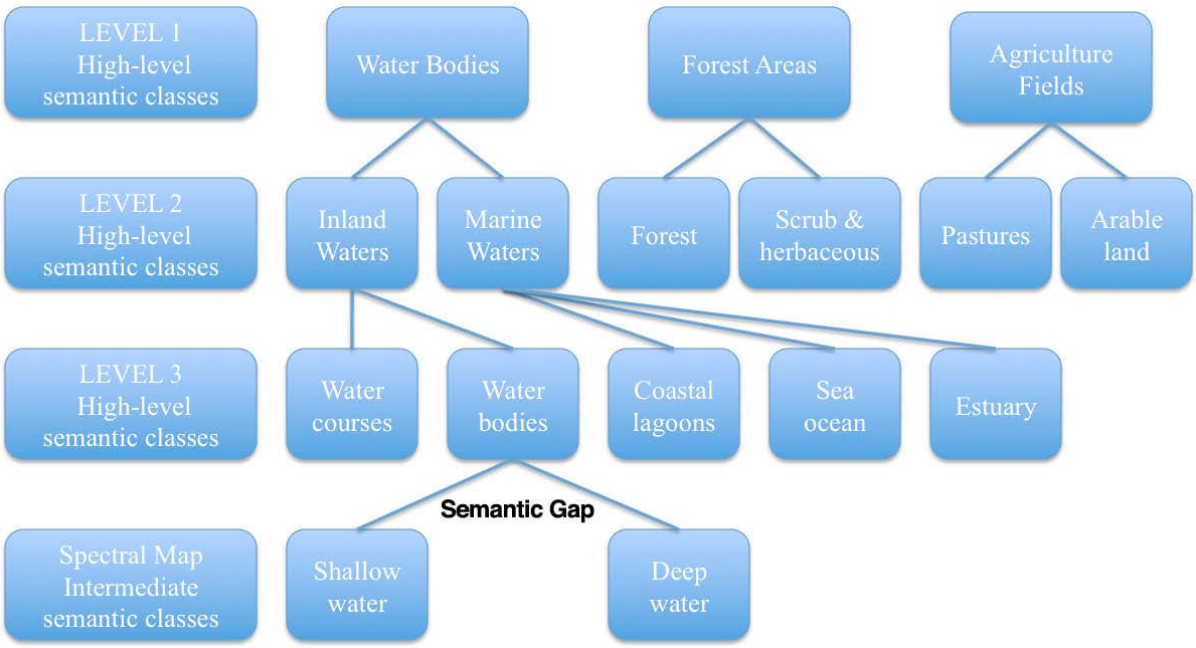


Figure 5.3 – The correspondence between high-level human-centered semantic classes and intermediate-level semantic classes generated by the unsupervised classifier

First-level computation of the spectral types consists of:

- Spectral fuzzy rules generated from spectral reflectance curves extracted from remote sensing for Earth Observation literature and partitioned into different portions of the electromagnetic spectrum. These rules are implemented as logical expressions of scalar numerical variables combined with relational operators (e.g. >; <) and logical operators (e.g. AND, OR)

- Feature extraction and fuzzy-sets computation started from a calibrated set of spectral bands.

The second-level processing step of the spectral categories consists of a hierarchy of logical expressions of binary variables. These binary variables are the outputs of spectral rules and FSs computed during the first-level processing. The output product of the proposed image mapping system is a discrete map consisting of kernel spectral layers, equivalent to a preliminary map. A brief description of the system's architecture is presented in figure 5.4

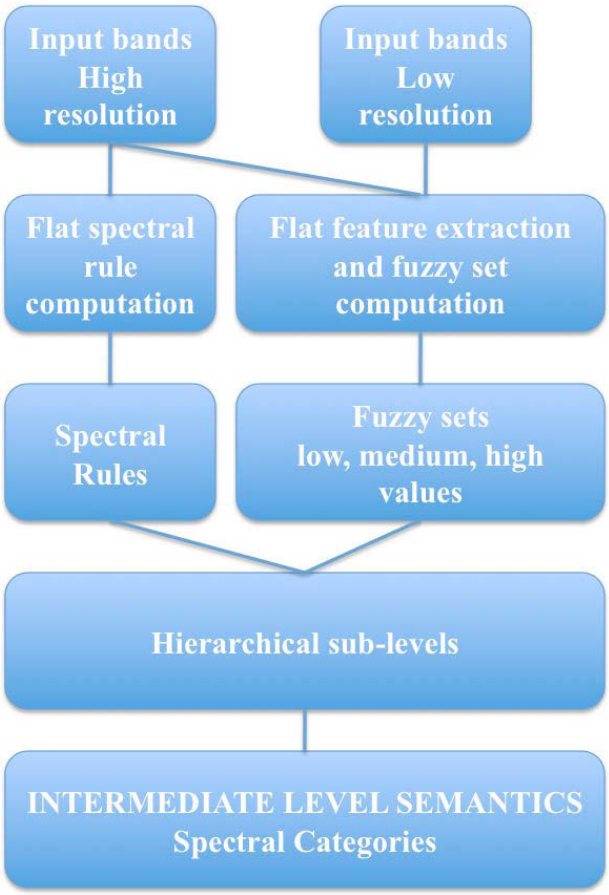


Figure 5.4: Physical model architecture & workflow. The system requires as input all the spectral bands from the sensor and provides as output a map with classes having intermediate-level semantic meaning.

A digital image is a two-dimensional array of pixels. Each pixel has an intensity value represented by a digital number (DN) and a location address referenced by its row and column numbers. According to the classification scheme adopted and the input sensor, the model maps all the pixels in the satellite image to a reduced number of spectral classes. A newer version of the application tries to adapt the mapping schemes to a fixed number of classes (57) disregarding the input sensor. Summarized below are the main characteristics of the physical model [63]:

- It requires as input a sensor-specific Radiometric Calibration and Correction algorithms to transform the digital numbers in each band of the remotely sensed image into values of planetary reflectance and at-satellite temperature. Provided with this, the pre-mapping system works as an independent multi-platform physical model, applicable to a large array of imaging sensors.
- It performs a fully automatic classification, requiring no training data or supervision
- As output it generates a preliminary spectral map whose pixels are labelled with intermediate level semantic meaning, between the low-level meaning of pixels, clusters or segments values and the high-level semantic meaning of land cover classes (e.g. Corine Land Cover 2000)

The physical model generates three layers of maps, with various number of spectral classes. A large set of 85 classes, an intermediate set of 41 and a small set of 16. The most recent version of the application permits also an intermediate set of 27 classes. This finite set of discrete values becomes the input vocabulary in the training stage of the workflow. In the final phase, this vocabulary explains the rules that link the pixels in the spectral map to the high-level cartographic layers of CLC 2000.

5.3 Map Label Learning Using Latent Dirichlet Allocation Model

The algorithm described in this section bridges the semantic gap between the low-level features and the cartographic products with high-level semantics. In the previous section we presented the classifier used to obtain a reduced map of spectral classes with intermediate-level semantics. These classes are used in this section to create the vocabulary of visual-words that will describe the maps' ontologies.

5.3.1 Latent Dirichlet Allocation

The Latent Dirichlet Model (LDA) is employed to correlate the heterogeneous pixels in a spectral map and to describe the corresponding information classes, by following the one-to-one or many-to-one rules. As described in the previous chapter, LDA is a generative probabilistic model for collections of discrete data. Generative models are random sources that can generate infinite sequences of samples according to a probability distribution. LDA was created to describe large collections of digital text documents and recently applied in classification and semantic annotation of satellite images [10], [217].

5.3.2 Document Definition – Matching Images And Words

LDA is a three-level hierarchical Bayesian model, in which each document in the collection is modelled as a finite random mixture over a latent set of topics. Each topic in turn is modelled as a probability distribution over a set of words in the vocabulary. Figure 5.5 shows the layering of information used within LDA by introducing a set of latent topics to describe documents in the text collection.

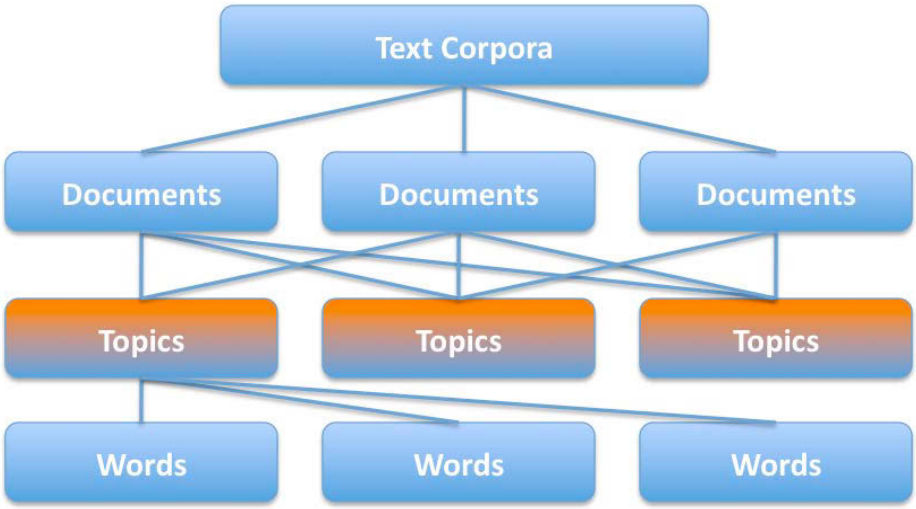


Figure 5.5 – LDA model finds latent topics to describe documents in a text collection. The application of a model for natural language processing on satellite images requires the definition of an analogy between the terminologies used. The text-based model is using the following three levels of data descriptors: text corpus (the collection of documents), documents and words in the vocabulary:

- Word = basic unit w defined to be an item from a vocabulary
- Document d = a sequence of N words denoted by $\underline{W}_d$
- Corpus = collection of M documents denoted by $\underline{W}$

The following correlations with the image domain are created:

- A visual-word w is a spectral class obtained in the pre-mapping phase performed by the physical model. The number of visual-words in the vocabulary is strictly related to the classification scheme adopted. Pixel values are indices to a table provided with intermediate-level semantic meaning. For case studies performed on Landsat images the vocabulary has 16, 27, 41 and 85 visual-words.

- A visual-document is a window of the image (tile). Its size was chosen similar to the size of the minimum mapping unit (MMU) of the land use land cover vector system to be described or generated. For studies performed on CLC 2000, the minimum mapping unit is 25 ha which corresponds to a tile of 15 X 15 pixels (Landsat) and 50 X 50 pixels (SPOT).

The visual-corpus is the satellite image to be annotated. All the documents yield the corpus. The LDA model works under the bag-of-words assumption, in which the order of words in the document is ignored and the image is represented as a random sequence of N visual words. However, extracting and separately analyzing the documents, i.e. image tiles, implies a spatial delimitation. Thus, the results depend on the implicit contextual distribution of the words in the document; this component is explained exclusively by the size of the patches. Spatial relationships have been considered in a transformed LDA model [218] applied for color images with relatively homogeneous structures. However, this method is not suitable to remote sensing data due to the difficulties in geometrically delimiting the image objects [219]. Each visual-document is described by a distribution over visual-words in the form of

frequency-count vector (histogram). LDA models each word in a document as a sample from a mixture model, where the mixture components can be viewed as representations of latent topics. Each document is described as a probability distribution over latent topics and each topic in turn is a distribution over a fixed set of words from the vocabulary. The LDA model learns the latent topics structure without use of background knowledge.

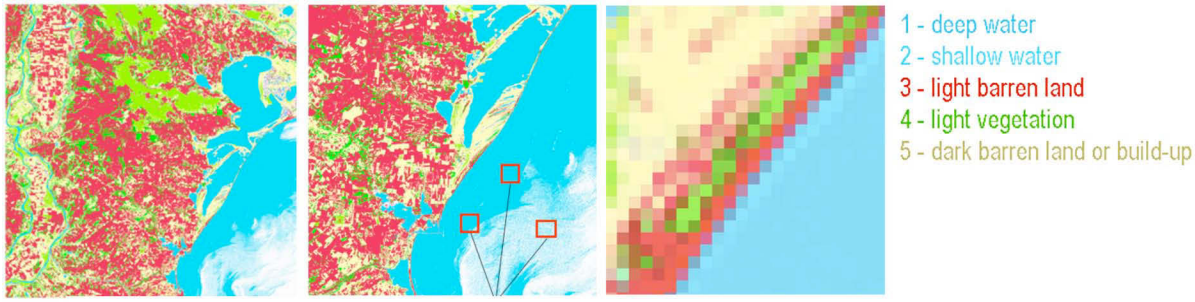


Figure 5.6 - Text – Image domains analogy

5.3.3 LDA Generative Process for Image Annotation

The visual documents are represented as a sequence of N visual-words from the visual vocabulary (i.e. spectral map). The LDA model discovers K latent topics and assigns one of the inferred topics to each pixel in the image. The number of latent topics is equal to the number of high-level semantic classes in CLC 2000. Therefore, by introducing a new layer of information between the words-level and the documents-level, each visual-document is represented as a probability distribution over the set of topics and each topic in turn as a probability distribution over the visual words in the vocabulary. The number of topics to describe the image is pre-defined by the user.

LDA assumes the following generative process for each image tile in the satellite scene:

1) Choose a K-dimensional Dirichlet random variable $\theta_i \sim Dir(\alpha)$, where K is the number of topics in the collection $K = \{5, 15, 44\}$ to correspond with the number of classes in the hierarchical semantic level of CLC data, θ_d is the topic distribution for document d and $i \in \{1,...,M\}$

2) Choose $\phi_k \sim Dir(\beta)$, where $k \in \{1...K\}$ and φ_k is the word distribution for topic k

3) For each of the word positions:

- Choose a topic $z_{d,w} \sim Multinomial\,(\theta_d)$
- Choose a word $w_{d,w} \sim Multinomial\left(\phi_{Z_{d,w}}\right)$

The multinomial distribution here is the multinomial with only one trial (categorical), as explained in the Appendix. The joint distribution of all known and hidden variables given the hyperparameters is:

$$p\left(\underline{W},\underline{Z},\underline{\theta}_d,\underline{\phi}\,|\underline{\alpha},\underline{\beta}\right)=\prod_{w=1}^{N_d}p\left(w_{d,w}\,|\,\underline{\phi}_{Z_{d,w}}\right)p\left(z_{d,w}\,|\,\underline{\theta}_d\right)\cdot p\left(\underline{\theta}_d\,|\,\underline{\alpha}\right)\cdot p\left(\underline{\phi}\,|\,\underline{\beta}\right) \tag{5.1}$$

This joint distribution is useful as the basis for other derivations. The probability that a word $w_{d,w}$ instantiates a particular term *t* from the vocabulary given the LDA parameters is obtained by marginalizing $z_{d,w}$ from the word plate and omitting the parameter distributions:

$$p\left(w_{d,w}=t\,|\,\underline{\theta}_d,\underline{\phi}\right)=\sum_{k=1}^{K}p\left(w_{d,w}=t\,|\,\underline{\phi}_k\right)p\left(z_{d,w}=k\,|\,\underline{\theta}_d\right) \tag{5.2}$$

The likelihoods of a document $\underline{W}_d$ and of the corpus $\underline{W}=\left\{\underline{W}_d\right\}_{d=1}^{M}$ are the joint likelihoods of the independent events of the token observations $w_{d,w}$:

$$p\left(\underline{W}\,|\,\underline{\theta},\underline{\phi}\right)=\prod_{d=1}^{M}p\left(\underline{W}_d\,|\,\underline{\theta}_d,\underline{\phi}\right)=\prod_{d=1}^{M}\prod_{w=1}^{N_d}p\left(w_{d,w}\,|\,\underline{\theta}_d,\underline{\phi}\right) \tag{5.3}$$

LDA has the flexibility to assign probabilities to documents outside the training corpus thus allowing supervised classification procedures over previously unseen documents. This property will be used to classify the entire satellite image and other images in the dataset. To infer the latent structures we used the software package described in [222].

5.3.4 Semantic Learning

In text and multimedia applications, the LDA model is used to infer latent topics from the distribution of words in the vocabulary or objects in the image as to obtain a small number of descriptors for the dataset. The objective is to bridge the semantic gap and is achieved by grouping the pixels in the satellite image so that the output map matches the classes with high-level semantics of the cartographic products (e.g. CLC 2000). The LDA model assigns pixels that follow a similarity criterion to a number of latent topics. By choosing the number of topics to be equal to the number of information classes, the pixels in the spectral map that belong to the same concept are ‘attracted’ to the same information class in CLC.

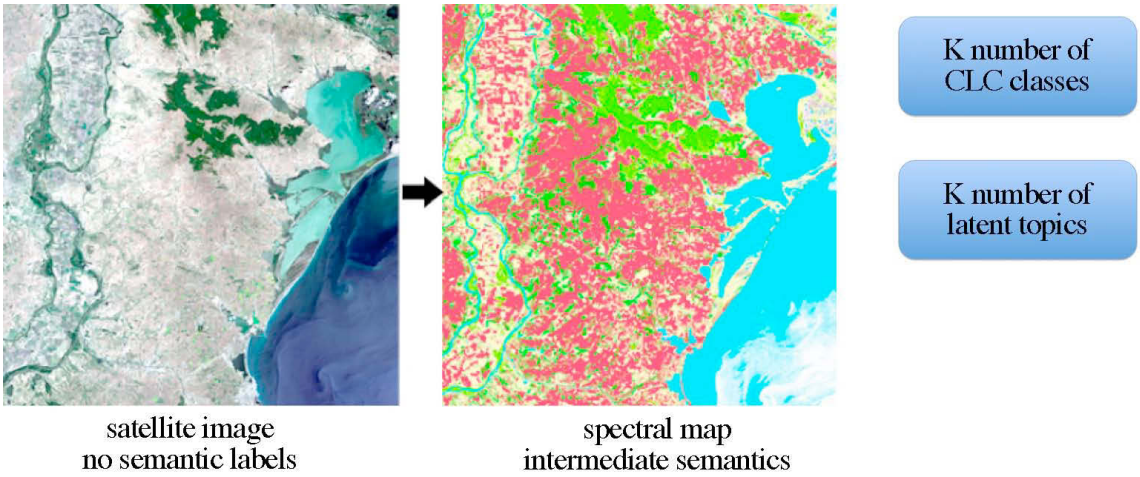


Figure 5.7: Correspondence latent topics – CLC information classes. The number of concepts in the image should be equal to the number of classes in the cartographic product.

During the training phase we defined a number of visual documents to learn the topics observed in the satellite image. Due to restrictions in terminology, a number of 5, 15, 44 topics were chosen in consent with the number of information classes found in the technical specifications of CLC 2000, for the three hierarchical levels, as shown in figure 5.7.

For precise assessment of the training documents, the vector data was overlapped on the satellite image and the spectral map to choose image patches underlying each of the CLC classes. The size of the image tiles follows the rule of the minimum mapping unit of the cartographic data. For CLC 2000, the minimum mapping unit is 25 ha corresponding to a visual document of 15 X 15 pixels (Landsat images) and 50 X 50 pixels (SPOT images). For optimum results, the number of training histograms fed to LDA should be balanced and the user should provide closely the same number of training data sets for each topic.

20 visual documents for each CLC class were enough for learning. Figure 5.8 describes one step of the learning process. The user selects the training documents belonging to the desired topic to annotate and the LDA model generates the distribution of that topic.

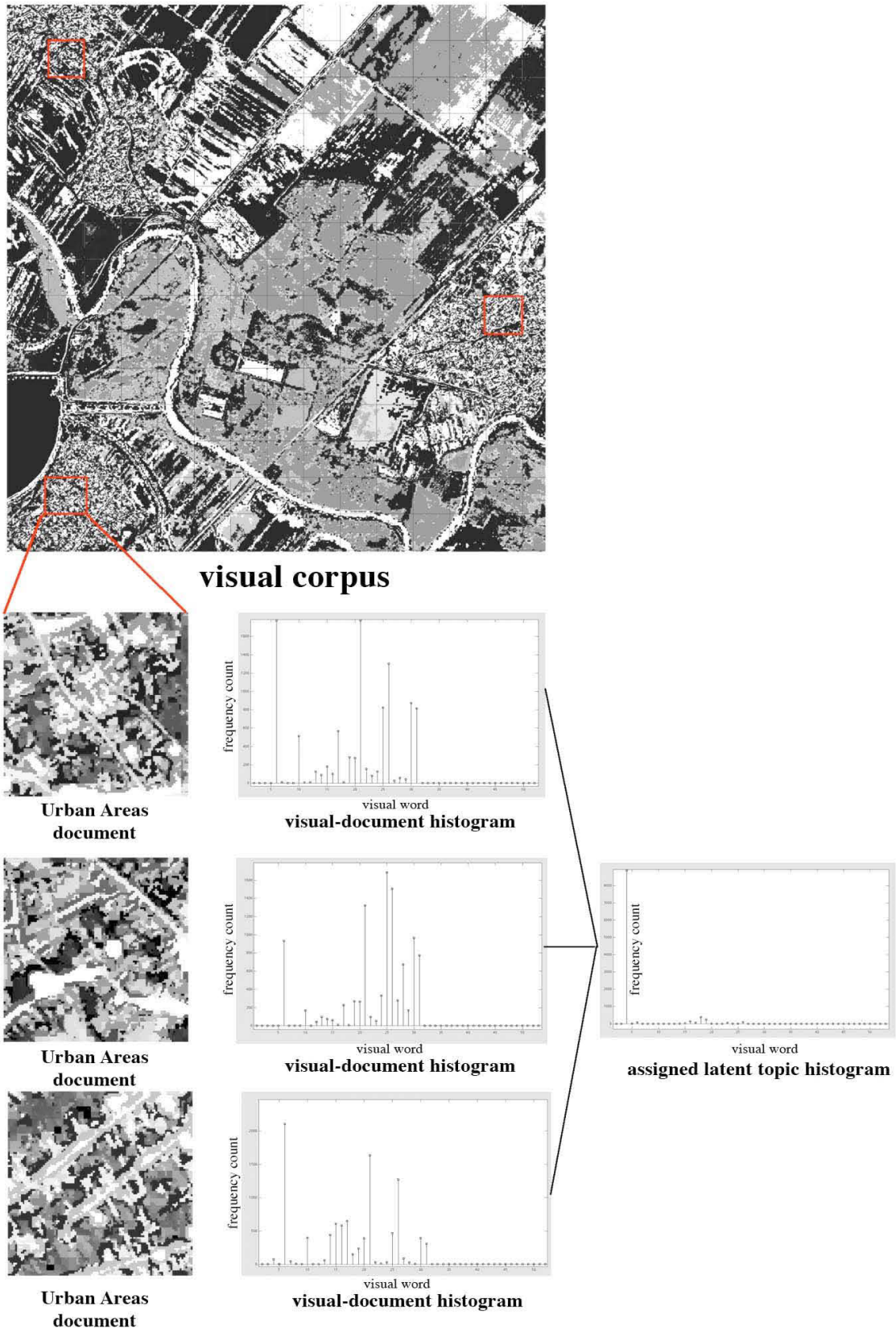


Figure 5.8: Learning step of *Urban Areas* class. The histograms are used as input to the LDA model and a latent topic is generated from the documents.

In text classification the purpose is to classify each document into two or more mutually exclusive topics and to use the topics to index the large collection of documents. The annotation of the satellite image consists of classifying each document to one of the classes in CLC and to generate several descriptors that bridge the semantic gap. Variations of this method may be applied to create descriptors for large collections of multimedia images or application-specific image databases, such as forensics or medical.

Learning using LDA achieves a model that best represents the distribution of visual words for each class (topic). With LDA it is possible to assign a probability to each image tile in the dataset, corresponding to the likelihood of that image tile for a specific class. The classification of the entire dataset is achieved via Maximum Likelihood. The algorithm assigns the image tile to the class that maximizes the likelihood $s = \text{argmax}\, p(z_d \mid \alpha, \beta)$

Collapsed Gibbs sampling [222] is used to infer the latent structures and the probability that bridges the semantic gap between the image clusters and the high-level semantic classes .

5.4 Composition Rules For Bridging The Semantic Gap

In the training phase a number of 20 visual documents / class are selected from the satellite image to represent every high-level semantic class that is to be described by the latent topics. The histograms of the training documents are used as input to the LDA model and global parameters α and β are inferred. Parameter α yields the distribution of topics over the corpus and parameter β the distribution of visual words over each latent topic. To link the spectral map to the output of the LDA model, we use only β that provides information on the probability of each topic to generate a specific word in the vocabulary.

$$\text{word } i \rightarrow \max(p(z_d \mid w_i)) \tag{5.4}$$

Figure 5.9 shows the distribution of words in the latent topics. A number of five topics were inferred from the satellite image and the visual representation of the latent aspects over the image is shown in figure 5.10.

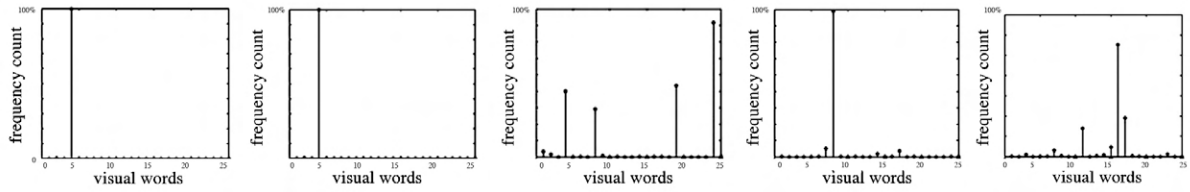


Figure 5.9: Distribution of visual-words in the latent topics 1 to 5

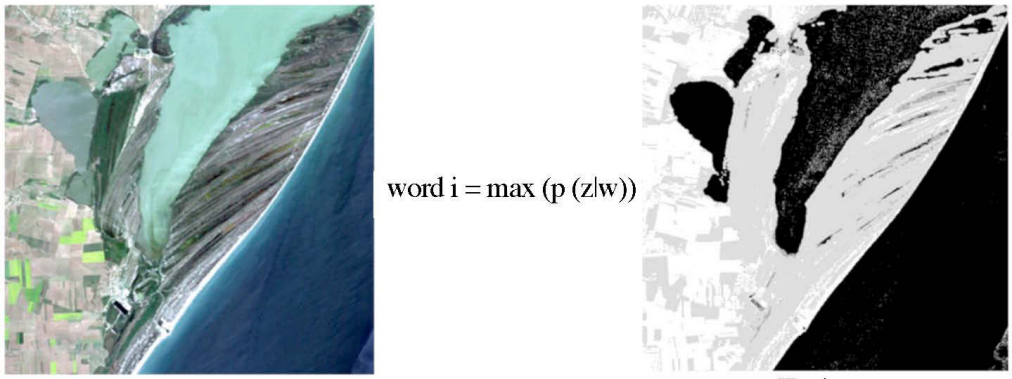


Figure 5.10: Mapping all pixels in the satellite image to one of the five latent topics

The goal was to discover the semantic rules that link the topic map to the high-level semantic classes in cartographic products. Experiments were performed on multiple sensors (e.g. Landsat ETM+ and SPOT 5) that discovered the semantic rules that bridge the semantic gap with the ontology of CLC 2000.

In the first case studies, the goal was to describe the first hierarchical level of CLC 2000 vector data. This level has 5 classes: *Water Bodies, Wetlands, Agricultural Areas, Artificial Surfaces, Forests and Semi-natural Areas*. Initially, the spectral map was generated with Soilmapper and 100 training visual-documents (20 for each semantic class) were chosen. The training histograms were used as input to LDA. 1600 visual-documents in the image were classified to one of the latent topics. The dimension of the Dirichlet random variable was set to match the number of vector classes in CLC (e.g. 5, 15, 44). The latent topic map was generated by replacing each pixel with the topic that was assigned to it with the maximum probability $\max(p(z_d \mid w_i))$

In the final step a majority filter was applied to exclusively assign each document to a single topic. We performed tests on a number of Landsat and SPOT images, yielding different complexity structures in land use and land cover classes. The images to be annotated were chipsets from Landsat ETM7+ (600 X 600 pixels) and documents of 15 X 15 pixels. With SPOT 5, the chipsets selected had 2000 X 2000 pixels, with 50 X 50 pixels / document.

5.5 Case Studies: Rules For Bridging Machine And Human Languages

In the following case studies, a visual word corresponds to an index from the preliminary spectral map obtained in the pre-classification phase using the physical model. Each value in the visual documents corresponds to an index from a semantic table with intermediate level meaning (e.g. strong vegetation, barren land, shallow water, snow), value able to describe the high-level semantics in existent cartographic data (e.g. CLC: *Agriculture Land, Forests and Natural Areas, Water Bodies*). The number of visual words was established by the classification scheme adopted in the physical model.

CASE STUDY 1 – Semantic rules for the first hierarchical level in CLC 2000, Landsat

This study was performed on Landsat ETM 7+ image (600 X 600 pixels) from Romania aiming to infer the semantic rules that link the spectral map to the CLC 2000 map. Figure 5.11 presents the results. The information in the multispectral image was reduced to the spectral map with 27 visual-words and then a number of five latent topics have been discovered in the data. Each pixel in the spectral map was assigned with the topic of maximum probability and the index map generated. In the end, a majority filter was applied to assign each document (15 X 15 pixels) to an exclusive topic. Figure 5.12 shows the distribution of visual words within each of the five estimated latent topics.

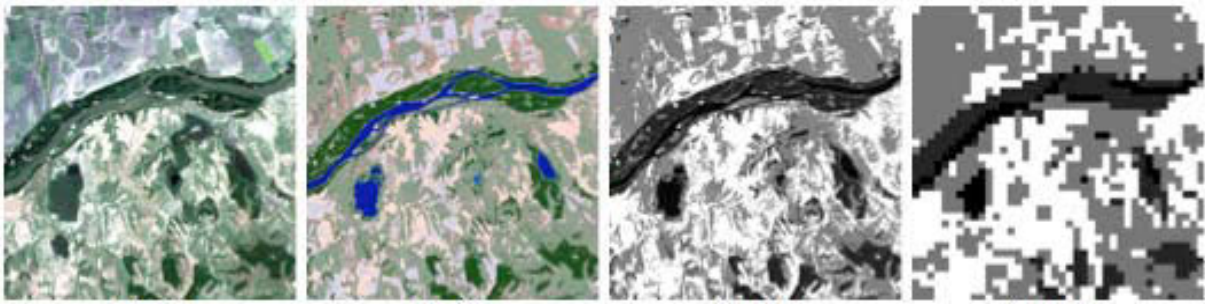


Figure 5.11a, 5.11b, 5.11c, 5.11d – Fig. 5.11a shows the Landsat image 600 X 600 pixels, fig. 5.11b shows the spectral map with 27 visual words, fig. 5.11c shows how each pixel is classified to one of the latent topics (topics map); fig. 5.11d shows how each document is classified to the one of the topics.

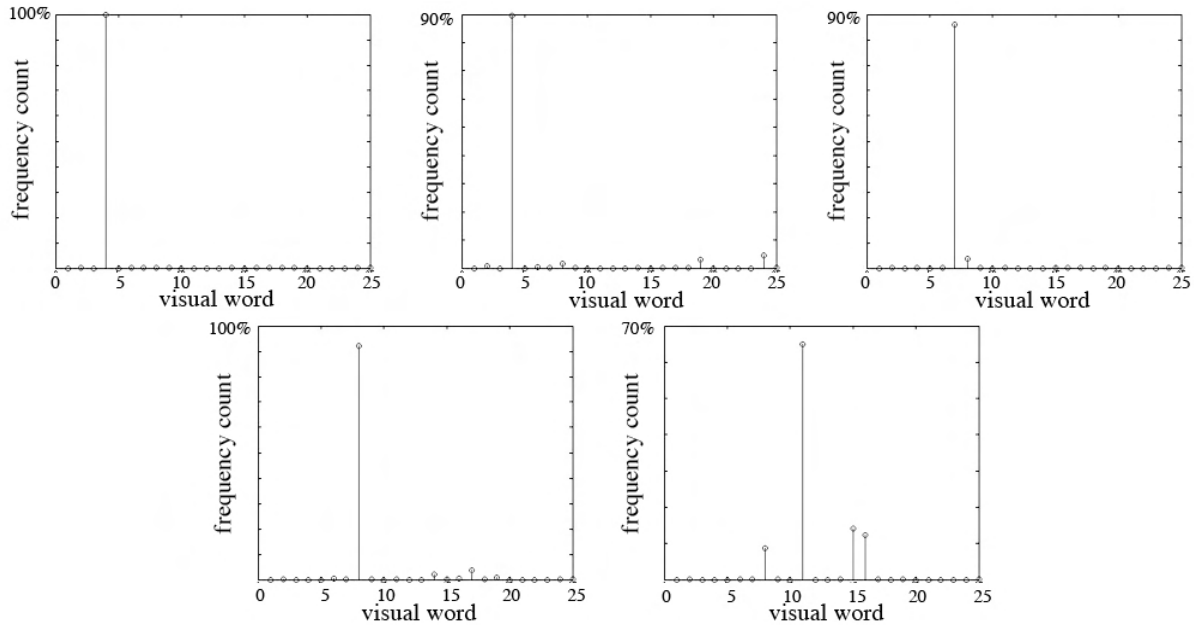


Figure 5.12 – The five latent topics estimated from the satellite image are distributions over words in the vocabulary

	Water bodies	Agricultural areas	Artificial Land	Forest and natural areas	Wetlands
Topic 1	73%	0%	0%	0%	0%
Topic 2	13%	0%	0%	0%	0%
Topic 3	0%	2%	0%	54%	15%
Topic 4	14%	53%	32%	36%	80%
Topic 5	0%	45%	68%	10%	5%

Table 5.1: Semantic Rules Discovery

Topic 1: 100% deep clear water
Topic 2: 90% deep clear water, 10% turbid water
Topic 3: 100% peat bogs
Topic 4: 100% average vegetation
Topic 5: 70% average shrub land, 20% bright barren land,
20% strong barren land, 10% average vegetation.

	Agriculture	Artificial	Forest	Water
Agriculture	97%	25%	0%	0%
Artificial	0%	75%	1%	0%
Forest	3%	0%	99%	1%
Water	0%	0%	0%	99%

Table 5.2: Confusion matrix for the semantic rules

Table 5.1 shows the distribution of latent topics over the classes in CLC 2000 and the semantic rules that bridge the intermediate-level semantics with the high-level information classes. The ideal case would be that each topic exclusively generates a single CLC class. Validation of results is assessed with the confusion matrix in Table 5.2 and the mean overall accuracy obtained is 92.5%.

The *Water* class is generated with a precision of 99% from the first two topics (Topic 1 - Deep Clear Water and Topic 2 – Deep Clear Water and Turbid Water). The *Forest* class is generated with 99% precision from the last three topics in table 5.1. Confusions between the *Agriculture* and *Artificial Land* classes arise due to the fact that both of these high-level semantic classes are described with similar intermediate visual-words (e.g. light barren land, average barren land, average vegetation).

CASE STUDY 2 - Semantic rules for the first hierarchical level in CLC 2000, Landsat

This study was performed on a Landsat image (600 X 600 pixels) from Romania to find the rules that explain the first hierarchical level of CLC 2000. Figure 5.13 shows the results at different steps in the workflow. The information in the image was reduced to the spectral map with 27 visual-words and a number of five latent topics was estimated, as shown in figure 5.14. Each of the pixels in the spectral map was assigned with a topic of maximum probability and the index map generated. In the end, a majority filter assigns each document (15 X 15 pixels) to an exclusive topic.

Figure 5.13a, 5.13b, 5.13c, 5.13d – Fig. 5.13a shows Landsat image 600 X 600 pixels, fig.13b shows the spectral map with 27 visual words, fig. 5.13c presents how each pixel is classified into one of the latent topics and fig. 5.13d shows how each document is classified into one of the topics.

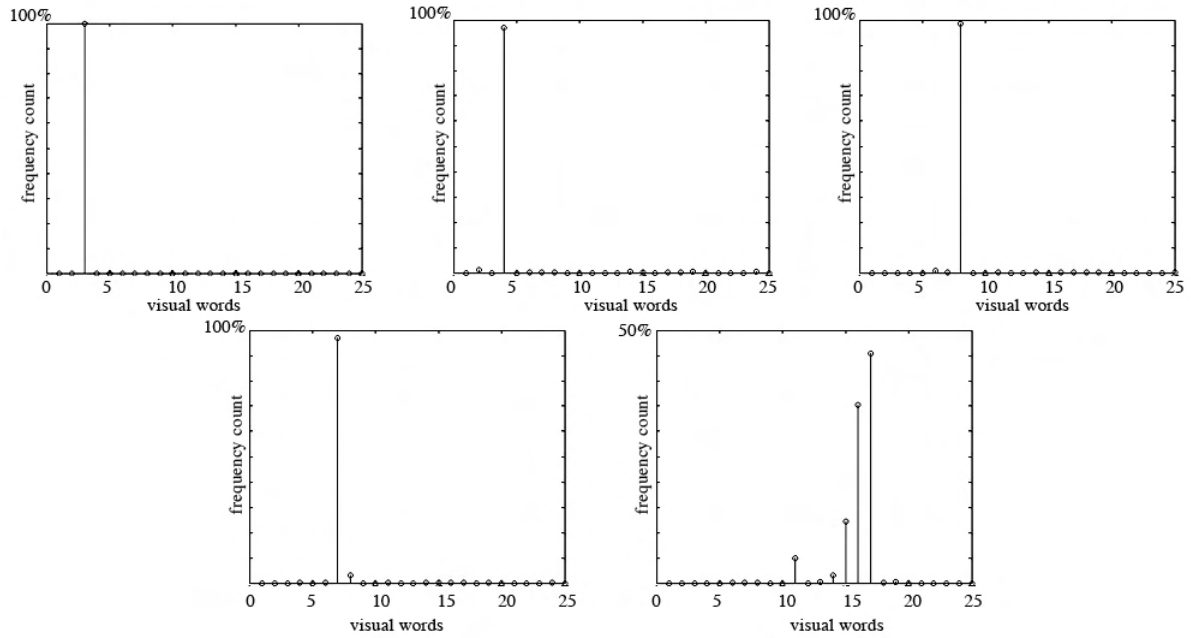


Figure 5.14: distribution of visual words on the 5 inferred topics

	Agriculture	Artificial Land	Forest	Water	Wetlands
Topic 1	0%	0%	0%	0%	0%
Topic 2	0%	0%	0%	88%	0%
Topic 3	52%	25%	51%	12%	80%
Topic 4	3%	0%	49%	0%	20%
Topic 5	45%	75%	0%	0%	0%

Table 5.3: Semantic rules discovery

Topic 1: 100% ice or snow
Topic 2: 100% deep clear water
Topic 3: 100% average vegetation
Topic 4: 100% strong vegetation
Topic 5: 50% average barren land, 40% strong barren land, 10% light barren land

Table 5.3 shows how each class in the first hierarchical level in CLC 2000 is described by latent topics as inferred from the LDA model. Table 5.4 shows the confusion matrix for the mapping results. There is a major error with Topic 1 because the physical model misclassified *Water* pixels as Ice & Snow but the LDA model didn't employ it in any of the following steps. *Agriculture* class is easily described by a half-half combination of Topic 3 (average vegetation) and Topic 5 (barren land - average, light and strong). The *Forest* class is explained by Topics 3 and 4 (Average and Strong Vegetation). *Water* class is described with an accuracy of 88% by Topic 2 (labelled as Deep Clear Water) and the errors come from the fact that 15 X 15 pixels documents don't overlap perfectly with the CLC vector class.

	Agriculture	Artificial	Forest	Water	Wetlands
Agriculture	97%	25%	0%	0%	80%
Artificial	0%	75%	1%	0%	0%
Forest	3%	0%	99%	1%	20%
Water	0%	0%	0%	99%	0%
Wetlands	0%	0%	0%	0%	0%

Table 5.4 – Confusion matrix for semantic rules

CASE STUDY 3 - Semantic rules for the first hierarchical level in CLC 2000, Landsat

This experiment was performed on a Landsat image (600 X 600 pixels) from Romania. Figure 5.15 shows the results at different steps of the workflow. Initially the information in the multispecrtal image was reduced to the spectral map with 27 visual-words and then a number of five latent topics was estimated for the classes in the first hierarchical level of CLC 2000. Each of the pixels in the spectral map was assigned to a topic of maximum probability and the index map generated. In the end, a majority filter assigns each document (15 X 15 pixels) to an exclusive topic.

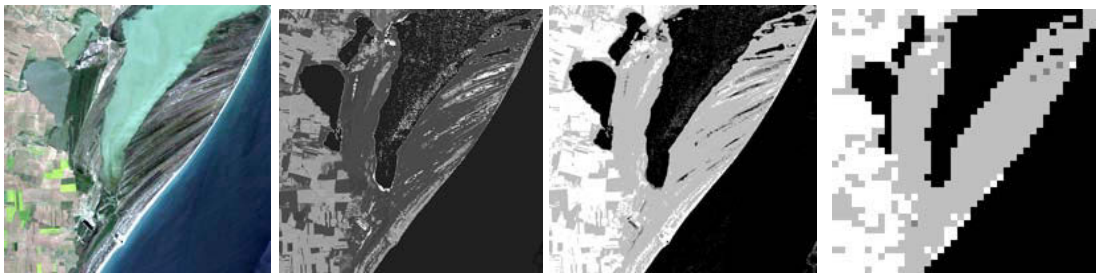


Figure 5.15a, 5.15b, 5.15c, 5.15d –Fig. 5.15a - Landsat image 600 X 600 pixels, fig. 5.15b shows the spectral map with 27 visual words, fig. 5.15c presents how each pixel is classified into one of the latent topics and fig. 5.15d shows how each document is classified to one of the topics.

	Agriculture	Artificial	Forest	Water	Wetlands
Topic 1	0%	0%	0%	93%	0%
Topic 2	0%	0%	0%	0%	0%
Topic 3	1%	4%	4%	5%	6%
Topic 4	27%	54%	82%	2%	88%
Topic 5	72%	42%	14%	0%	6%

Table 5.5: Semantic rules discovery

Topic 1: 100% deep clear water
Topic 2: 100% deep clear water
Topic 3: 40% turbid water, 25% shadow vegetation, 25% deep clear water
Topic 4: 90% average vegetation, 10% strong vegetation
Topic 5: 60% strong barren land, 20% average barren land

Table 5.5 presents how each class in the first hierarchical level in CLC 2000 is described by the latent topics inferred by the LDA model. Table 5.6 shows the confusion matrix computed to validate the classification results. The *Water* class was generated using topic 1 (Deep Clear Water 93%) and topic 3 (Turbid Water, Shadow Vegetation and Deep Clear Water). The *Forest* and *Agriculture* classes are described by topics 4 and 5 (Strong vegetation, Average vegetation, Strong and Average barren land). There are some confusions with the *Wetlands* class because it was generated from the same topics and visual words as the *Forest* class (water and vegetation).

	Agriculture	Artificial	Forest	Water	Wetlands
Agriculture	99%	0%	0%	0%	0%
Artificial	0%	96%	14%	0%	0%
Forest	0%	0%	82%	2%	88%
Water	1%	4%	4%	98%	6%
Wetlands	0%	0%	0%	0%	6%

Table 5.6 – confusion matrix for semantic rules

CASE STUDY 4 - Semantic rules for the second hierarchical level in CLC 2000, Landsat

This experiment was performed on a Landsat image (600 X 600 pixels) from Romania. Figure 5.16 presents the results at different steps in the workflow. The process is similar to the previous ones. The information in the satellite image is reduced to a spectral map with 27 visual words and then a number of 15 latent topics estimated. The number of topics corresponds to the number of classes in the second hierarchical level of CLC 2000. Each of the pixels in the spectral map was assigned with a topic of maximum probability and the index map generated. In the end a majority filter assigns each document (15 X 15 pixels) to an exclusive topic. The semantic rules are presented in table 5.7.

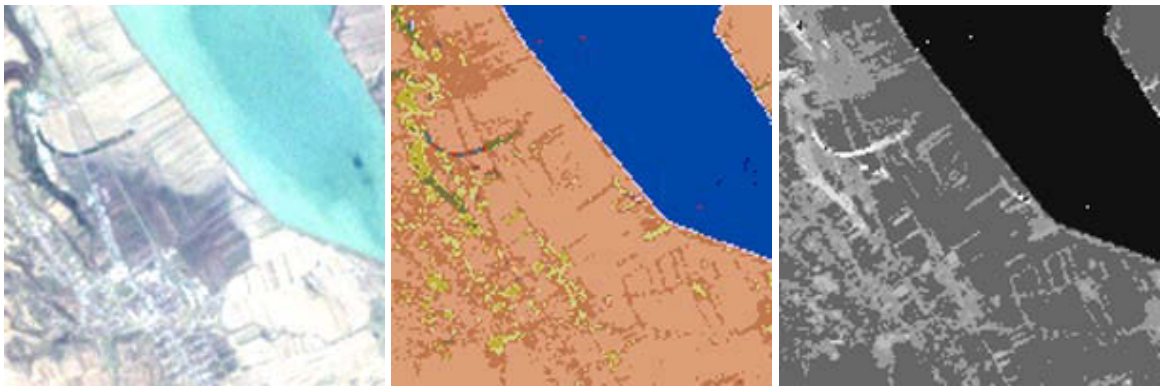


Figure 5.16a, 5.16b, 5.16c – Fig 5.16a - Landsat image tile, fig. 5.16b shows the spectral map with 27 visual words, fig. 5.16c presents how each pixel is classified into one of the 'latent' topics.

	Arable Land	Forests	Heterogeneous Areas	Industrial Commercial	Inland Waters	Inland Wetlands	Pastures	Permanent crops	Urban fabric
1	0%	0%	0%	0%	98%	4%	0%	0%	0%
2,5	0.4%	0%	0%	0%	1%	19%	0%	0%	0%
6,7	74.5%	13%	56%	30%	0.5%	36%	33%	48%	48%
8	0%	0%	0%	0%	0%	0%	0%	0%	0%
9	20%	25%	32%	41%	0.5%	27%	65%	35%	41%
10	2.6%	17%	4%	20%	0%	4%	1%	9%	8%
11	0.8%	20%	6%	7%	0%	0%	0%	8%	3%
12	0%	0%	0%	0%	0%	0%	0%	0%	0%
13	0.7%	25%	2%	2%	0%	10%	0%	0%	0%
14	0.6%	0%	0%	0%	0%	0%	0%	0%	0%
15	0.4%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.7: Semantic rules discovery

Topic 1: 100% Shadow or turbid water
Topic 2: 40% Thin Cloud on water areas, 30% strong barren land or built-up, 20% shadow or turbid water, 10% average barren land or built-up
Topic 3: 90% strong barren land or built-up, 10% average barren land or built-up.
Topic 4: 100% strong barren land or built-up
Topic 5: 100% strong barren land or built-up
Topic 6: 100% strong barren land or built-up
Topic 7: 70% strong barren land or built-up, 30% average barren land or built-up
Topic 8: 60% average barren land or built-up, 40% strong barren land or built-up
Topic 9: 70% average barren land or built-up, 20% weak rangeland leaf, 10% strong barren land or built-up
Topic 10: 40% weak rangeland leaf, 25% average barren land or built-up, 25% average barren land or built-up, 25% average shrub rangeland, 10% strong barren land or built up.
Topic 11: 50% average shrub rangeland, 20% strong barren land, 20% average barren land or built-up, 10% weak rangeland leaf
Topic 12: 55% average barren land or built-up, 30% wetland
Topic 13: 45% wetland or dark rangeland leaf, 25% average barren land or built-up, 15% strong barren land or built-up, 10% weak rangeland leaf
Topic 14: 55% dark barren land or built-up, 30% average barren land or built-up, 15% strong barren land or built-up.
Topic 15: 55% dark barren land or built-up, 30% average barren land or built-up, 15% strong barren land or built-up.

The second hierarchical level in CLC 2000 contains 15 semantic classes, generated from the first hierarchical level with only five. The results show less acurracy than in the previous cases due to the increased complexity of the vector classes to be generated from a limited vocabulary of 27 words.

Conclusions

This chapter offered a solution to bridge the semantic gap and discover the rules between the output of the state-of-the-art automatic classifiers and the high-level semantics of manually defined terminologies of cartographic data. Using a purely-spectral rule-based fully automatic classifier to define the basic visual vocabulary, the method provides a hybrid approach to automatically understand and describe the semantic rules that connect existent mapping data with different specifications (e.g. CORINE LAND COVER) to the end-results of unsupervised information mining methods.

Using a tool initially developed for statistical text modelling in large documents collections – Latent Dirichlet Allocation LDA we discovered a correspondence between the text and image domains and linked the low-level spectral and spatial features to the spectral map with intermediate-level semantic labels (e.g. strong vegetation, barren land, deep water) and to the vector maps with application specific semantic labels.

6

Spectral Band Discovery for Advancing Multispectral Satellite Image Analysis and Photo-Interpretation

Almost all remote sensing applications involve several steps of visual inspection – data quality assessment, operation-oriented area/object search and analysis, algorithm learning, information mining evaluation, etc. Even if today's geo-information software packages available offer satisfactory end products, most tasks are still manually performed and evaluated by a human operator. Multiple domains require highly accurate analysis of satellite images and the demands of users working in these areas are so challenging that machines have not yet reached the quality standards required for these applications. For this reason, data analysis is performed through extensive visual interpretation.

Because the manual processes performed by experts to extract information from images are currently too complex to be applied systematically on even a small subset of the acquired scenes [242], next-generation software tools will have to be designed to support the human operator in his work, put control into his/her hands and optimize visual investigations.

This chapter presents a novel sensor-independent, spectral-based, one-sample based training, spectrally and spatially balanced, application-free, fast response, low cost, information-based spectral band selector that automatically enhances visualization of target classes for image analysis and photo-interpretation [226]. Computer assisted visual analysis is an extremely important approach to information discovery, extracting, mapping and reporting.

6.1 Exploratory Visual Analysis of Satellite Images

Multispectral and hyperspectral images contain multiple types of signatures that allow analysts to identify a wide range of activities (e.g. environmental, urban, marine) and objects based on their unique spectral "fingerprints". Hyperspectral, recent high-resolution (e.g. ESA Sentinel-2, WorldView-2) and the classic Landsat multispectral data offer flexibility to visualize and extract information about many classes, objects and targets of interest using these fingerprints.

The goal of the method described hereafter may be viewed from two perspectives: (1) the algorithm selects from the available spectral bands only the ones containing the highest amount of information relevant only to the target class and (2) its final goal is to maximize contrast and color difference to the surrounding classes.

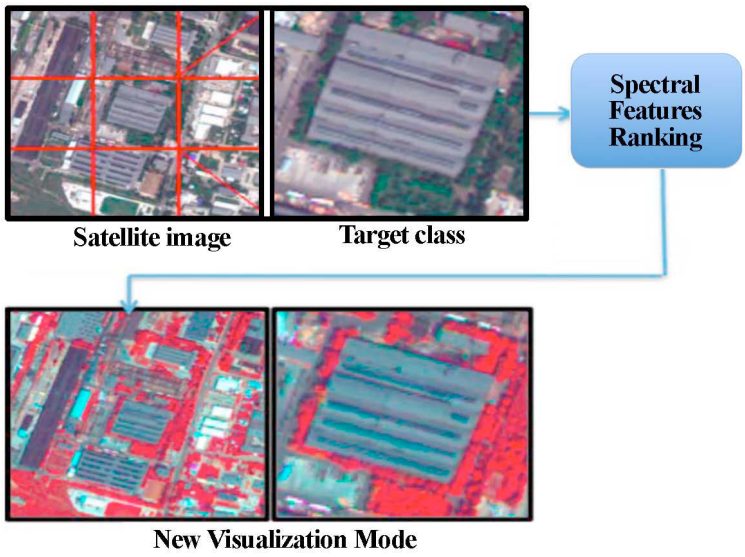


Figure 6.1 – Concept and workflow for the feature selection algorithm in advancing satellite image visualization

The modus operandi of the feature selection algorithm is presented in figure 6.1. The satellite image is imported into the system and a regular grid is overlaid on top of it. The equally sized grid solves the problem of training samples imbalance by forcing the algorithm to operate in a spectrally and spatially balanced mode [230]. This ensures training classes of equal dimensions and also focuses the algorithm on the enhancement of local color and contrast for visual exploration. The user simply clicks once on the image patch labeled as target class (one-sample learning). Measures of mutual information are automatically employed between the target class and the available spectral bands (information-based). The spectral features are ranked based on their capability to represent the target class in a mining algorithm and to enhance its visual separability from other classes. The bands are ranked using the minimum-redundancy-maximum-relevance (mRMR) criterion [231, 232]. The mRMR criterion evaluates the statistical dependency of the target class to the spectral bands, measured in terms of mutual information, and simultaneously minimizes interband redundancy. The image analyst can choose any area of the image as target class and discover the optimum spectral bands that enhance visualization for that region or object (application free). Although the mRMR criterion is based on complex measures of mutual information, the computation time and effort are reduced because the system operates on discrete image histograms (fast response, low cost). The number of spectral bands that can be evaluated is

unlimited and the workflow is identical for any multi-band sensor data (sensor independent). The top three bands evaluated and ranked with the mRMR scores are automatically displayed for preview. Before integrating the methodology into the scientific and operational framework, several definitions are required.

The primary purpose of satellite imagery is to describe, assess and visually depict physical features and activities on Earth [243]. Direct visualization of objects and classes of interest leads to discovery of information through analysis and interpretation for various tasks [242]. An important goal of satellite image understanding approaches is to present a comprehensive visual depiction of the imaged scenes: "A primary purpose of geospatial products has been to provide visualization of operational spaces and patterns of all sizes and scales ranging from global to regional level, to cities and buildings. A picture is simply the fastest way to communicate spatial information" [243, p.24].

However, the complexity of data recorded by satellite imaging sensors is so high that discovery of knowledge through visual analysis is a challenging cognitive activity. For this reason, visualization tools should be capable of assisting the human operator in understanding the data through optimum representations and to offer cognitive support in discovering relevant information in the scenes [244]. The process of exploratory visual analysis has been studied across many scientific fields [245-259] but further research is required in this direction [260].

6.2 Contextual Information Integration For Spectral Feature Selection

In the paper "Perceptual principles for effective visualizations", Rheingans and Landreth [261] evaluate the effect of contextual information to the perception and understanding of an object in a scene. They conclude that (1) perception of size of an object may be influenced by its colour; (2) perception of color hue may be influenced by saturation; (3) perception of color saturation may be influenced by hue; (4) perception of color of an object may be influenced by the color of surrounding objects.

To integrate human-derived contextual knowledge into the automatic feature selection algorithm, the methodology presented in this paper utilizes patch-level analysis. In meter and sub-meter resolution images, image patches interconnect complex structures (objects) with high diversity of spectral information [262], [10], [226].

Figure 6.2 – Contextual information implies studying the spectral statistics of a spatially delimited environment. An image patch contains a high diversity of spectral information.

Patch-level analysis implies investigating sub-scenes of equal size (regular grid) using the spectral statistics within a local spatial context, as shown in figure 6.2. The variance of the spectral information of recorded compact objects increases with the spatial resolution of the image. On one hand for example, in urban analysis, the roof of a building may present regions of different colors, depending on illumination, shadows and other factors, although at conceptual level it is a single object. On the other hand, the spectral characteristics of urban land cover classes may present high similarity and cannot be separated using only spectral information [213] – e.g. roads and buildings with similar signatures. People perceive and integrate contextual information to recognize objects in the scene [212] and in some cases the objects are detected only by using the contextual information even though the appearance of the objects themselves is withheld. This effect is called blind recognition. Satellite images have high complexity of structures in the scenes and the higher the complexity the greater the likelihood of it benefitting from the context [211]. A scene is a full satellite multiband product covering hundreds of square kilometres.

Patch analysis implies extracting and separately analyzing the spectral signatures in a spatially limited environment. Because the mRMR criterion discovers the most relevant spectral features using complex measures of mutual information, it requires a window of analysis large enough to contain a well-defined histogram and small enough in order for the distribution to characterize only the target class. If the analyzing window is too large, the probability density function becomes uniform, decreasing the capability of the feature selection algorithm to rank the spectral bands relevant to the specific target class.
This approach lines up to the way EO emergency centers create cartographic products for multiple applications (e.g. maps for emergency response, geo-intelligence). Another motivation for the patch-level analysis of satellite images is the similarity of the method to the quadrant analysis of maps, widely used in GIS. The quadrant analysis works by dividing the area of interest into cells of equal size and analyzing the statistics of each patch independently [210].
Image analysis and interpretation are essential processes through which information is obtained from satellite images. Several researchers support the contextual information integration reasoning included in this paper: “the degree of accuracy and completeness of image interpretation are in large measure dependent upon the experience base of the observer with reference to the context within which the interpretation is occurring” [35, p. 987].

The single most important parameter of the patch-level analysis that influences the accuracy of the feature selection and mining algorithms operating at patch-level is the size of the analyzing window. The patch size and the resolution of the information within the patch are critical to understand the level of contextual detail that will be modelled. The size of the tile is a function of the spatial resolution of the image and is directly related to the user's choice for the target class. In order to fully benefit from the integration of contextual information in a feature selector and an automatic classifier, the size of the patch must be chosen so that it contains only the concept class of interest. If the size is too large and the patch contains information about other irrelevant classes, this information will be automatically evaluated as relevant and the accuracy will decrease. The size of the grid cells is 50 X 50 pixels for images with spatial resolution under 5 meters (WorldView-2, Quickbird, RapidEye, GeoEye-1) and 100 X 100 pixels for images with medium resolution (Spot-5, Landsat ETM 7+) to ensure the homogeneity of concept classes inside each cell. The user can define the size of the analyzing window using a slide-bar and operate this method for multiple spatial resolutions, multiple sensors and multiple applications. Figure 6.3 depicts an optimal cell (a) and a suboptimal cell (b).

Figure 6.3 – Grid cells size: (a) optimal, (b) suboptimal
(a) the patch contains solely the concept class of interest
(b) the patch contains information about other classes

6.3 Minimum-Redundancy-Maximum-Relevance Criterion For Feature Selection

In information-mining applications, feature selection is a critical step in optimizing a decision condition - in our case maximizing visualization for the target class. Having available the data set D, described by M features $X = \{x_i, i = 1..M\}$ and the target class C, the feature selection problem is to discover a subspace of three features in the spectral space R^M that best displays C with respect to its surroundings. The optimal characterization condition – i.e. what features best represent the target class – most of the times implies an extreme value for the decision function. For visualization, the function is evaluated both qualitatevely by the human operator and quantitavely using the color difference and contrast between the target area and its neighbours.

The feature selector consists of two steps [227] – the choice of a suitable criterion to evaluate the effectiveness of each spectral band (feature ranking) and the automatic display of the top three bands in order: R, G, B channels. Feature ranking assigns features with a score by a metric measure and eliminates all features that do not achieve the score over an user-defined threshold. The criterion used to rank the features in rapport with the target class is the minimum-redundancy-maximum-relevance (mRMR).

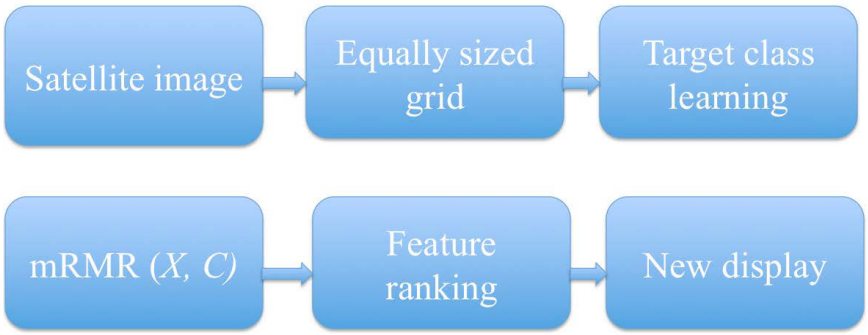


Figure 6.4 – Hybrid spectral feature selection workflow for advancing satellite image visualization

Figure 6.4 depicts the spectral feature selection workflow for automatic discovery of the optimum spectral bands. Spectral features are ranked based on their capability to represent the target class in a mining algorithm and to enhance its visual separability from other classes. In information theory, these properties translate to maximization of statistical dependency of the target class C to the available features.

The spectral bands are regarded as random variables and their histograms are the probability density functions. Mutual information is a good indicator of relevance between two random variables. If the mutual information between two random variables (i.e. in our case between the spectral bands and the target class) is large it means the two variables are closely related. Thus, mutual information between the target class and the spectral features is a relevant measure of statistical dependency.

$$Dependency(X,C) = I((x_i, i = 1...M), C\}) \tag{6.1}$$

A common way to obtain maximal dependency is to use the maximum relevance (MR) criterion – ranking the features by their statistical relevance to the target class C. The relevance of features R^m to the features in C is defined in terms of mutual information. Maximum relevance criterion (MR) is employed in practical applications because maximum dependency is often hard to implement even for discrete random variables. The MR criterion evaluates the dependency between multiple random variables by approximating $\max(I(x_i, i = 1...M), C)$ with the mean value of all mutual information values between individual features x_i and class C. | X | represents the number of features available.

$$D(X,C) = \frac{1}{|X|}\sum_{x_i} I(x_i, C) \tag{6.2}$$

The spectral bands of satellite images may be correlated. When two or more random variables have a rich content of mutual information, the discrimination power will decrease. To select only the mutually exclusive features and to reduce the dimensionality of the data, the redundancy (mR) scores between every pair of features are computed with:

$$R(X) = \left(\frac{1}{|X|^2} \sum_{x_i, x_j} I(x_i, x_j) \right) \tag{6.3}$$

To rank the features yielding minimum redundancy (mR) between them with the simultaneous maximization of the discriminating power (MR) for the target class, the above steps (6.2) and (6.3) are combined into the minimum-redundancy-maximum-relevance scores (mRMR) (6.4), where *D* is the maximum dependency and *R* is the minimum relevance:

$$mRMR = \max(D - R) \tag{6.4}$$

The mRMR spectral feature ranking criterion is implemented as a Matlab package, using (6.5):

$$\max \left[I(x_j, C) - \left(\frac{1}{m-1} \right) \sum_{x_i \in S_{m-1}} I(x_j, x_i) \right] \tag{6.5}$$

The calculus gives the mRMR score for each available feature with respect to the learned target class. The top three bands in the mRMR score - yielding maximal dependency to the target and minimum redundancy among them - are automatically displayed in the R, G, B channels. The algorithm is implemented using this pseudocode:

Input: Spectral bands for the scene, X = 1 to M

Spectral bands for the target image patch C = 1 to M

for *i* = 1 to *M*

D = 2-D mutual information $\left(X_i, C_i\right)$

R = 2-D mutual information $\left(X_i, X_j\right)_{i \neq j}$

$Score(i) = mRMR = \max(D - R)$

end

Output: order, top-down $Score(i)$

Display, top-three spectral bands in R,G,B channels

6.4 Objective Evaluation Of Subjective Visual Information

For visualization to be scientific it has to be able to generate a collection of methods, techniques and tools developed to fulfill a technical request for which standard measures apply. Visualization has to be effective and efficient [233]. Image I is perceived by a user, with an increase in knowledge K as a result. The amount of knowledge gain is a function of the image I, the a priori knowledge K_0 of the user and the particular properties of the perceptual and cognitive abilities P of the user. How can one evaluate the increase in knowledge K that this new display creates, as compared to the standard R-G-B display? The goal of visualization methods – either by image processing techniques or by selecting specific spectral bands – is to maximize the response in the human visual system and increase the saliency of the object / area of interest. The reader may refer to [35, p. 988-992] and [268] for a detailed description of the mechanics of human vision.

6.4.1 Visual Image Analysis – Elements of Processing

Perceiving and sensing are distinct. Sensing implies only the recording of specific flows of radiation while perceiving also integrates and is subject to the influence of learning. Space borne sensors record the reflected, emitted, transmitted and scattered energy across several regions of the electromagnetic sensor. The data are represented unto a 2D space (the image plane) in many colors, shapes, sizes and scales. The elements within an image that provide direct access to detection, identification and measurement tasks of advanced image analysis include: tone/color, size, shape, texture, pattern, height, shadow, site and association [225]. Figure 6.5 shows the elements related to the image interpretation process, as a function of their degree of complexity.

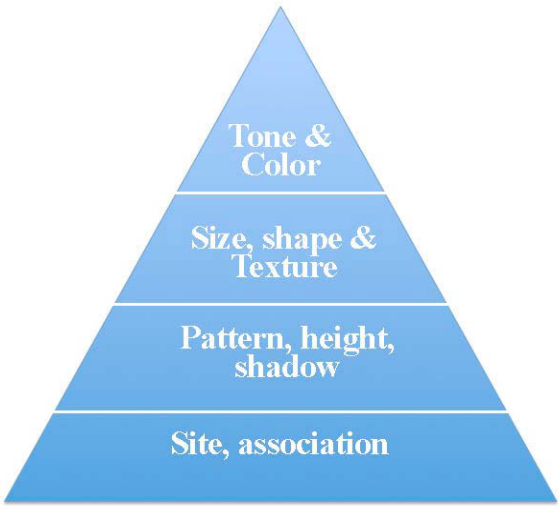


Figure 6.5 - Primary ordering of image elements fundamental to image analysis process (adapted from [169])

An important statement in [35, p. 993] reminds us that tone, as expressed in shades of gray and color as expressed in hue, value and chrome convey more information than any other single element of interpretation. In almost all cases, it is the difference in tone or color between objects or between an object and the background that is important. Size and shape image elements represent geometric arrangements of the tone and color of pixels making up a given object or phenomenon.

The first step in all image interpretation tasks is to detect and identify important phenomena. Since the R-G-B display does not always provide an optimum preview for certain target classes/objects, a visual enhancement method is mandatory. This will allow characteristics of objects invisible to the human eye to be discovered and displayed in an easy-to-understand way.

Color perception is an important element of awareness of the environment. Different objects reflect, emit, and transmit different amounts and wavelengths of energy that are recorded by the sensor as tonal, color or density variations. True color images often facilitate interpretation by providing a familiar perception of the objects but for many cases, the R-G-B display fails to provide users the necessary contrast to discover and analyze classes under investigation. In these cases, the 'false color' display is successfully used to improve interpretability by enhancing object-to-object or object-to-background contrast. In [35 p.1001] the authors advise: "in almost all cases, it is the difference in tone or color between objects, or between and object and a background that is important." This statement will prove highly valuable in the last part of this section, where we demonstrate the operational usefulness of this approach. Since almost all remote sensing image analyses appear to utilize tone and color, the method described here aims to offer users exactly this possibility – to discover the optimum spectral bands of a satellite image that maximize the difference and contrast between an object and the surrounding background.

6.4.2 Color Metrics For Satellite Image Analysis

Considering the satellite scene in figure 6.6a, the image analyst is interested in evaluating the lake in the center-right side of the image. A quick qualitative visual analysis reveals that the R-G-B display doesn't offer much contrast between the target class and the surroundings.

The automatic color mapping is computed scene by scene. The human operator selects only a tile that represents the target class and the full scene is displayed with the top mRMR score bands. If the user decides to choose another target class, the mRMR is applied again and the entire scene is displayed with the new bands. The workflow to automatically discover the spectral bands that enhance visualization for this area is the following:

1. The image is imported into the system and the grid is automatically overlaid (50X50 pixels)
2. The analyst clicks on the tile depicting the target class thus training the system
3. The mRMR scores are automatically computed for each spectral band – figure 6.7
4. The top three features yielding the highest scores are automatically displayed in the R, G, B channels

Once the mRMR calculus is performed for each spectral feature, the system automatically displays the image in 'false colors', with the top three bands mapped to the R, G, B channels. In this case:

* R channel - Nir-1 (X)
* G channel - Red Edge (Y)
* B channel - Nir-2 (Z)

It can be easily observed that with the standard natural color display (bands 532), the area of interest is barely contrasting the surroundings. With the automatic 'false color' X-Y-Z display (bands 768) as shown in figure 6.8b, the difference in tone and color has increased, thus forwarding the visualization and interpretation.

Because the human visual system is more sensitive to the green region of the electromagnetic spectrum, with a peak in sensitivity at 550 nm, another possibility to display the top three bands is to use the maximum score mRMR band in the green channel. In this case, the new visualization is:

* R channel - Red Edge (Y)
* G channel - Nir-1 (X)
* B channel - Nir-2 (Z)

Figure 6.9 shows the three visualizations for comparison: natural color (bands 532), X-Y-Z (bands 768) and Y-X-Z (bands 678) of the WorldView-2 full scene. An important point is that the method selects the three spectral bands with the maximum amount of information relevant to the target class to enhance the visual differences to the surrounding classes - interclass variation. The new spectral features simultaneously maximize intra-class variation.

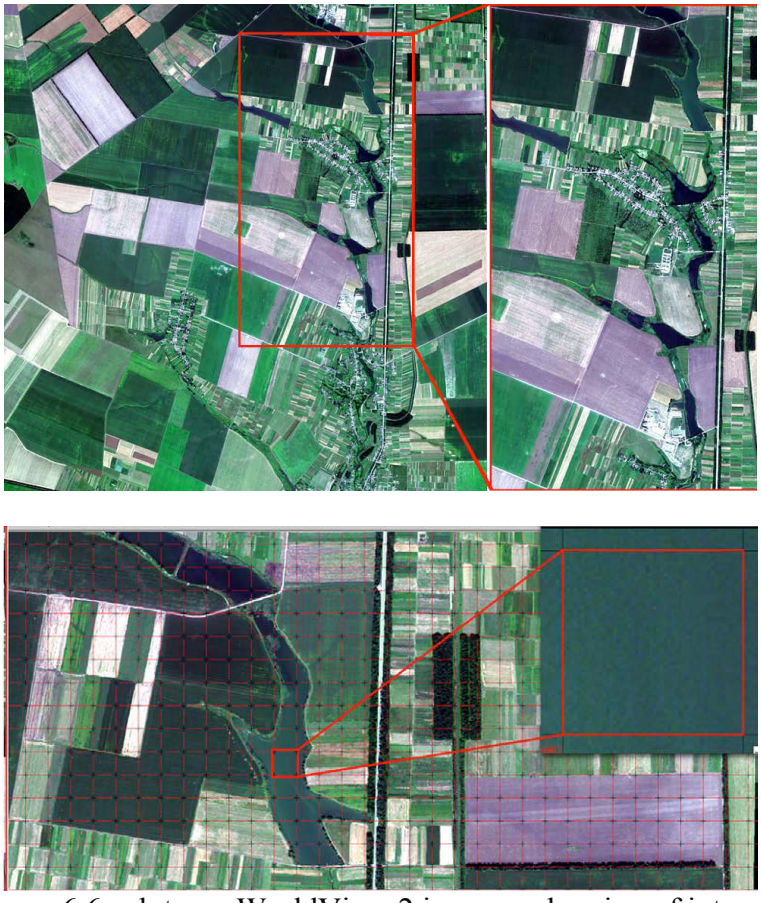


Figure 6.6 a, b top – WorldView-2 image and region of interest;
6.6 c, d bottom – region of interest and target class selection
(Image Credits: DigitalGlobe)

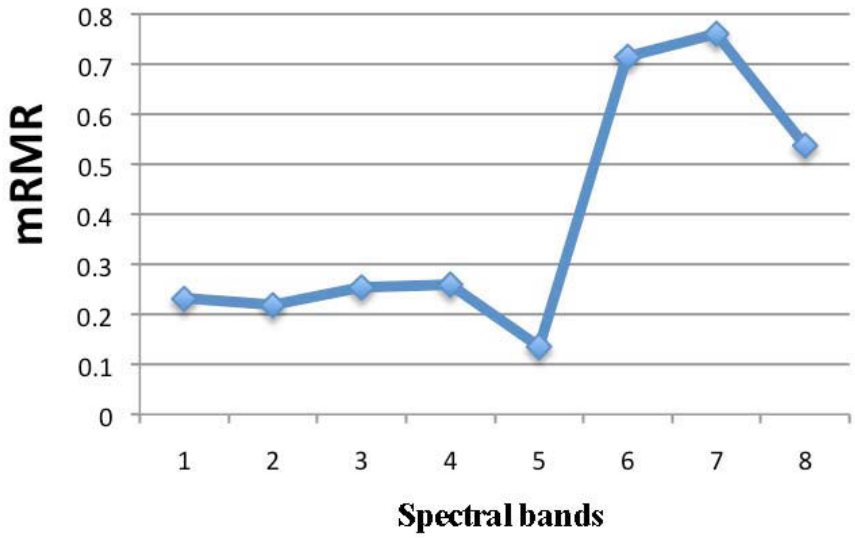


Figure 6.7 – Measures of mutual information (*X, C*) between the target class and the available features

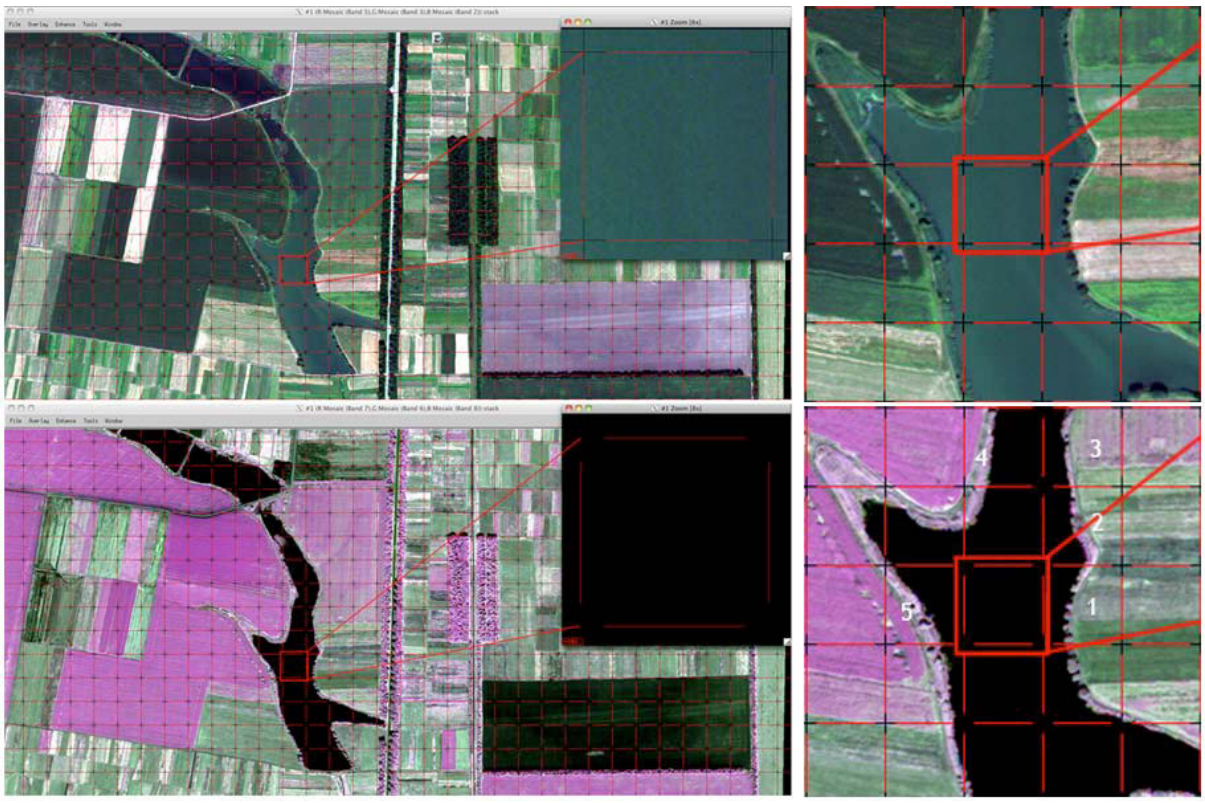


Figure 6.8a top – R-G-B display large scene & detail (bands 532);
Figure 6.8b bottom - X-Y-Z display, large scene & detail (bands 768)

Figure 6.9 – R-G-B, X-Y-Z (bands 768), Y-X-Z (bands 678)

6.4.3 Color Models For Satellite Image Analysis

Vision sciences and psychological tests measure two kinds of factors: stimuli and responses. In image interpretation, the stimuli are the physical variations in tone, texture, pattern, configuration and the responses are the elements of perception of the individual user. The stimuli characteristics of images and the interpreter's ability to respond to them are the critical determinants of the analysis performance. At the response level, the difference between objects and background is evaluated solely by the human visual system, automatically upon seeing the image and has only individual empirical value. While the human operator easily understands this difference at mind level (responses), the evaluation method should also provide numerical results to prove the usefulness of the tool, besides the users' mutual agreement. How should the improvement in contrast and difference between an object and the background be analyzed and described numerically? While multiple users may accept the contrast and difference enhancement evaluated at the response level, the difference should also be evaluated at stimuli level.

Color difference (distance) between two colors is a very useful measure in color science. It allows the human operators to quantify a notion that would otherwise be described with adjectives and subjective terms. For satellite image analysis, the analysts aim at enhancing the difference between the target class (i.e. lake in this example) and its neighbours .

The color distance between two colors (i.e. target class and neighbours in various displays) can be computed using the Euclidean distance between color vectors in device dependent or independent color models.

6.4.4 Quantitative Evaluation Using Color Distances

To continue with the same example, once the three top features in the mRMR scores are displayed, the target class (figure 6.10, center) has higher contrast values in terms of color and intensity to the surrounding neighbours. The response of the human visual system is enhanced in the new 'false color' display and the improvement in both stimuli and response can be computed and evaluated. While the response is personal for each human operator, the improvement of stimuli is calculated in terms of color distance between the target class and the neighbours, for both displays. The evaluation consists in assessing the color difference between center and neighbours for the standard R-G-B display (532 bands) and the same

difference for the X-Y-Z display (768 bands in this case). Figure 6.10 shows the R-G-B (532) and X-Y-Z (displays), the target class in the center and the neighbours numbered 1 to 5.

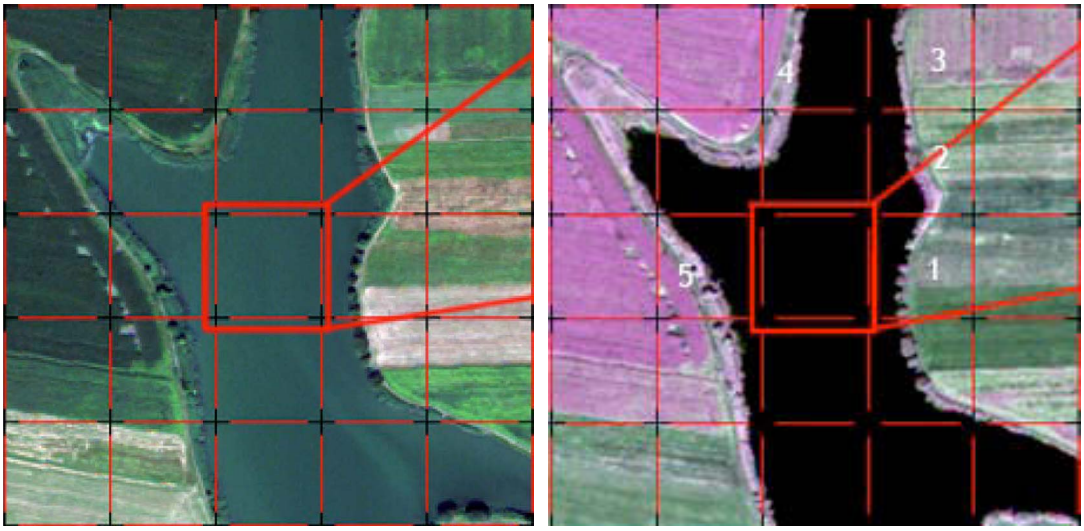


Figure 6.10: R-G-B display and X-Y-Z (NIR-1-RedEdge-NIR-2) display. The color differences computed between the target class (center) and the neighbours numbered 1 to 5

As shown in table 6.1, the color difference between the target class (center) and the surrounding neighbours has a much higher value for the automatically detected display X-Y-Z than for the standard R-G-B display. The stimuli distances between center and surroundings were computed within the CIE1976, CIE1994, CIE2000, CMC and CMC.2 color models. The properties and differences of these color models are described in detail in the appendix. These models achieve close similarity to the responses of the human visual system, meaning that a difference in color, measured in these color spaces is equivalent to the same difference recorded in the human visual system. The mathematic models of these color spaces are described in detail in the appedix of this dissertation.

Neighbour i	Difference (Center - Neighbour i) R-G-B display	Difference (Center - Neighbour i) X-Y-Z display
1	CIE1976=20.639767 CIE1994=14.928122 CIE2000=15.551529 CMC=20.545436 CMC.2=16.589518	CIE1976=54.488531 CIE1994=54.488531 CIE2000=27.964263 CMC=106.289434 CMC.2=54.055677
2	CIE1976=34.885527 CIE1994=32.229948 CIE2000=29.429358 CMC=39.384806 CMC.2=25.808988	CIE1976=56.044625 CIE1994=56.044625 CIE2000=42.638862 CMC=108.258960 CMC.2=57.832300
3	CIE1976=17.117243 CIE1994=14.579045 CIE2000=12.939496 CMC=17.269852 CMC.2=13.298139	CIE1976=45.188494 CIE1994=45.188494 CIE2000=32.216115 CMC=88.299434 CMC.2=44.503044
4	CIE1976=13.89244 CIE1994=11.378142 CIE2000=10.387977 CMC=13.084284 CMC.2=11.291230	CIE1976=67.572184 CIE1994=67.572184 CIE2000=51.545117 CMC=127.77398 CMC.2=75.089728

5	CIE1976=33.970576 CIE1994=26.883862 CIE2000=23.010282 CMC=31.706614 CMC.2=25.582071	CIE1976=60.860496 CIE1994=60.860496 CIE2000=46.998762 CMC=117.124749 CMC.2=63.686887

Table 6.1: evaluation of color difference enhancement

The values in table 6.1 evaluate objectively how the selection of three variables from the signal space (spectral bands) enhances the response in the space of human perception. Each of the color models used to measure the color difference (before and after analysis) simulate the human visual system and give a reliable measure of visual improvement at both stimuli and response levels.

6.5 Experiments And Results

CASE STUDY 1 – Visual Enhancement of Urban Objects

In this study, we evaluate the capabilities of the mRMR feature selection method on a more complicated case of visual analysis using a WorldView-2 image with 2m/pixel spatial resolution. In this instance, the image analyst aims at investigating the target class – Building X. The workflow to automatically discover the spectral bands that enhance visualization for this target is the same as in the previous section:

1. the image is imported into the system and the grid is automatically overlaid (50X50 pixels)
2. the analyst clicks on the tile depicting the target class thus training the system
3. the mRMR scores are automatically computed for each spectral band
4. the top three features yielding the highest scores are displayed in the R,G,B channels

Figure 6.11 shows the satellite image used in test, the target class and the mRMR scores. Once the mRMR calculus is performed for each spectral feature, the system automatically displays the image in 'false colors', with the top three bands in the R, G and B channels.
In this case:

* R channel - Nir-2 (X)
* G channel - Coastal (Y)
* B channel - Red (Z)

While using the standard R-G-B display (bands 532) the area of interest is barely contrasting the surroundings; with the automatic 'false color' X-Y-Z display (bands 815) the difference in tone and color has increased manifold. Another possibility to visualize the results is to feed the maximum mRMR score band Nir-2 to the Green channel of the display. In this case, the new visualization is:

* R channel – Coastal (Y)
* G channel - Nir-2 (X)
* B channel - Red (Z)

Figure 6.12 depicts the R-G-B and X-Y-Z full scene displays.

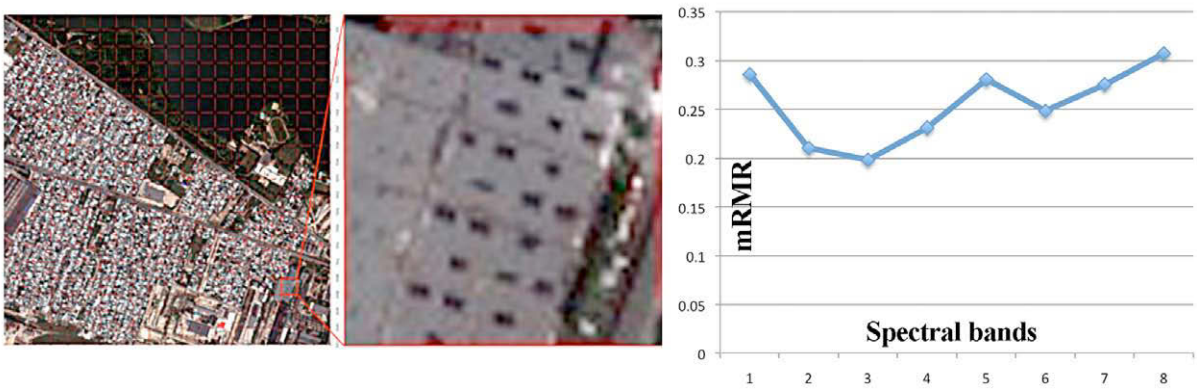


Figure 6.11 – Satellite image, target class, mRMR spectral features scores.
The top three bands that will be displayed automatically are Nir-2, Coastal and Red

Figure 6.12 – Satellite image full scene R-G-B and X-Y-Z (Nir-2, Coastal, Red)

Once the three top spectral bands in the mRMR scores are automatically displayed, the target class (figure 6.13, center) has higher contrast values in terms of color and intensity with respect to the surrounding neighbours. The response of the human visual system is enhanced in the new 'false color' display and the improvement in both stimuli and response is computed and evaluated. While the response is personal for each human operator, the improvement of stimuli is calculated in terms of color distance between the target class and the neighbours for both displays.

The evaluation consists in assessing the difference between center and neighbours for the standard R-G-B display (532 bands) and the same difference for the X-Y-Z display (815 bands in this case). Figure 6.13 shows the R-G-B (bands 532), X-Y-Z (bands 815) and Y-X-Z (bands 185) displays, with the target class in the center and the neighbours numbered 1 to 7.

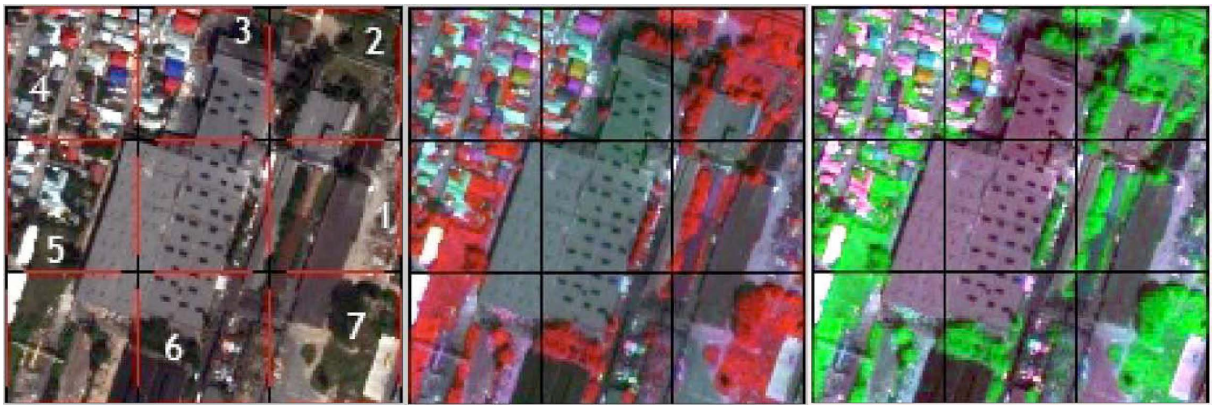


Figure 6.13 – Evaluation scene R-G-B, X-Y-Z and Y-X-Z displays. The color differences are computed between the target class (center) and the neighbours numbered 1 to 4. Color distances are shown in table 2 for the R-G-B and X-Y-Z displays.

The color difference between the target class (center) and the surrounding neighbours has a higher value for the automatically detected display X-Y-Z than for the standard R-G-B display. The stimuli distances between center and surroundings were computed within the CIE1976, CIE1994, CIE2000, CMC and CMC.2 color models (Table 6.2).

Neighbor i	Difference (Center - Neighbor i) R-G-B display	Difference (Center - Neighbor i) X-Y-Z display
1	CIE1976=36.523965 CIE1994=36.183517 CIE2000=31.311818 CMC=36.696567 CMC.2=23.354890	CIE1976=90.448881 CIE1994=69.249540 CIE2000=38.734551 CMC=85.208034 CMC.2=85.190009
2	CIE1976=32.939338 CIE1994=32.832303 CIE2000=26.606823 CMC=32.317841 CMC.2=18.295922	CIE1976=69.649121 CIE1994=54.288439 CIE2000=36.291654 CMC=67.234029 CMC.2=67.2283188
3	CIE1976=40.049969 CIE1994=40.035893 CIE2000=29.649214 CMC=38.518898 CMC.2=19.359649	CIE1976=64.373908 CIE1994=51.417370 CIE2000=38.788487 CMC=63.031829 CMC.2=62.148550
4	CIE1976=34.014703 CIE1994=34.010560 CIE2000=26.020360 CMC=32.703813 CMC.2=16.381461	CIE1976=76.478755 CIE1994=59.559259 CIE2000=39.879707 CMC=73.304748 CMC.2=73.047668

Table 6.2: Evaluation of color difference enhancement

Assessment of the effectiveness and efficiency of different visualization methods and techniques is mandatory from a technological point of view. The science of visualization should be empirical in the sense that concrete measurements of the phenomena under investigation are done and verified. Because the value of visualization is ultimately determined by the perceptual abilities of end users, their knowledge on the data presented and

the value they assign to various insights, both qualitative and quantitative evaluation methods should be employed in assessing the results of a visualization technique.

CASE STUDY 2 – Water Class Visualization

This scenario studies the capabilities of the method presented in this paper to enhance visualization of the concept class Water to the surrounding classes. The experiment is performed on the WorldView-2 satellite scene depicted in figure 6.14a. The human operator trains the system with a single image-patch representing the desired class Water (figure 6.14b) and the optimum three spectral bands were automatically detected and displayed on the the screen – figure 6.14d.

The new display is:
* R channel – Coastal (X)
* G channel – Yellow (Y)
* B channel – Blue (Z)

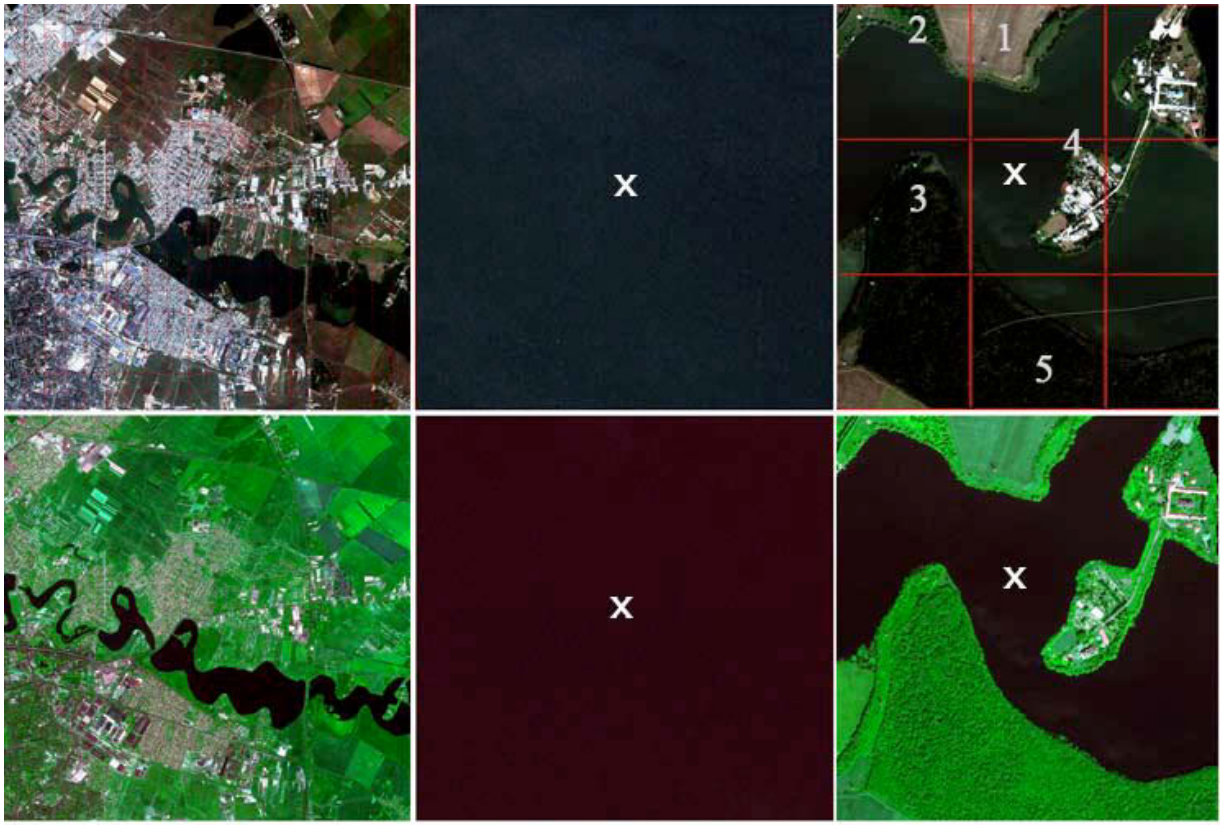


Figure 6.14

(a) WorldView-2 satellite image R-G-B,
(b) Target class "Water" displayed in R-G-B,
(c) Detail of the target class and neighbours displayed in R-G-B – top row
(d) Satellite image X-Y-Z,
(e) Target class "Water" displayed in X-Y-Z bands
(f) Detail of the target class and neighbours displayed in X-Y-Z - bottom row. While the R-G-B display barely reveals the water and the tree line in the bottom of the image, the new visualization maximizes the color difference between these classes for further analysis.

Figure 6.15 shows the R-G-B (bands 532), X-Y-Z (bands 142) and Y-X-Z (bands 412) displays for comparison.

Figure 6.15 - R-G-B, X-Y-Z and Y-X-Z full scene displays

An expert-driven qualitative visual analysis reveals that the R-G-B depiction (figure 6.14c) shows little contrast between the concept class Water and the surrounding neighbours – the tree line north to the water body and the forest to the south. The new display (figure 6.14f) creates a powerful contrast for the target class. The enhancement offered by the new display is evaluated also in the CIE2000 color space. While in the R-G-B, the color difference between the target class and the neighbours has a mean value of CIELAB2000 = 3.75, in the X-Y-Z display, the color difference is 10 times higher, with a value of CIELAB2000 = 37.5. In the images, X marks the target class and the numbers indicate the neighbours. Table 6.3 evalutes the color differences between the target class and the neighbours, in the standard R-G-B display and the new X-Y-Z display. Figures 6.16a - 6.16f represent sample scenes extracted from figure 6.14a and 6.14d. To evaluate the effectiveness of the algorithm presented in this paper, an important step is to assess its generalization capabilities across the entire scene.

Neighbor i	Difference (Center - Neighbor i) R-G-B display	Difference (Center - Neighbor i) X-Y-Z display
1	CIE1976=41.4246 CIE1994=22.2769 CIE2000=30.3129 CMC=79.5065 CMC.2=41.5417	CIE1976=71.9027 CIE1994=66.4310 CIE2000=51.5410 CMC=112.7743 CMC.2=78.1052
2	CIE1976=36.5650 CIE1994=34.0719 CIE2000=22.5088 CMC=54.0473 CMC.2=44.6151	CIE1976=93.6269 CIE1994=84.4601 CIE2000=69.8385 CMC=137.1914 CMC.2=102.206
3	CIE1976=2.4494 CIE1994=1.6401 CIE2000=2.0666 CMC=4.3230 CMC.2=2.6832	CIE1976=46.6261 CIE1994=41.7054 CIE2000=31.7161 CMC=64.7258 CMC.2=52.9081
4	CIE1976=10.488 CIE1994=10.4171 CIE2000=6.9019 CMC=19.9953 CMC.2=10.6109	CIE1976=93.3059 CIE1994=85.6344 CIE2000=59.8364 CMC=144.34172 CMC.2=100.732

5	CIE1976=4.6904 CIE1994=4.5929 CIE2000=4.5016 CMC=7.5784 CMC.2=5.6119	CIE1976=78.1472 CIE1994=69.9134 CIE2000=43.7152 CMC=110.827 CMC.2=86.3392

Table 6.3: Evaluation of color difference enhancement

A qualitative visual analysis reveals that the R-G-B display does not provide a powerful discrimination between the target class and surroundings. While 6.16d, 6.16e and 6.16f enhance the visualization and contrast with neighbours, the displays in 6.16e and 6.16f create a powerful contrast also with the plants in the water, undistinguishable in R-G-B. While in the R-G-B, the color difference between Water and the plants (figure 6.16b) has a mean value of CIELAB2000 = 8.90, in the X-Y-Z display (figure 6.16e), the color difference has a value of CIELAB2000 = 50.

Figure 6.16

(a) R-G-B detail shows possible confusion between the river and the forest in the right upper part of the image; (b) R-G-B detail shows that the phytoplankton is not distinuishable in the water; (c) R-G-B detail shows possible confusion between the water and the forest in the right and left lower parts of the image – top row. (d) X-Y-Z detail shows the border between water and forest classes, (e) X-Y-Z detail reveals clearly the phytoplankton in the water; (f) X-Y-Z detail – bottom row shows the border between the water and the forest classes with enhanced contrast.

CASE STUDY 3 – Forest Concept Visualization

This scenario studies the capabilities of this algorithm to enhance visualization of the concept class Forest to the surrounding areas in a WorldView-2 image from Romania. The experiment is performed on the satellite scene depicted in figure 6.17a. The human operator trained the system with a single image-patch representing the desired class (figure 6.17b) and the optimum three spectral bands were automatically detected and displayed on the the screen, figure 6.17d. The new display is:

* R channel – Red Edge (X)
* G channel – Coastal (Y)
* B channel – Yellow (Z)

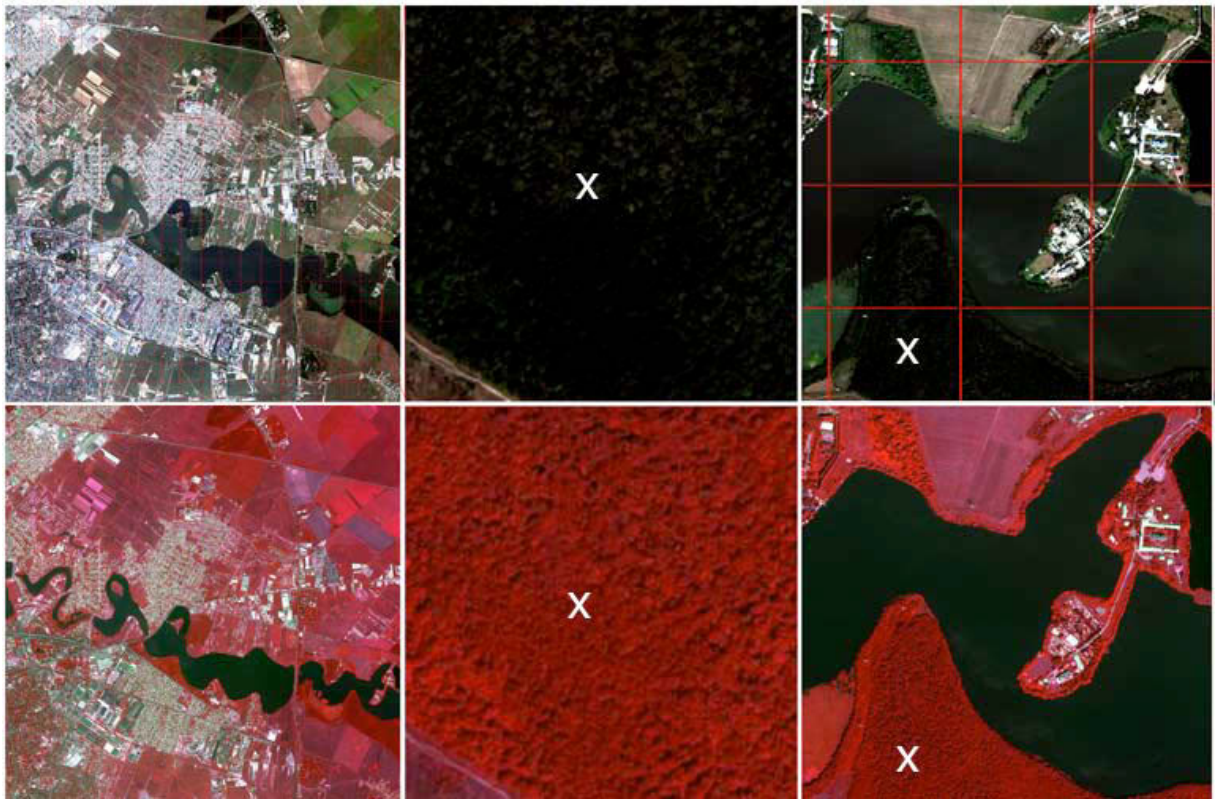


Figure 6.17

(a) WorldView-2 satellite image R-G-B, (b) target class "Forest" displayed in R-G-B,
(c) Detail of the target class and neighbours displayed in R-G-B – top row
(d) Satellite image X-Y-Z, (e) target class "Forest" displayed in X-Y-Z bands; (f) detail of the target class and neighbours displayed in X-Y-Z - bottom. While the R-G-B display barely reveals the water and the tree line in the bottom of the image, the new visualization maximizes the color difference between these classes. Figure 6.18 shows the R-G-B (bands 532), X-Y-Z (bands 614) and Y-X-Z (bands 164) full scene displays for visual comparison.

Figure 6.18 - R-G-B (bands 532), X-Y-Z (bands 614) and Y-X-Z (bands 164) full scene

The qualitative visual analysis reveals that the R-G-B depiction (figure 6.17c) shows little contrast between the target class Forest and the surrounding neighbours – water bodies. The new visualization (figure 6.17f) creates a powerful contrast for the target class. The enhancement offered by the new display is evaluated also in the CIE2000 color space. In the R-G-B display, the color difference between the target class and the neighbors has a mean value of CIELAB2000 = 4, in the X-Y-Z display, the mean color difference was evaluated at CIELAB2000 = 35.

A qualitative visual analysis reveals that the R-G-B display does not provide a powerful discrimination between the target class and surroundings. While figure 6.19d and 6.19e enhance the visualization and contrast with the neighbours, the display in 6.19f creates a powerful contrast also for the plants in the water, undistinguishable in the R-G-B. This contrast is assessed also by color difference measures in CIE2000 color space. While the difference in color between plants in the water and the water in R-G-B (figure 6.19c) is CIE2000 = 5.25, the X-Y-Z display (figure 6.19f) gives a color difference value of CIE2000 = 36.5. In the images, X marks the target class and the numbers indicate the neighbours. Table 6.4 evalutes the color differences between the target class and the neighbours, in the standard R-G-B display and the new display.

Figure 6.19

(a) R-G-B detail shows low contrast between the river and the forest in the right upper part of the image; (b) R-G-B detail; (c) R-G-B detail - the phytoplankton is not distinuishable in the water - top row; (d) X-Y-Z detail shows increased contrast between the forest and the water classes; (e) X-Y-Z detail shows improvement in the details of the target class; (f) X-Y-Z detail reveals clearly the phytoplankton in the water - bottom row

Neighbor i	Difference (Center - Neighbor i) R-G-B	Difference (Center - Neighbor i) X-Y-Z
1	CIE1976=18.0554 CIE1994=12.2068 CIE2000=13.8524 CMC=13.2309 CMC.2=13.1219	CIE1976=65.5896 CIE1994=33.6991 CIE2000=41.4472 CMC=46.1940 CMC.2=41.9247
2	CIE1976=45.5741 CIE1994=43.3749 CIE2000=36.3868 CMC=83.0572 CMC.2=42.8006	CIE1976=47.6864 CIE1994=17.3999 CIE2000=22.7026 CMC=23.1756 CMC.2=21.5127
3	CIE1976=85.44 CIE1994=84.4315 CIE2000=81.6416 CMC=164.672 CMC.2=82.7683	CIE1976=91.1756 CIE1994=63.8974 CIE2000=69.3035 CMC=84.4547 CMC.2=65.0367
4	CIE1976=20.1494 CIE1994=15.5861 CIE2000=14.8240 CMC=23.3942 CMC.2=16.1210	CIE1976=63.0713 CIE1994=29.8463 CIE2000=38.8948 CMC=41.2587 CMC.2=38.8078

Table 6.4: Evaluation of color difference enhancement

CASE STUDY 4 – Water Class Visualization

This scenario studies the capabilities of the method to enhance visualization of the concept class Water to the surrounding areas in another satellite scene from Romania. The experiment is performed on the WorldView-2 satellite scene depicted in figure 6.20a. The human operator trained the system with a single image-patch representing the desired class (figure 6.20b) and the optimum three spectral bands were automatically detected and displayed on the the screen – figure 6.20d. The new display is:

* R channel – Blue (X)
* G channel – Nir-2 (Y)
* B channel – Red (Z)

Figures 6.20g - 6.20i depict the same areas but with the maximum mRMR score band in the green channel of the display. For these image the new display is:

* R channel – Nir-2 (Y)
* G channel – Blue (X)
* B channel – Red (Z)

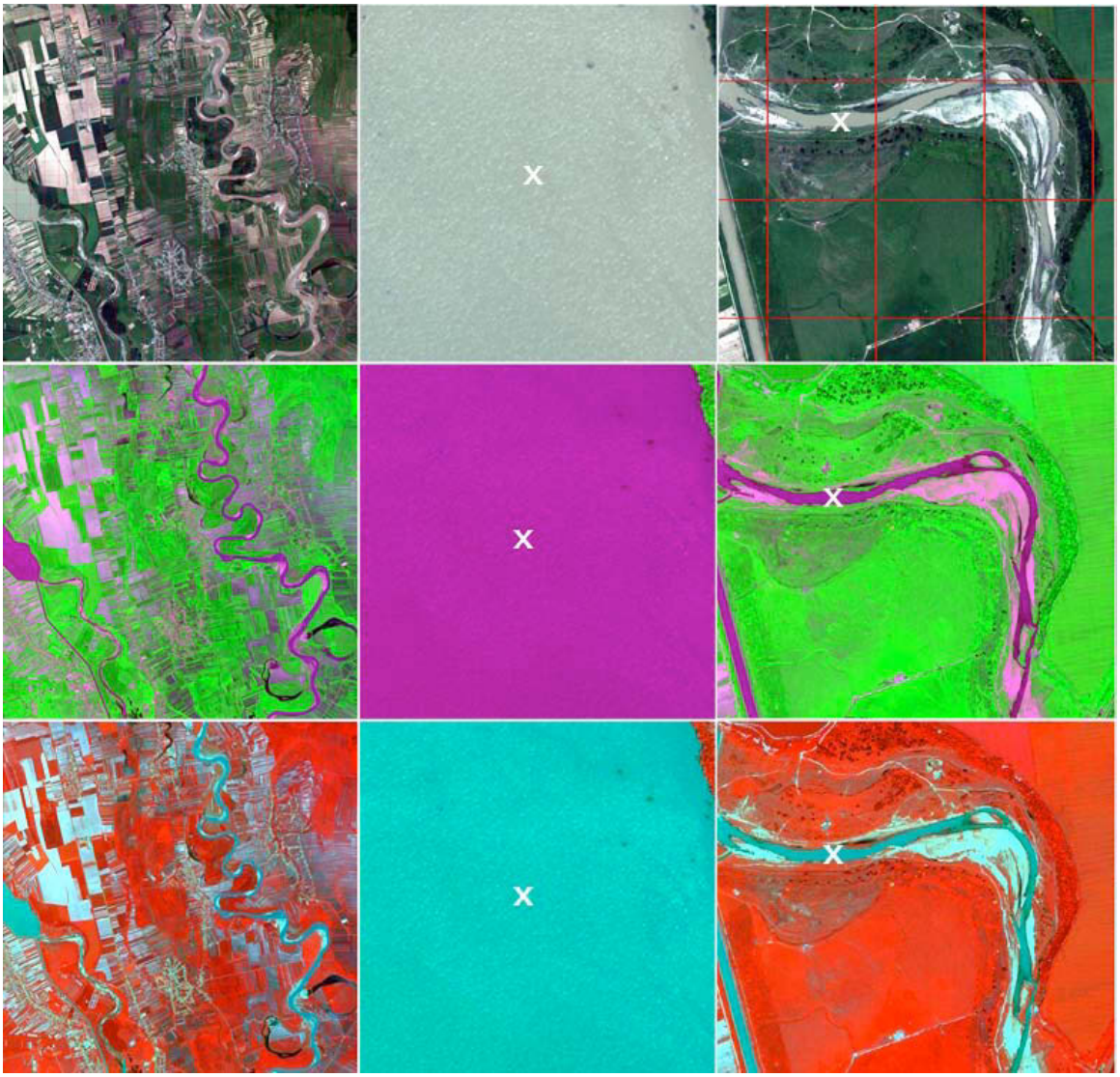


Figure 6.20

(a) Satellite image R-G-B; (b) target class (R-G-B), (c) R-G-B detail – top row
(d) Satellite image X-Y-Z; (e) target class X-Y-Z, (f) X-Y-Z detail shows improved contrast for the class river compared to the surroundings – middle row
(g) Satellite image Y-X-Z; (h) target class Y-X-Z; (i) Y-X-Z detail – bottom row

The qualitative visual analysis reveals that the R-G-B depiction (figure 6.20c) shows little contrast between the target class Water and the surrounding neighbours – agriculture fields, tree lines, etc. The new display in figure 6.20f creates a powerful contrast for the target class. The enhancement offered by the new display is evaluated also in the CIE2000 color space. While in the R-G-B, the color difference between the target class and the neighbours has a mean value of CIELAB2000 = 39, in the X-Y-Z display, the color difference has a value of CIELAB2000 = 99. Figures 6.21a - 6.21f represent sample scenes extracted from figure 6.20a and 6.20d. While figure 6.21c displays the target class in R-G-B with little color difference from the surroundings, figure 6.21f depicts a powerful contrast for concept class Water. The difference in color between water and the agricutlure field in R-G-B (figure 6.21c) is CIE2000 = 36, the X-Y-Z display (figure 6.21f) gives a color difference value of CIE2000 = 94.

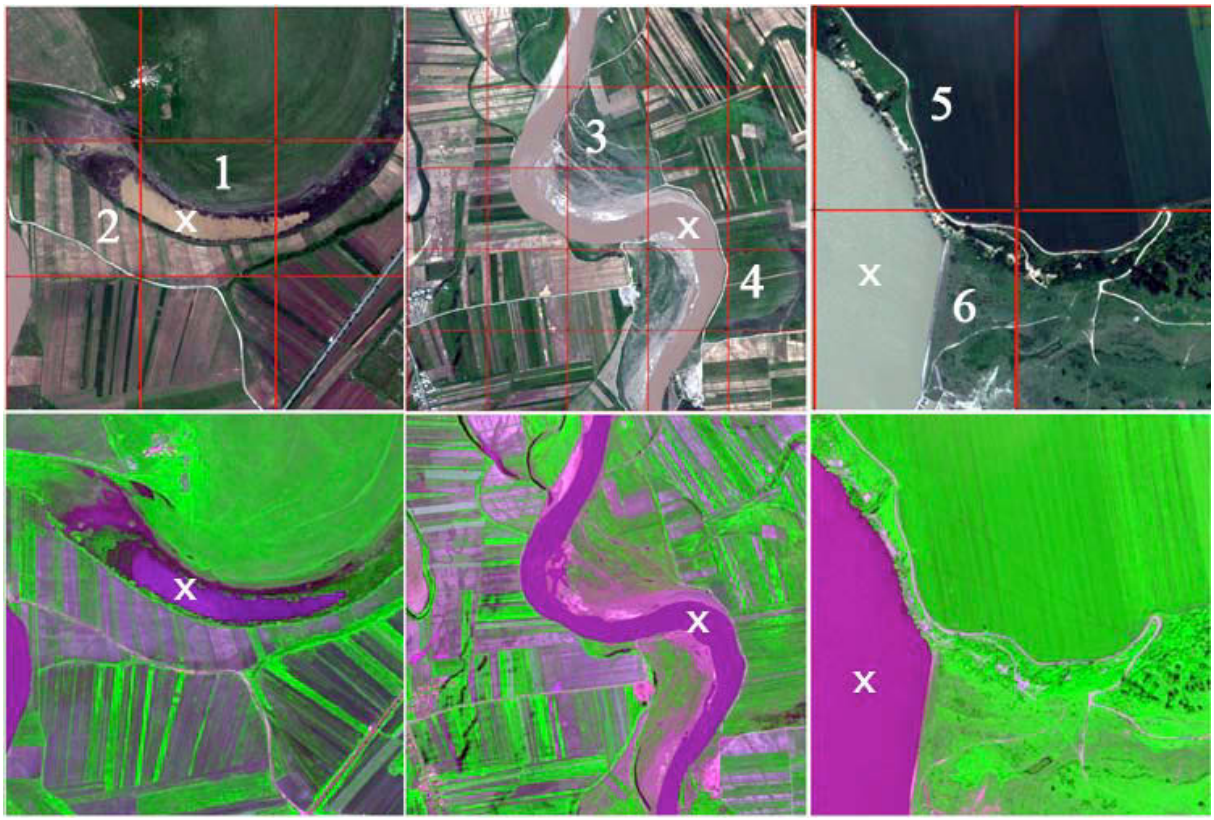


Figure 6.21

(a) R-G-B detail; (b) R-G-B detail; (c) R-G-B detail – top row
(d) X-Y-Z detail reveals the fields that contain high amounts of water, undetectable in the R-G-B display; (e) X-Y-Z detail shows the river with increased contrast to the surrounding areas and the crops that contain high amount of water. The new combination of bands clearly differentiates the visual signatures of the river, crops with high moisture level and crops with normal moisture level; (f) detail X-Y-Z shows improved contrast and color distance between the river and its neighbours – bottom row

Table 6.5 evalutes the color differences between the target class and the neighbours , in the R-G-B and X-Y-Z displays.

Neighbor i	Difference (Center - Neighbor i) R-G-B display	Difference (Center - Neighbor i) X-Y-Z display
1	CIE1976=47.6759 CIE1994=45.9082 CIE2000=45.3277 CMC=46.9951 CMC.2=35.9724	CIE1976=175.356 CIE1994=82.99 CIE2000=98.069 CMC=90.0785 CMC.2=82.496295
2	CIE1976=37.6031 CIE1994=35.6820 CIE2000=38.0183 CMC=37.205 CMC.2=29.079	CIE1976=46.9361 CIE1994=10.8456 CIE2000=12.9679 CMC=15.7088 CMC.2=14.8345

3	CIE1976=40.4103 CIE1994=37.4051 CIE2000=38.2101 CMC=42.9525 CMC.2=36.1426	CIE1976=115.95 CIE1994=74.11 CIE2000=90.8987 CMC=80.2562 CMC.2=74.4494
4	CIE1976=45.0333 CIE1994=42.5835 CIE2000=44.2472 CMC=49.008 CMC.2=41.0801	CIE1976=171.201 CIE1994=81.298 CIE2000=96.907 CMC=88.3293 CMC.2=80.5826
5	CIE1976=59.5399 CIE1994=57.3698 CIE2000=50.137 CMC=57.7571 CMC.2=43.5156	CIE1976=180.820 CIE1994=82.613 CIE2000=100.32 CMC=83.7809 CMC.2=85.161
6	CIE1976=38.0131 CIE1994=35.0173 CIE2000=37.8534 CMC=42.7139 CMC.2=37.1394	CIE1976=160.433 CIE1994=76.0737 CIE2000=92.5973 CMC=82.7365 CMC.2=75.3796

Table 6.5: Evaluation of color difference enhancement

Calculus of color distances reveal that the automatically-generated visualizations enhance the target class in rapport with the surrounding regions. These measurements allow this method to be integrated into a scientific workflow and its results to be verified and compared to other approaches.

CASE STUDY 5 – Smoke Plume Visualization

This scenario studies the capabilities of the method to enhance visualization of the concept class Smoke Plume to the surrounding areas. The experiment is performed on the Landsat ETM 7+ satellite scene depicted in figure 6.22a. The human operator trained the system with a single image-patch representing the desired class (figure 6.22b) and the optimum three spectral bands were automatically detected and displayed on the the screen – figure 6.22d. The new display is:

* R channel – band 1 (X)
* G channel – band 6-2 (Y)
* B channel – band 3 (Z)

Figure 6.23 shows the mRMR scores for this target class. Another possibility of displaying the first three bands is by feeding the maximum score band to the Green channel. The new display in this case is presented in figures 6.23c and 6.23f:

* R channel – band 6-2 (Y)
* G channel – band 1 (X)
* B channel – band 3 (Z)

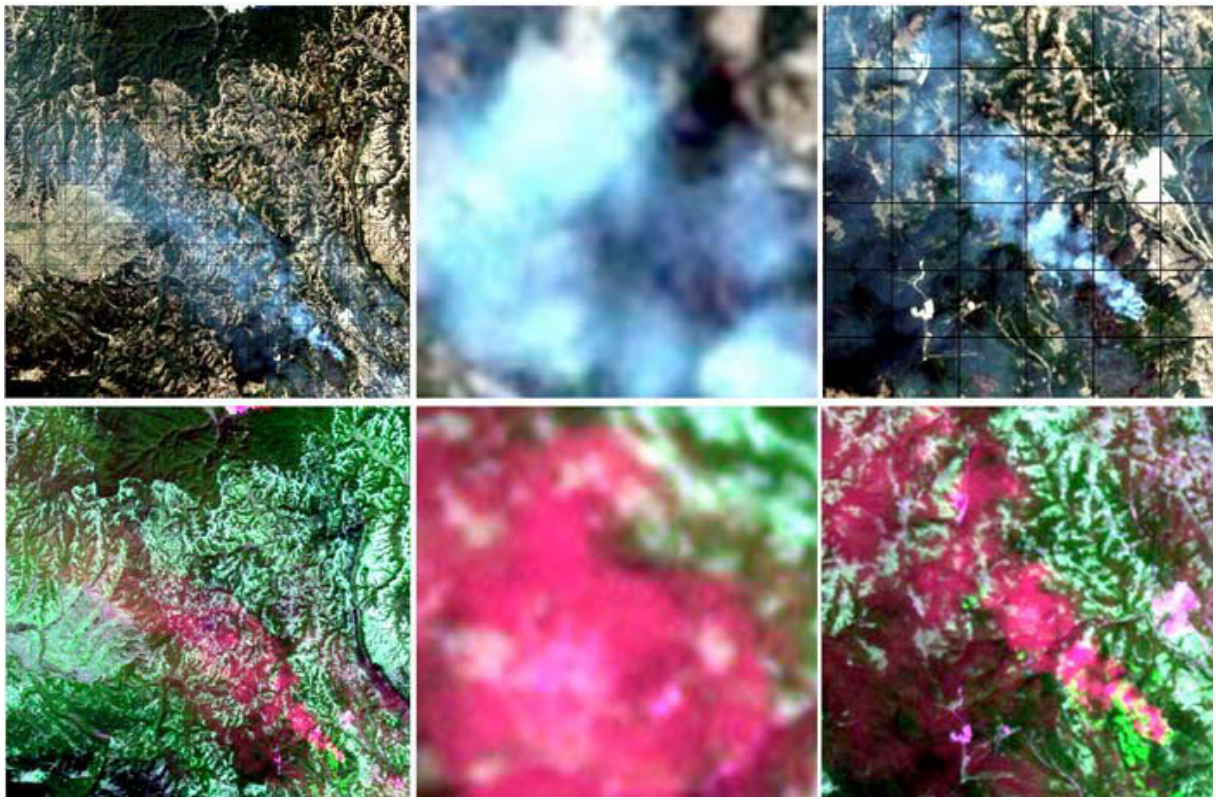


Figure 6.22

(a) Satellite image R-G-B shows the smoke coming from a forest fire and its direction in the atmosphere
(b) Target class chosen by the user depicted in R-G-B
(c) R-G-B detail of the smoke plume and its surroundings - top row
(d) Satellite image X-Y-Z shows clearly the direction of the smoke plume and the affected areas
(e) Target class X-Y-Z,
(f) X-Y-Z detail shows areas affected by the smoke plum that were not visible in the R-G-B display. The upper right and the lower left areas of the image show clearly the signature of the smoke plume, undistinguishable in the R-G-B display - bottom row.

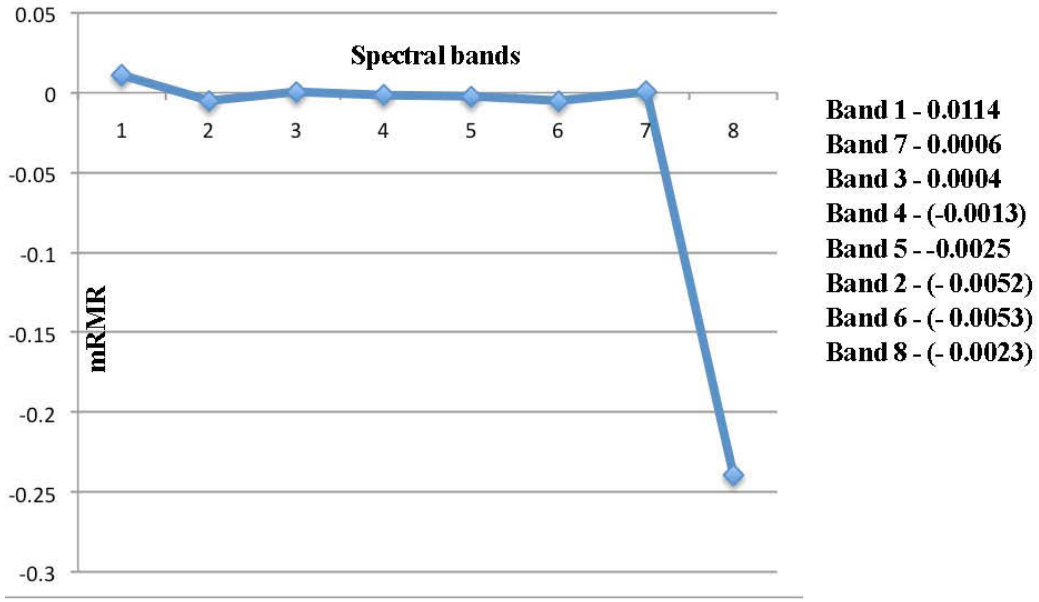


Figure 6.23 – mRMR scores

The qualitative visual analysis of figure 6.22 shows that the R-G-B depiction (figure 6.22a, 6.22c) displays little contrast between the target class Smoke Plume and the surrounding areas. The new display in figures 6.22c and 6.22f creates a powerful visualization for the target class, allowing detection of the smoke plume in areas where R-G-B display makes it undistinguishable. An example of this is given in figure 6.24a, where the plume is barely observable in R-G-B. The new visualizations (figures 6.24b, 6.24c) highlight the target class allowing for precise detection and evaluation of the smoke extent.

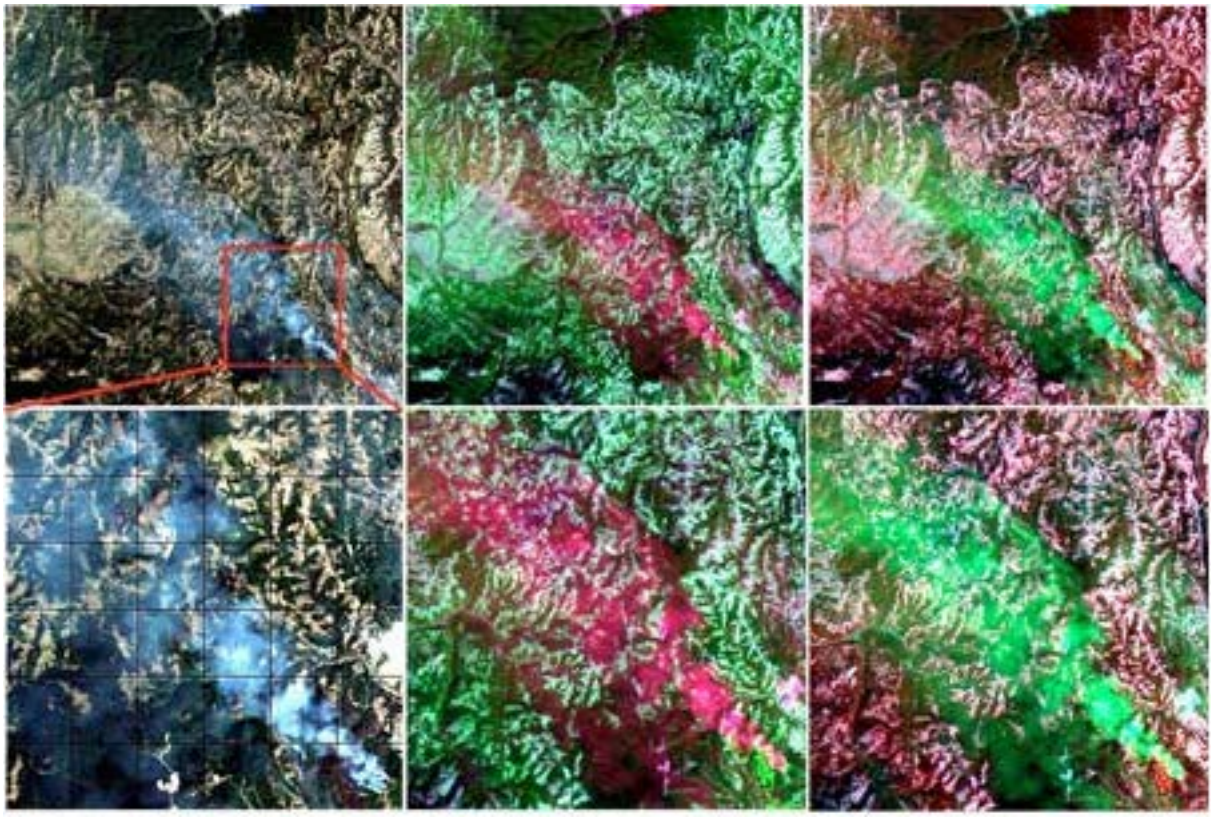


Figure 6.24 (a) R-G-B (b) X-Y-Z (c) Y-X-Z full scene – upper row
(d) R-G-B, (e) X-Y-Z, (f) Y-X-Z detail – bottom row

6.6 Discussion

6.6.1 Comparison and evaluation: mRMR, PCA, ICA

In 2004, Johnson [223] wrote in a review article for the IEEE Journal of Computer Graphics and Applications that scientific visualization is still a relatively new discipline and visualization researchers are not necessarily accustomed to undertaking strict examinations in their work. In trying to take the science of visualization to the technological level of exact science, new tools and methods have to be measurable, effective and efficient.
This chapter presented an adaptive visualization technology used for enhancing visual analysis of multi-band satellite imagery. Since several researchers used PCA and ICA in their studies, in this section we evaluate the new visualizations created using our method to PCA and ICA. Figure 6.25a shows the R-G-B display of a WorldView-2 image, 6.25b shows the X-Y-Z display with the bands discovered by the mRMR criterion, 6.25c shows the first three principal components and 6.25d the first three independent components.

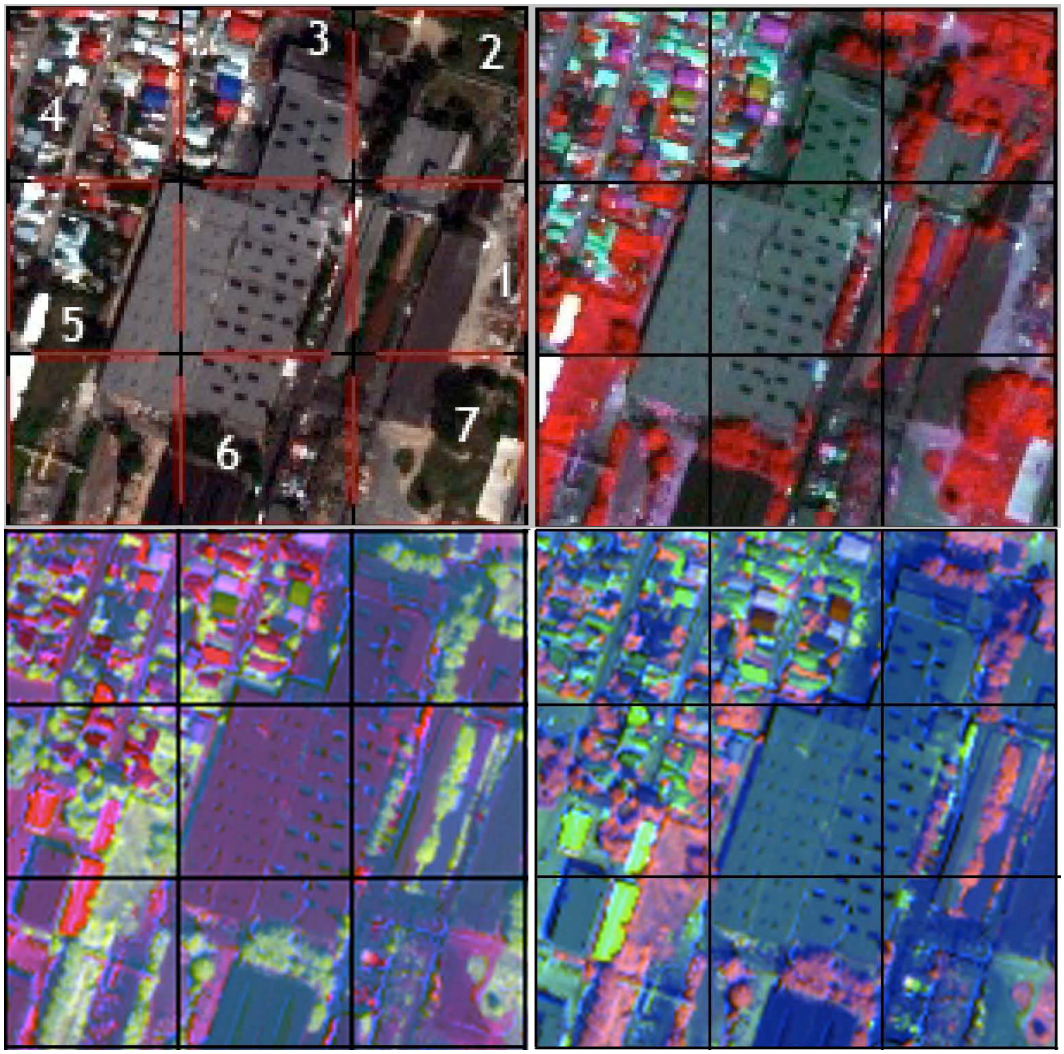


Figure 6.25

(a) R-G-B display (bands 5-3-2); (b) X-Y-Z display (bands 8-1-5) - top row
(c) PCA – first three components; (d) ICA – first three components - bottom row

The evaluation consists in assessing the color differences between center and neighbours for the standard R-G-B display (5-3-2 bands), for the X-Y-Z display, for PCA and ICA displays. The following graphics in figures 6.26, 6.27 and 6.28 show the results for the first three neighbours marked with 1, 2, 3 in the R-G-B image. The analysis reveals that the maximum increase in contrast is given by the mRMR criterion, followed by R-G-B, ICA and PCA. For CIE1994 and CIE2000, PCA visualization shows a negative trend, i.e. a loss in the quality of visualization, as compared to the standard R-G-B display.

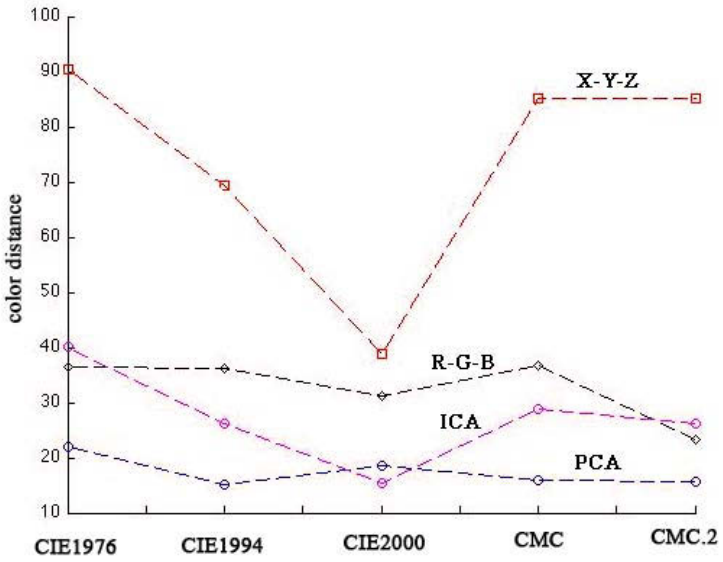


Figure 6.26 – Color distance evaluation for neighbour number 1.
The graph clearly shows that the X-Y-Z display creates the best visual enhancement, when compared to R-G-B, PCA and ICA

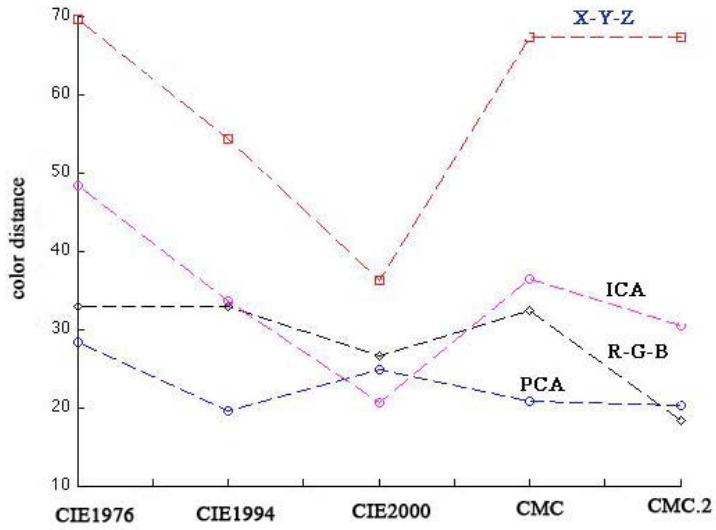


Figure 6.27 – Color distance evaluation for neighbour number 2.
The graph clearly shows that the X-Y-Z display creates the best visual enhancement, when compared to R-G-B, PCA and ICA

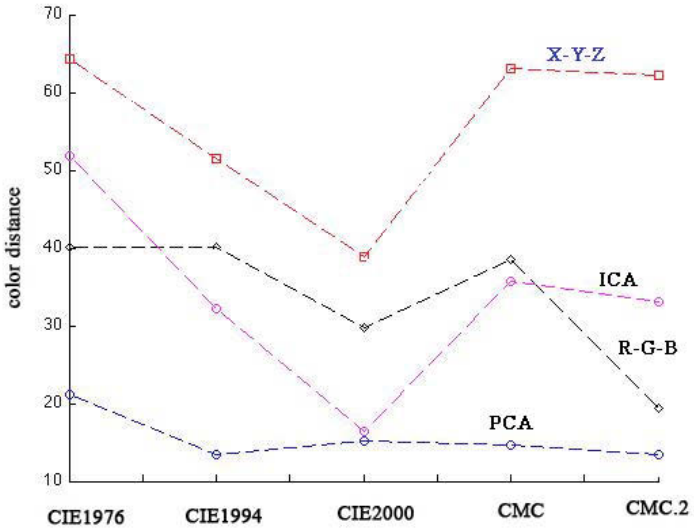


Figure 6.28 – color distance evaluation for neighbour number 3.
The graph clearly shows that the X-Y-Z display creates the best visual enhancement, when compared to R-G-B, PCA and ICA.

If results are clear for homogeneous classes like water bodies, forest areas and even large objects, several questions arise when applying this method on a heterogeneous class from an urban environment. The following example evaluates the performance of the method on a class containing several small-scaled objects, i.e. buildings. The spatial resolution of the image in this case is 2m/pixel and an image tile (50 X 50 pixels) covers an area of 100 m X 100 m. Although the tiles might contain several objects, it is likely that these objects belong to the same class (i.e. residential, commercial, etc). Figure 6.29 depicts the WorldView-2 satellite image in test and the target class containing small-scale objects. Figure 6.30 shows the mRMR scores evaluating the eight spectral bands. In this case, the new visualization is:

* Red channel - Red Edge (X)
* Green channel - Nir-2 (Y)
* Blue channel - Nir-1 (Z)

Figure 6.29 - WorldView-2 satellite image and heterogeneous target class

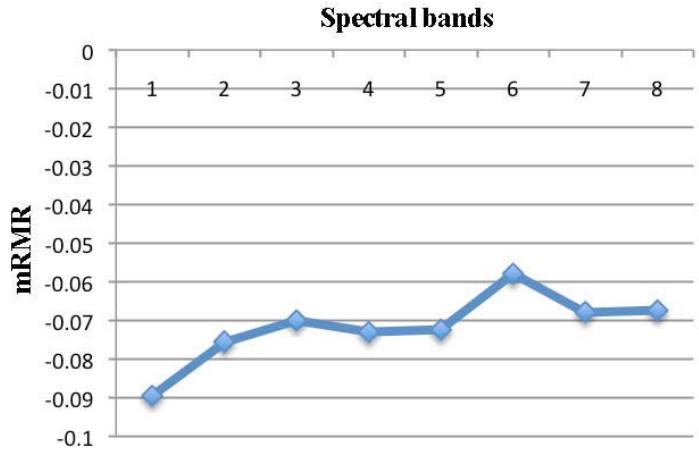


Figure 6.30 - mRMR spectral feature evaluation scores.
The top three bands that will be displayed automatically are Red Edge, Nir-2 and Nir-1

This case study evaluates the performance of the method to enhance visualization of the target class containing multiple small-scale objects and compares the results to the visualizations created using the first components of the PCA and ICA transformations. Figure 6.31 shows the R-G-B, X-Y-Z, Y-X-Z, PCA and ICA displays for visual comparison. The first column contains the R-G-B (bands 532) scene, the second column shows the mRMR, PCA and ICA top three features displayed in R,G,B channels and the last column shows the mRMR, PCA and ICA top three feature displayed in G,R,B order - i.e. the top feature displayed in the Green channel.

Figure 6.32 shows a detailed view of the target class and its surrounding neighbours. The first column depicts the R-G-B (bands 532) displays, the second column shows the mRMR, PCA and ICA top three features displayed in R,G,B channels and the last column shows the mRMR, PCA and ICA top three feature displayed in G,R,B order. A qualitative visual analysis reveals that the finest spatial details are maintained only by displaying the spectral bands of the satellite image. Although the first three components PCA and ICA contain more information than the first three mRMR bands, the transformations reduce the spatial detail, as described in [208]. This reduction of high frequency details can be observed especially at the corners and sides of the buildings.

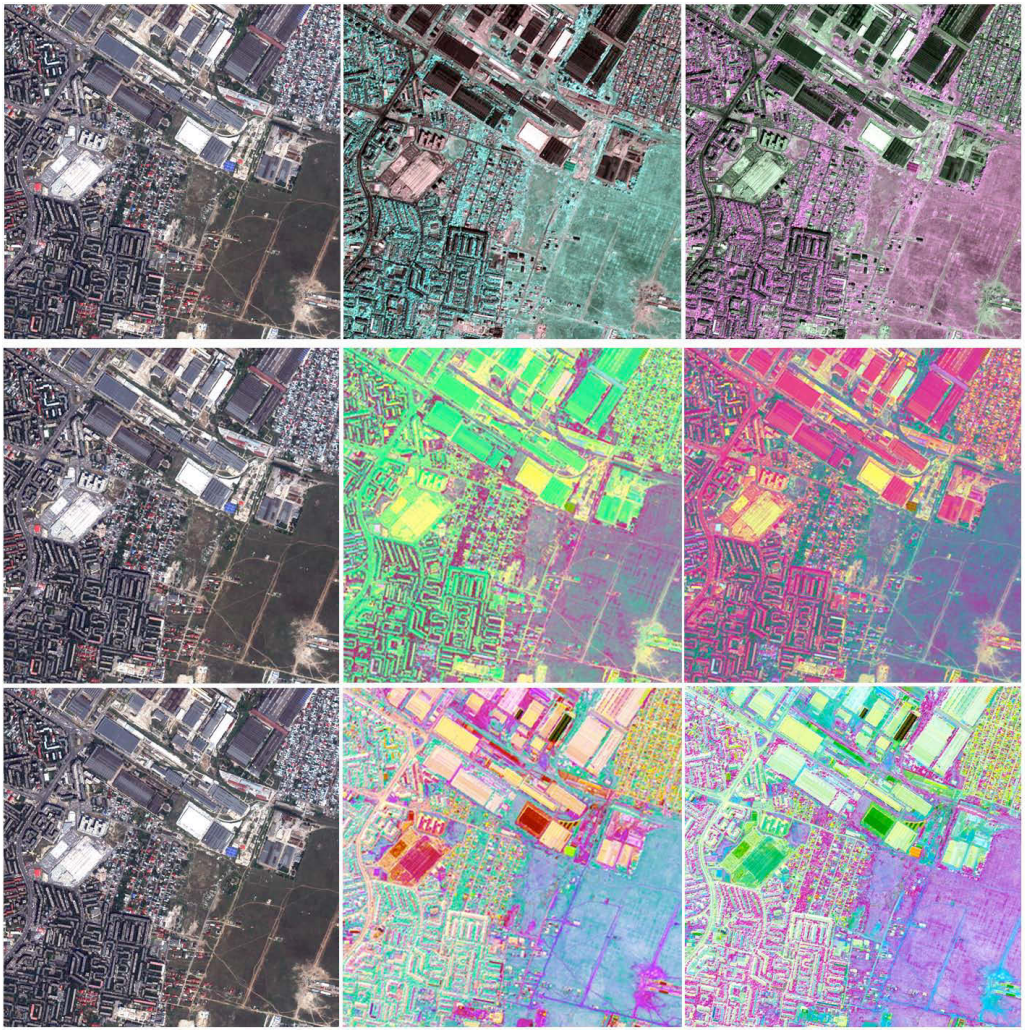


Figure 6.31 - The first column depicts the R-G-B (bands 532) displays, the second column shows the mRMR, PCA and ICA top three features displayed in R,G,B channels and the last column shows the mRMR, PCA and ICA top three feature displayed in G,R,B order.

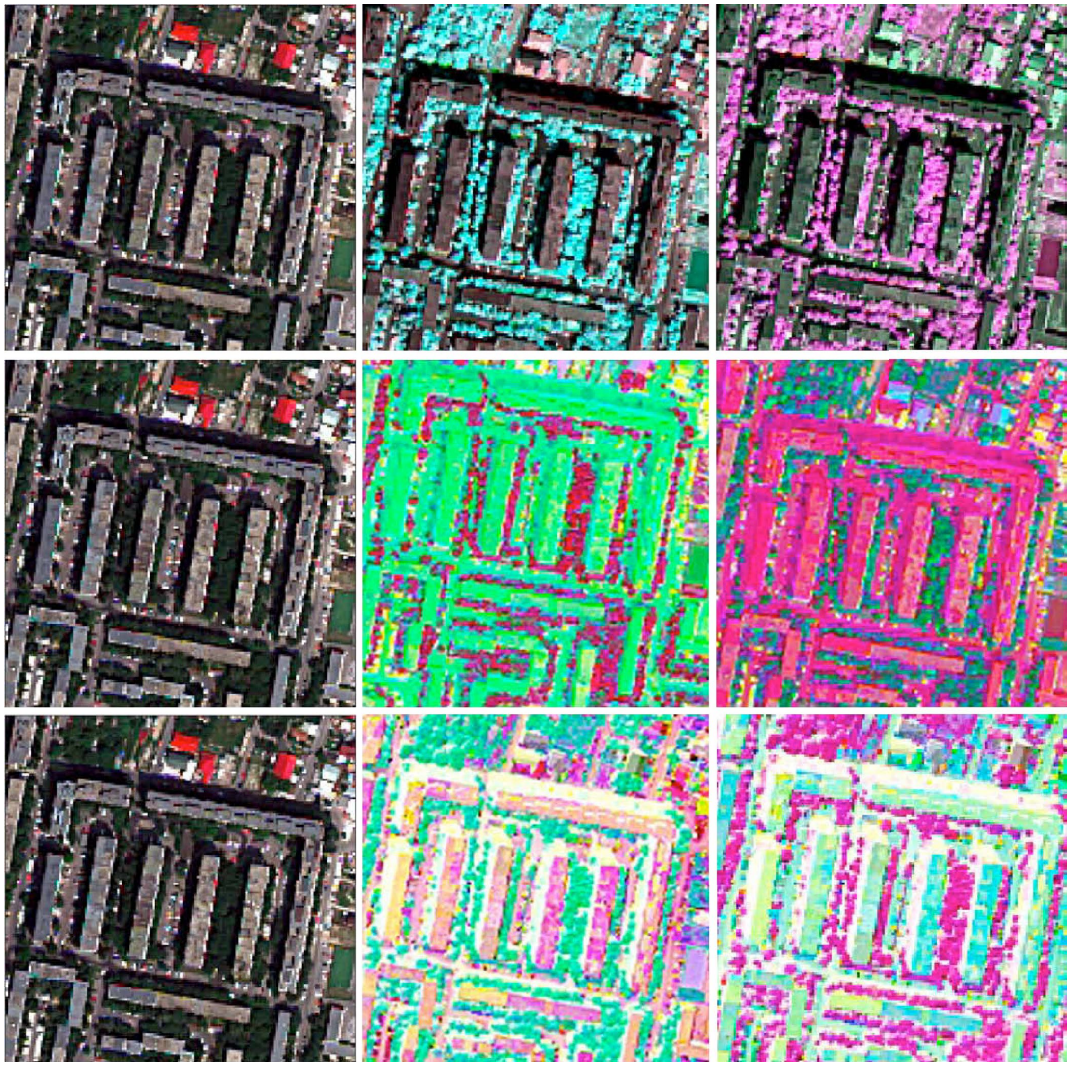


Figure 6.32 - Detail of the target class in the center and the eight surrounding tiles from the previous image. The first column depicts the R-G-B (bands 532) displays, the second column shows the mRMR, PCA and ICA top three features displayed in R, G, B channels and the last column shows the mRMR, PCA and ICA top three feature displayed in G, R, B order.

Besides the color distances calculated in multiple color models and the difference in entropy, the Kullback-Leibler divergence (KL) measure can be utilized to evaluate the increase of visual information in the new combinations of features. If p is the probability density function of the R-G-B display and q is the probability density function of the new display (top three components of mRMR, PCA, ICA), the KL divergence gives a measure of the difference between p and q. This measure is a quantitative evaluation of the increase / decrease of visual information between the standard R-G-B and the new display. For this case study we evaluate the KL divergence between the R-G-B display and each of the mRMR, PCA and ICA displays in figure 6.32. The X-Y-Z notation below represents the top three components of mRMR, PCA, ICA displayed in R, G, B order and Y-X-Z notation represents the top three components displayed in G, R, B order - the most relevant component in the green channel. The calculus shows that the X-Y-Z display contains a higher amount of information compared to the Y-X-Z display. As expected, the first three components of PCA and ICA contain more information than the top three features in the mRMR scores because they contain 99% of the information available in all the eight spectral bands. However, this higher

amount of information doesn't ensure a powerful visualization of objects in the target class and the surrounding areas for three reasons: the color distance and contrast are reduced in the PCA and ICA display, high-frequency spatial details are filtered by the PCA and ICA transforms and spectral artefacts emerge following the two transformations.

KL (R-G-B, mRMR_X-Y-Z) = 0.327553824203326
KL (R-G-B, mRMR_Y-X-Z) = 0.303522796330313
KL (R-G-B, PCA_X-Y-Z) = 1.383523583171621
KL (R-G-B, PCA_Y-X-Z) = 0.999025574181783
KL (R-G-B, ICA_X-Y-Z) = 1.573529476403733
KL (R-G-B, ICA_Y-X-Z) = 2.103526699090990

6.6.2 mRMR score similarity statistics & physical modelling

This is an information-based method that ranks the spectral bands of a satellite image according to the amount of information relevant to a user-defined target class, while simultaneously reducing the inter-feature redundancy. Two important questions that need to be addressed is what is the physical meaning of the results and if they can be verified by remote sensing literature studies.

We evaluated the capabilities of the mRMR criterion to discover the top three spectral bands in 64 satellite images recorded by various sensors, using the same target class Forest Vegetation. Results reveal a high similarity between the top three bands for the same class of interest. Because the physical responses of vegetation in remote sensing imagery have been extensively studied during the last 30 years, we verified if physical models support the results. We discovered that independent of the sensor, the NIR, MIR, SWIR bands (where applicable) were ranked among the first mRMR top three bands.

Table 6.6 presents a statistical analysis of results across 64 satellite images. Testing the ranking capabilities of the mRMR criterion on 20 Landsat ETM+ images (Forest Vegetation class) the following conclusions were drawn: spectral band 6-2 (SWIR 2100-2135 nm) was ranked in the top three spectral bands in 90% of the cases and band 6-1 (TIR 1040-1250 nm) was ranked in the top three bands in 95% of the cases. When testing 17 SPOT-5 images, band 4 (MIR 1580-1750 nm) was ranked among the first three in 100% of the cases and band 3 (NIR 780-890 nm) in 76% of the cases. With Quickbird and GeoEye-1 data, band 4 (NIR 760-900nm / 780-920 nm) was ranked in the top three spectral bands in 100% of the cases. When testing MERIS satellite data, bands 10 (750-760 nm) and 13 (845-885 nm) were ranked in 100% of the cases among the top three bands and bands 14 (880-900 nm) and 15 (890-910nm) in 50% of the cases. Conclusions show that independent of the sensor, specific wavelength intervals are relevant for specific target classes and estimations and predictions can be made based on this generalization capability.

Studies performed on WorldView-2 images show interesting results: band 4 (Yellow 585-625 nm) was ranked among the first three bands in 100% of the cases and bands 1 (Coastal 400-450 nm) and 7 (NIR-1 770-895 nm) in 66% of the cases. These results were expected because the Yellow band is highly important for vegetation applications. Plants' spectral responses in this spectral domain are directly correlated to their health status. This band is used to evaluate individual tree crown cover and leaf health. The Coastal band also optimizes vegetation

identification and analysis upon its chlorophyll penetration characteristics [157]. Measurements on RapidEye data ranked band 3 (Red 630-685 nm) among the first three features in 87% of the cases, followed by band 2 (Green 520-590 nm) and band 4 (Red Edge 690-730 nm) in 50% of the cases. The Red Edge band measures the chlorophyll production and plant health status. Researchers have demonstrated that the Red Edge band can better discriminate healthy and damaged trees, can reveal the differences between young and mature plants and even discriminate species [181-183].

These results conclude that the mRMR criterion has the potential to discover the optimum wavelength intervals that hold the maximum amount of information relevant to a specific target class and to simultaneously minimize spectral correlation among the available features. In order to rank the features only by the amount of relevant information without taking into account the inter-band redundancy, the maximum-relevance (MR) criterion yields reliable results.

Sensor	Number of Images	Statistics
Landsat ETM+	20	Band 6-2 – 90% Band 6-1 – 95% Band 7 – 40% Band 2 – 20% Band 4 – 15%
SPOT-5	17	Band 4 – 100% Band 3 – 76% Band 2 – 76% Band 1 – 47%
Quickbird	4	Band 4 – 100% Band 2 – 100% Band 3 – 100%
GeoEye-1	2	Band 4 – 100% Band 3 – 100% Band 2 – 50% Band 1 – 50%
WorldView-2	3	Band 6 – 100% Band 1 – 66% Band 7 – 66% Band 8 – 33%
MERIS	2	Band 10 – 100% Band 13 – 100% Band 14 – 50% Band 15 – 50%
RapidEye	16	Band 5 – 50% Band 4 – 62% Band 3 – 87% Band 2 – 81%

Table 6.6

The mRMR criterion was employed to discover the top three spectral bands in 64 satellite images recorded by various sensors, using the same target class Forest Vegetation (table 6.7). Results reveal high similarity between the top three bands for the same class of interest. The mRMR criterion has the potential to discover the optimum wavelength intervals that contain the maximum amount of information relevant for a specific target class and to simultaneously minimize spectral correlation among the available features.

Another possibility to assess the physical meaning of results is to investigate the visualization generated by the mRMR criterion against a spectral map. For this purpose we use the model described in [63] - chapter 5 - and compare the spectral classes retrieved from satellite images to the new visualization.

The model is a purely spectral per-pixel rule-based classifier, based solely on the spectral domain and prior knowledge retrieved from the remote sensing literature. It requires no training and performs a fully unsupervised preliminary classification over multiple sensors' imagery calibrated into planetary reflectance. The degree of user supervision required to detect spectral rule-based categories is the same as unsupervised data clustering and far inferior to reference sample selection required by supervised classifiers. The classifier is based on prior spectral knowledge and in the following paragraph we briefly summarize its characteristics. Pattern recognition is based exclusively on known spectral signatures of target classes taken from remote sensing literature and adapted as fuzzy data templates. This implies that the classification system is pixel-based (context-insensitive) and purely spectral. It uses a set of spectral rules and the mapping system employs no supervised data learning mechanism to dynamically generate new rules.
For a throughout description of the system, the reader may refer to [63]. Using the Landsat ETM+ image depicted in figure 6.33a, we generated a set of 12 spectral categories using the physical model (figure 6.33b) and a new visualization with the mRMR criterion (figure 6.33c). In this case the target class is "Forest Vegetation" to continue the previous tests and the new visualization is:

* R channel - TIR
* G channel - SWIR
* B channel - NIR

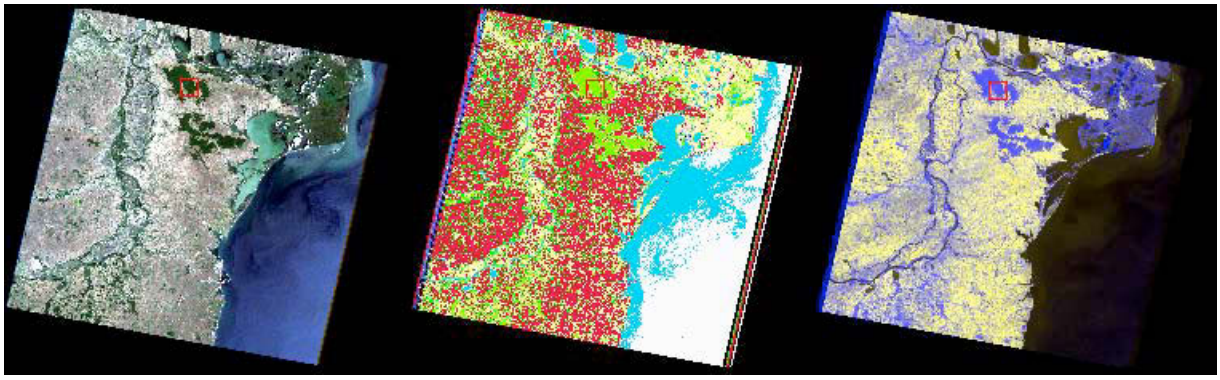
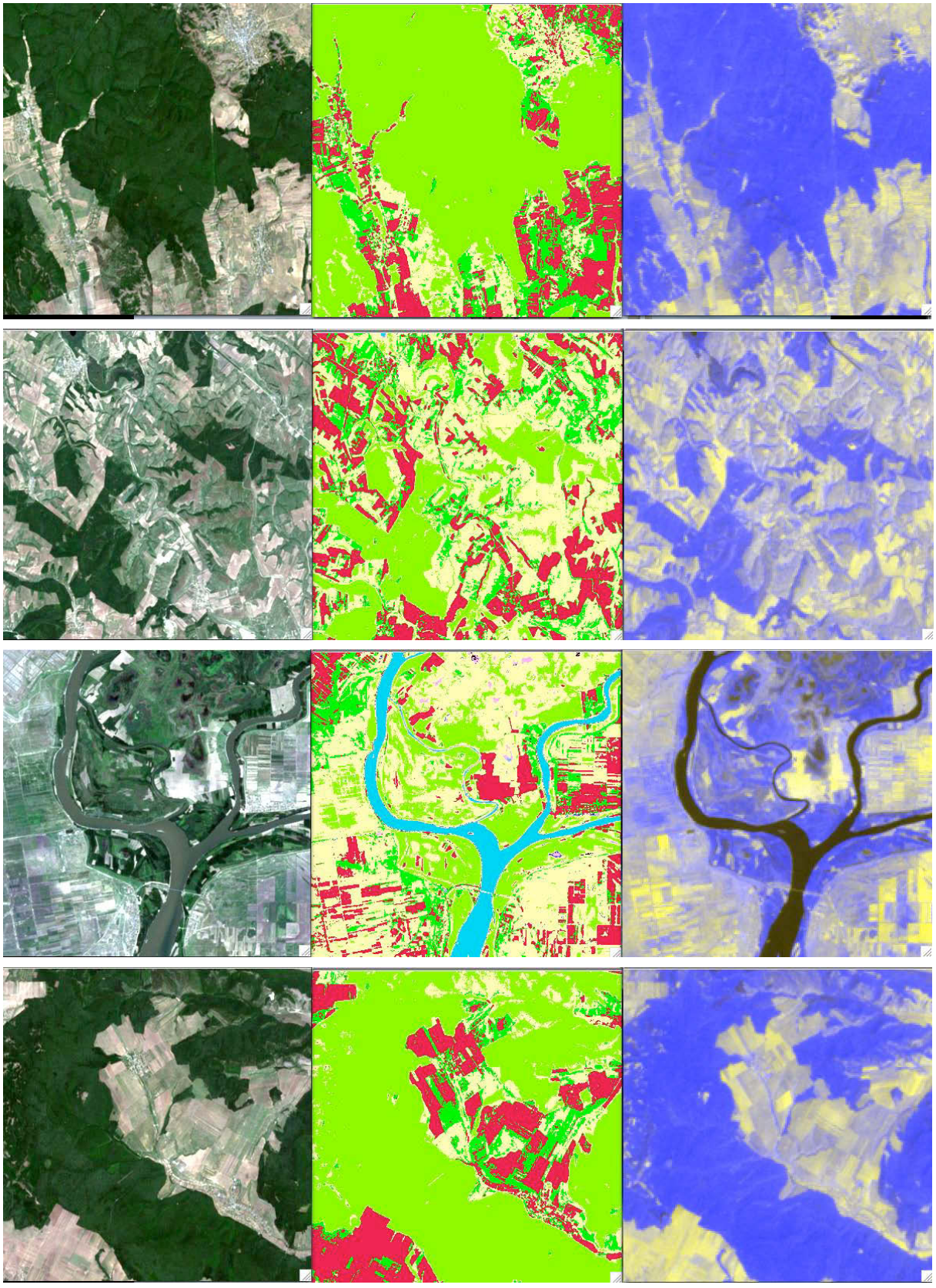


Figure 6.33 - (a) Landsat ETM+ satellite image
(b) Map with 12 spectral categories
(c) New visualization generated by the mRMR criterion



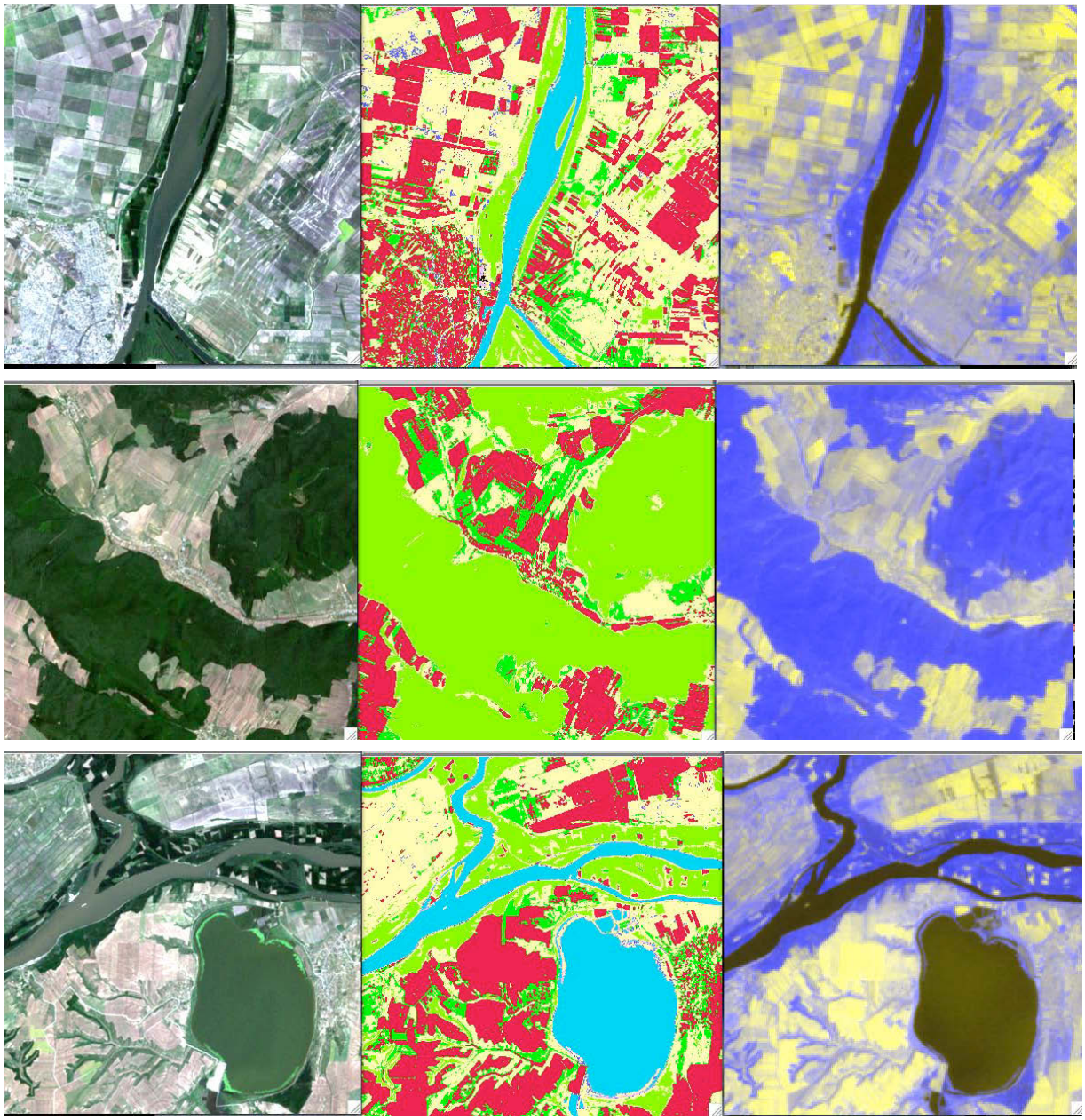


Figure 6.34 - The left column shows the details from Landsat ETM+ satellite image (R-G-B), the middle column shows the corresponding details in the spectral map and the target class depicted in green and the right column the new visualization (X-Y-Z display) with the target class depicted in blue.

Figure 6.34 reveals a high correlation between the spectral class "high vegetation" generated by the physical model and the corresponding regions in the new visualization mode. The left column shows the details from Landsat ETM+ satellite image in R-G-B display, the middle column shows the corresponding details in the spectral map with 12 categories and the target class depicted in green and the right column the new visualization (X-Y-Z display) with the target class depicted in blue. These examples confirm the fact that the physical meaning of results is maintained in the new mRMR visualization.

The visualization technique can also improve the accuracy assessment procedures of different automatic algorithms and support different users to discover classification errors in their results. Figure 6.35 shows a detail from a Landsat image, the corresponding spectral

map generated by the physical model and the same detail depicted using the new visualization. The model misclassified a large area in the image as "high vegetation" (figure 6.35b), while the new visualization (figure 6.35c) clearly shows the same region belonging to a different class. Although the R-G-B display (figure 6.35a) can be used to evaluate the classification results, the new display increases the confidence of the visual investigation.

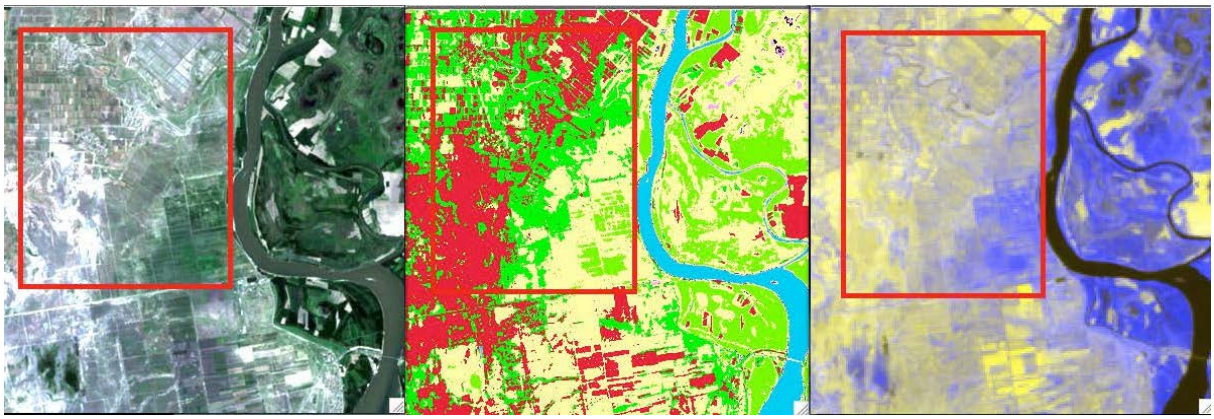


Figure 6.35 - (a) Landsat image R-G-B, (b) spectral map showing misclassification of Barren Land as Forest (c) new visualization X-Y-Z showing increased confidence in assessing the accuracy of classification maps.

The following case studies and results how the algorithm can be integrated in operational activities relying on multispectral satellite images. These examples present only the qualitative visual analysis and all have been performed on WorldView-2 images.

Figure 6.36 shows a visual comparison between the natural color display (bands 5, 3, 2) and the new visualization (bands 6, 4, 5) discovered with this band ranking method in the case of a flooding event. The new visualization clearly outlines the flooded areas in blue. In the natural color display the flooded areas present high similarity to the vegetation areas.

Figure 6.37 shows a visual comparison between the natural color display (bands 5, 3, 2) and the new visualization (bands 8, 7, 6) used for an agricultural study. The new visualization clearly indicates the healthy areas (depicted in white) and the areas with a lower health status (the irregular patches depicted in blue). This differentiation is almost invisible in the natural color display.

Figure 6.38 shows a visual comparison between the natural color display (bands 5, 3, 2) and the new visualization (bands 8, 7, 6) for the evaluation of a large forested area. Barely noticeable in the natural color combination, various vegetation types are clearly revealed in the new visualization in multiple variations of yellow.

Figure 6.39 shows the visual comparison between the natural color display (bands 5, 3, 2) and the new visualization (bands 8, 7, 6) for investigating a large forest area under fire. The algorithm selects the three features that allow visualization of areas under the smoke plumes.

Figure 6.40 takes the previous study a step further and discovers the optimum combination of spectral bands for visualizing smoke plumes (bands 1, 3, 2). The new visualization shows enhanced contrast between the target class and the surrounding areas, thus allowing a more efficient mapping of smoke plumes.

Figure 6.41 shows the visual comparison between the natural color display (bands 5, 3, 2) and the new visualization (bands 4, 1, 7) discovered for the investigation of man-made objects in the desert. The new combination of features depicts enhanced contrast between the target class and the surrounding areas.

Figure 6.42 shows the visual comparison between the natural color display (bands 5, 3, 2) and the new visualization (bands 4, 8, 6) employed for studying the toxic waters resulting from mining activities in the Australian deserts. The new combination of features depicts enhanced contrast between the target class and the surrounding areas. Results show that the polluted waters blend with the background in the natural color display and clearly stand out in the new visualization mode.

Conclusions

This chapter presented an interactive technique to discover the optimum combination of three spectral features of a multi-band satellite image applied to enhance visualization of learned targets and phenomena of interest. The software prototype is designed to support Earth Observation service providers to improve and testify the accuracy and reliability of their products and guarantee the quality of information offered to end users. The method is designed to assist the geospatial operator in understanding the satellite image through optimum representations and to offer cognitive support in discovering relevant information in the scenes.

A key point of this approach is its broad applicability, it can be used in all civil, commercial, defence and security operations to generate increased confidence in information products and services. The method facilitates an easier and more reliable interpretation of data and generates increased amount of information derived directly from EO images. The system models the human visual system and guarantees the correlation between the output of the machine and the understanding of the human operator. For this reason, users from commercial and academic environments, working in fields like natural resources and conflicts, nuclear and treaties monitoring, crisis management and assessment, marine safety, marine and coastal environmental monitoring, emergency response, spatial planning, agriculture, water management, forestry, natural resources can benefit from the capabilities of this visualization prototype. Through various examples and experiments, we demonstrated its operational efficiency and inter-operational adaptability using scenes from around the world on a wide variety of sensors and EO-based applications.

Figure 6.36 - WorldView-2, Flooding events
Natural Color (bands 5,3,2) & New Visualization (bands 6,4,5)

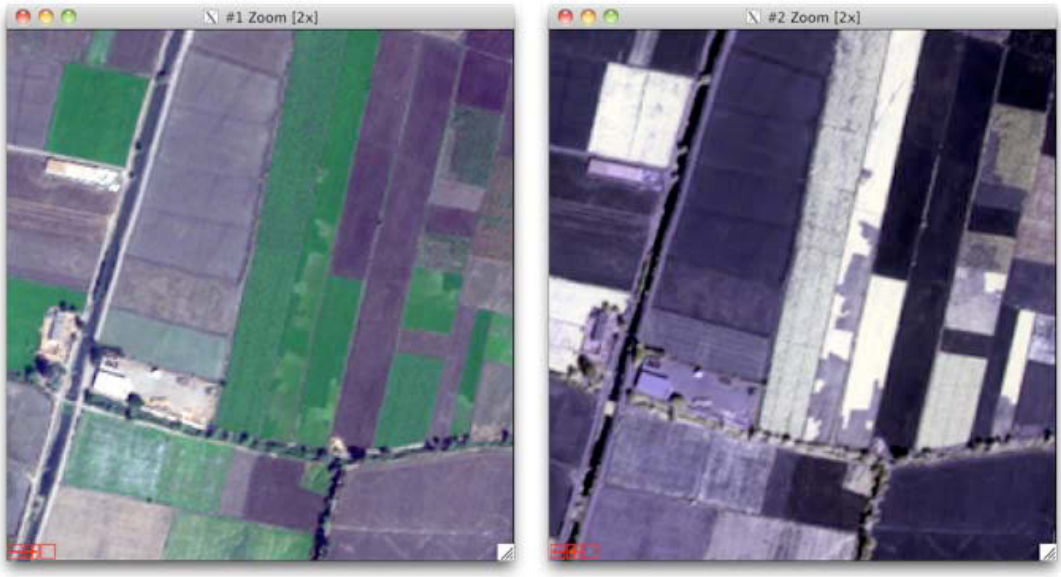


Figure 6.37 - WorldView-2, Agriculture, crop health status
Natural Color (bands 5,3,2) & New Visualization (bands 8,7,6)

Figure 6.38 - WorldView-2, Forest areas
Natural Color (bands 5,3,2) & New Visualization (bands 8,7,6)

Figure 6.39 -WorldView-2, Forest areas – note how the algorithm allows visualization under plumes, Natural Color (bands 5,3,2) & New Visualization (bands 8,7,6)

Figure 6.40 WorldView-2, Smoke plumes – note how the algorithm reveals areas covered by plumes, Natural Color (bands 5,3,2) & New Visualization (bands 1,3,2)

Figure 6.41 - WorldView-2, Man-made objects in the desert
Natural Color (bands 5,3,2) & New Visualization (bands 4,1,7)

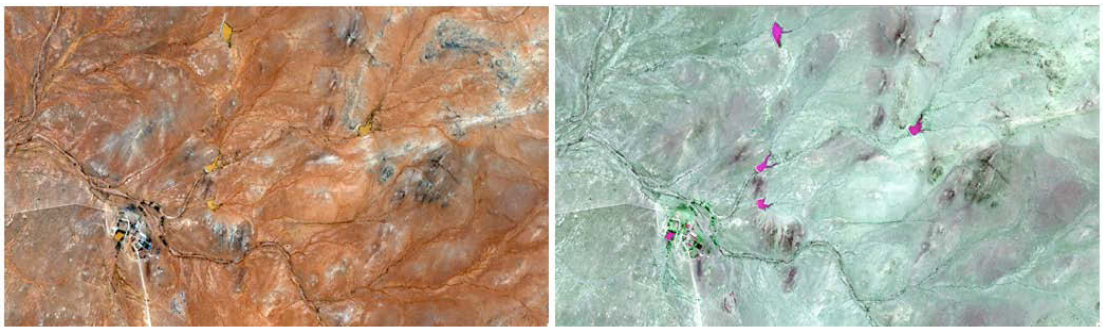


Figure 6.42 - WorldView-2, Toxic waters, mining activity
Natural Color (bands 5,3,2) & New Visualization (bands 4,8,6)

7

Conclusions

Although the availability of data has increased and its cost decreased, statistics show that less than 5% of the satellite images available in archives are actually downloaded, analyzed and applied in operational scenarios. Studies show that users require the information that can be derived from remote sensing images but the process of obtaining that information is still prohibitive for main stream communities for multiple reasons:

(1) Data are stored in large collections and access is available only via specific metadata (e.g. geographic location, acquisition date, sensor) that does not always offer a user-friendly interface. The operator accessing the collection must have knowledge of how to operate a geospatial database, how to evaluate the query results using only a visual inspection and finally, how to generate information from the raw files of a satellite image. Most of the times, this workflow can be approached only by a limited number of experts and end users depend on their availability and capability to access, process and translate satellite imaging data into information ready to be implemented in projects.

(2) Reliable information can be obtained from the data only through highly complex processes that require expert knowledge, dedicated software tools and extended periods of interactive analyses.

(3) The output information products (i.e. maps, reports) are subject to operator bias. The link between data and knowledge is not standardized and this may lead to misunderstandings between users working in various fields.

The recently developed cartographic products (e.g. Corine Land Cover, Urban Atlas) offer land use land cover inventories of the Earth using a standard nomenclature, clearly explained to be applied by users in different areas. However, these information products are very limited due to the standard spatial resolution (minimum mapping unit), to the standard nomenclature, not leaving users the possibility to derive their custom semantic concepts and to the extended amount of time and high cost necessary to be produced. In some areas, the land cover changes and the layers are outdated before they reach the final users.

In recent years, the fields of IIM and CBIR have been investigating new solutions for querying collections by image content described using semantic concepts that can be used, understood and disseminated by multiple users. These state-of-the-art systems employ advanced algorithms from image processing, data dimensionality reduction, text processing, data compression and mining and semantic knowledge discovery. To become fully operational, these algorithms need to be implemented in user-friendly software packages.

7.1 The Value of the Contributions

In this dissertation we combined knowledge from IIM and CBIR - signal processing, image analysis, pattern recognition, artificial intelligence, machine learning, information theory, databases, semantic knowledge discovery - to design the concept of a new system that brings solutions to query large image archives directly by content, to bridge the gap of understanding between machine and human languages and between science and operations. This thesis also brings several solutions to the IIM domain by developing new image processing algorithms for latent information discovery in satellite images and visual data mining based on models of the human visual system. To summarize the value of the contributions, we enumerate the following key aspects:

(1) Semantic Rules Discovery - we bring a solution to one of the most challenging puzzles scientists have to solve: bridging the semantic gap between the low-level image features and the high-level semantic concepts. Using our approach, users can generate customized vocabularies, discover the semantic rules that link images with cartographic data and create new maps that have semantic meaning. By discovering the set of rules that explain semantic classes in available cartographic systems, we introduce the prototype of an interactive learning loop that uses the concept of direct semantics applied on satellite imagery.

(2) Visual Mining of Satellite Images - because the final decision in almost all EO-based applications is made by a human operator, we developed a visual mining module that is essential for an IIM system. This module can be implemented as a stand-alone tool used for advanced visual analysis and also offers the possibility to be integrated into commercial software packages. It can be applied in operational tasks to verify the quality of data, to search and analyze for objects and classes of interest, to improve selection of learning areas for mining algorithms and to validate the output of automatic image understanding methods.

(3) Theoretical concepts - we adapted the LDA model to bridge the semantic gap between low-level features of an image and high-level semantic concepts attached by a user or available in standard cartographic data. The LDA model was initially developed to annotate large collections of text documents but studies conclude that future work should focus on developing a simplified version of this algorithm. Text documents contain a more complex vocabulary of words and the satellite images contain more complex spatial and contextual structures. For this reason, a more simple vocabulary of visual words is required to describe these structures. Future work will aim at developing a simplified LDA model that can operate optimally with complex spatial structures.

This dissertation presented novel concepts and methods that support users to access and discover latent information in large image collections, visualize, analyze and interpret satellite scenes in an interactive, human-centered, result-driven workflow.

APPENDIX

1. Color Science

The first part of the Appendix describes the principles of color science, the mathematics of color perception and the color models used in this dissertation. For thorough explanations and further details, the read is advised to refer to [264-275], [70-71], [144], [243-250], [253-256].

Color vision is the ability of an organism or machine to distinguish objects based on the wavelengths of the light they reflect, emit, or transmit. Colors can be measured and quantified in various ways. The human perception of colors is a subjective process whereby the brain responds to the stimuli that are produced when incoming light reacts with the several types of cone photoreceptors in the eye. In essence, different people may see the same illuminated object or light source in different ways. Perception of color begins with specialized retinal cells containing pigments with different spectral sensitivities, known as cone cells. In humans, there are three types of cones sensitive to three different spectra, resulting in trichromatic color vision. The cones are conventionally labeled according to the ordering of the wavelengths of the peaks of their spectral sensitivities: short (S), medium (M), and long (L) cone types. These three types do not correspond well to particular colors as we know them. Rather, the perception of color is achieved by a complex process that starts with the differential output of these cells in the retina and it will be finalized in the visual cortex and associative areas of the brain. The peak response of human cone cells varies, even among individuals with 'normal' color vision.

Two complementary theories of color vision are the trichromatic theory and the opponent process theory. The trichromatic theory proposed in the 19th century states that the retina's three types of cones are preferentially sensitive to blue, green, and red. Ewald Hering proposed the opponent process theory in 1872. It states that the visual system interprets color in an antagonistic way: red vs. green, blue vs. yellow, black vs. white. Both theories are now accepted as valid, describing different stages in visual physiology.

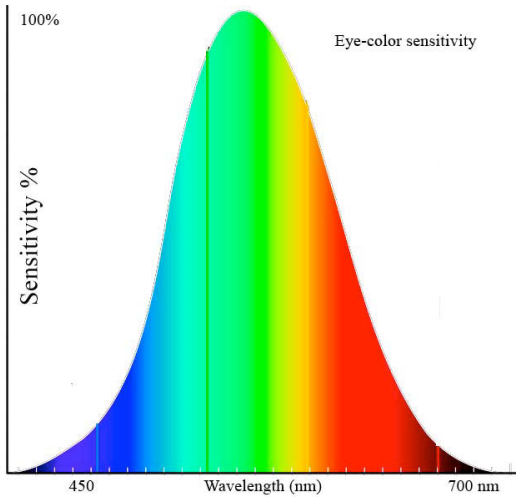


Figure A1.1 – Eye Color sensitivity

Nothing categorically distinguishes the visible spectrum of electromagnetic radiation from invisible portions of the broader spectrum. In this sense, color is not a property of electromagnetic radiation, but a feature of visual perception by an observer. Furthermore, there is an arbitrary mapping between wavelengths of light in the visual spectrum and human experiences of color. The possibility of a clean dissociation between color experience from properties of the world reveals that color is a subjective psychological phenomenon.

Mathematics of Color Perception

A "physical color" is a combination of pure spectral colors in the visible range. Since there are, in principle, infinitely many distinct spectral colors, the set of all physical colors may be thought of as an infinite-dimensional vector space. We call this space H-color. An element C of H-color space is a function from the range of visible wavelengths—considered as an interval of real numbers $[\lambda_{min},\lambda_{max}]$ to the real numbers, assigning to each wavelength in $[\lambda_{min},\lambda_{max}]$ its intensity $C(\lambda)$. A humanly perceived color may be modeled as three numbers: the extents to which each of the 3 types of cones is stimulated. Thus a perceived color may be thought of as a point in 3-dimensional Euclidean space. This space is called the R3-color.

Since each wavelength λ stimulates each of the 3 types of cone cells to a known extent, these extents may be represented by 3 functions $S(\lambda)$, $M(\lambda)$ and $L(\lambda)$ corresponding to the response of the *S, M,* and *L* cone cells, respectively.

Finally, since a beam of light can be composed of many different wavelengths, to determine the extent to which a physical color *C* in H-color stimulates each cone cell, we must calculate the integral (with respect to λ), over the interval $[\lambda_{min},\lambda_{max}]$ of $C(\lambda)\bullet S(\lambda)$ of $C(\lambda)\bullet M(\lambda)$ and of $C(\lambda)\bullet L(\lambda)$. The triple of resulting numbers associates to each physical color *C* to a particular perceived color which is a single point in R3-color. Many different elements in the H-color space can result in the same single perceived color in R3-color, so a perceived color is not unique to one physical color.

Thus human color perception is determined by a specific, non-unique linear mapping from the infinite-dimensional space H-color to the 3-dimensional Euclidean space R3-color. The image of the mathematical cone over the simplex whose vertices are the spectral colors, by

linear mapping, is also a mathematical cone in R3-color. Moving directly away from the vertex of this cone represents maintaining the same chromaticity while increasing its intensity. Taking a cross-section of this cone yields a 2-D chromaticity space. Both the 3-D cone and its projection or cross-section are convex sets; that is, any mixture of spectral colors is also a color. In practice, it would be quite difficult to physiologically measure an individual's three cone responses to various physical color stimuli. Instead, a psychophysical approach is taken. Three specific benchmark test lights are typically used; let us call them *S, M* and *L*. To calibrate human perceptual space, scientists allowed human subjects to try to match any physical color by turning dials to create specific combinations of intensities (*IS, IM, IL*) for the *S, M*, and *L* lights, resp., until a match was found. This needed only to be done for physical colors that are spectral, since a linear combination of spectral colors will be matched by the same linear combination of their (*IS, IM, IL*) matches.

By considering all the resulting combinations of intensities (*IS, IM, IL*) as a subset of 3-space, a model for human perceptual color space is formed. A color model is an abstract mathematical model describing the way colors can be represented as numbers, typically as three or four values or color components. When this model is associated with a precise description of how the components are to be interpreted (viewing conditions, etc.), the resulting set of colors is called color space. This section of the Appendix describes ways in which human color vision can be modeled.

CIE 1931 Color Space

The CIE XYZ color space encompasses all color sensations that humans can experience. It serves as a standard reference against which many other color spaces are defined. A set of color-matching functions, like the spectral sensitivity curves of the LMS space but not restricted to be nonnegative sensitivities, associates physically-produced light spectra with specific tristimulus values.

Most wavelengths will not stimulate just one type of cone cell only, because the spectral sensitivity curves of the three types of cone cells overlap. Certain tristimulus values are thus physically impossible. And LMS tristimulus values for pure spectral colors would, in any normal trichromatic additive color space (e.g. RGB color spaces), imply negative values for at least one of the three primaries, since the chromaticity would be outside the color triangle defined by the primary colors. To avoid these negative RGB values, and to have one component that describes the perceived brightness, "imaginary" primary colors and corresponding color-matching functions have been formulated. The resulting tristimulus values are defined by the CIE 1931 color space, in which they are denoted *X, Y*, and *Z*.

When judging the relative luminance (brightness) of different colors in well-lit situations, humans tend to perceive light within the green parts of the spectrum as brighter than red or blue light of equal power. The luminosity function that describes the perceived brightness of different wavelengths is thus roughly analogous to the spectral sensitivity of M cones. The CIE model capitalizes on this fact by defining Y as luminance. Z is quasi-equal to blue stimulation, or the S cone response, and X is a mix (a linear combination) of cone response curves chosen to be nonnegative. The XYZ tristimulus values are thus analogous to, but not equal to, the LMS cone responses of the human eye. Defining Y as luminance has the useful result that for any given Y value, the XZ plane will contain all possible chromaticities at that luminance.

Color Matching Functions

The CIE's color matching functions $\bar{x}(\lambda),\bar{y}(\lambda),\bar{z}(\lambda)$are the numerical description of the chromatic response of the human observer. They can be regarded as the spectral sensitivity curves of three linear light detectors yielding the CIE tristimulus values *X, Y, Z*. The tristimulus values for a color with a spectral power distribution $I(\lambda)$ are given by:

$$X = \int_{380}^{780} I(\lambda)\bar{x}(\lambda)d\lambda$$

$$Y = \int_{380}^{780} I(\lambda)\bar{y}(\lambda)d\lambda \qquad \text{(A1.1)}$$

$$Z = \int_{380}^{780} I(\lambda)\bar{z}(\lambda)d\lambda$$

where λ is the wavelength of the equivalent monochromatic light measured in nanometers.

CIE Chromaticity Diagram & CIE Color Space

Since the human eye has three types of color sensors that respond to different ranges of wavelengths, a full plot of all visible colors is a three-dimensional image. The concept of color can be divided into two parts: brightness and chromaticity. The CIE XYZ color space was deliberately designed so that the *Y* parameter was a measure of the brightness or luminance of a color. The chromaticity of a color was then specified by the two derived parameters *x* and *y*, two of the three normalized values which are functions of all three tristimulus values *X, Y*, and *Z*:

$$x = \frac{X}{X+Y+Z}$$

$$y = \frac{Y}{X+Y+Z} \qquad \text{(A1.2)}$$

$$z = \frac{Z}{X+Y+Z} = 1-x-y$$

The derived color space specified by *x, y,* and *Y* is known as the **CIE xyY** color space and is widely used to specify colors in practice. The *X* and *Z* tristimulus values can be calculated back from the chromaticity values *x* and *y* and the *Y* tristimulus value:

$$X = \frac{Y}{y}x \qquad \text{(A1.3)}$$

$$Z = \frac{Y}{y}(1-x-y) \qquad \text{(A1.4)}$$

The chromaticity diagram is an instrument specifying how the human eye will experience light with a given spectrum. The outer curved boundary is the spectral locus, with the wavelengths shown in nanometers.

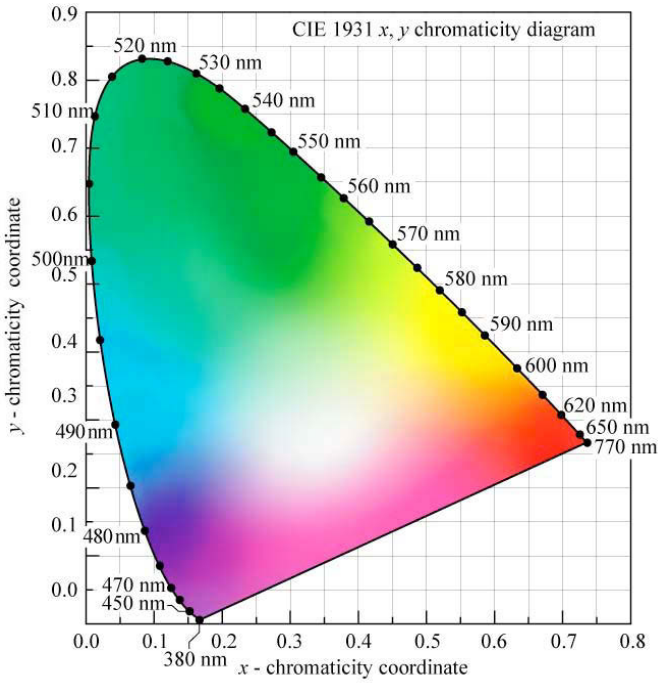


Figure A1.2 – CIE 1931 XY chromaticity diagram

The chromaticity diagram illustrates a number of properties of the CIE XYZ color space:

1. The diagram represents all of the chromaticities visible to the human being - the gamut of human vision. The gamut of all visible chromaticities on the CIE plot shown in figure A1.2. The curved edge of the gamut is called the spectral locus and corresponds to monochromatic light (each point representing a pure hue of a single wavelength), with wavelengths listed in nanometers. The straight edge on the lower part of the gamut is called the line of purples. These colors, although they are on the border of the gamut, have no counterpart in monochromatic light. Less saturated colors appear in the interior of the figure with white at the center.

2. It is seen that all visible chromaticities correspond to non-negative values of x, y, and z and therefore to non-negative values of X, Y, and Z.

3. If one chooses any two points of color on the chromaticity diagram, then all the colors that lie in a straight line between the two points can be formed by mixing these two colors. It follows that the gamut of colors must be convex in shape. All colors that can be formed by mixing three sources are found inside the triangle formed by the source points on the chromaticity diagram

4. An equal mixture of two equally bright colors will not generally lie on the midpoint of that line segment. In more general terms, a distance on the xy chromaticity diagram does not correspond to the degree of difference between two colors. In the early 1940s, David MacAdam studied the nature of visual sensitivity to color differences, and summarized his results in the concept of a MacAdam ellipse. Based on the work of MacAdam, the CIE 1960,

CIE 1964, and CIE 1976 color spaces were developed, with the goal of achieving perceptual uniformity (have an equal distance in the color space correspond to equal differences in color). Although they were a distinct improvement over the CIE 1931 system, they were not completely free of distortion.

CIE RGB Color Space

The CIE RGB color space is one of many RGB color spaces, distinguished by a particular set of monochromatic primary colors. The standardized CIE RGB color matching functions $\bar{r}(\lambda),\bar{g}(\lambda),\bar{b}(\lambda)$ are obtained using three monochromatic primaries at standardized wavelengths of 700 nm (red), 546.1 nm (green) and 435.8 nm (blue). The color matching functions are the amounts of primaries needed to match the monochromatic test primary. These functions are depicted in figure A1.3.

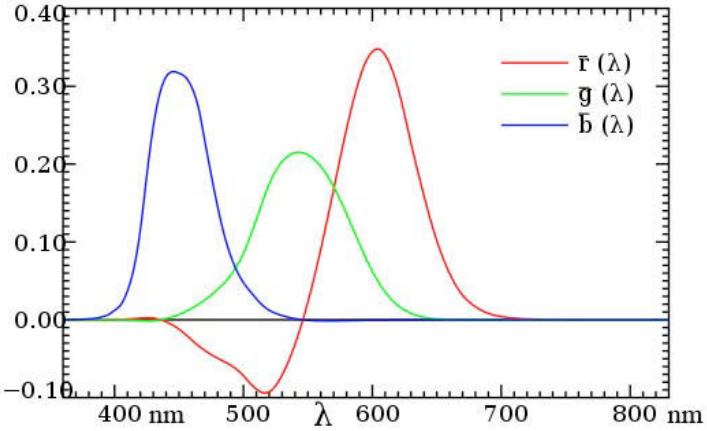


Figure A1.3 - The CIE 1931 RGB Color matching functions. The color matching functions are the amounts of primaries needed to match the monochromatic test primary at the wavelength shown on the horizontal scale.

The primaries with wavelengths 546.1 nm and 435.8 nm were chosen because they are easily reproducible monochromatic lines of a mercury vapor discharge. The 700 nm wavelength, which in 1931 was difficult to reproduce as a monochromatic beam, was chosen because the eye's perception of color is rather unchanging at this wavelength, and therefore small errors in wavelength of this primary would have little effect on the results. The curves are normalized to have constant area beneath them, fixed to a particular value by specifying that:

$$\int_0^\infty \bar{r}(\lambda)d\lambda = \int_0^\infty \bar{g}(\lambda)d\lambda = \int_0^\infty \bar{b}(\lambda)d\lambda \qquad \text{(A1.5)}$$

The resulting normalized color matching functions are then scaled in the r:g:b ratio of 1:4.5907:0.0601 for source luminance and 72.0962:1.3791:1 for source radiant power to reproduce the true color matching functions. By proposing that the primaries be standardized, the CIE established an international system of objective color notation.

Given these scaled color matching functions, the RGB tristimulus values for a color with a spectral power distribution $I(\lambda)$ would then be given by:

$$R = \int_0^\infty I(\lambda)\bar{r}(\lambda)d\lambda$$
$$G = \int_0^\infty I(\lambda)\bar{g}(\lambda)d\lambda \qquad \text{(A1.6)}$$
$$B = \int_0^\infty I(\lambda)\bar{b}(\lambda)d\lambda$$

The CIE RGB space can be used to define chromaticity the usual way:

$$r = \frac{R}{R+G+B}$$
$$g = \frac{G}{R+G+B} \qquad \text{(A1.7)}$$

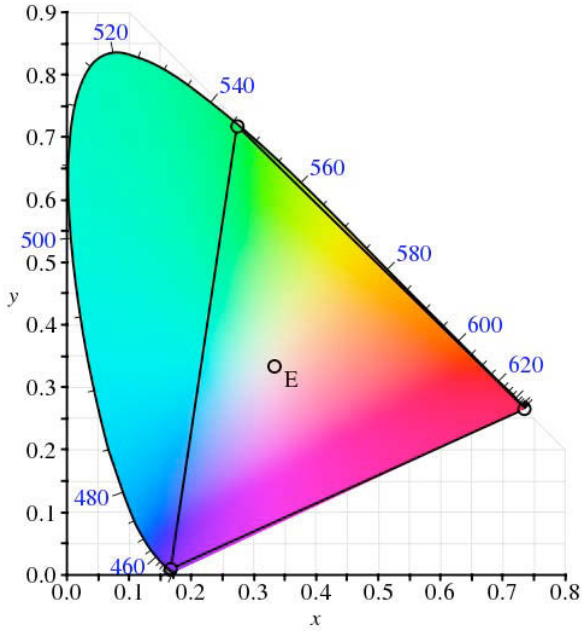


Figure A1.4 – Gamut of the CIE RGB primaries and location on the CIE 1931 *xy* chromaticity diagram

The standard transformation between the CIE RGB and XYZ spaces is described by:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{1}{b_{21}} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \frac{1}{0.17697} \begin{bmatrix} 0.49 & 0.31 & 0.20 \\ 0.17687 & 0.81240 & 0.01063 \\ 0.00 & 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad \text{(A1.8)}$$

While the above matrix is exactly specified in standards, going the other direction uses an inverse matrix that is not exactly specified, but is approximately defined by:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 0.41847 & -0.15866 & -0.082835 \\ -0.091169 & 0.25243 & 0.015708 \\ 0.00092090 & -0.0025498 & 0.17860 \end{bmatrix} \bullet \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad \text{(A1.9)}$$

Color Difference

The International Commission on Illumination (CIE) calls their distance metric ΔE^*_{ab}. Different studies have proposed different ΔE^*_{ab} values that have a JND - just noticeable difference. Recently the standard has been set to a JND of $\Delta E^*_{ab} = 2.3$. However, perceptual non-uniformities in the underlying CIELAB color space prevent this and have led to the CIE's refining their definition over the years, leading to the superior CIE1994 and CIE2000 formulas. These non-uniformities are important because the human eye is more sensitive to certain colors than others. A good metric should take this into account in order for the notion of a JND to have meaning. Otherwise, a certain ΔE^*_{ab} that may be insignificant between two colors that the eye is insensitive to may be conspicuous in another part of the spectrum.

CIE1976

In colorimetry, the CIE 1976 (L*, u*, v*) color space, commonly known by its abbreviation CIELUV, is a color space adopted by the International Commission on Illumination CIE in 1976, as a simple-to-compute transformation of the 1931 CIE XYZ color space, that attempted perceptual uniformity.

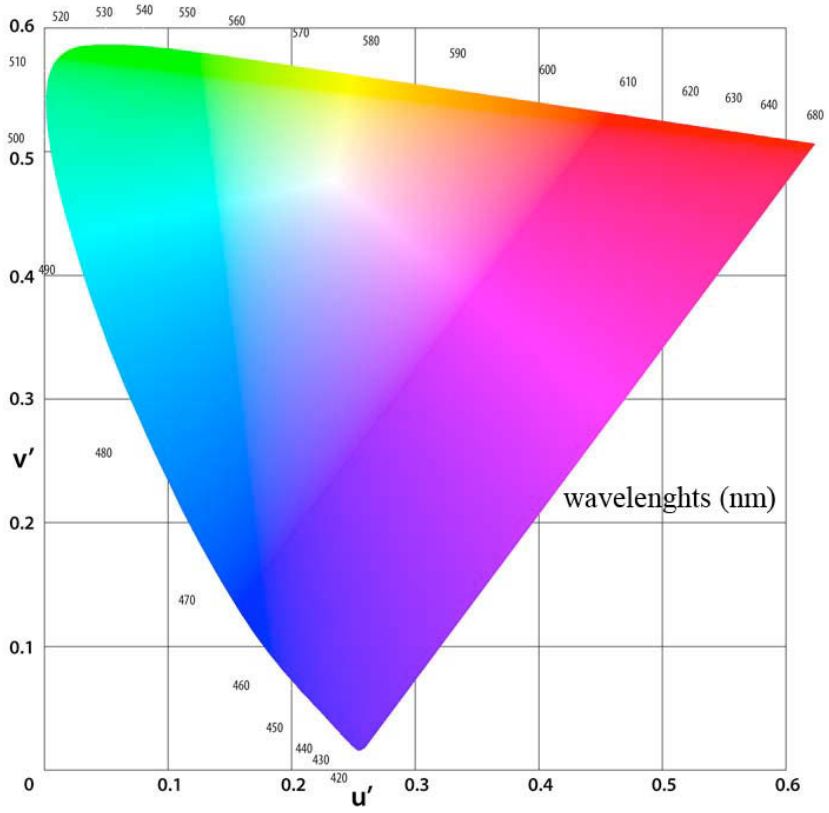


Figure A1.5 - CIE 1976 uniform chromaticity scale diagram.

CIELUV attempts to define an encoding with uniformity in the perceptibility of color differences. The non-linear relations between CIE XYZ and CIELUV are expressed with:

$$L^* = \begin{cases} \left(\frac{29}{3}\right)^3 Y/Y_n, & Y/Y_n \le \left(\frac{6}{29}\right)^3 \\ 116\left(Y/Y_n\right)^{1/3} - 16, & Y/Y_n > \left(\frac{6}{29}\right)^3 \end{cases} \tag{A1.10}$$

$$u^* = 13L^* \bullet \left(u' - u_n'\right)$$
$$v^* = 13L^* \bullet \left(v' - v_n'\right) \tag{A1.11}$$

The quantities u_n' and v_n' are the (u′, v′) chromaticity coordinates of the white point – and Y_n is its luminance. In reflection mode, this is often but not always taken as the (u′, v′) of the perfect reflecting diffuser under that illuminant. Equations for u′ and v′ are given below:

$$u' = \frac{4X}{X + 15Y + 3Z} = \frac{4x}{-2x + 12y + 3}$$

(A1.12)

$$v' = \frac{9Y}{X + 15Y + 3Z} = \frac{9y}{-2x + 12y + 3}$$

The transformation from (*u',v'*) to (*x, y*) is the following:

$$x = \frac{9u'}{6u' - 16v' + 12}$$
$$y = \frac{4v'}{6u' - 16v' + 12} \tag{A1.13}$$

The transformation from CIELUV to CIE XYZ is defined with:

$$Y = \begin{cases} Y_n \bullet L^* \bullet \left(\frac{3}{29}\right)^3, & L^* \le 8 \\ Y_n \bullet \left(\frac{L^* + 16}{116}\right)^3, & L^* > 8 \end{cases} \tag{A1.14}$$

$$X = Y \bullet \frac{9u'}{4v'} \tag{A1.15}$$

$$Z = Y \bullet \frac{12 - 3u' - 20v'}{4v'} \tag{A1.16}$$

Cylindrical Representation

The cylindrical version of CIELUV is known as CIE LCh_{uv} where C^*_{uv} is the chroma and h_{uv} is the hue, defined with:

$$C^*_{uv} = \sqrt{(u^*)^2 + (v^*)^2}$$
$$h_{uv} = a\tan 2(v^*, u^*) \tag{A1.17}$$

where $a\tan 2$ function computes the polar angle from a Cartesian coordinate pair.
The saturation can be defined as:

$$s_{uv} = \frac{C^*}{L^*} = 13\sqrt{\left(u' - u'_n\right)^2 + \left(v' - v'_n\right)^2} \tag{A1.18}$$

Color and Hue Difference

The color difference can be calculated using the Euclidean distance of the L^*, u^*, v^* coordinates with $\sqrt{(\Delta u')^2 + (\Delta v')^2}$. The Euclidean metric can be used in CIE LCh_{uv} with that component of ΔE^*_{uv} attributable to the difference in hue as $\Delta H^* = \sqrt{C^*_1 C^*_2} \bullet 2\sin(\Delta h/2)$ where $\Delta h = h_2 - h_1$.

CIE 1994

The 1976 definition was extended to address perceptual non-uniformities, while retaining the L*a*b* color space. Δ*E*1994 is defined in the L*C*h* color space with difference in lightness, chroma and hue calculated from L*a*b* coordinates. Given a reference color $\left(L_1^*, a_1^*, b_1^*\right)$ and another color $\left(L_2^*, a_2^*, b_2^*\right)$, the difference is:

$$\Delta E_{94}^* = \sqrt{\left(\frac{\Delta L^*}{k_L S_L}\right)^2 + \left(\frac{\Delta C_{ab}^*}{k_c S_c}\right)^2 + \left(\frac{\Delta H_{ab}^*}{k_H S_H}\right)^2} \tag{A1.19}$$

where

$$\Delta L^* = L_1^* - L_2^* \tag{A1.20}$$

$$C_1^* = \sqrt{a_1^{*2} + b_1^{*2}} \tag{A1.21}$$

$$C_2^* = \sqrt{a_2^{*2} + b_2^{*2}} \tag{A1.22}$$

$$\Delta H_{ab}^* = \sqrt{\Delta E_{ab}^{*\,2} - \Delta L^{*2} - \Delta C_{ab}^{*\,2}} = \sqrt{\Delta a^{*2} + \Delta b^{*2} - \Delta C_{ab}^{*\,2}} \tag{A1.23}$$

$$\Delta a^* = a_1^* - a_2^* \tag{A1.24}$$

$$\Delta b^* = b_1^* - b_2^* \tag{A1.25}$$

$$S_L = 1 \tag{A1.26}$$

$$S_C = 1 + K_1 C_1^* \tag{A1.27}$$

$$S_H = 1 + K_2 C_1^* \tag{A1.28}$$

and where K_c and K_H are usually unity and the weighting factors $K_L = 1$, $K_1 = 0.045$ and $K_2 = 0.015$ for graphic arts.

CIE2000

Since the 1994 definition did not adequately resolve the perceptual uniformity issue, the CIE refined their definition, adding five corrections:

- A hue rotation term R_T to deal with the problematic blue region - hue angles in the
- neighborhood of 275°
- Compensation for neutral colors, the primed values in the *L*C*h* differences
- Compensation for lightness S_L
- Compensation for chroma S_C
- Compensation for hue S_H

$$\Delta E_{00}^{*}=\sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2+\left(\frac{\Delta C'}{k_C S_C}\right)^2+\left(\frac{\Delta H'}{k_H S_H}\right)^2+R_T\frac{\Delta C'}{k_C S_C}\frac{\Delta H'}{k_H S_H}} \tag{A1.29}$$

$$\Delta L'=L_2^{*}-L_1^{*} \tag{A1.30}$$

$$\overline{L}=\frac{L_1^{*}+L_2^{*}}{2} \tag{A1.31}$$

$$\overline{C}=\frac{C_1^{*}+C_2^{*}}{2} \tag{A1.32}$$

$$a_1^{'}=a_1^{*}+\frac{a_1^{*}}{2}\left(1-\sqrt{\frac{\overline{C}^7}{\overline{C}^7+25^7}}\right) \tag{A1.33}$$

$$a_2^{'}=a_2^{*}+\frac{a_2^{*}}{2}\left(1-\sqrt{\frac{\overline{C}^7}{\overline{C}^7+25^7}}\right) \tag{A1.34}$$

$$\overline{C}'=\frac{C_1^{'}+C_2^{'}}{2} \tag{A1.35}$$

$$\Delta C'=C_2^{'}-C_1^{'} \tag{A1.36}$$

$$C_1^{'}=\sqrt{a_1^{'2}+b_1^{*2}} \tag{A1.37}$$

$$C_2^{'}=\sqrt{a_2^{'2}+b_2^{*2}} \tag{A1.38}$$

$$\begin{aligned} h_1^{'}&=a\tan 2\left(b_1^{*},a_1^{'}\right)\bmod 360\\ h_2^{'}&=a\tan 2\left(b_2^{*},a_2^{'}\right)\bmod 360 \end{aligned} \tag{A1.39}$$

$$\Delta h'=\begin{cases} h_2^{'}-h_1^{'} & for \quad \left|h_2^{'}-h_1^{'}\right|\le 180^{\circ}\\ h_2^{'}-h_1^{'}+360^{\circ} & for \quad \left|h_2^{'}-h_1^{'}\right|>180^{\circ} \quad ,h_2^{'}\le h_1^{'}\\ h_2^{'}-h_1^{'}+360^{\circ} & for \quad \left|h_2^{'}-h_1^{'}\right|>180^{\circ},h_2^{'}>h_1^{'} \end{cases} \tag{A1.40}$$

$$\Delta H'=2\sqrt{C_1^{'}C_2^{'}}\sin\left(\Delta h'/2\right) \tag{A1.41}$$

$$\overline{H}'=\begin{cases} \left(h_1^{'}+h_2^{'}+360^{\circ}\right)/2 & for \quad \left|h_1^{'}-h_2^{'}\right|>180^{\circ}\\ \left(h_1^{'}+h_2^{'}\right)/2 & for \quad \left|h_1^{'}-h_2^{'}\right|\le 180^{\circ} \end{cases} \tag{A1.42}$$

$$T=1-0.17\cos\left(\overline{H}'-30^{\circ}\right)+0.24\cos\left(2\overline{H}'\right)+0.32\cos\left(3\overline{H}'+6^{\circ}\right)-0.20\cos\left(4\overline{H}'-63^{\circ}\right) \tag{A1.43}$$

$$S_L = 1 + \frac{0.015\left(\bar{L} - 50\right)^2}{\sqrt{20 + \left(\bar{L} - 50\right)^2}} \tag{A1.44}$$

$$S_C = 1 + 0.045\bar{C}' \tag{A1.45}$$

$$S_H = 1 + 0.015\bar{C}'T \tag{A1.46}$$

$$R_T = -2\sqrt{\frac{\bar{C}'^7}{\bar{C}'^7 + 25^7}} \sin\left[60^\circ \exp\left(-\left[\frac{\bar{H}' - 275^\circ}{25^\circ}\right]^2\right)\right] \tag{A1.47}$$

CMC l:c

In 1984, the Color Measurement Committee of the Society of Dyers and Colorists defined a new difference measure, also based on the *L*C*h* color model. Named after the developing committee, their metric is called CMC l:c. The quasimetric has two parameters: lightness (*l*) and chroma (*c*), allowing the users to weight the difference based on the ratio of l:c that is deemed appropriate for the application. Commonly used values are 2:1 for acceptability and 1:1 for the threshold of imperceptibility.

The distance of a color $\left(L_2^*, C_2^*, h_2\right)$ to a reference $\left(L_1^*, C_1^*, h_1\right)$

$$\Delta E_{CMC}^* = \sqrt{\left(\frac{L_2^* - L_1^*}{lS_L}\right)^2 + \left(\frac{C_2^* - C_1^*}{cS_c}\right)^2 + \left(\frac{\Delta H_{ab}^*}{S_H}\right)^2} \tag{A1.48}$$

$$S_L = \begin{cases} 0.511 & L_1^* < 16 \\ \dfrac{0.040975L_1^*}{1 + 0.01765L_1^*} & L_1^* \geq 16 \end{cases} \tag{A1.49}$$

$$S_C = \frac{0.063C_1^*}{1 + 0.0131C_1^*} + 0.638 \tag{A1.50}$$

$$S_H = S_C\left(FT + 1 - F\right) \tag{A1.51}$$

$$F = \sqrt{\frac{C_1^{*4}}{C_1^{*4} + 1900}} \tag{A1.52}$$

$$T = \begin{cases} 0.56 + \left|0.2\cos\left(h_1 + 168^\circ\right)\right| & 164^\circ \leq h_1 \leq 345^\circ \\ 0.36 + \left|0.4\cos\left(h_1 + 35^\circ\right)\right| & \textit{otherwise} \end{cases} \tag{A1.53}$$

2. Principal Component Analysis

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This goal is a achieved by transforming the data to a new set of variables - the principal components - which are uncorrelated and which are ordered so that the first few retain most of the variation present in all the original variables. For further details the reader is adviced to refer to [276]. This part of the Appendix has been adapted from [276].

Notations

- X is a vector of p random variables
- P_k is a vector of p constants
- $P_k^{'}X = \sum_{j=1}^{p} P_{kj}X_j$ (A2.1)

Procedural description

- Find linear function of X: $P_1^{'}X$ with maximum variance
- Next, find another linear function of X, $P_2^{'}X$ uncorrelated with $P_1^{'}X$ with maximum variance
- Iterate

In general, the goal is that most of the variation in X will be accounted for by the first m principal components, where $m << p$.

Derivation of PCA

- Σ is the known covariance matrix for the random variable X
- Σ will be replaced with S - the sample covariance matrix, when Σ is unknown

Solution

- For $k = 1, 2\ldots, p$ the k-th principal component is given by $z_k = P_k^{'}X$ where P_k is an eigenvector of Σ corresponding to its k-th largest eigenvalue of λ_k
- If P_k is chosen to have unit length (i.e. $P_k^{'}P_k = 1$) then $Var(z_k) = \lambda_k$

First Step

- Find $P_k^{'}X$ that maximizes $Var(P_k^{'}X) = P_k^{'}\Sigma P_k$
- Without constraint, select P_k
- Choose normalization constraint $P_k^{'}P_k = 1$, unit length vector

Constrained maximization – The method of Lagrange multipliers

- To maximize $P_k^{'}\Sigma P_k$ subject to $P_k^{'}P_k = 1$, the Lagrange multipliers technique is used.

  The function $P_k^{'}\Sigma P_k - \lambda\left(P_k^{'}P_k - 1\right)$ is maximized with regard to P_k.

- Resulting in the following:

$$\frac{d}{dP_k}\left(P_k^{'}\Sigma P_k - \lambda_k(P_k^{'}P_k - 1)\right) = 0$$
$$\Sigma P_k - \lambda_k P_k = 0 \qquad \text{(A2.2)}$$
$$\Sigma P_k = \lambda_k P_k$$

This should be recognizable as an eigenvector equation where P_k is an eigenvector of $\Sigma_b f$ and λ_k is the associated eigenvalue. Which eigenvector should we choose? If we recognize that the quantity to be maximized is $P_k^{'}\Sigma P_k = P_k^{'}\lambda_k P_k = \lambda_k P_k^{'}P_k = \lambda_k$ then we should choose λ_k to be as big as possible.

Calling λ_1 the largest eigenvector of Σ and P_1 the corresponding eigenvector, then the solution to $\Sigma P_1 = \lambda_1 P_1$ is the first principal component of *X*. In general, P_k will be the *k*-th principal component of *X* and $Var(P'X) = \lambda_k$. We will demonstrate this for $k = 2$. The second principal component $P_2 X$ maximizes $P_2^{'}\Sigma P_2$ subject to being uncorrelated with $P_1 X$.

The un-correlation constraint can be expressed by using the equations:

$$\text{cov}(P_1^{'}X, P_2^{'}X) = P_1^{'}\Sigma P_2 = P_2^{'}\Sigma P_1 = P_2^{'}\lambda_1 P_1^{'} = \lambda_1 P_2^{'}P_1 = \lambda_1 P_1^{'}P_2 = 0 \qquad \text{(A2.3)}$$

We can choose a Lagrangian to maximize

$$P_2^{'}\Sigma P_2 - \lambda_2(P_2^{'}P_2 - 1) - \varphi P_2^{'}P_1 \qquad \text{(A2.4)}$$

Differentiation of this quantity with regard to P_2 and setting the result equal to zero gives:

$$\frac{d}{dP_2}\left(P_2^{'}\Sigma P_2 - \lambda_2(P_2^{'}P_2 - 1) - \varphi P_2^{'}P_1\right) = 0 \qquad \text{(A2.5)}$$
$$\Sigma P_2 - \lambda_2 P_2 - \varphi P_1 = 0$$

$$P_1^{'}\Sigma P_2 - \lambda_2 P_1^{'}P_2 - \varphi P_1^{'}P_1 = 0 \qquad \text{(A2.6)}$$

Therefore ϕ must be zero, leading to $\Sigma P_2 - \lambda_2 P_2 = 0$. This equation is another eigenvalue equation and the same strategy of choosing P_2 to be the eigenvector associated with the second largest eigenvalue yields the second principal component of *X,* namely $P_2' X$.

This process can be repeated for k = 1, …, p yielding up to p different eigenvectors of Σ along with the corresponding eigenvalues $\lambda_1,...,\lambda_p$. Furthermore, the variance of each principal component is given by $Var[P_k' X] = \lambda_k$.

Properties of PCA

For any integer q, $1 \le q \le p$, consider the ortho-normal linear transformation $y = B' X$ where y is a q-element vector and B is a $q \times p$ matrix and let $\Sigma_y = B' \Sigma B$ be the variance-covariance matrix for y. Then the trace of Σ_y, denoted $tr(\Sigma_y)$ is maximized by taking $B = A_q$, where A_q consists of the first q columns of A. Therefore, if you want to choose a lower dimensional projection of X, B is a good option. It maximizes the retained variance of the resulting variables. Since projections are not correlated, the percentage of variance accounted for by retaining the first q principal components is given by:

$$\frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \times 100 \qquad \text{(A2.7)}$$

The sample covariance matrix – an unbiased estimator for the covariance matrix of X is given by $S = \frac{1}{(n-1)} X' X$ where X is a $(n \times p)$ matrix with $(i,j) - th$ element $\left(x_{ij} - \bar{x}_j\right)$.

The matrix A is formed by combining the p eigenvectors of S, then we can define a matrix of principal components scores $Z = XA$. Given the sample covariance matrix $S = \frac{1}{(n-1)} X' X$, the most straightforward way of computing the principal component analysis loading matrix is to use the singular value decomposition of $S = A' \Lambda A$ where A is a matrix consisting of the eigenvectors of S and Λ is a diagonal matrix whose diagonal elements are the eigenvalues corresponding to each eigenvector. Creating a reduced dimensionality projection of X is accomplished by selecting the q largest eigenvalues in Λ and retaining the q corresponding eigenvectors from A.

Limitations of PCA

- PCA assumes approximate normality of the input space distribution – PCA may still be able to produce a "good" low dimensional projection of the data even if the data isn't normally distributed
- PCA assumes that the input data is real and continuous

3. Independent Component Analysis

Independent component analysis (ICA) is a statistical method, with the goal to decompose multivariate data into a linear sum of non-orthogonal basis vectors with coefficients being statistically independent. ICA generalizes a widely-used subspace analysis method such as principal component analysis (PCA) and factor analysis, allowing latent variables to be non-Gaussian and basis vectors to be non-orthogonal in general. ICA is a density estimation method where a linear model is learned such that the probability distribution of the observed data is optimally captured, while factor analysis aims at best modeling the covariance structure of the observed data. For further details the reader may refer to [277-280]. This part of the Appendix has been adapted from [280].

We consider a linear generative model where *m*-dimensional observed data $x \in R^m$ is assumed to be generated by a linear combination of *n* basis vectors $\{a_i \in R^m\}$

$x = a_1 s_1 + a_2 s_2 + ... + a_n s_n$ where $\{s_i \in R\}$ are encoding variables representing the extent to which each basis vectors is used to reconstruct the data vector. Given *N* samples, the model can be then written in a compact form: $X = AS$, where $X = [x(1),...x(N)] \in R^{m \times N}$ is a data matrix, $A = [a_1,...,a_n] \in R^{m \times n}$ is a basis matrix and $S = [s(1),...,s(N)] \in R^{n \times N}$ is an encoding matrix with $s(t) = [s_1(t),...,s_n(t)]^T$

A strong application of ICA is a problem of blind source separation – the goal of which is to restore sources *S* without the knowledge of *A,* given the data matrix *X.* ICA and blind source separation have often been treated as identical problems since they are closely related to each other. In blind source separation, the matrix *A* is referred to as mixing matrix. In practice, we find a linear transformation *W* referred to as demixing matrix such that the rows of the output matrix $Y = WX$ are statistically independent. In this case, *WA* becomes a transparent transformation when the rows of *Y* are statistically independent. The transparent transformation is given by $WA = P\Lambda$ where *P* is a permutation matrix and Λ is a nonsingular diagonal matrix involving scaling. This transparent transformation reflects two indeterminacies in ICA: (1) scaling ambiguity and (2) permutation ambiguity.

Principal component analysis PCA has been used for dimensionality reduction and feature extraction. Having a data matrix $X \in R^{m \times N}$, the covariance matrix R_{XX} is defined by $R_{XX} = \frac{1}{N} XHX^T$ where $H = I_{N \times N} - \frac{1}{N} 1_N 1_N^T$ is the centering matrix where $I_{N \times N}$ is the $N \times N$ identity matrix and $1_N = [1,...1]^T \in R^N$. The approximation of the covariance matrix R_{XX} is $R_{XX} \approx U \Lambda U^T$, where $U \in R^{m \times n}$ contains *n* eigenvectors associated with *n* largest eigenvalues of R_{XX} in its columns and the corresponding eigenvalues are in the diagonal entries of Λ. The principal components are determined by projecting data points $x(t)$ onto the eigenvectors - $Z = U^T X$.

ICA is a generalization of PCA in the sense that the latent components are non-Gaussian and *A* can be a non-orthogonal transformation. PCA assumes only orthogonal transformation and considers the Gaussian variables. Figure A3.1 shows the main difference between PCA and ICA.

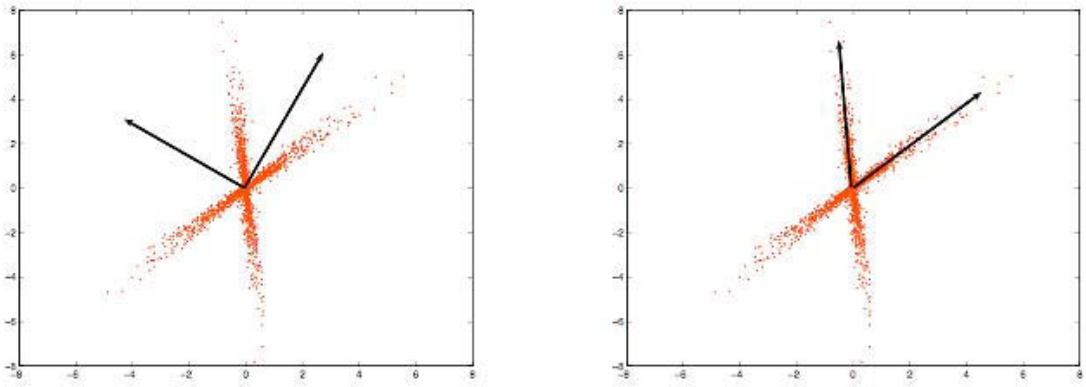


Figure A3.1 - Two-dimensional data with two main arms are fitted by two different basis vectors: (a) PCA makes the implicit assumption that the data have a Gaussian distribution and determines the optimal basis vectors that are orthogonal, which are not efficient at representing non-orthogonal distributions; (b) ICA does not require that the basis vectors be orthogonal and considers non-Gaussian distributions, which is more suitable in fitting more general types of distributions (Credits: "Independent Component Analysis", Seungjin Choi)

Theorem – Let $\{s_1,...,s_n\}$ be a set of independent random variables. Consider two random variables x_1 and x_2 which are linear combinations of $\{s_i\}$

$$\begin{aligned} y_1 &= \alpha_1 s_1 + ... + \alpha_n s_n \\ y_2 &= \beta_2 s_2 + ... + \beta_n s_n \end{aligned} \tag{A3.1}$$

where $\{\alpha_i\}$and $\{\beta_i\}$ are real constants. If y_1 and y_2 are statistically independent, then each variable s_i for which $\alpha_i \beta_i \neq 0$ is Gaussian. Considering $m = n$, we define the global transformation $G = WA$, where A is the mixing matrix and W is the demixing matrix. The output is written as $y(t) = Wx(t) = Gs(t)$. If A and W are nonsingular, G is nonsingular. If $\{y_i(t)\}$ are mutually independent non-Gaussian signals, in this G has the following decomposition $G = P\Lambda$, justifying the fact that ICA performs blind source separation.

The task of ICA is to estimate the mixing matrix A or its inverse $W = A^{-1}$ such that the elements of the estimate $y = A^{-1}x = Wx$ are independent. There are multiple techniques to estimate ICA, including maximum likelihood estimation, mutual information minimization, information maximization. In this Appendix we present only the maximum likelihood estimation.

Maximum likelihood estimation

Suppose that sources *s* are independent, with marginal distributions $q_i s_i$, $q(s) = \prod_{i=1}^{n} q_i(s_i)$. In the linear model $x = As$, a single factor in the likelihood function is given by:

$$p(x \mid A,q) = \int p(x \mid s,A)q(s)ds = \int \prod_{j=1}^{n} \delta\left(x_j - \sum_{i=1}^{n} A_{ji} s_i\right) \prod_{i=1}^{n} q_i(s_i)ds = |\det A|^{-1} \prod_{i=1}^{n} q_i\left(\sum_{j=1}^{n} A_{ij}^{-1} x_j\right) \tag{A3.2}$$

The log-likelihood is written as $\log p(x \mid A,q) = -\log|\det A| + \log q\left(A^{-1}x\right)$

$\log p(x \mid W,q) = \log|\det W| + \log p(y)$

Where $W = A^{-1}$ and y is the estimate of s with the true distribution q replaced by a hypothesized distribution p. Because sources are assumed statistically independent, the previous equation can be written as $\log p(x \mid W,q) = \log|\det W| + \sum_{i=1}^{n} \log p_i(y_i)$.

The demixing matrix W is determined by:

$$\hat{W} = \arg\max_{W} \left\{ \log|\det W| + \sum_{i=1}^{n} \log p_i(y_i) \right\} \tag{A3.3}$$

Maximum likelihood estimation is equivalent to Kullback matching where the optimal model is estimated by minimizing the Kullback Leibler (KL) divergence between the empirical distribution and the model distribution. If $\tilde{p}(x)$ is the empirical distribution and $p_\theta(x) = p(x \mid A,q)$ is the model distribution, then the KL divergence is defined as:

$$KL\left[\tilde{p}(x) \,\|\, p_\theta(x)\right] = \int \tilde{p}(x) \log \frac{\tilde{p}(x)}{p_\theta(x)} dx = -H(\tilde{p}) - \int \tilde{p}(x) \log p_\theta(x) dx \tag{A3.4}$$

where $H(\tilde{p}) = -\int \tilde{p}(x) \log \tilde{p}(x) dx$ is the entropy of $\tilde{p}$. Given a set of data points $\{x_1,...,x_N\}$ drawn from the underlying distribution $p(x)$, the empirical distribution $\tilde{p}(x)$ puts probability $1/N$ on each point in the data:

$\tilde{p}(x) = \frac{1}{N} \sum_{t=1}^{N} \delta(x - x_t)$, leading to $\arg\min_{\theta} KL[\tilde{p}(x) \,\|\, p_\theta(x)] = \arg\max_{\theta} \left\langle \log p_\theta(x) \right\rangle_{\tilde{p}}$ where $\langle \bullet \rangle_{\tilde{p}}$ represents the expectation with respect to the distribution $\tilde{p}$. In conclusion:

$$\left\langle \log p_\theta(x) \right\rangle_{\tilde{p}} = \frac{1}{N} \int \sum_{t=1}^{N} N\delta(x - x_t) \log p_\theta(x) dx = \frac{1}{N} \sum_{t=1}^{N} \log p_\theta(x) \tag{A3.5}$$

Maximum likelihood estimation is obtained from minimizing the KL divergence.

4. Information Theory

4.1 Relationship Between Entropy and Mutual Information

$$I(X,Y) = \sum_{i,j} p_{XY}\left(\xi_i,\rho_j\right) \log \frac{p_{XY}\left(\xi_i,\rho_j\right)}{p_X(\xi_i)p_Y(\rho_i)}$$

$$= \sum_{i,j} p_{XY}\left(\xi_i,\rho_j\right) \log \frac{p_{X|Y}\left(\xi_i \mid \rho_j\right)}{p_X(\xi_i)}$$

$$= -\sum_{i,j} p_{XY}\left(\xi_i,\rho_j\right) \log p_X\left(\xi_i\right) + \sum_{i,j} p_{XY}\left(\xi_i,\rho_j\right) \log p_{X|Y}\left(\xi_i \mid \rho_j\right) \qquad \text{(A4.1)}$$

$$= -\sum_{i} p_X(\xi_i) \log p_X(\xi_i) - \left(-\sum_{i,j} p_{XY}\left(\xi_i,\rho_j\right) \log p_{X|Y}\left(\xi_i \mid \rho_j\right)\right)$$

$$= H_X - H_{X|Y}$$

The mutual information is the reduction in the uncertainty of X due to the knowledge of Y. By symmetry, it follows that $I(X,Y) = H_Y - H_{Y|X}$. The random variable X contains the same amount of information about Y that Y contains about X.

Since $H_{X,Y} = H_X + H_{Y|X}$, the following applies: $I(X,Y) = H_X + H_Y - H_{X,Y}$ (A4.2)

Finally we note that $I(X,Y) = H_X - H_{X|X} = H_X$ - the mutual information of a random variable with itself is the entropy of the random variable.

Theorem – the relationship between entropy and mutual information

$$I(X,Y) = H_X - H_{X|Y}$$
$$I(X,Y) = H_Y - H_{Y|X}$$
$$I(X,Y) = H_X + H_Y - H_{X,Y} \qquad \text{(A4.3)}$$
$$I(X,Y) = I(Y,X)$$
$$I(X,X) = H_X$$

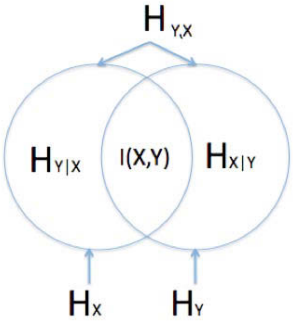


Figure A4.1 – Relationship between entropy and mutual information

The relationship between H_X, $H_Y, H_{X,Y}$, $H_{X|Y}, H_{Y|X}$ and $I(X,Y)$ is expressed in the Venn diagram, figure A4.1. The mutual information corresponds to the intersection of the information in *X* and the information in *Y*.

4.2 Chain Rules for Entropy, Relative Entropy and Mutual Information

Definition – Let $X_1,...,X_n$ be drawn according to $p_{X_1,...,X_n}(\xi_1,...,\xi_n)$, then:

$$H_{X_1...X_n} = \sum_{i=1}^{n} H_{X_i|X_{i-1},...,X_1} \tag{A4.4}$$

By repeating the two-variable expansion rule for entropies, we have:

$$\begin{aligned} &H_{X_1X_2} = H_{X_1} + H_{X_2|X_1} \\ &H_{X_1X_2X_3} = H_{X_1} + H_{X_2|X_1} + H_{X_3|(X_2,X_1)} \\ &... \end{aligned} \tag{A.45}$$

$$H_{X_1...X_n} = \sum_{i=1}^{n} H_{X_i|X_{i-1},...,X_1} \tag{A4.6}$$

The conditional mutual information as the reduction in the uncertainty of *X* due to knowledge of *Y* when *Z* is given. The conditional mutual information of random variables *X, Y, Z* is defined by $I(X,Y \mid Z) = H_{X|Z} - H_{X|Y,Z}$

Mutual information also satisfies the chain rule for information:

$$\begin{aligned} I(X_1,...,X_n;Y) &= \sum_{i=1}^{n} I\left(X_i;Y \mid X_{i-1},X_{i-2}...,X_1\right) \\ &= \sum_{i=1}^{n} H_{X_i|X_{i-1},...,X_1} - \sum_{i=1}^{n} H_{X_i|X_{i-1},...,X_1,Y} \\ &= \sum_{i=1}^{n} I\left(X_i;Y \mid X_{i-1},X_{i-2}...,X_1\right) \end{aligned} \tag{A4.7}$$

4.3 Jensen's Inequality and Its Consequences

Definition – A function $f(x)$ is said to be convex over an interval (*a, b*) if for every $x_1, x_2 \in (a,b)$ and $0 \le \lambda \le 1$,

$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \le \lambda f(x_1) + (1-\lambda) f(x_2) \tag{A4.8}$$

A function *f* is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$

Definition – A function *f* is concave if (- *f*) is convex. A function is convex if it always lies below any chord. A function is concave if it always lies above any chord.

Theorem (Jensen's Inequality) – If *f* is a convex function and *X* is a random variable, then $Ef(X) \ge f(EX)$, where *E* = expectation. If *f* is strictly convex, then $X = EX$, with probability 1 and *X* is a constant.

From a two mass point distribution, the inequality becomes:

$p_1 f\left(x_1\right) + p_2 f(x_2) \ge f(p_1 x_1 + p_2 x_2)$ which follows directly from the definition of convex functions. Suppose the theorem is true for distributions with (k – 1) mass points. Then, writing $p_i' = p_i/(1 - p_k)$ for i = 1, 2, …, (k – 1):

$$\begin{aligned}
&\sum_{i=1}^{k} p_i f\left(x_i\right) = p_k f(x_k) + (1-p_k)\sum_{i=1}^{k-1} p_i' f(x_i) \\
&\ge p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \\
&\ge f\left(p_k x_x + (1-p_k)\sum_{i=1}^{k-1} p_i' x_i\right) \\
&= f\left(\sum_{i=1}^{k} p_i x_i\right)
\end{aligned} \tag{A4.9}$$

The following theorem is of fundamental importance for entropy and relative entropy.

Theorem – Let $p_X(\xi)$, $q_X(\xi)$ be two probability mass functions

Then $KL(p \| q) \ge 0$ with equality if and only if $p_X(\xi) = q_X(\xi)$ for all ξ.

Let $A = \{\xi : p_X(\xi) > 0\}$ be the support set of $p_X(\xi)$.

Then:

$$
\begin{aligned}
-KL(p \,\|\, q) &= -\sum_{\xi \in A} p_X(\xi) \log \frac{p_X(\xi)}{q_X(\xi)} \\
&= \sum_{\xi \in A} p_X(\xi) \log \frac{q_X(\xi)}{p_X(\xi)} \\
&\le \log \sum_{\xi \in A} p(x) \frac{q_X(\xi)}{p_X(\xi)} \\
&= \log \sum_{\xi \in A} q_X(\xi) \\
&\le \log \sum_{\xi \in L} q_X(\xi) \\
&= \log 1 = 0
\end{aligned}
\qquad \text{(A4.10)}
$$

Since log t is a strictly concave function of t, we have the above equation valid if and only if $q_X(\xi)/p_X(\xi) = 1$ everywhere. Thus, $KL(p \,\|\, q) = 0$ if and only if $p_X(\xi) = q_X(\xi)$ for all ξ.

Corollary – For any two random variables *X, Y,* $I(X,Y) \ge 0$ with equality if and only if X and Y are independent. $I(X,Y) = KL(p_{XY}(\xi,\rho) \,\|\, p_X(\xi)p_Y(\rho)) \ge 0$ with equality if and only if $p_{XY}(\xi,\rho) = p_X(\xi,)p_Y(\rho)$, i.e. X and Y are independent.

Corollary - $KL\big(p_{Y|X}(\rho \mid \xi) \,\|\, q_{Y|X}(\rho \mid \xi)\big) \ge 0$ with equality if and only if $p_{Y|X}(\rho \mid \xi) = q_{Y|X}(\rho \mid \xi)$ for all ξ and ρ with $p_X(\xi) > 0$

Corollary - $I(X,Y \mid Z) \ge 0$ with equality if and only if X and Y are conditionally independent given Z.

Theorem – Conditioning reduces entropy: $H_{X|Y} \le H_X$ with equality if and only if X and Y are independent. $0 \le I(X,Y) = H_X - H_{X|Y}$. The theorem express that knowing another random variable Y can only reduce the uncertainty in X. This is true only on the average. Specifically, $H_{X|Y=\rho}$may be greater than or less than or equal to H_X, but on the average $H_{X|Y} = \Sigma_\rho p_Y(\rho) H_{X|Y=\rho} \le H_X$. For example, in a court case new evidence could increase uncertainty, but on the average evidence decreases uncertainty.

Theorem – Independence bound on entropy. Let $X_1,...,X_n$ be drawn according to $p_{X_1,...X_n}(\xi_1,...,\xi_n)$, then $H_{X1,...,Xn} \le \sum_{i=1}^{n} H_{Xi}$ with equality if and only if the X_i are independent.

For further reading the reader is adviced to refer to [263].

5. Estimation Theory

5.1 The Naïve Bayes Model for Classification

Naïve Bayes is an important model for binary classification. We start with the training set $\left(\underline{\xi}^{(i)},\rho^{(i)}\right)$ for $i = 1\ldots n$, where each $\underline{\xi}^{i}$ is a vector and each $\rho^{(i)} \in (1,...,k)$. k is an integer specifying the number of classes. In the case of a multiclass classification problem, the goal is to map each input vector $\underline{\xi}$ to a label ρ that can take any one of the k possible values. When $k = 2$, the classification problem is binary.

Each vector $\underline{\xi}$ is in the set $\{-1,+1\}^d$ for an integer value d specifying the number of features in the model. Each component ξ_j for $j = 1\ldots d$ can take one of the two possible values. Because this dissertation focused on classifying text and image documents, we use an example from the text domain to illustrate this. Consider the problem of classifying newspaper articles into k different categories – e.g. ρ =1 corresponds to a *sport* class, $\rho = 2$ corresponds to *entertainment* class, $\rho = 3$ to *politics,* etc). The label $\rho^{(i)}$ represents of the category of the i-*th* article / document in the collection. Each component $\xi_j^{(i)}$ for $j = 1\ldots d$ might represent the presence or the absence of a particular word. We can define $\xi_1^{(i)}$ to be (+1) if the i-*th* document contains the word *football*, or (-1) otherwise; $\xi_2^{(i)}$ to be (+1) if the i-*th* document contains the word *government,* or (-1) otherwise.

Assuming the random variables Y and $X_1 \ldots X_d$ corresponding to the label ρ and the vector components $\xi_1,...,\xi_d$, our goal with the Naïve Bayes is to model the joint discrete probability $P(Y = \rho, X_1 = \xi_1,...,X_d = \xi_d)$ for any label ρ paired with attribute values $\xi_1,...,\xi_d$. The following assumption is fundamental in the Naïve Bayes model:

$$P(Y = \rho, X_1 = \xi_1,...,X_d = \xi_d) = P(Y = \rho)\prod_{j=1}^{d} P\left(X_j = \xi_j \mid Y = \rho\right) \tag{A5.1}$$

First, by the chain rule, any joint distribution over $Y, X_1,...,X_d$ can be factored as:

$$P(Y = \rho, X_1 = \xi_1,...,X_d = \xi_d) = P(Y = \rho) \times P(X_1 = \xi_1,...,X_d = \xi_d \mid Y = \rho) \tag{A5.2}$$

$$\begin{aligned} &P(X_1 = \xi_1,...,X_d = \xi_d \mid Y = \rho) = \\ &= \prod_{j=1}^{d} P\left(X_j = \xi_j \mid X_1 = \xi_1,...,X_{j-1} = \xi_{j-1} \mid Y = \rho\right) \qquad \text{(A5.3)} \\ &= \prod_{j=1}^{d} P\left(X_j = \xi_j \mid Y = \rho\right) \end{aligned}$$

By the chain rule, the first equality is exact. The second equality – by the Naïve Bayes assumption – follows that for all $j = 1\ldots d$ the value for the random variable X_j is independent of all other attribute values $X_{j'}$ for all $j' \neq j$, when conditioned on the identity of the label Y. This assumption dramatically reduces the number of parameters, yet keeping the model effective.

The Naïve Bayes model has two types of parameters:

- $q(\rho)$ for $\rho \in \{1,..,k\}$, with $P(Y=\rho)=q(\rho)$
- $q_j(\xi \mid \rho)$ for $j = 1\ldots d$, $\xi \in \{-1,+1\}$, $\rho \in \{1,..,k\}$, $P(X_j=\xi \mid Y=\rho)=q_j(\xi \mid \rho)$

Resulting $p(\rho,\xi_1,...,\xi_d)=q(\rho)\prod_{j=1}^{d} q_j(\xi_j \mid \rho)$ (A5.4)

The Naïve Bayes model has k = the number of labels, d = the number of attributes and the following parameters explained:

- $q(\rho)$ for $\rho \in \{1,..,k\}$ - the probability of seeing the label ρ
- $q_j(\xi \mid \rho)$ for $j = 1\ldots d$, $\xi \in \{-1,+1\}$, $\rho \in \{1,..,k\}$ - the probability of attribute j taking value ξ, conditioned on the underlying label being ρ

The probability for any $\rho,\xi_1,...,\xi_d$ is defined as: $p(\rho,\xi_1,...,\xi_d)=q(\rho)\prod_{j=1}^{d} q_j(\xi_i \mid \rho)$. In the following paragraph we describe the estimation of the parameters using training samples. After the parameters have been estimated, the output of the Naïve Bayes classifier for a new test example $\underline{\xi}=\langle \xi_1,...,\xi_d \rangle$ is:

$$\arg\max_{\rho \in \{1,...k\}} p(\rho,\xi_1,...,\xi_d) = \arg\max_{\rho \in \{1,...k\}} \left(q(\rho)\prod_{j=1}^{d} q_j(\xi_j \mid \rho) \right) \quad \text{(A5.5)}$$

5.2 Maximum Likelihood Estimation for the Naïve Bayes Model

The question is how to estimate the parameters $q(\rho)$and $q_j(\xi \mid \rho)$ from the data.

The training set is $\left(\underline{\xi}^{(i)},\rho^{(i)}\right)$ for $i = 1\ldots n$, where each $\underline{\xi}^{i}$ is a d-dimensional vector. We write $\xi_j^{(i)}$ for the value of the j-*th* component of $\underline{\xi}^{(i)}$ and $\xi_j^{(i)}$ have the values {+1, -1}. The maximum-likelihood estimate for the parameters $q(\rho)$, with $\rho \in \{1,...,k\}$ is:

$$q(\rho)=\frac{\sum_{i=1}^{n}[[q^{(i)}=q]]}{n}=\frac{count(q)}{n} \quad \text{(A5.6)}$$

Defining $[[q^{(i)}=q]]=1$ if $q^{(i)}=q$ and 0 otherwise. Therefore, $\sum_{i=1}^{n}[[q^{(i)}=q]]=count(q)$ represents the number of times the label q is in the training set.

The ML estimates for the $q_j(\xi \mid \rho)$ parameters will have the following form:

$$q_j(\xi \mid \rho)=\frac{\sum_{i=1}^{n}[[\rho^{(i)}=\rho,\xi_j^{(i)}=\xi]]}{\sum_{i=1}^{n}[[\rho^{(i)}=\rho]]}=\frac{count_j(\xi \mid \rho)}{count(\rho)} \quad \text{(A5.7)}$$

$$count_j(\xi \mid \rho)=\sum_{i=1}^{n}[[\rho^{(i)}=\rho,\xi_j^{(i)}=\xi]] \quad \text{(A5.8)}$$

ML simply counts the number of times label ρ is observed with ξ_j taking value ξ, counts the number of times the label ρ is observed in total and compute the ratio of the two.

5.3 Maximum Likelihood Estimates

In this section we describe how the ML estimates are derived.

Given the training set $(\xi^{(i)},\rho^{(i)})$, the log-likelihood function is:

$$\begin{aligned} L(\underline{\theta}) &= \sum_{i=1}^{n} \log p\left(\xi^{(i)},\rho^{(i)}\right) \\ &= \sum_{i=1}^{n} \log\left(q\left(\rho^{(i)}\right)\prod_{j=1}^{d} q_j\left(\xi_j^{(i)} \mid \rho^{(i)}\right)\right) \\ &= \sum_{i=1}^{n} \log q\left(\rho^{(i)}\right) + \sum_{i=1}^{n} \log\left(\prod_{j=1}^{d} q_j\left(\xi_j^{(i)} \mid \rho^{(i)}\right)\right) \\ &= \sum_{i=1}^{n} \log q\left(\rho^{(i)}\right) + \sum_{i=1}^{n}\sum_{j=1}^{d} \log q_j\left(\xi_j^{(i)} \mid \rho^{(i)}\right) \end{aligned} \tag{A5.9}$$

$\underline{\theta}$ is the parameter vector having the values for all parameters $q(\rho)$ and $q_j(\xi \mid \rho)$ in the model. The log-likelihood function is a function of the parameter values and the training samples. The log-likelihood function $L(\underline{\theta})$ is a measure of how well the parameter values fit the training data. Therefore, we need to find the parameter value that maximize $L(\underline{\theta})$.

5.3.1 ML Estimation for Naïve Bayes Models

The maximum-likelihood estimates are the parameter values $q(\rho)$ for $q = \{1...k\}$, $q_j(\xi \mid \rho)$ for $j = 1\ldots d$, $\rho \in \{1...k\}$, $\xi \in \{-1,+1\}$ that maximize:

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log q\left(\rho^{(i)}\right) + \sum_{i=1}^{n}\sum_{j=1}^{d} \log q_j\left(\xi_j^{(i)} \mid \rho^{(i)}\right) \tag{A5.10}$$

with the following conditions:

- $q(\rho) \geq 0$ for all $\rho \in \{1...k\}$. $\sum_{\rho=1}^{k} q(\rho) = 1$
- For all ρ, j, ξ, $q_j(\xi \mid \rho) \geq 0$; for all $\rho \in \{1...k\}$, for all $j = 1\ldots d$, $\sum_{\xi \in \{-1,+1\}} q_j(\xi \mid \rho) = 1$

The ML estimated parameters for Naïve Bayes models take the following form:

$$q(\rho) = \frac{\sum_{i=1}^{n} [[q^{(i)} = q]]}{n} = \frac{count(q)}{n} \tag{A5.11}$$

$$q_j(\xi \mid \rho) = \frac{\sum_{i=1}^{n} [[\rho^{(i)} = \rho, \xi_j{}^{(i)} = \xi]]}{\sum_{i=1}^{n} [[\rho^{(i)} = \rho]]} = \frac{count_j(\xi \mid \rho)}{count(\rho)} \tag{A5.12}$$

Proof

$$\begin{aligned} L(\underline{\theta}) &= \sum_{i=1}^{n} \log q(\rho_i) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log q_i(\xi_{i,j} \mid \rho_i) \\ &= \sum_{\rho} count(\rho) \log q(\rho) + \sum_{j=1}^{d} \sum_{\rho} \sum_{\xi \in \{-1,+1\}} count_j(\xi \mid \rho) \log q_j(\xi \mid \rho) \end{aligned} \tag{A5.13}$$

Considering that:

$$count(q) = \sum_{i=1}^{n} [[q^{(i)} = q]] \tag{A5.14}$$

$$count_j(\xi \mid \rho) = \sum_{i=1}^{n} [[\rho^{(i)} = \rho, \xi_j{}^{(i)} = \xi]] \tag{A5.15}$$

We can write:

$$\begin{aligned} &\sum_{i=1}^{n} \log q(\rho^{(i)}) = \sum_{i=1}^{n} \sum_{\rho=1}^{k} [[\rho^{(i)} = \rho]] \log q(\rho) \\ &= \sum_{\rho=1}^{k} \sum_{i=1}^{n} [[\rho^{(i)} = \rho]] \log q(\rho) \\ &= \sum_{\rho=1}^{k} \log q(\rho) \sum_{i=1}^{n} [[\rho^{(i)} = \rho]] \\ &= \sum_{\rho=1}^{k} (\log q(\rho)) \times count(\rho) \end{aligned} \tag{A5.16}$$

and similarly $\sum_{i=1}^{n} \sum_{j=1}^{d} \log q_j(\xi_{i,j} \mid \rho_i) = \sum_{j=1}^{d} \sum_{\rho} \sum_{\xi \in \{-1,+1\}} count_j(\xi \mid \rho) \log q_j(\xi \mid \rho).$ (A5.17)

Going back to the term

$$\sum_{\rho} count(\rho)\log q(\rho) + \sum_{j=1}^{d}\sum_{\rho}\sum_{\xi\in\{-1,+1\}} count_j(\xi \mid \rho)\log q_j(\xi \mid \rho) \tag{A5.18}$$

Results that maximizing with regard to the $q(\rho)$ parameters implies maximizing only the term $\sum_{\rho} count(\rho)\log q(\rho)$, considering the constraints previously defined. The value of $q(\rho)$ that maximize the term under the constraints is defined as:

$$q(\rho) = \frac{count(\rho)}{\sum_{i=1}^{k} count(\rho)} = \frac{count(\rho)}{n} \tag{A5.19}$$

Similarly for $\sum_{\xi\in\{-1,+1\}} count_j(\xi,\rho)\log q_j(\xi \mid \rho)$, (A5.20)

we can find the value of $q_j(\xi \mid \rho)$ maximizing this term: $q_j(\xi \mid \rho) = \frac{count_j(\xi \mid \rho)}{\sum_{\xi\in\{-1,+1\}} count_j(\xi \mid \rho)}$

5.3.2 ML Estimation for Multinomial Distributions

This section describes the ML estimation method for multinomial distributions. Considering a finite set Λ, the distribution over Λ is a vector q with components q_ρ for each $\rho \in \Lambda$, expressing the probability of observing element *q*. P_ρ is the set of all distributions over the set Λ.

$$P_\rho = \left\{ q \in R^{|\Lambda|} : \forall q \in \Lambda, q_Y \geq 0, \sum_{q\in\Lambda} q_\rho = 1 \right\} \tag{A5.21}$$

There exists a vector *c* with components c_ρ for each $\rho \in \Lambda$, each $c_\rho \geq 0$.

- c_ρ will usually be a count value derived from the data – i.e. the number of times element *q* is observed.
- There is at least one $\rho \in \Lambda$ such that c_ρ is strictly positive

The estimation problems means finding the distribution *q** that maximizes:

$$q^* = \arg\max_{q\in P_\rho} \sum_{\rho\in\Lambda} c_\rho \log q_\rho \tag{A5.22}$$

Vector *q** has the components $q_\rho^* = \frac{c_\rho}{N}$ for all $\rho \in \Lambda$, where $N = \sum_{\rho\in\Lambda} c_\rho$.

Proof

The goal is to maximize the function $\sum_{\rho\in\Lambda} c_\rho \log q_\rho$, conditioned by $q_\rho \geq 0$ and $\sum_{\rho\in\Lambda} q_\rho = 1$. For simplicity, in this case all c_ρ are strictly positive.

We introduce a Lagrange multiplier $\lambda \in R$, corresponding to the constraint $\sum_{\rho\in\Lambda} q_\rho = 1$:

$$g(\lambda,q) = \sum_{\rho\in\Lambda} c_\rho \log q_\rho - \lambda\left(\sum_{\rho\in\Lambda} q_\rho - 1\right) \tag{A5.23}$$

The solution q_ρ^* must satisfy the following conditions:

- $\frac{d}{dq_\rho} g(\lambda,q) = 0$, for all ρ (A5.24)
- $\sum_{\rho\in\Lambda} q_\rho = 1$ (A5.25)

$$\frac{d}{dq_\rho} g(\lambda,q) = \frac{c_\rho}{q_\rho} - \lambda \tag{A5.26}$$

Setting this derivative to zero, results that $\frac{c_\rho}{\lambda} = q_\rho$ and $q_\rho = \frac{c_\rho}{\sum_{\rho\in\Lambda} c_\rho}$ (A5.27)

5.4 Expectation-Maximization (EM)

This section presents the general form of the EM algorithm. We start with the following definitions:

- The sets Γ and Λ, with $\Lambda = \{1,...,k\}$. The model $p(\xi,\rho;\underline{\theta})$ assigns a probability to each (ξ,ρ) such that $\xi\in\Gamma$ and $\rho\in\Lambda$ under parameters $\underline{\theta}$ - includes all the parameters in the model.
- Ω refers to the set of all valid parameter settings in the model
- The training set is $\xi^{(i)}$, $i = (1,\ldots,n)$, with $\xi^{(i)} \in \Gamma$.
- The log-likelihood function is: $L(\underline{\theta}) = \sum_{i=1}^{n} \log p\left(\xi^{(i)};\underline{\theta}\right) = \sum_{i=1}^{n} \log \sum_{\rho\in\Lambda} p\left(\xi^{(i)},\rho;\underline{\theta}\right)$
- The maximum likelihood estimates are: $\underline{\theta}^* = \arg\max_{\underline{\theta}\in\Omega} L(\underline{\theta})$

In general, computing the ML estimates in this approach is intractable. The EM algorithm is an iterative algorithm that defines parameter settings $\underline{\theta}^0,\underline{\theta}^1..\underline{\theta}^T$ and is driven by updating $\underline{\theta}^t = \arg\max_{\underline{\theta}\in\Omega} Q\left(\underline{\theta},\underline{\theta}^{t-1}\right)$ for $t = 1\ldots T$.

The function $Q\left(\underline{\theta},\underline{\theta}^{t-1}\right)$ is defined as $Q\left(\underline{\theta},\underline{\theta}^{t-1}\right)=\sum_{i=1}^{n}\sum_{\rho\in\Lambda}\delta(\rho\mid i)\log p\left(\xi^{(i)},\rho,\underline{\theta}\right)$ (A5.28)

Where $\delta(\rho\mid i)=p\left(\rho\mid\xi^{(i)};\underline{\theta}^{t-1}\right)=\frac{p\left(\xi^{(i)},\rho,\underline{\theta}^{t-1}\right)}{\sum_{\rho\in\Lambda}p\left(\xi^{(i)},\rho,\underline{\theta}^{t-1}\right)}$ (A5.29)

The idea is to fill in the $\delta(\rho\mid i)$ values using the conditional distribution, under the previous parameter values $\delta(\rho\mid i)=p(\rho\mid\xi^{(i)};\underline{\theta}^{t-1})$. The EM algorithm is the following

Initialization

* Set $\underline{\theta}^{0}$ to an initial value in the set Ω - a random initial value conditioned by $\underline{\theta}\in\Omega$

Algorithm

For $t=1\ldots T$, $\underline{\theta}^{t}=\arg\underset{\underline{\theta}\in\Omega}{max}\,Q\left(\underline{\theta},\underline{\theta}^{t-1}\right)$

Where $Q\left(\underline{\theta},\underline{\theta}^{t-1}\right)=\sum_{i=1}^{n}\sum_{\rho\in\Lambda}\delta(\rho\mid i)\log p\left(\xi^{(i)},\rho,\underline{\theta}\right)$

And $\delta(\rho\mid i)=p\left(\rho\mid\xi^{(i)};\underline{\theta}^{t-1}\right)=\frac{p\left(\xi^{(i)},\rho,\underline{\theta}^{t-1}\right)}{\sum_{\rho\in\Lambda}p\left(\xi^{(i)},\rho,\underline{\theta}^{t-1}\right)}$

Output: Parameters $\underline{\theta}^{T}$

6. Latent Dirichlet Allocation

6.1 Gibbs Sampling - The key inferential problem in Latent Dirichlet Allocation that requires solution is computing the posterior distribution of the latent variables. This section explains the Gibbs sampling approximation method.

The total probability of the LDA model is:

$$P(\underline{W},\underline{Z},\underline{\theta},\underline{\phi};\alpha,\beta) = \prod_{i=1}^{K} P(\phi_i;\beta)\prod_{j=1}^{M} P(\theta_j;\alpha)\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)P(W_{j,t}\mid\phi_{j,t}) \tag{A6.1}$$

$\underline{\theta}$ and $\underline{\phi}$ need to be integrated out.

$$\begin{aligned} P(\underline{Z},\underline{W};\alpha,\beta) &= \int_{\theta}\int_{\phi} P(\underline{W},\underline{Z},\underline{\theta},\underline{\phi};\alpha,\beta)\,d\underline{\phi}\,d\underline{\theta} \\ &= \int_{\phi}\prod_{i=1}^{K} P(\phi_i;\beta)\prod_{j=1}^{M}\prod_{t=1}^{N} P(W_{j,t}\mid\phi_{Z_{j,t}})\,d\underline{\phi}\int_{\theta}\prod_{j=1}^{M} P(\theta_j;\alpha)\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)\,d\underline{\theta} \end{aligned} \tag{A6.2}$$

All $\underline{\theta}$ and $\underline{\phi}$ are independent to each other and we treat them separately.

1. Integrating $\underline{\theta}$

$$\int_{\theta}\prod_{j=1}^{M} P(\theta_j;\alpha)\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)\,d\underline{\theta} = \prod_{j=1}^{M}\int_{\theta_j} P(\theta_j;\alpha)\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)\,d\theta_j \tag{A6.3}$$

By focusing on only one θ, we obtain the following:

$$\int_{\theta_j} P(\theta_j;\alpha)\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)\,d\theta_j \tag{A6.4}$$

In the above equation we replace the probabilities with the actual expression of the distribution and we obtain:

$$\int_{\theta_j} P(\theta_j;\alpha)\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)\,d\theta_j = \int_{\theta_j}\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_{j,i}^{\alpha_i-1}\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)\,d\theta_j \tag{A6.5}$$

Let $n_{j,r}^{i}$ be the number of word tokens in the j-th document with the same word symbol (the r-th word in the vocabulary) assigned to the i-th topic. $n_{j,r}^{i}$ is three dimensional. If any of these three dimensions is not limited to a specific value, we adopt the notation (•). To exemplify this, $n_{j,(\cdot)}^{i}$ denotes the number of word tokens in the j-th document assigned to the i-th topic. Thus, we can write

$$\prod_{t=1}^{N} P\left(Z_{j,t} \mid \theta_j\right) = \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i}} \tag{A6.6}$$

The integration formula becomes:

$$\begin{aligned} &\int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \prod_{i=1}^{K} \theta_{j,i}^{\alpha_i - 1} \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i}} d\theta_j \\ &= \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i} + \alpha_i - 1} d\theta_j \end{aligned} \tag{A6.7}$$

The expression inside the integration has the same form as the Dirichlet distribution.

$$\int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(n_{j,(\cdot)}^{i} + \alpha_i\right)} \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i} + \alpha_i - 1} d\theta_j = 1 \tag{A6.8}$$

Thus

$$\begin{aligned} &\int_{\theta_j} P\left(\theta_j; \alpha\right) \prod_{t=1}^{N} P\left(Z_{j,t} \mid \theta_j\right) d\theta_j = \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i} + \alpha_i - 1} d\theta_j \\ &= \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \frac{\prod_{i=1}^{K} \Gamma\left(n_{j,(\cdot)}^{i} + \alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)} \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(n_{j,(\cdot)}^{i} + \alpha_i\right)} \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i} + \alpha_i - 1} d\theta_j \\ &= \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \frac{\prod_{i=1}^{K} \Gamma\left(n_{j,(\cdot)}^{i} + \alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)} \end{aligned} \tag{A6.9}$$

2. Integrating $\underline{\phi}$

$$
\begin{aligned}
&\int_{\underline{\phi}} \prod_{i=1}^{K} P(\phi_i;\beta) \prod_{j=1}^{M} \prod_{t=1}^{N} P\left(W_{j,t} \mid \phi_{Z_{j,t}}\right) d\underline{\phi} \\
&= \prod_{i=1}^{K} \int_{\phi_i} P(\phi_i;\beta) \prod_{j=1}^{M} \prod_{t=1}^{N} P\left(W_{j,t} \mid \phi_{Z_{j,t}}\right) d\phi_i \\
&= \prod_{i=1}^{K} \int_{\phi_i} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \prod_{r=1}^{V} \phi_{i,r}^{\beta_r - 1} \prod_{r=1}^{V} \phi_{i,r}^{n_{(\cdot),r}^{i}} d\phi_i \qquad \text{(A6.10)} \\
&= \prod_{i=1}^{K} \int_{\phi_i} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \prod_{r=1}^{V} \phi_{i,r}^{n_{(\cdot),r}^{i} + \beta_r - 1} d\phi_i \\
&= \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \frac{\prod_{r=1}^{V} \Gamma\left(n_{(\cdot),r}^{i} + \beta_r\right)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)}
\end{aligned}
$$

The final equation with $\underline{\theta}$ and $\underline{\phi}$ integrated out is:

$$
P(\underline{Z},\underline{W};\alpha,\beta) = \prod_{j=1}^{M} \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \frac{\prod_{i=1}^{K} \Gamma\left(n_{j,(\cdot)}^{i} + \alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)} \times \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \frac{\prod_{r=1}^{V} \Gamma\left(n_{(\cdot),r}^{i} + \beta_r\right)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)} \qquad \text{(A6.11)}
$$

The goal of Gibbs sampling is to approximate the distribution $P(\underline{Z} \mid \underline{W};\alpha,\beta)$. $P(\underline{W};\alpha,\beta)$ is invariable for any Z, Gibbs sampling equations can be directly derived from $P(\underline{Z},\underline{W};\alpha,\beta)$. The key point is to derive the following conditional probability:

$$
P\left(Z_{(m,n)} \mid \underline{Z}_{-(m,n)},\underline{W};\alpha,\beta\right) = \frac{P\left(Z_{(m,n)},\underline{Z}_{-(m,n)},\underline{W};\alpha,\beta\right)}{P\left(\underline{Z}_{-(m,n)},\underline{W};\alpha,\beta\right)} \qquad \text{(A6.12)}
$$

where $Z_{(m,n)}$ denotes the Z hidden variable of the n-th word token in the m-th document and we make the assumptions that the word symbol of it is the v-th word in the vocabulary. $Z_{-(m,n)}$ denotes all the Zs except for $Z_{(m,n)}$. Gibbs sampling only requires a sample value for $Z_{(m,n)}$. It doesn't require an exact value of $P\left(Z_{(m,n)} \mid \underline{Z}_{-(m,n)},\underline{W};\alpha,\beta\right)$ but the ratios among the probabilities that $Z_{(m,n)}$ can take value. Therefore, the above equation becomes:

$$P\left(Z_{(m,n)}=k \mid \underline{Z}_{(m,n)},\underline{W};\alpha,\beta\right) \propto P\left(Z_{(m,n)}=k,\underline{Z}_{(m,n)},\underline{W};\alpha,\beta\right)$$

$$=\left(\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma\left(\alpha_i\right)}\right)^{M}\prod_{j\neq m}\frac{\prod_{i=1}^{K}\Gamma\left(n_{j,(\cdot)}^{i}+\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K}n_{j,(\cdot)}^{i}+\alpha_i\right)}$$

$$\times\left(\frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma\left(\beta_r\right)}\right)^{K}\prod_{i=1}^{K}\prod_{r\neq v}\Gamma\left(n_{(\cdot),r}^{i}+\beta_r\right) \tag{A6.13}$$

$$\times\frac{\prod_{i=1}^{K}\Gamma\left(n_{m,(\cdot)}^{i}+\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i}+\alpha_i\right)}\prod_{i=1}^{K}\frac{\Gamma\left(n_{(\cdot),v}^{i}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i}+\beta_r\right)}$$

$$\propto\frac{\prod_{i=1}^{K}\Gamma\left(n_{m,(\cdot)}^{i}+\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i}+\alpha_i\right)}\prod_{i=1}^{K}\frac{\Gamma\left(n_{(\cdot),v}^{i}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i}+\beta_r\right)}$$

To conclude, let $n_{j,r}^{i,-(m,n)}$ have the same meaning as $n_{j,r}^{i}$ with $Z_{(m,n)}$ excluded. In the above equation, we regard terms that are not dependent on *k* as constants:

$$\propto\frac{\prod_{i\neq k}\Gamma\left(n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)}{\Gamma\left(\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)+1\right)}\prod_{i\neq k}\frac{\Gamma\left(n_{(\cdot),v}^{i,-(m,n)}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i,-(m,n)}+\beta_r\right)}$$

$$\times\Gamma\left(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1\right)\frac{\Gamma\left(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1\right)}{\Gamma\left(\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)+1\right)}$$

$$\propto\frac{\Gamma\left(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1\right)}{\Gamma\left(\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)+1\right)}\frac{\Gamma\left(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1\right)}{\Gamma\left(\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)+1\right)}$$

$$=\frac{\Gamma\left(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\right)\left(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\right)}{\Gamma\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)}\frac{\Gamma\left(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\right)\left(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)} \tag{A6.14}$$

$$\alpha\frac{\left(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\right)}{\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)}\frac{\left(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\right)}{\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)}$$

$$\propto\left(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\right)\frac{\left(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\right)}{\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)}$$

LDA – Implementation

The implementation of LDA using a collapsed Gibbs sampler involves setting up the requisite count variables, randomly initializing them and then running a loop over the desired number of iterations, where on each loop a topic is sampled from each word instance in the corpus. Post to the Gibbs iterations, the counts can be used to compute the latent distributions θ_d and ϕ_k. The only required count variables include $n_{d,k}$ - the number of words assigned to the topic *k* in document *d* - and $n_{k,w}$ - the number of times word *w* is assigned to the topic *k*. For the algorithm to be more efficient, a running count of n_k - the total number of times any word is assigned to topic *k* – is kept. Another array **z** is necessary to keep the current topic assignment for each of the *N* words in the corpus.

Because the Gibbs sampling method involves a sampling from distributions conditioned on all *other* variables (all other topic assignments, except the current one), before building a distribution from equation 4.83 the current assignment must be removed from the equation. This assignment can be removed by decrementing the counts associated with the current assignment because the topic assignments in LDA are exchangeable, i.e. the joint probability distribution is invariant to permutation. Then the probability of each topic assignment is calculated using equation 4.83. This discrete distribution is then sampled from and the chosen topic is set in the **z** array and the appropriate counts are then incremented. The following algorithms explains the full LDA Gibbs sampling procedure.

Input: words **w** $\in$ documents **d**
Output: topic assignments **z** and counts $n_{d,k}$, $n_{k,w}$, and n_k
Begin
Randomly initialize **z** and increment counters
For each *iteration* **do**

for $i = 0 \rightarrow (N-1)$ **do**
$word \leftarrow w[i]$
$topic \leftarrow z[i]$
$n_{d,topic^-} = 1$; $n_{word,topic^-} = 1$; $n_{topic^-} = 1$
for $k = 0 \rightarrow (K-1)$ **do**

$$p(z = k \mid \bullet) = (n_{d,k} + \alpha_k)\frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$$

end
$topic \leftarrow$ sample from $p(z \mid \bullet)$
$z[i] \leftarrow topic$
$n_{d,topic^+} = 1$, $n_{word,topic^+} = 1$, $n_{topic^+} = 1$
end
end
return z, $n_{d,k}$, $n_{k,w}$, n_k
end

Categorical Distribution – a generalized Bernoulli distribution is the probability distribution that describes the result of a random event that can have one of the *K* possible outcomes, and the probability of each outcome is specified. The *K*-dimensional categorical (discrete) distribution is the most general distribution over an event with *K* outcomes.

The categorical distribution is the generalization of the Bernoulli distribution for a discrete variable with more than two possible outcomes. It is common in machine learning and natural language processing to use "multinomial distribution" for to refer to a categorical distribution. In the Dirichlet multinomial distribution which comes as the results of a Gibbs sampling where Dirichlet distributions are collapsed out of a Hierarchical Bayesian model, it is very important to distinguish categorical from multinomial.

The sample space is a finite sequence of integers - $\{1,2,...,k\}$.

and $\underline{P} = \{p_1,...,p_K\}$

The probability mass function is $f_X(\xi = i \mid \underline{P}) = \prod_{i=1}^{k} p_i^{[\xi=i]}$ (A6.15)

With $[\xi = i]$ is 1 for $\xi = i$ and 0 otherwise. This representation of the probability mass function shows that the Dirichlet distribution is the conjugate prior of the categorical distribution.

The categorical distribution can also be treated as a special case of the multinomial distribution, in which the parameter *n* (the number of sampled items) is set to 1. The sample space is now regarded as the set of 1-of-*K* encoded random vectors *X* of dimension *k*, with the property that exactly one element has the value 1 and the others equal 0. The particular element having the value 1 indicates which category has been chosen. In this case, $f_X(\xi \mid \underline{P}) = \prod_{i=1}^{k} p_i^{\xi_i}$ where p_i represents the probability of seeing element *i* and $\sum_i p_i = 1$.

- The distribution is completely expressed by the probabilities associated with each number *i:* $p_i = P(X = i)$, with $i \in \{1,2,...,k\}$ and $\sum_i p_i = 1$
- For $k = 2$, the categorical distribution is reduced to the Bernoulli distribution
- The sufficient statistic from *n* independent observations is the set of counts in each category, where the total number of trials *n* is fixed.
- The conjugate prior distribution of a categorical distribution is a Dirichlet distribution. In a model consisting of a data point having a categorical distribution with unknown vector parameter $\underline{P}$, if we treated $\underline{P}$ as a random variable and give it a prior distribution defined using a Dirichlet distribution, then the posterior distribution of the parameter after incorporating the knowledge gained from observed data is also a Dirichlet.

Given a model:

$\underline{\alpha} = (\alpha_1,...,\alpha_K)$ concentration hyper-parameter

$\underline{P} \mid \underline{\alpha} = (p_1,...,p_K) \sim Dir(K,\alpha)$

$X \mid \underline{P} = (\xi_1,...,\xi_N) \sim Cat(K,\underline{P})$

The following holds:

$\underline{C} = (c_1,...,c_K)$ number of occurrences of category $i = \sum_{j=1}^{N} [\xi_j = i]$

$$\underline{P} \mid X,\underline{\alpha} \sim Dir(K,\underline{C} + \underline{\alpha}) = Dir(K,c_1 + \alpha_1,...,c_K + \alpha_K) \qquad \text{(A6.16)}$$

This relationship is used to estimate the underlying parameter $\underline{P}$ of a categorical distribution given a collection of N samples. The expected value of the posterior distribution is:

$$E[p_i,\underline{\alpha}] = \frac{c_i + \alpha_i}{N + \sum_k \alpha_k} \qquad \text{(A6.17)}$$

Multinomial Distribution – is a generalization of the binomial distribution. For n independent trials, each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

The binomial distribution is the probability distribution of the number of successes for one of just two categories in n independent Bernoulli trials, with the same probability of success on each trial. In a multinomial distribution, each trial results in exactly one of the fixed finite k possible outcomes, with probabilities $p_1,...,p_k$, with $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. If the random variables X_i indicate the number of times outcome i is observed over the n trials, the vector $\underline{X} = (X_1,...,X_k)$ follows a multinomial distribution with parameters n and $\underline{\boldsymbol{P}}$ where $\underline{P} = (p_1,...,p_k)$.

The **probability mass function** of the multinomial distribution is:

$$f_{\underline{X}}(\xi_1,...,\xi_n;n,p_1,...,p_k) = \Pr(X_1 = \xi_1 \,\&\, ...\,\&\, X_1 = \xi_1)$$

$$= \begin{cases} \dfrac{n!}{\xi_1!...\xi_k!} p_1^{\xi_1}...p_k^{\xi_k}, & when \sum_{i=1}^{k} \xi_i = n \\ 0 & otherwise \end{cases} \quad \text{for } \xi_i \geq 0, \text{ integers} \qquad \text{(A6.18)}$$

Properties

- The expected number of times the outcome *i* was observed over *n* trials is $E(X_i) = np_i$
- The covariance matrix is described as follows
  - (1) each diagonal entry is the variance of a binomially distributed random variable $\text{var}(X_i) = np_i(1-p_i)$
  - (2) the off-diagonal entries are the covariances $\text{cov}(X_i, X_j) = -np_i p_j \mid i \neq j$
- The correlation matrix is described with:

$$\rho(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)\,\text{var}(X_j)}} = \frac{-p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} = \sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}} \qquad \text{(A6.19)}$$

Dirichlet Distribution – is a family of continuous multivariate probability distributions parameterized by a vector $\boldsymbol{\alpha}$ of positive real numbers. Dirichlet is the multivariate generalization of the beta distribution. Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution – its probability density function returns the belief that the probabilities of *K* rival events are ξ_i given that each event has been observed $\alpha_i - 1$ times.

The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, .., \alpha_k > 0$ has the following probability density function:

$$f_X\left(\xi_1, ..., \xi_{K-1}; \alpha_1, ..., \alpha_K\right) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \xi_i^{\alpha_i - 1} \qquad \text{(A6.20)}$$

for all $\xi_i > 0$; satisfying the condition $\xi_1 + ... + \xi_{K-1} < 1$ and $\xi_K = 1 - \xi_1 - ... - \xi_{K-1}$

and $B(\alpha) = \dfrac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$

The domain of the Dirichlet distribution is a *K*-dimensional discrete distribution. The set of points in the support of a *K*-dimensional Dirichlet distribution is the *K* – 1 simplex, a generalization of a triangle embedded in the next-higher dimension.

The density function of a **symmetric Dirichlet distribution –** where all elements in the parameter vector $\underline{\boldsymbol{\alpha}}$ are equal – is parameterized by a single scalar value α:

$$f_X(\xi_1, .., \xi_{K-1}; \alpha) = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} \xi_i^{\alpha - 1} \qquad \text{(A6.21)}$$

For α =1, the Dirichlet distribution is similar to a uniform distribution.

Properties of the Dirichlet distribution

For $X = (X_1,...,X_K) \sim Dir(\alpha)$; $X_K = 1 - X_1 - ... - X_{K-1}$ and $\alpha_0 = \sum_{i=1}^{K} \alpha_i$:

Mean $E[X_i] = \frac{\alpha_i}{\alpha_0}$ (A6.22)

Variance $Var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$ (A6.23)

Covariance $Cov[X_i, X_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}; i \neq j$ (A6.24)

Marginal distributions = the beta distributions $X_i \sim Beta\left(\alpha_i, \left(\sum_{j=1}^{K} \alpha_j\right) - \alpha_i\right)$ (A6.25)

Dirichlet multinomial distribution – In a model where a Dirichlet prior distribution is placed over a set of categorical-valued observations, the marginal joint distribution of the observations is a Dirichlet multinomial distribution. In Gibbs sampling or variational Bayes Dirichlet prior distributions are often marginalized out.

Entropy – If $X \sim Dir(\alpha)$ is a random variable, the exponential family differential identities can be used to get an analytic expression for the expectation of $\log(X_i)$.

$$\begin{aligned} E[\log(X_i)] &= \psi(\alpha_i) - \psi(\alpha_0) \\ Cov[\log(X_i), \log(X_j)] &= \psi'(\alpha_i)\delta_{ij} - \psi'(\alpha_0) \end{aligned}$$ (A6.26)

where ψ is the digamma function, ψ' is the trigamma function and δ_{ij} is the Kronecker delta (not to be confused with Dirac delta)

Digamma function $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$ (A6.27)

Gamma function $\Gamma(n) = (n-1)!$ (A6.28)

Trigamma function $\psi'(x) = \frac{d^2}{dx^2} \ln \Gamma(x) = \sum_{n=0}^{\infty} \frac{1}{(x+n)^2}$ (A6.29)

Kronecker delta $\delta_{ij} = \begin{cases} 0 & if \quad i \neq j \\ 1 & if \quad i = j \end{cases}$ (A6.30)

Entropy - $H(X) = \log B(\alpha) + (\alpha_0 - K)\psi(\alpha_0) - \sum_{j=1}^{K} (\alpha_j - 1)\psi(\alpha_j)$ (A6.31)

Dirichlet distributions are most commonly used as the prior distribution of categorical variables or multinomial variables in Bayesian mixture models and hierarchical Bayesian models. Inference over hierarchical Bayesian models is often done using Gibbs sampling, and in such a case, instances of the Dirichlet distribution are typically marginalized out of the model by integrating out the Dirichlet random variable. This causes the various categorical variables drawn from the same Dirichlet random variable to become correlated, and the joint distribution over them assumes a Dirichlet-multinomial distribution, conditioned on the hyper-parameters of the Dirichlet distribution.

For further reading the reader is adviced to refer to [281-286]

Acronym List

ASTER – Advanced Spaceborne Thermal Emission and Reflection Radiometer
CBIR – Content-based Image Retrieval
CBVIR – Content-based visual information
CIE – Commission Internationale De L'Eclairage
CLC – Corine Land Cover
CMC – Color Management Committee
DLR - Deutsche Luft und Raumfahrt, German Aerospace Center
DN – Digital number
EEA – European Environmental Agency
EO – Earth Observation
ETM – Enhanced Thematic Mapper
ESA – European Space Agency
GEO – Group on Earth Observation
GEOSS - Global Earth Observation System of Systems
GIS – Geographic Information System
GMES – Global Monitoring for Environment and Security
IFOV – Instantaneous Field of View
IIM – Image Information Mining
KDD – Knowledge Database Discovery
KEO – Knowledge-centered Earth Observation
KES – Knowledge Enabled Services
KIM – Knowledge Driven Information Mining
LiDAR – Light Detection and Ranging
MSS – Multispectral Scanner
QBIC – Query by Image Content
RADAR – Radio Detection and Ranging
ROSA – Romanian Space Agency
SOA - Service-Oriented Architecture
SONAR – Sound Detecting and Ranging
SPOT – System Pour L'Observation de la Terre
TM – Thematic Mapper
USGS – United States Geological Survey

Acknowledgements

I am deeply grateful to my supervisor, Prof. Dr. Mihai Datcu. He has been and will continue to be an inspiring mentor and leader for me as well as a model to follow throughout my career. I would like to take this opportunity to thank him for the time, effort, understanding and patience he showed me in these three years. He encouraged me to follow my own ideas and my personal road in life and guided my steps along the way. Without his help and guidance I wouldn't be where I am today. Prof. Datcu opened for me doors in life that I had no idea they exist, made me believe in my own power and thinking and created wonderful opportunities for me to develop my career in space sciences. Thank you!

I would like to thank Prof. Dr. Otmar Loffeld for creating the opportunity for international students to join the IPP research team. It is an honour to be part of the IPP professional group of scientists and researchers from which I had a lot to learn. Thank you for giving me the chance to present this dissertation at University of Siegen.

Deep gratitude goes to Prof. Dr. Marius Piso, who offered me the chance to work for the Romanian Space Agency ROSA, to speak and present my research work at international conferences around the world and to be part of a wonderful team. I would like to thank him for mentoring me along these years, challenging me to follow ever growing dreams and ideas. I also want to thank my colleagues Ion Nedelcu, Alina Radutu, Ioana Vlad, Anca Popescu, Corina Vaduva, Carmen Patrascu, Daniela Faur for the understanding and support they have given me and for the positive, friendly working environment they created at ROSA.

Very special thanks go to my girlfriend Alina, who has been a true friend and a real support in all these years. Thank you for your patience and understanding you have showed me when I had to travel without you, when I had to spend nights researching, thank you for believing in me and my dreams and for being there for me every step of the way.

In special, aceasta teza este dedicata parintilor mei. Va multumesc ca ati fost alaturi de mine si va multumesc pentru tot ce ati facut pentru mine de-a lungul anilor. Intotdeauna am stiut ca pot conta pe voi si asta inseamna foarte mult pentru mine. Va multumesc!

Bibliography

[1] www.esa.int

[2] DLR EOWEB – Earth Observation Data Service
http://eoweb.dlr.de:8080/servlets/template/welcome/entryPage.vm

[3] Alexandria Digital Library www.alexandria.ucsb.edu

[4] USGS Global Visualization Viewer http://glovis.usgs.gov/

[5] USGS Earth Explorer http://earthexplorer.usgs.gov/

[6] M. Lew, “Content-based Multimedia Information Retrieval: State-of-the-Art and Challenges”, ACM Transactions on Multimedia Computing, Communications and Applications, Pp. 1- 19, 2006

[7] U. Fayyad, G. P. Shapiro, P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine, Vol. 17, No. 3, 1996

[8] Globetrotter http://clients.alexandria.ucsb.edu/globetrotter/

[9] Gazetteer - Alexandria digital library gazetteer server client http://webclient.alexandria.ucsb.edu/client/gaz/adl/index.jsp

[10] D. Bratasanu, I. Nedelcu, 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, No. 1, Pp. 193-204, 2011

[11] B. S. Manjunath, W. Y. Ma., "Texture Features for Browsing and Retrieval of Image Data", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No.8, Pp. 837–842, 1996.

[12] A. Rosenfeld, A. Kah, "Digital Picture Processing", Vol. 1, Academic Press, 1982

[13] M. Levine, "Vision in Man and Machine", McGraw-Hill, 1985

[14] A. Materka, M. Strzelecki, "Texture Analysis Methods - A Review", Technical University of Lodz, Institute of Electronics, COST B11 report, Brussels, 1998

[15] R. Haralick, K. Shanmugam, I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man and Cybernetics, Vol. 3, No. 6, Pp. 610-621, 1973

[16] J. Serra, "Image Analysis and Mathematical Morphology", Academic Press, 1982

[17] Y. Chen, E. Dougherty, "Grey-scale Morphological Granulometric Texture classification", Optical Engineering, Vol. 33, No. 8, Pp. 2713-2722, 1994

[18] J. Weszka, C. Myers, W. Boyne, "A Max-Min Measure For Image Texture Analysis", IEEE Transactions on Computers, Pp. 404-414, 1977

[19] H. Niemann, "Pattern Analysis", Springer-Verlag, 1981

[20] R. Lerski, K. Straughan, L. Shad, D. Boyce, S. Bluml, I. Zuna, "MR Image Texture Analysis - An Approach to Tissue Classification", Magnetic Resonance Imaging, Vol 11, Pp. 873-887, 1993

[21] M. Strzelecki, "Segmentation of Textured Biomedical Images Using Neural Networks", PhD thesis, Technical University of Lodz, Poland, 1995

[22] K. Valkealathi, E. Oja, "Reduced Multidimensional Co-occurrence Histograms in Texture Classification", IEEE Transactions on Pattern Analysis and machine Intelligence, Vol. 20, No. 1, Pp. 90-94, 1998

[23] G. Cross, A. Jain, "Markov Random Field Texture Models", IEEE Transactions on Pattern Recognition and Machine Analysis, Vol. 5, No. 1, Pp. 25-39, 1983

[24] A. Pentland, "Fractal-based Description of Natural Scenes", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, No. 6, Pp. 661-674, 1984

[25] R. Chellappa, S. Chatterjee, "Classification of Textures Using Gaussian Markov Random Fields", IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 33, No. 4, Pp. 959-963, 1985

[26] H. Derin, H. Elliot, "Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields", IEEE Transactions on Pattern Recognition and Machine Analysis, Vol. 9, No. 1, Pp. 39-55, 1987

[27] B. Manjunath, R. Chellappa, "Unsupervised Texture Segmentation Using Markov Random Fields ", IEEE Transactions on Pattern Recognition and Machine Analysis, Vol. 13, No. 5, Pp. 478-482, 1991

[28] M. Strzelecki, A. Materka, "Markov Random Fields as Models of Textured Biomedical Images", Proceedings 20th National Conference on Circuit Theory and Electronic Networks, KTOiUE, 1997

[29] B. Chaudhuri, N. Sarkar, "Texture Segmentation using Fractal Dimension", IEEE Transactions on Pattern Recognition and Machine Analysis, Vol. 17, No. 1, Pp. 72-77, 1995

[30] L. Kaplan, C.C. Kuo, "Texture Roughness Analysis and Synthesis via Extended Self-similar ESS Model", IEEE Transactions on Pattern Recognition and Machine Analysis, Vol. 17, No. 11, Pp. 1043 - 1056, 1995

[31] P. Cichy, A. Materka, J. Tuliszkiewicz, "Computerized Analysis of X-ray Images for Early Detection of Osteoporotic changes in the bone", Proceedings Conference of Information Technology in Medicine, TIM 1997

[32] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space", Philosophical Magazine, Vol. 2, No. 11, Pp 559–572, 1901

[33] S. Kullback, R.A. Leibler, "On Information and Sufficiency". Annals of Mathematical Statistics, Vol. 22, No. 1, Pp 79–86, 1951

[34] S. Kullback, “Information theory and statistics”, John Wiley and Sons, NY, 1959

[35] “Manual of Remote Sensing”, Volume 1, 2nd Edition, American Society for Photogrammetry and Remote Sensing, 1960

[36] OMAR, www.radiantblue.com

[37] Y. Yang, L. Ming, “A Survey on Content-based Video Retrieval”, Hong Kong University of Science and Technology

[38] N. Rea, R. Dahyot, A, Kokaram, “Semantic Event Detection in Sports Through Motion Understanding”, Proceedings of the 3rd International Conference on Image and Video Retrieval, 2004

[39] H.J. Zhang, S.W. Smoliar, J.H. Wu, “Content-based Video Browsing Tools”, SPIE Conference on Multimedia Computing and Networking, 1995

[40] N. Sebe, M.S. Lew, X. Zhou, T.S. Huang, E.M. Bakker, “The State-of-the-Art in Image and Video Retrieval”, Proceedings of the International Conference on Image and Video Retrieval, Pp. 1-8, 2003

[41] S. Marchand-Maillet, “Content-based Video Retrieval: An Overview”, 2000

[42] C. Faloutsos, R. Barber, M. Flicknet, J. Hafner, W. Niblack, D. Petkovic, W. Equitz, “Efficient and Effective Querying by Image Content”, Journal of Intelligent Information Systems, Vol. 3, Pp. 231-262, 1994

[43] H. Zhang, J. Wu, D. Zhong, S.W. Smoliar, “An integrated system for content-based video retrieval and browsing”, Pattern Recognition, Vol. 30, No. 4, Pp. 643-658, 1997

[44] The MediaMill TRECVID 2011, Semantic Video Search Engine
Proceedings of the TRECVID Workshop, 2011

[45] A. Rosenfeld, J. Weszka, "Picture Recognition", Digital Pattern Recognition, K. Fu Editor, Springer-Verlag, 135-166, 1980

[46] J. Daugman, "Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two Dimensional Visual Cortical Filters", Journal of the Optical Society of America, Vol. 2, Pp. 1160-1169, 1985

[47] A. Bovik, M. Clark, W. Giesler, "Multichannel Texture Analysis Using Localized Spatial Filters", IEEE Transactions on Pattern Recognition and Machine Analysis, Vol. 12, Pp. 55-73, 1990

[48] S. Mallat, "Multifrequency Channel Decomposition of Images and Wavelet Models", IEEE Transactions, Acoustic, Speech and Signal Processing, Vol. 37, No. 12, Pp. 2091-2110, 1989

[49] A. Laine, J. Fan, "Texture Classification by Wavelet Packet Signatures", III Transactions on Pattern Recognition and Machine Analysis, Vol. 15, No. 11, Pp. 1186-1191, 1993

[50] C. Lu, P. Chung, C. Chen, "Unsupervised Texture Segmentation via Wavelet Transform ", Pattern Recognition, Vol. 30, No. 5, Pp. 729-742, 1997

[51] A. C. She Y. Rui and T. S. Huang. "A Modified Fourier Descriptor for Shape Matching in MARS", Image Databases and Multimedia Search, Series on Software Engineering and Knowledge Engineering, S. K. Chang Editor, 1998

[52] D. Comaniciu and P. Meer. "Mean Shift: A Robust Approach Toward Feature Space Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 5, 2002

[53] D. Marr, H. Nishihara, “Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes,” Proceedings of Royal Society of London, Biological Sciences, Vol. 200, No. 1140, Pp.269-194, 1978

[54] M. Brady, “Criteria for Representations and of shape, Human and Machine Vision,” Academic Press. New York, Pp. 39-84, 1993

[55] A. Soffer and H. Samet ”Pictorial Query Specification for Browsing Through Spatially-referenced Image Databases”, Journal of Visual Languages and Computing, Vol. 9, No. 6, Pp.567-596, 1998

[56] A. Folkers and H. Samet ”Content-based Image Retrieval Ising Fourier Descriptors on a Logo Database”, Proceedings of the 16th International conference on Pattern Recognition, Pp. 521-524, 2002.

[57] C. T. Zahn, R. Z. Roskies, “Fourier Descriptors for Plane Closed Curves,” IEEE Transactions on Computers, Pp.269-281, 1972

[58] J. R. Bennet, J.S. McDonald, “On the Measurement of Curvature in a Quantized Environment,” IEEE Transactions Computer Vol. 24, Pp.803-820, 1975

[59] W.N. Lie, Y.C. Chen, “Shape Representation and Matching Using the Polar Signature,” Proceedings of the International Computer Symposium, Tainan, Taiwan, pp. 710-718, 1986

[60] Thomas Bernier, J. A. Landry, “A New Method for Representing and Matching Shapes of Natural Objects,” Pattern Recognition, Vol. 36 , Pp. 1711 - 1723, 2003

[61] S. Arivazhagan, L. Ganesan, S. Selvanidhyananthan, "Image Retrieval Using Shape Feature", International Journal of Imaging Science and Engineering, Vol. 1, No. 3, 2007

[62] A. K. Ray T. Acharya. "Image Processing, Principles and Applications", Wiley, 2005

[63] A. Baraldi, V. Puzzolo, P. Blonda, L. Bruzzone, C. Tarantino, "Automatic Spectral Rule-Based Preliminary Mapping of Calibrated Landsat TM and ETM+ Images", IEEE Transactions on Geoscience and Remote Sensing, Vol. 44, No. 9, Pp. 2563-2586, 2006

[64] D. C. Burr and M. C. Morrone, “A nonlinear model of feature detection,” in Nonlinear Vision: Determination of Neural Receptive Fields, Functions, and Networks, R. B. Pinter and N. Bahram, Editors. Baca Raton, FL CRC, Pp. 309–327, 1992

[65] R. Irish, ”Landsat 7 automatic cloud cover assessment (ACCA),” in Proceedings of SPIE—Algorithms Multispectral, Hyperspectral, and Ultraspectral Imagery VI, S. S. Shen and M. R. Descour Editors, Vol. 4049, Pp. 348–355, 2000

[66] M. Nagao and T. Matsuyama, "A Structural Analysis of Complex Aerial Photographs", New York: Plenum, 1980.

[67] P. Roy, S. Miyatake, and A. Rikimaru, "Biophysical Spectral Response Modelling Approach for Forest Density Stratification", 1992

[68] R.B. Myneni, F.G. Hall, P.J. Sellers, A.L. Marshak, "The Interpretation of Spectral Vegetation Indexes", IEEE Transactions on Geoscience and Remote Sensing, Vol. 33, No. 2, Pp. 481-486, 1995

[69] J.D. van Leeuwen Willem, Barron J. Orr, "Spectral Vegetation Indices and Uncertainty: Insights from a user's perspective", IEEE Transactions on Geoscience and Remote Sensing, Vol. 44, No. 7, Pp. 1931-1933, 2006

[70] M.S. Lew, "Principles of Visual Information Retrieval", Springer-Verlag, UK, 2001

[71] T. Gevers, "Color-based Retrieval", in Principles of Visual Information Retrieval, M.S. Lew Ed., Springer-Verlag, 2001

[72] T. Ojala, M. Pietikainen. D. Harwood, "Comparative Study of Texture Measures with Classification Based on Feature Distributions", Pattern Recognition, Pp. 51-59, 1996

[73] K. Jafari-Khouzani, H. Soltanian-zadeh, "Radon Transform Orientation Estimation for Rotation Invariant Texture Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 6, Pp. 1004-1008, 2005

[74] K. Koperski, G. Marchisio, S. Aksoy, C. Tusk, "VisiMine: Interactive Mining in Image Databases", IEEE IGARSS 2002, Vol. 3, Pp. 1810-1812

[75] P. Wu, Y. Choi, Y.M. Ro., C.S. Won, "Mpeg-7 Texture Descriptors", International Journal of Image and Graphics, Vol. 1, No.3, Pp. 547-563, 2001

[76] A. Srivastava, S.H. Joshi, W. Mio, X. Liu, "Statistical Shape Analysis: Clustering, Learning and Testing", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 4, Pp. 590-602, 2005

[77] S. Zhong and J. Ghosh, "A Unified Framework for Model-based Clustering", Machine Learning Research, Vol. 4, Pp. 1001–1037, 2003

[78] D. Aloise, A. Deshpande, P. Hansen, P. Popat, "NP-hardness of Euclidean sum-of-squares clustering", Machine Learning, Vol. 75, Pp. 245–249, 2009

[79] S. Dasgupta, Y. Freund, "Random Projection Trees for Vector Quantization", IEEE Transactions on Information Theory, Vol. 55, 2009

[80] G. Hamerly, C. Elkan, "Alternatives to the k-means Algorithm that Find Better Clusterings", Proceedings of the 11th International Conference on Information and Knowledge Management, 2002

[81] A.D. Peterson, A. P. Ghosh and R. Maitra., "A Systematic Evaluation of Different Methods for Initializing the K-means Clustering Algorithm", 2010

[82] L.A. Rowe, R. Jain, "ACM SIGMM Retreat Report on Future Directions in Multimedia Research", ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 1, No 1, Pp. 3-13, 2005

[83] D.H. Ballard, C.M. Brown, "Computer Vision", Prentice Hall, USA, 1982

[84] M. Levine, "Vision in Man and Machine", Mcgraw Hill, Columbus, 1985

[85] R.M. Haralick, L.G. Shapiro, "Computer and Robot Vision", Addison-Wesley, New York, USA, 1993

[86] J.R. Smith, S.F. Chang, "Visually Searching the Web for Content", IEEE Multimedia, Vol. 4, No. 3, Pp. 12-20, 1997

[87] C. Frankel, M.J. Swain, V. Athitsos, "Webseer: An Image Search Engine for the World Wide Web", University of Chicago Technical Report, USA, 1996

[88] R. Bliujute, S. Saltenis, G. Slivinskas, C.S. Jensen, "Developing a datablade for a new index", Proceedings of the IEEE International Conference on Data Engineering, IEEE, Sydney, Pp. 314-323, 1999

[89] R. Egas, N. Huijsmans, M.S. Lew, N. Sebe, "Adapting k-d Trees to Visual Retrieval", Proceedings of the International Conference on Visual Information Systems, Pp. 533-540, 1999

[90] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based Image Retrieval at the End of Early Years", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, Pp. 1349-1380, 2000

[91] A. Jaimes, N. Sebe, "Multimodal Human-computer Interaction: A Survey, Computer Vision and Image Understanding", 2006

[92] K. Rodden, K. Wood, "How Do People Manage Their Digital Photographs?", Proceedings of the ACM Conference on Human Factors in Computing Systems, ACM Press, New York, Pp. 409-416, 2005

[93] H. Rowley, S. Baluja, K. Kanade, "Human Face Detection in Visual Scenes", Advances in Neural Information Processing Systems 8, USA, Pp. 875-881, 1996

[94] M.S. Lew, N. Huijsmans, "Information Theory and Face Detection", Proceedings of the International Conference on Pattern Recognition, Po. 601-605, 1996

[95] M.S. Lew, "Next Generation Web Searches For Visual Content", IEEE Computer, Vol. 33, No. 11, Pp. 46-53, 2000

[96] M. S. Lew, "Content-based Multimedia Information Retrieval: State-of-the-art and challenges", ACM Transactions on Multimedia Computing, Communications and Applications, Pp. 1-19, 2006

[97] K. Rodden, W. Basalaj, D. Sinclair, K. Wood, "Does Organization by Similarity Assist Image Browsing?", Proceedings of the SIGCHI conference on human factors in computing systems, Pp. 190-197, 2001

[98] D. Frohlich, A. Kuchinsky, C. Pering, A. Don, S. Ariss, "Requirements for Photoware", Proceedings of the ACM Conference on CSCW, ACM Press, NY, Pp. 7-14, 2002

[99] J.H. Lim, Q. Tian, P. Mulhelm, "Home Photo Content Modeling for Personalized Event-based Retrieval", IEEE Multimedia, Vol. 10. No. 4, Pp. 28-37, 2003

[100] A. Graham, H. Garcia-Molina, A. Paepcke, T. Winograd, "Time as the Essence for Photo Browsing Through Personal Digital Libraries", Proceedings of the Joint Conference on Digital Libraries ACM Press, NY, Pp. 326-335, 2002

[101] M. Worring, T. Gevers, "Interactive Retrieval of Color Images", International Journal of Image and Graphics", Vol. 1, No. 3, Pp. 387-414, 2001

[102] M. Worring, G.P. Nguyen, L. Hollink, J.C. Gemert, D.C. Koelma, "Accessing Video Archives Using Interactive Search", Proceedings of IEEE International Conference on Multimedia and Expo, IEEE Taiwan, 2004

[103] S. Mongy, F. Bouali, C. Djeraba, "Analyzing User's Behaviour on a Video Database", Proceedings of ACM MDM KDD, 2005

[104] R. Jain, "A Game Experience in Every Application: Experiential Computing", Communications of the ACM, Vol. 46, No. 7, Pp. 48-54, 2003

[105] R. Jain, P. Kim, Z. Li, "Experiential Meeting System", Proceedings of ACM SIGMM, Workshop on Experiential Telepresence, USA, Pp. 1-12, 2003

[106] B. Gong, R. Singh, R. Jain, "ResearchExplorer: Gaining Insights through Exploration in Multimedia Scientific Data", Proceedings of the 6th International Workshop on Multimedia Information Retrieval, NY, 2004

[107] R.W. Picard, "Affective computing", MIT Press, Cambridge, USA, 2000

[108] N.B. Berthouze, T. Kato, "Towards a Comprehensive Integration of Subjective Parameters in Database Browsing", Advanced Database Systems for Integration of Media and User Environments, World Scientific, Pp. 227-232, 1998

[109] A. Hanjalic, R.L. Lagendijk, J. Biemond, "A New Method for Key Frame Based Video Content Representation", Image Databases and Multimedia Search, World Scientific, Pp. 97-107, 1997

[110] N. Sebe, M.S. Lew, "Robust Shape Matching", Proceedings of the 1st International Conference on Image and Video Retrieval, Springer-Verlag London, Pp. 17-28, 2002

[111] W. Wang, Y. Yu, J. Zhang, "Image Emotional Classification: Static vs. Dynamic", Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Pp. 6407-6411, 2004

[112] A. Salway, M. Graham, "Extracting Information About Emotions in Films", Proceedings of the ACM International Conference on Multimedia, USA, Pp. 299-302, 2003

[113] C.W. Therrien, "Decision, Estimation and Classification", Wiley, USA, 1989

[114] P. Winston, "Artificial Intelligence", Addison-Wesley, NY, USA, 1992

[115] C. Djeraba, "Content-based Multimedia Indexing and Retrieval", IEEE Multimedia, Vol. 9, Pp. 18-22, 2002

[116] C. Djeraba, "Association and Content-based Retrieval", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 1, Pp. 118-135, 2003

[117] R. Krishnapuram, S. Medasani, S.H. Jung, Y.S Choi, R. Balasubramaniam, "Content-based Image Retrieval Based on a Fuzzy Approach", IEEE Transactions on Knowledge and Data Engineering, Vol . 16, No. 10, Pp. 1185-1199, 2004

[118] H. Greenspan, J. Goldberger, A. Mayer, "Probabilistic Space-Time Video Modelling via Piecewise GMM", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 3, Pp. 384-396, 2004

[119] C. Tusk, G. Marchisio, S. Aksoy, K. Kopersky and J. C. Tilton, “Learning Bayesian Classifiers for Scene Classification with a Visual Grammar", IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, No. 3, Pp 581–589, 2005

[120] L. Fei-Fei, P. Perona. "A Bayesian Hierarchical Model for Learning Natural Scene Categories", Califorina Institute of Technology, USA.

[121] F. J. Seinstra, C. G. M. Snoek, J-M. Geusebroek and A. W. M. Smeulders, "The Semantic Pathfinder: Using an Authoring Metaphor for Generic Multimedia Indexing" IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, No. 10, 2006.

[122] N. Maillot, C. Hudelot and M. Thonnat. "Symbol Grounding for Semantic Image Interpretation: From Image data to Semantics", Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV’05), 2005

[123] M.H. Yang, N. Ahuja, "Detecting Faces in Images: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 1, Pp. 34-58, 2002

[124] W. Chang G. Sheikholeslami and A. Zhang. "Semquery: Semantic Clustering and Querying on Heterogeneous Features for Visual Data", IEEE Transactions on Knowledge and Data Engineering, 14, No.5, 2002.

[125] J. Fan, Y. Gao, H. Luo, "Multilevel Annotation of Natural Scenes Using Dominant Image Components and Semantic Concepts", Proceedings of the ACM International Conference on Multimedia ACM, Pp. 540-547, 2004

[126] J. Li, J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 9, Pp. 1075-1088, 2003

[127] Rocchio, "Relevance Feedback in Information Retrieval", The Smart Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, 1971

[128] S. Sclaroff, M. La Cascia, S. Sethi, L. Taycher, "Mix and Match Features in the Image Rover Search Engine", Principles of Visual Information Retrieval, Springer-Verlag, London, Pp. 259-277, 2001

[129] Y. Rui, T.S. Huang, "Relevance Feedback Techniques in Image Retrieval", Principles of Visual Information Retrieval, Springer-Verlag, Pp. 219-258, 2001

[130] Y. Li and T. Bretschneider. "Remote Sensing Image Retrieval using a Context-Sensitive Bayesian Network with Relevance Feedback" Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS), Vol. 5, Pp. 2461–2464, 2006.

[131] Y. Chen, X.S. Zhou, T.S Huang, "One class SVM for Learning in Image Retrieval", Proceedings of the IEEE International Conference on Image Processing, Greece, Pp. 815-818, 2001

[132] X. He, W.Y. King, O. Li, H. Zhang, "Learning and Inferring a Semantic Space from User's Relevance Feedback for Image Retrieval", Proceedings of the ACM Multimedia, ACM New York, Pp. 343-347, 2002

[133] P. Y. Yin, B. Bhanu, K.C. Chang, A. Dong, "Integrating Relevance Feedback Techniques for Image Retrieval Using Reinforcement Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 10, Pp, 1536-1551, 2005

[134] H. Muller, W. Muller, S. Marchand-Maillet, T. Pun, D. Squire, "Strategies for Positive and Negative Relevance Feedback in Image Retrieval", Proceedings of 15th International Conference on Pattern Recognition, Pp. 1043-1046, 2000

[135] W. Niblac J. Ashley Q. Huang B. Dom M. Gorkani J. Hafner D. Lee D. Petkovic, D. Steele M. Flickner, H. Sawhney and P. Yanker. "Query by Image
and Video Content: The QBIC system", IBM Almaden Research Center, http://wwwqbic.almaden.ibm.com/, 1995.

[136] R. W. Picard A. Pentland and S. Sclaroff. "Photobook: Content-based Manipulation of Image Databases", SPIE Storage and Retrieval Image and Video Databases II, No. 2185, February 1994.

[137] I. J. Cox, T. V. Papathomas, M. L.Miller, T. P. Minka, and P. N. Yianilos. "The Bayesian Image Retrieval System Pichunter: Theory, Implementation, and Psychophysical Experiments", IEEE Transactions on Image Processing, 9 No.1:20–37, 2000.

[138] Y. Chen , V. Roussev , G. G. Richard Iii , Y. Gao, "Content-Based Image Retrieval For Digital Forensics", Proceedings of the First International Conference on Digital Forensics, 2005

[139] T. Deserno, L. R. Long, H. Greenspan, "Content-based Image Retrieval: Major Challenges for Biomedical Applications"

[140] T. M. Deserno, S. Antani, & R. Long, "Ontology Of Gaps In Content-Based Image Retrieval", Journal of Digital Imaging, 2008

[141] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, "A Review Of Content-Based Image Retrieval Systems In Medical Applications: Clinical Benefits And Future Directions". International Journal of Medical Informatics. February, Vol. 73, No. 1, Pp. 1-23, 2004

[142] V. Ogle and M. Stonebraker, “Chabot: Retrieval From a Relational Database of Images,” IEEE Computer, Vol. 28, No. 9, Pp. 40-48, 1995.

[143] C. Carson, S. Belongie, H. Greenspan, and J. Malik, “Blobworld: Image Segmentation Using Expectation-Maximization and its Application to Image Querying,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 8, Pp.1026-1038, 2002.

[144] A. Gupta and R. Jain, “Visual Information Retrieval,” ACM Communications, Vol. 40, No. 5, Pp. 70-79, 1997.

[145] W. Y. Ma and B. Manjunath, “NeTra: A Toolbox for Navigating Large Image Databases,” Proceedings of IEEE International Conference on Image Processing, Pp. 568-571, 1997

[146] S. Mehrotra, Y. Rui, M. Ortega-Binderberger, and T. S. Huang, “Supporting Content-Based Queries over Images in MARS,” Proceedings of IEEE International Conference on Multimedia Computing and Systems, Pp. 632-633, 1997.

[147] T. Gevers and A. W. M. Smeulders, “PicToSeek: Combining Color and Shape Invariant Features for Image Retrieval,” IEEE Transactions on Image Processing, Vol. 9, No. 1, Pp. 102-119, 2000.

[148] J. Z. Wang, G. Wiederhold, O. Firschein, and X. W. Sha, “Content-Based Image Indexing and Searching Using Daubechies’ Wavelets,” International Journal Digital Libraries, Vol. 1, No. 4, Pp. 311-328, 1998

[149] J. Z. Wang, J. Li, and G. Wiederhold, “SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 9, Pp. 947-963, 2001

[150] J. Li and R. M. Narayanan. "Integrated Spectral And Spatial Information Mining In Remote Sensing Imagery", IEEE Transactions on Geoscience and Remote Sensing, Vol. 42, No. 3, 2004

[151] ESA-EUSC 2004 Conference: Theory and with Focus on Earth Observation Applications of Knowledge driven Image Information Mining.
http://earth.esa.int/rtd/events/esa-eusc 2004/. EUSC, Madrid, Spain, 2004

[152] ESA-EUSC 2005 Conference: Image Information Mining Theory and Applications to Earth Observation. http://earth.esa.int/rtd/events/esa-eusc 2005/. ESRIN, Frascati, Italy, 2005

[153] Chi-Ren Shyu; Klaric, M.; Scott, G.J.; Barb, A.S.; Davis, C.H.; Palaniappan, K.; "GeoIRIS: Geospatial Information Retrieval and Indexing System—Content Mining, Semantics Modeling, and Complex Queries", IEEE Transactions on Geosciences and Remote Sensing, Vol. 45, No. 4, Pp. 839-852, 2008

[154] M. Datcu and K. Seidel. "New Concepts For Remote Sensing Information Dissemination: Query By Image Content And Information Mining". Proceedings of IEEE Geoscience and Remote Sensing Symposium (IGARSS), Vol. 3, Pp. 1335–1337, 1999

[155] A. Pelizzari M. Quartulli A. Galoppo A. Colapicchioni M. Pastori K. Seidel, P. G. Marchetti M. Datcu, H. Daschiel and S. D'Elia. "Information Mining In Remote Sensing Images Archives - Part A: System Concepts", IEEE Transactions. on Geoscience and Remote Sensing, Vol. 41, No. 12, Pp. 2923–2936, 2003

[156] ESA-EUSC 2006 Conference: Image Information Mining for Security and Intelligence. http://earth.esa.int/rtd/events/esa-eusc 2006/. EUSC, Madrid Spain, November 27-29 2006.

[157] DigitalGlobe www.digitalglobe.com

[158] ESA-EUSC 2008 Conference: Image Information Mining: pursuing automation of geospatial intelligence for environment and security.
http://earth.esa.int/rtd/events/esa-eusc 2008/index.html. ESRIN, Frascati (Italy), 2008

[159] K. Seidel M. Schroeder, H. Rehrauer and M. Datcu. "Interactive Learning And Probabilistic Retrieval In Remote Sensing Image Archives", IEEE Transactions on Geoscience and Remote Sensing, Vol. 38, Pp. 2288–2298, 2000

[160] H. Liu, L. Yu, "Toward Integrating Feature Selection Algorithms For Classification And Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol .17, No. 4, Pp. 491-502, 2005

[161] L. Talavera, "An evaluation of filter and wrapper methods for feature selection in categorical clustering", Advances in Intelligent Data Analysis VI, Pp. 440-451, 2005

[162] H. Yuan, S-S. Tseng, W. Gangshan, Z. Fuyan, "A Two-Phase Feature Selection Method Using Both Filter And Wrapper", IEEE International Conference on Systems, Man and Cybernetics, Vol. 2, Pp. 132-136, 1999

[163] S. Boutemedjet, N. Bouguila, D. Ziou, "A Hybrid Feature Extraction Selection Approach For High-Dimensional Non-Gaussian Data Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 8, Pp. 1429-1443, 2009

[164] H. Liu, H. Motoda, R. Setiono, Z. Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining", JMKR Workshop on Feature Selection in Data Mining, Pp. 4-13, 2010

[165] G. Brown, "A New Perspective for Information Theoretic Feature Selection", Proceedings of the International Conference on Artificial Intelligence and Statistics AISTATS, 2009

[166] C.E. Shannon, "A Mathematical Theory of Communication", The Bell System Technical Journal, Vol. 27, Pp. 379-423, 1948

[167] H.E. Maia, A. Hammouch, M. Bakrim, "Color Texture Feature Selection by MIFS for Image Classification", Proceedings of IV Communications and Mobile Network, 2010

[168] N. Kwak, C.Choi, "Input Feature Selection by Mutual Information based on Parzen Window", IEEE Pattern Analysis and Machine Intelligence, Vol. 24, No. 12, 2002

[169] R.N. Colwell (Ed.), "Manual of Photographic Interpretation", ASPRS, 1960

[170] R.W. Dahlberg, J.R. Jensen, "Education for Cartography and Remote Sensing in the Service of an Information Society: The United States Case", American Cartographer, Vol. 13, No. 1, Pp. 51-71, 1986

[171] P.F. Fisher, R.E. Lindenberg, "On Distinctions Among Cartography, Remote Sensing And Geographic Information Systems", Photogrammetric Engineering and Remote Sensing, Vol. 55, No. 10, Pp. 1431-1434, 1989

[172] J.E. Estes, E.J. Hajic, L.R. Tinney, "Fundamentals of Image Analysis: Visible and Thermal Infrared Data", Manual of Remote Sensing, R.N. Conwell Editor, ASPRS, Pp. 897-1125, 1983

[173] J.R.G. Townshend, C.O. Justice, "Toward Operational Monitoring Of Terrestrial Systems By Moderate Resolution Remote Sensing", Remote Sensing of Environment, Vol. 83, Pp. 351-359, 2002

[174] M. Kraak, "Geovisualization Illustrated", ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 70, No. 1, Pp. 390-399, 2003

[175] J. Donnay, M.J. Barnsley, P.A. Longley, "Remote Sensing and Urban Analysis", NY, Taylor and Francis, Pp. 268, 2001

[176] J.D. Bossler, J.R. Jensen, R.B. and C. Rizos, "Manual of Geospatial Science and Technology", London, Taylor and Francis, Pp. 623, 2002

[177] J.R. Jensen, A. Saalfeld, F. Broome, D. Cowen, K. Price, D. Ramsey, L. Lapine, E.L. Usery, "Chapter 2: Spatial Data Acquisition and Integration", R.B. McMaster and E.L. Usery Editors, A Research Agenda for Geographic Information Science, CRC Press, Pp. 17-60, 2005

[178] J.A. Richards, X. Jia, "Remote Sensing Digital Image Analysis: An Introduction", 4th Edition, Springer, 2006

[179] J. Jensen, "Remote Sensing for the Environment. An Earth Resource Perspective", K. Clarke Editors, Pearson Prentice Hall, 2007

[180] V. Solomonson, "Landsat 4 and 5 Status and Results from Thematic Mapper Data Analyses", Proceedings Machine Processing of Remotely Sensed Data, W. Lafayette: Lab. for the Applications of Remote Sensing, Pp. 13-18, 1984

[181] C.J. Tucker, "Red And Photographic Infrared Linear Combinations For Monitoring Vegetation", Remote Sensing of the Environment, Vol. 8, Pp. 127-150, 1979

[182] C.J. Tucker, "Remote Sensing Of Leaf Water Content In The Near Infrared", Remote Sensing of the Environment, Vol.10, Pp. 23-32, 1980

[183] T.R. Loveland, T.L. Sohl, S.V. Stehman, A.L. Gallant, K.L. Sayler, D.E. Napton, "A Strategy For Estimating The Rates Of Recent United States Land-Cover Changes", Photogrammetry & Remote Sensing, Vol. 68, Pp. 1091-1099, 2002

[184] The MediaMill TRECVID 2007 Semantic Video Search Engine

[185] http://www.esa.int/esaLP/SEMM4T4KXMF_LPgmes_0.html (15.11.11)

[186] http://www.eea.europa.eu/data-and-maps/data/urban-atlas (15.11.11)

[187] http://www.eea.europa.eu/publications/COR0-landcover (15.11.11)

[188] http://gmes.info/

[189] http://www.emergencyresponse.eu/

[190] M.H. Degroot, M.J. Schervish, "Probability and Statistics", 4th Edition, Addison Wesley, 2011

[191] Pierre Comon, "Independent Component Analysis: a new concept?", Signal Processing, Elsevier, No. 36, Vol. 3:287--314, 1994

[192] J. M. Ponte and W. B. Croft,"A Language Modeling Approach To Information Retrieval", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pp. 275–281. ACM Press, 1998

[193] K. Sparck Jones, W. Walker, and S. Robertson. "A Probabilistic Model Of Information Retrieval: Development And Comparative Experiments, Parts 1 & 2", Information Processing & Management, Vol. 36, Pp. 779–840, 2000

[194] A. F. Smeaton, W. Kraaij, P. Over, “TRECVID 2003 - An Introduction” TRECVID 2003 Workshop, Gaithersburg, MD, USA, 2003

[195] S. Robertson, K. S. Jones, “Relevance Weighting Of Search Terms”, Journal of the American Society for Information Science, Vol. 27, Pp. 129–146, 1976.

[196] C. Zhai, J. Lafferty, "A Study Of Smoothing Methods For Language Models Applied To Ad Hoc Information Retrieval", Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pp. 334–342. ACM Press, 2001

[197] H. Luo, J. Fan, J. Xiao, and X. Zhu. "Semantic Principal Video Shot Classification Via Mixture Gaussian", IEEE International Conference on Multimedia and Expo (ICME), 2003

[198] H. Greenspan, J. Goldberger, A. Mayer, "Probabilistic Space-Time Video Modelling Via Piecewise GMM", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No.3, Pp. 384–396, 2004

[199] T. Westerveld, "Using Generative Probabilistic Models For Multimedia Retrieval", PhD thesis, University of Twente, 2004

[200] H. Greenspan, J. Goldberger, L. Ridel, "A Continuous Probabilistic Framework For Image Matching", Computer Vision and Image Understanding, Vol. 84, No.3, Pp. 384–406, 2001

[201] D. Hiemstra, “A Linguistically Motivated Probabilistic Model Of Information Retrieval”, Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Vol. 513 of Lecture Notes in Computer Science, Pp. 569–584. Springer-Verlag, 1998.

[202] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. "Indexing By Latent Semantic Analysis", Journal of the American Society of Information Science, Vol. 41, No.6, Pp. 391–407, 1990.

[203] G. Salton and M. McGill, editors. "Introduction to Modern Information Retrieval", McGraw-Hill,1983

[204] T. Hofmann, J. Puzicha, and M. I. Jordan "Unsupervised Learning From Dyadic Data", Advances in Neural Information Processing Systems, Vol. 11, 1999

[205] L. Saul, F. Pereira, "Aggregate And Mixed-Order Markov Models For Statistical Language Processing", Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing, 1997

[206] M. Lew, N. Sebe, C. Djeraba, R. Jain, "Content-based Multimedia Information Retrieval: State-of-the-art and Challenges", ACM Transactions on Multimedia Computing, Communication, and Applications, Vol 2, No. 1, Pp. 1-19, 2006

[207] R. Datta, D. Joshi, J. Li, J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, Vol. 40, Pp. 1-60, 2008

[208] N. Jacobson, Gupta, "Design Goals And Solutions For Display Of Hyperspectral Images", IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, Pp. 2684-2692, 2005

[209] T. Landauer, S. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge", Psychological Review, Pp. 211-240, 1997

[210] A. Mitchell, "The ESRI Guide To GIS Analysis. Volume 2: Spatial measurements and statistics", ESRI Press, 2005

[211] M. Datcu, K. Seidel, "Human Centered Concepts For Exploration And Understanding Of Earth Observation Images", IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, No. 3, Pp. 1226-1238, March 2005

[212] A. Torralba, R. Fergus, W. Freeman, "Tiny images", MIT Technical Report, 2007

[213] R. Bellens, K. Douterloigne, S Gautama, W, Philips, "Per-pixel Contextual Information For Classification Of VHR Images", IEEE IGARSS, 2008

[214] M. Erickson, L. Cooper, "Time Distortion in Hypnosis", OTC Publishing Corp, 2004

[215] R. Duda, P. Hart, D. Stork, "Pattern Classification", Wiley-Interscience, 2nd Edition, 2000

[216] D. Blei, A. Yang, M. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3, Pp. 993-1022, 2003

[217] M. Lienou, H. Maitre, M. Datcu, “Semantic Annotation of Satellite Images using Latent Dirichlet Allocation”, IEEE Geoscience and Remote Sensing Letters, Vol. 7, No. 1, Pp. 28-32, 2010

[218] E. Sudderth, B. Antonio, Torralba, W.T. Freeman, A.S. Willsky, “Describing Visual Scenes Using Transformed Objects and Parts”, International Journal of Computer Vision, Vol. 77, Pp. 291-330, 2008

[219] S. Aksoy, "Spatial Techniques for Image Classification," in C.H. Chen, ed., Signal and Image Processing for Remote Sensing, Taylor & Francis, Pp. 491-513, 2006

[220] Z. Geradts, J. Bijhold, “Content Based Information Retrieval in Forensic Image Databases”, Journal of Forensic Science, Vol. 47, Pp.40-47, 2002

[221] C. Pavlopoulou, A. C. Kak and C. Brodley, "Content-based Image Retrieval for Medical Imagery", Proceedings SPIE Medical Imaging: PACS and Integrated Medical Information Systems, San Diego CA, 2003

[222] D. Mochihashi, “LDA: a ‘Latent’ Dirichlet Allocation Package” ATR Spoken Language Communication Research Laboratories, Kyoto, Japan. 2004

[223] C. Johnson, “Top Ten Scientific Visualization Research Problems”, IEEE Computer Graphics and Applications, Vol. 24, No. 4, Pp. 13-17, August 2004

[224] P.C. Smits, “Comparison Of Some Feature Subset Selection Methods For Use In Remote Sensing Image Analysis”, IEEE IGARSS, Vol. 1, Pp. 530-533, 2001

[225] J. Bertin, “Graphical Semiology”, USA, University of Wisconsin Press, 1985

[226] D. Bratasanu, I. Nedelcu, M. Datcu, “Interactive Spectral Band Discovery for Exploratory Visual Analysis of Satellite Images”, IEEE Journal on Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 5, 2011

[227] H. Liu, “Evolving Feature Selection”, IEEE Intelligent Systems, Vol. 5, Pp. 154-1672, 2005

[228] L. Wang, “Feature Selection with Kernel Class Separability”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 9, Pp. 1534-1546, 2008

[229] L. Bruzzone, C. Persello, “A Novel Approach To The Selection Of Spatially Invariant Features For The Classification Of Hyperspectral Images With Improved Generalization Capability”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 47, No. 9, Pp. 3180-3191, 2009

[230] X. Chen, T. Fang, H. Huo, D. Li, “Graph-based Feature Selection for Object-Oriented Classification in VHR Airborne Imagery”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 49, No. 1, Pp. 353-366, 2011

[231] H.C. Peng, F. Long, C. Ding, "Feature Selection Based On Mutual Information: Criteria Of Max-Dependency, Max-Relevance, And Min-Redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, Pp. 1226–1238, 2005

[232] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", 2nd IEEE Computer Society Bioinformatics Conference, Stanford, USA. Pp. 523-529, 2003

[233] A. Csinger, “The Psychology of Visualization”, Technical report series. Department of Computer Science, University of British Columbia, 1992

[234] C. F. Barnes, “Image-Driven Data Mining For Image Content Segmentation, Classification And Attribution”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 45, No. 9, Pp. 2964-2978, 2007

[235] M. Kerroum, A. Hammouch, D. Aboutajdine, A. Bellaachia, “Using the Maximum Mutual Information Criterion to Textural Feature Selection for Satellite Image Classification”, IEEE Symposium on Computers and Communications, Pp. 1005-1009, 2008

[236] B. Waske, S. Linden, J.A. Benediktsson, A. Rabe, P. Hostert, “Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 48, No. 7, Pp. 2880-2889, 2010

[237] G. Camps-Valls, J. Mooij, B. Scholkopf, “Remote Sensing Feature Selection by Kernel Dependence Measures”, IEEE Geoscience and Remote Sensing Letters, Vol. 7, No. 3, Pp. 587-591, 2010

[238] P. Estevez, M. Tesmer, C. Perez, J. Zurada, “Normalized Mutual Information Feature Selection”, IEEE Transactions on Neural Networks, Vol. 20, No. 2, Pp. 189-201, 2009

[239] R. Archibald, G. Fann, “Feature Selection and Classification of Hyperspectral Images With Support Vector Machines”, IEEE Geoscience and Remote Sensing Letters, Vol. 4, No. 4, Pp. 673-677, 2007

[240] P.E. Meyer, C. Schretter, G. Bontempi, “Information Theoretic Feature Selection in Microarray data using variable complementarity”, IEEE Journal of Selected Topics in Signal Processing, Vol. 2, No. 3, Pp. 261-274, 2008

[241] T. Eriksson, S. Kim, H-G. Kang, C. Lee, “An Information Theoretic Perspective on Feature Selection In Speaker Recognition”, IEEE Signal Processing Letters, Vol. 12, No. 7, Pp. 500-503, 2005

[242] European Union Satellite Center website, www.eusc.org

[243] National System for Geospatial Intelligence – “Geospatial Intelligence GEOINT Basic Doctrine”, NGA, 2006

[244] M. Tory, T. Moller, “Human Factors in Visualization Research”, IEEE Transactions on Computer Graphics, Vol. 10, No. 1, Pp. 72-84, June 2004

[245] A. Walenstein, “Cognitive Support In Software Engineering Tools: A Distribution Cognition Framework”, PhD dissertation, Computer Science Department, Simon Fraser University, Canada, 2002

[246] Y. Waern, “On The Dynamics Of Mental Models”, Mental Models and Human Computer Interaction, D. Ackermann and M.J. Tauber, Pp. 73-93, New York, Elsevier Science, 1990

[247] P. Rheingans, “Are we there yet? Exploring with dynamic visualization”, IEEE Computer Graphics and Applications, Vol. 22, No. 1, Pp. 6-10, 2002

[248] M.N. Gahegan, “Visualization Strategies For Exploratory Spatial Analysis”, Proceedings of Third International Conference on GIS and Environmental Modeling, USA, 1996

[249] M. Gahegan, “Four Barriers To The Development Of Effective Exploratory Visual Analysis”, International Journal of Geographical Information Science - GIS, Vol. 13, No. 4, Pp. 289-309, 1999

[250] E.R. Tufte, “Envisioning Information”, Cheshire Connecticut: Graphics Press”, 1990

[251] M.S. Monmonier, “Strategies For The Interactive Exploration Of Geographic Correlation”, Proceedings of the 4th International Symposium on Spatial Data Handling, editors K. Brassel, H. Kishimoto, Pp. 381-389, 1990

[252] A.M. MacEachren, J.H. Ganter, “A Pattern Identification Approach To Cartographic Visualization”, Cartographica, Pp. 64-91, 1990

[253] P. Robertson, J.F. O'Callaghan, "The Application of Perceptual Color Spaces to the Display of Remotely Sensed Imagery", IEEE Transactions on Geoscience and Remote Sensing, Vol. 26, No. 1, January 1988

[254] D. DiBiase, A. MacEachren, J. Krygier, C. Reever, "Animation And The Role Of Map Design In Scientific Representations", Cartography and Geography Information Systems, Vol. 19, No. 4, Pp. 201-214, 1992

[255] C. Beshers, S. Feiner, "Autovisual: Rule-Based Design Of Interactive Multivariate Visualizations" IEEE Computer Graphics and Applications, Vol. 13, No. 4, Pp. 41-49, July 1993

[256] P.K. Robertson, "A Methodology for Scientific Data Visualization: Choosing Representations Based on a Natural Scene Paradigm", Proceedings of the First IEEE Conference on Visualization, Pp. 114-123, 2002

[257] B.E. Rogowitz and L.A. Treinish, "An Architecture For Rule-Based Visualization", IEEE Conference on Visualization, Pp. 236-243, 2002

[258] H. Senay, E. Igantius, "Compositional Analysis And Synthesis Of Scientific Data Visualization Techniques", Scientific visualization of Physical Phenomena, Springer-Verlag, Hong Kong, 1991

[259] J.D. Mackinlay, "Automating The Design Of Graphical Presentation Of Relational Information", ACM Transactions on Graphics, Vol. 5, Pp. 110-141, 1986

[260] National Visualization and Analytics Center, "Illuminating the Path: The R&D Agenda for Visual Analytics", 2008

[261] P. Rheingans, C. Landreth, "Perceptual Principles For Effective Visualizations", Perceptual Issues in Visualization, Springer-Verlan, Pp. 59-69, 1995

[262] H. Daschiel, 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

[263] Maybeck P.S., "Stochastic model estimation and control", Academic Press, 1979

[264] Burchett, K. E. (2002). Color harmony. Color Research and Application

[265] Pointer, M. R. & Attridge, G.G. (1998). The number of discernible colors. Color Research and Application

[266] Mahnke, F, "Color, environment and human response", New York: John Wiley & Sons, 1996

[267] Albers, J. "Interaction of Color. Revised and Expanded Edition" Yale University Press, 1996

[268] Judd, Deane B.; Wyszecki, G., "Color in Business, Science and Industry", Wiley Series in Pure and Applied Optics 3rd Edition "New York, 1975

[269] Hermann von Helmholtz, "Physiological Optics – The Sensations of Vision" 1866, as translated in Sources of Color Science, David L. MacAdam, ed., Cambridge MIT Press, 1970

[270] M.D. Fairchild, "Color Appearance Models", 2nd Ed., Wiley, Chichester, 2005

[271] William David Wright, "50 years of the 1931 CIE Standard Observer", Die Farbe, 1981

[272] Fraser, B., Murphy, C., Bunting, F. "Real World Color Management", Peachpit Press, 2005

[273] Sharma, G. "Digital Color Imaging Handbook", CRC Press, 2003

[274] Klein, G.A., "Industrial Color Physics", Springer Series in Optical Sciences, 2010

[275] Lindbloom, B.J., "Delta E", Brucelindbloom.com

[276] Joliffe, I.T., "Principal Component Analysis" 2nd Edition, Springer Series in Statistics, 2002

[277] Stone, J.V, "Independent Component Analysis: a tutorial introduction", Cambridge MIT Press, 2004

[278] Hyvarinen A., Karhunen, J., Oja, E., "Independent Component Analysis", 1st Edition, New York, Wiley, 2001

[279] Choi, S, Cichicjum A, Park, H.M., Lee, S.Y., "Blind Source Separation and Independent Component Analysis – A Review", Neural Information Processing Letters and Reviews, Vol. 6. No. 1, January 2005

[280] Choi, S., "Independent Component Analysis", Handbook of Natural Computing, Pp. 435-459, 2012

[281] Minka, T. "Bayesian Inference, Entropy and the Multinomial Distribution", Technical Report MIT, 2003

[282] Bishop, C., “Pattern Recognition and Machine Learning”, Springer, 2006

[283] Johnson, N.L., Kotz, S., Balakrishnan, N., “Discrete Multivariate Distributions”, Wiley, Pp. 105, 1997

[284] Agresti, A., “An Introduction to Categorical Data Analysis”, Wiley Interscience, Pp.25, 2007

[285] Evans, M., Hastings, N., Peacock, B., “Statistical Distributions”, New York, Wiley, Pp. 134-136, 2000

[286] Johnson, N.L., Kotz, S., Balakrishnan, N., “Continuous Multivariate Distributions. Volume 1: Models and Applications”, New York, Wiley, Pp. 488, 2000