TECHNICAL UNIVERSITY OF CRETE

ELECTRONIC AND COMPUTER ENGINEERING DEPARTMENT

Hyper-spectral imaging and spectral segmentation algorithms for the non-destructive analysis of El Greco's paintings

Diploma thesis by George Epitropou

Committee: Assoc. Professor Kostas Balas (supervisor)

Professor Michalis Zervakis

Assist. Professor Pagona-Noni Maravelaki-Kalaitzaki

December 2008, Chania

Acknowledgements

I wish to express my sincere appreciation and thanks to Professor Costas Balas for his guidance and support during the implementation of this diploma thesis.

I am also grateful to Gregory Antonopoulos, PhD, Athanasios Tsapras, PhD Candidate and Georgios Papoutsoglou, PhD Candidate as well my friend Apostolis Tsivitis for their guidance and suggestions for the improvement of the current diploma thesis.

Also I would like to thank Professors Michalis Zervakis and Pagona-Noni Maravelaki-Kalaitzaki for their participation in the presentation and the evaluation of this diploma thesis.

Last but not least, I wish to thank my family for their support and encouragement, as well as my friends for sharing my thoughts, worries and expectations during these years of my study.

Abstract

Imaging spectroscopy (also multi/hyper-spectral imaging or chemical imaging) is the application of reflectance spectroscopy to every pixel in a spatial image. Spectroscopy can be used to detect individual absorption features due to specific chemical bonds in a solid, liquid, or gas, providing a unique spectral profile for identifying different materials. Actual detection is dependent on the spectral coverage, spectral resolution, and signal-to-noise of the spectrometer, the abundance of the material and the strength of absorption features for the material in the wavelength region measured.

Application of modern spectral acquisition & processing systems for analyzing works of art and manuscripts refers only to the study of various colors' spectral responses composing an artwork, without physically extracting pigment materials. In this project, the problem of non-destructive analysis of artworks and especially the identification of color pigments in El Greco's paintings with the aid of MuSIS HS was addressed. MuSIS HS comprises an all optical imaging monochromator, operating as an electronically tunable optical filter in a wide wavelength range (370-1000 nm). The monochromator is coupled with a megapixel CCD sensor and MuSIS HS records light intensity as a function of both wavelength and location.

A spectral database of painting material replicas, with known chemical and structural characteristics, following their original development processes and corresponding to different artists and eras, was utilized for building a classifier suitable for pigment identification and mapping. Using training and testing data from the spectra of the painting materials acquired by MuSIS HS, several algorithms for spectral classification and segmentation of pigments have been tested. Results from the comparative evaluation of both unsupervised (k-Means, Isodata) and supervised (Maximum Likelihood, Expectation Maximization, Normalized Euclidean Distance, Spectral Angle Mapper, Spectral Correlation Mapper, Spectral Information Divergence, Spectral Gradient Mapper) algorithms applied in color pigments and paintings by El Greco are presented. Establishing an appropriate train set and therefore a list of classes of informational value, exhaustive and separable, several discriminality measures were employed such as Bhattacharyya distance, Relative Spectral Discriminality Power, Relative Spectral Discriminality rate and Relative Spectral Discriminality Entropy. A GUI application was developed for facilitating the classification processing.

Contents

Introduction

The interpretation of data acquired from hyper-spectral imaging methods uses techniques from a number of disciplines including pattern recognition, artificial intelligence, computer vision, image processing and statistical analysis. The methodology of pattern recognition applied to a particular problem depends on the data, the data model, and the information that one is expecting to find within the data (Bezdek, 1981). Statistical image classification techniques are ideally suited for data in which the distribution of the data within each of the classes can be assumed to follow a theoretical model. The most commonly used statistical classification methodology is based on maximum likelihood, a pixel-based probabilistic classification method which assumes that spectral classes can be described by a normal probability distribution in multispectral space (Swain and Davis, 1978). This traditional approach to classification is found to have some limitations in resolving interclass confusion if the data used are not normally distributed. As a result, in recent years, and following advances in computer technology, alternative classification strategies have been proposed. An expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated.

The performance of a classifier depends on the interrelationship between sample size, number of features, and classifier complexity. One of the important stages in image classification is that of collection of samples for training and testing the classifier. Sample size has an influence on the classification accuracy with which estimates of statistical parameters are obtained for statistical classifiers. Sample selection also depends on a number of factors which finally affect classification accuracy. The factors affecting sample selection are:

1. Number of training sites for sample collection.

2. Sampling method (random or systematic sampling).

3. Data source for labeling training sites

With high-dimensional data sets, such as those acquired by an imaging spectrometer, the training set size requirements for the correct application of a classification system may be too high. It is well known that the probability of misclassification of a decision rule does not increase as the number of features increases, as long as number of training samples is arbitrarily large. However, it has been observed in practice that additional features may degrade the performance of a classifier if the number of training samples that are used to design the classifier

is small relative to the number of features. This behavior is referred to as the "peaking phenomenon" (Raudys and Jain, 1991; Jain and Chandrasekaran, 1982). Distance metrics can use fewer samples for referencing a class, with an expected decrease on classification accuracy. However, classifiers incorporating distance metrics have been proven useful in applications where a training set of adequate size is not easy to acquire or due to the high dimensionality of the data. The Spectral Angle Mapper Classification (SAM) is an automated method for directly comparing image spectra to a known spectra (usually determined in a lab or in the field with a spectrometer) or an endmember. This method treats both (the questioned and known) spectra as vectors and calculates the spectral angle (angle distance) between them. SAM is insensitive to illumination since the algorithm uses only the vector direction and not the vector length. A series of classification algorithms for discriminating spectra using a generalized concept of distance involving statistical correlation, spectral gradients, information theory etc. can be used for discriminating materials.

As every material is formed by chemical bonds, has the potential for detection with spectroscopy. Painting materials used for tempera or oil-painting and particularly in early modern Europe, are pigments bound with a medium of oil (linseed oil) or egg (egg yolk). These pigments which were extracted from plants, bones, minerals etc. are essentially no different from the materials under study in a spectral imaging application. The work reported in this diploma focuses on employing classification algorithms for the non-destructive analysis of artworks and especially for the identification of color pigments in El Greco's paintings. A spectral database of painting material replicas, with known chemical and structural characteristics, following their original development processes and corresponding to different artists and eras, was utilized for building a classifier suitable for pigment identification and mapping. Using training and testing data from the spectra of the painting materials acquired by MuSIS HS, an all optical imaging monochromator, operating as an electronically tunable optical filter coupled with a megapixel CCD sensor, several algorithms for spectral classification and segmentation of pigments have been tested.

Chapter 1 & 2 present all the general concepts about spectral imaging, hyperspectral images and hyperspectral signal representation. A description of a general data analysis procedure for multi/hyper spectral imagery is also provided.

In ***chapter 3*** the classification process and various classification algorithms including unsupervised and supervised, parametric and non parametric classification techniques, are discussed in detail. Also the methodologies used to assess classification accuracy, such as the Kappa value and the confusion matrix are described.

Chapter 4 describes the methodology followed in order to achieve the objectives of this work and the results of classification. It also gives information about the hyperspectral images and the hyperspectral acquisition system that were used in this research.

Chapter 5 summarizes the major findings of this project, and provides a number of recommendations for future work using different classifiers.

Chapter 1

Introduction to imaging spectroscopy

Visual perception of scenes depends on appropriate illumination to visualize objects. The human visual system is limited to a very narrow portion of the spectrum of electromagnetic radiation, called light. In some cases natural sources, such as solar radiation, moonlight, lightning flashes, or bioluminescence, provide sufficient ambient light to navigate our environment. Because humankind was mainly restricted to daylight, one of the first attempts was to invent an artificial light source—fire (not only as a food preparation method).

Computer vision is not dependent upon visual radiation, fire, or glowing objects to illuminate scenes. As soon as imaging detector systems became available other types of radiation were used to probe scenes and objects of interest. Recent developments in imaging sensors cover almost the whole electromagnetic spectrum from x-rays to radiowaves. In standard computer vision applications illumination is frequently taken as given and optimized to illuminate objects evenly with high contrast. Such setups are appropriate for object identification and geometric measurements. Radiation, however, can also be used to visualize quantitatively physical properties of objects by analyzing their interaction with radiation.

Physical quantities such as penetration depth or surface reflectivity are essential to probe the internal structures of objects, scene geometry, and surface-related properties. The properties of physical objects therefore can be encoded not only in the geometrical distribution of emitted radiation but also in the portion of radiation that is emitted, scattered, absorbed or reflected, and finally reaches the imaging system. Most of these processes are sensitive to certain wavelengths and additional information might be hidden in the spectral distribution of radiation. Using different types of radiation allows taking images from different depths or different object properties. As an example, infrared radiation of between 3 and 5 μm is absorbed by human skin to a depth of < 1 mm, while x-rays penetrate an entire body without major attenuation. Therefore, totally different properties of the human body (such as skin temperature as well as skeletal structures) can be revealed for medical diagnosis.

1.1 Electromagnetic waves

Electromagnetic waves (figure 1.1) were first postulated by James Clerk Maxwell and subsequently confirmed by Heinrich Hertz. Maxwell derived a wave form of the electric and magnetic equations, revealing the wave-like nature of electric and magnetic fields, and their symmetry. Because the speed of EM waves predicted by the wave equation coincided with the measured speed of light, Maxwell concluded that light itself is an EM wave.

According to Maxwell's equations, a time-varying electric field generates a magnetic field and vice versa. Therefore, as an oscillating electric field generates an oscillating magnetic field, the magnetic field in turn generates an oscillating electric field, and so on. These oscillating fields together form an electromagnetic wave.

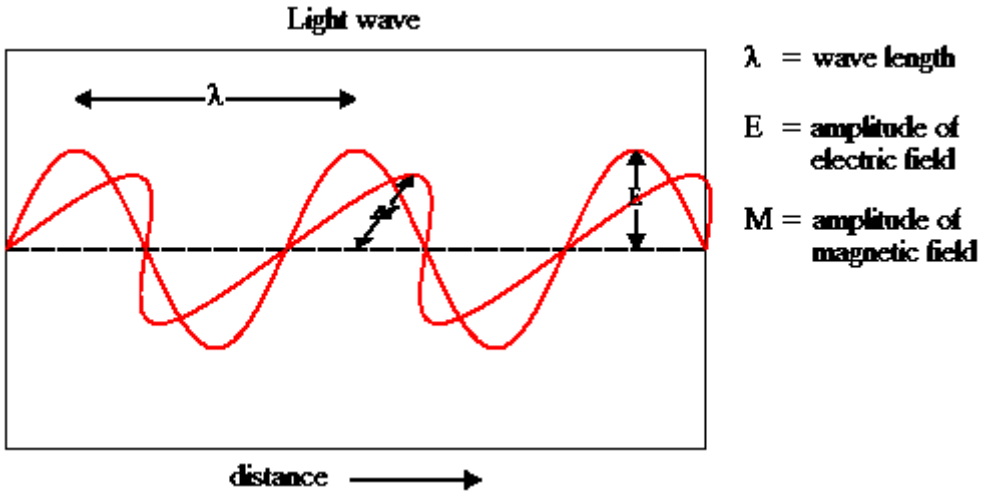


figure 1.1 : EM wave

In addition to electromagnetic theory, radiation can be treated as a flow of particles, discrete packets of energy called photons. One photon travels at the speed of light c and carries energy

$$e_p = h\nu = \frac{hc}{\lambda}$$

,

where h = 6.626 × 10−34 J s is Planck's constant. Therefore the energy content of radiation is quantized and can only be a multiple of hν for a certain frequency ν. While the energy per photon is given by e_p the total energy of radiation is given by the number of photons. It was this quantization of radiation that gave birth to the theory of quantum mechanics at the beginning of the twentieth century.

Although photons do not carry electrical charge this unit is useful in radiometry, as electromagnetic radiation is usually detected by interaction of radiation with electrical charges in sensors. In solid-state sensors, for example, the energy of absorbed photons is used to lift electrons from the valence band into the conduction band of a semiconductor. The bandgap energy Eg defines the minimum photon energy required for this process. As a rule of thumb the detector material is sensitive to radiation with energies Ev >Eg.

Electric and magnetic fields obey the properties of superposition, so fields due to particular particles or time-varying electric or magnetic fields contribute to the fields due to other causes. (As these fields are vector

fields, all magnetic and electric field vectors add together according to vector addition.) These properties cause various phenomena including refraction and diffraction. For instance, a travelling EM wave incident on an atomic structure induces oscillation in the atoms, thereby causing them to emit their own EM waves. These emissions then alter the impinging wave through interference.

Since light is an oscillation, it is not affected by travelling through static electric or magnetic fields in a linear medium such as a vacuum. In nonlinear media such as some crystals, however, interactions can occur between light and static electric and magnetic fields - these interactions include the Faraday effect and the Kerr effect.

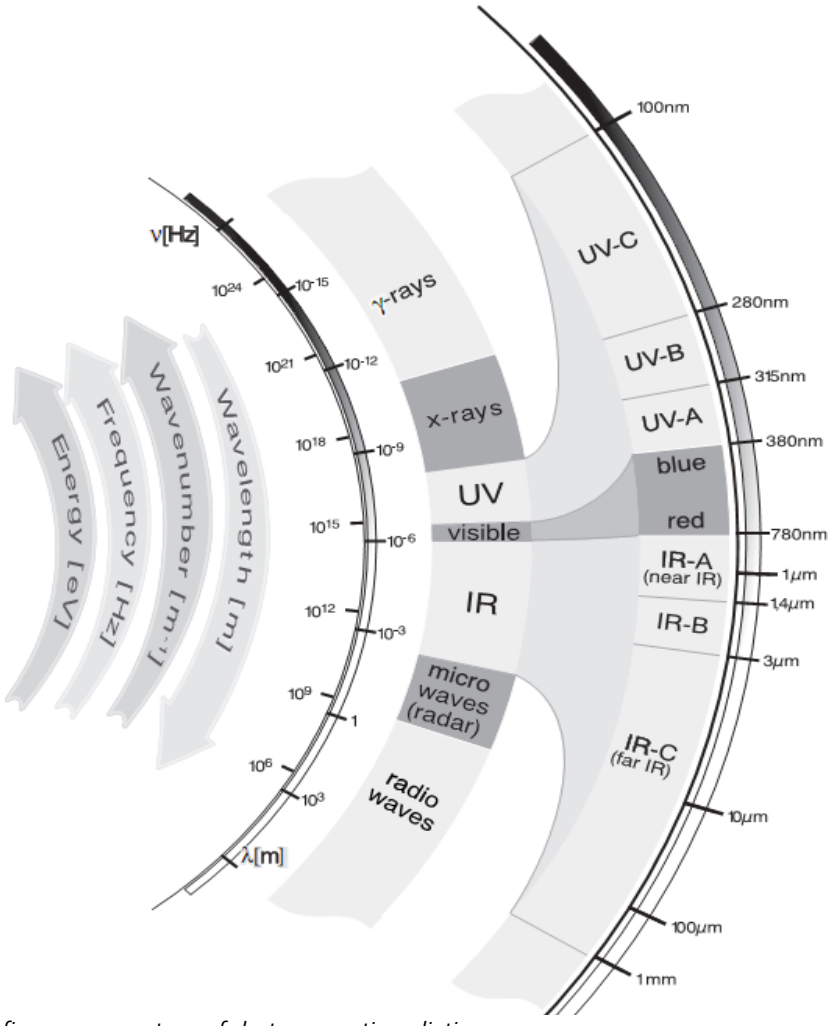


figure 1.2: spectrum of electromagnetic radiation

In refraction, a wave crossing from one medium to another of different density alters its speed and direction upon entering the new medium. The ratio of the refractive indices of the media determines the degree of refraction, and is summarized by Snell's law. Light disperses into a visible spectrum as light is shone through a prism because of the wavelength dependant refractive index of the prism material (Dispersion).

EM radiation exhibits both wave properties and particle properties at the same time .The wave characteristics are more apparent when EM radiation is measured over relatively large timescales and over large distances, and the particle characteristics are more evident when measuring small distances and timescales. Both characteristics have been confirmed in a large number of experiments.

Monochromatic radiation consists of only one frequency and wavelength. The distribution of radiation over the range of possible wavelengths is called spectrum or spectral distribution. Figure 1.2 shows the spectrum of

electromagnetic radiation together with the standardized terminology separating different parts. Electromagnetic radiation covers the whole range from very high energy cosmic rays with wavelengths in the order of 10^{-16} m(ν =1024 Hz) to sound frequencies above wavelengths of 106 m(ν =102 Hz). Only a very narrow band of radiation between 380nm and 780nm is visible to the human eye. Each portion of the electromagnetic spectrum obeys the same principal physical laws. Radiation of different wavelengths, however, appears to have different properties in terms of interaction with matter and detectability that can be used for wavelength selective detectors. There are experiments in which the wave and particle natures of electromagnetic waves appear in the same experiment, such as the diffraction of a single photon. When a single photon is sent through two slits, it passes through both of them interfering with itself, as waves do, yet is detected by a photomultiplier or other sensitive detector only once. Similar self-interference is observed when a single photon is sent into a Michelson interferometer or other interferometers (wave-particle duality).

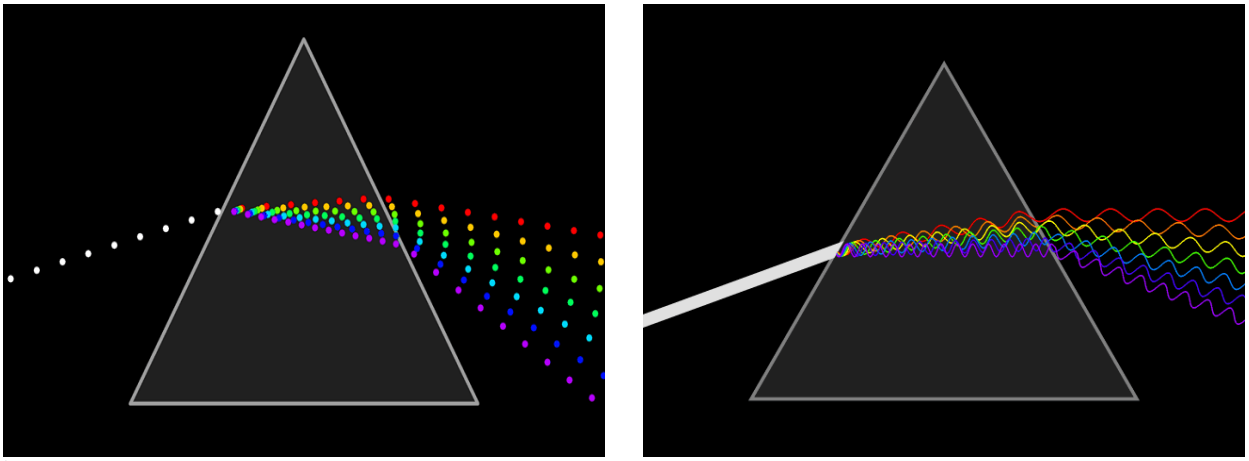


figure 1.3 wave-particle duality

1.2 Spectroscopy/spectrometry

Spectroscopy was originally the study of the interaction between radiation and matter as a function of wavelength (λ). In fact, historically, spectroscopy referred to the use of visible light dispersed according to its wavelength, e.g. by a prism. Later the concept was expanded greatly to comprise any measurement of a quantity as function of either wavelength or frequency. Thus it also can refer to interactions with particle radiation or to a response to an alternating field or varying frequency (ν). A further extension of the scope of the definition added energy (E) as a variable, once the very close relationship $E=h\nu$ for photons was realized. Spectrometry is the spectroscopic technique used to assess the concentration or amount of a given species. In those cases, the instrument that performs such measurements is a spectrometer or spectrograph.

Spectroscopy/spectrometry is often used in physical and analytical chemistry for the identification of substances through the spectrum emitted from or absorbed by them. Spectroscopy/spectrometry is also heavily

used in astronomy and remote sensing. Most large telescopes have spectrometers, which are used either to measure the chemical composition and physical properties of astronomical objects or to measure their velocities from the Doppler shift of their spectral lines.

1.2.1 Nature of excitation measured

The type of spectroscopy depends on the physical quantity measured. Normally, the quantity that is measured is an intensity, either of energy absorbed or produced.

Electromagnetic spectroscopy involves interactions of matter with electromagnetic radiation, such as light.

Electron spectroscopy involves interactions with electron beams. Auger spectroscopy involves inducing the Auger effect with an electron beam. In this case the measurement typically involves the kinetic energy of the electron as variable.

Mass spectrometry involves the interaction of charged species with magnetic and/or electric fields, giving rise to a mass spectrum. The term "mass spectroscopy" is deprecated, for the technique is primarily a form of measurement, though it does produce a spectrum for observation. This spectrum has the mass m as variable, but the measurement is essentially one of the kinetic energy of the particle.

Acoustic spectroscopy involves the frequency of sound.

Dielectric spectroscopy involves the frequency of an external electrical field

Mechanical spectroscopy involves the frequency of an external mechanical stress, e.g. a torsion applied to a piece of material.

1.2.2 Measurement process

Most spectroscopic methods are differentiated as either atomic or molecular based on whether or not they apply to atoms or molecules. Along with that distinction, they can be classified on the nature of their interaction:

Absorption spectroscopy uses the range of the electromagnetic spectra in which a substance absorbs. This includes atomic absorption spectroscopy and various molecular techniques, such as infrared spectroscopy in that region and nuclear magnetic resonance (NMR) spectroscopy in the radio region.

Reflectance/Emission spectroscopy uses the range of electromagnetic spectra in which a substance radiates (emits). The substance first must absorb energy. This energy can be from a variety of sources, which determines the name of the subsequent emission, like luminescence. Molecular luminescence techniques include spectrofluorimetry.

Scattering spectroscopy measures the amount of light that a substance scatters at certain wavelengths, incident angles, and polarization angles. The scattering process is much faster than the absorption/emission process. One of the most useful applications of light scattering spectroscopy is Raman spectroscopy.

1.3 Sensor types

We define a sensor as an instrument capable of measuring electromagnetic radiation. There are hand held sensors (e.g., field spectrometer) that can be used in the laboratory or during field work for reference measurements. Multispectral sensors contain multiple detectors that are sensitive to specific ranges of the EM spectrum (spectral bands). Typically, multispectral sensors have a few bands (< 10), superspectral sensors have many (> 10) bands, and hyperspectral sensors have up to a few hundred spectral bands (figure 1.4).

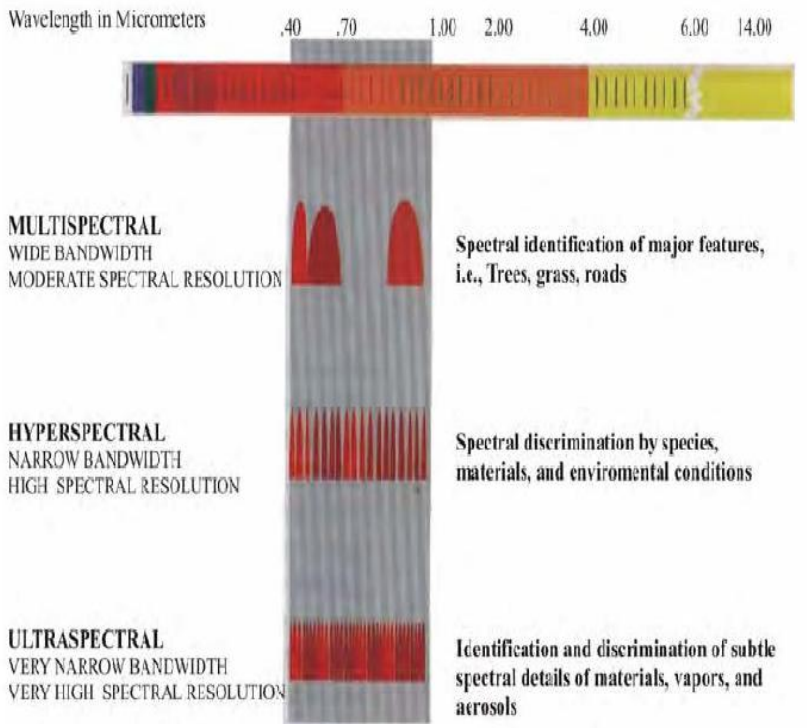


figure 1.4 spectral resolution

Spectrometers are a special type of hyperspectral sensors, where spectral bands are contiguous.

Some hyperspectral sensors have tunable bands that allow gaps in the spectral domain and thus are not spectrometers. The precise spectral information contained in a hyperspectral image enables better characterization and identification of targets.

Another classification scheme of sensors concerns scanning: whiskbroom and pushbroom scanners. Their introduction is important to understand some of the causes of errors in the image data we have to deal with in data processing. The whiskbroom scanner uses a rotating mirror to scan the target surface. It directs a narrow beam of energy onto the detector. An important factor of the whiskbroom is the instantaneous field of view (IFOV) of the scanner.

It is the field of view (or cone angle) of the mirror at the instant that the energy is sensed on the detector. Because there is only one detector for each spectral image, we do not have to deal with inter-calibration problems between neighboring pixels. As a drawback, there is the mechanical part of the rotating mirror, limiting resolution due to slower data rates.

The pushbroom scanner uses a wide-angle optical system that focuses on a strip across the whole of the scene onto a linear array of CCD (Charged Coupled Device) or CMOS (Complementary Metal Oxide Semi-conductor) detectors. The signal from each detector is sampled to create a record for the across track pixels. Each column of each spectral image is acquired with a different sensor, which can lead to striping artifacts. On the other hand, high data rates can be obtained, which make this technique popular for high resolution sensors.

Because sensors are a type of transducer, they change one form of energy into another. For this reason, sensors can be classified according to the type of energy transfer that they detect: thermal sensors, heat sensors, electromagnetic sensors, metal, chemical, optical radiation sensors.

Optical radiation sensors can be further classified:

- Light time-of-flight. Used in modern surveying equipment, a short pulse of light is emitted and returned by a retroreflector. The return time of the pulse is proportional to the distance and is related to atmospheric density in a predictable way.
- **Light sensors, or photodetectors**, including semiconductor devices such as photocells, photodiodes, phototransistors, CCDs, and Image sensors; vacuum tube devices like photo-electric tubes, photomultiplier tubes and mechanical instruments such as the Nichols radiometer. Solid-state detectors detect light by causing photons to excite electrons from immobile, bound states of the semiconductor (the valence band) to a state where the electrons are mobile (the conduction band). The mobile electrons in the conduction band and the vacancies, or "holes," in the valence band can be moved through the solid with externally applied electric fields, collected onto a metal electrode, and sensed as a photoinduced current. Microfabrication techniques developed for the integrated-circuit semiconductor industry are used to construct large arrays of individual photodiodes closely spaced together. The device, called a charge-coupled device (CCD), permits the charges that are collected by the individual diodes to be read out separately and displayed as an image.
- Infra-red sensor, especially used as occupancy sensor for lighting and environmental controls.
- Proximity sensor- A type of distance sensor but less sophisticated. Only detects a specific proximity. May be optical - combination of a photocell and LED or laser. Applications in cell phones, paper detector in photocopiers, auto power standby/shutdown mode in notebooks and other devices. May employ a magnet and a Hall effect device.
- Scanning laser- A narrow beam of laser light is scanned over the scene by a mirror. A photocell sensor located at an offset responds when the beam is reflected from an object to the sensor, whence the distance is calculated by triangulation.

- Focus. A large aperture lens may be focused by a servo system. The distance to an in-focus scene element may be determined by the lens setting.
- Binocular. Two images gathered on a known baseline are brought into coincidence by a system of mirrors and prisms. The adjustment is used to determine distance. Used in some cameras (called range-finder cameras) and on a larger scale in early battleship range-finders
- Interferometry. Interference fringes between transmitted and reflected lightwaves produced by a coherent source such as a laser are counted and the distance is calculated. Capable of extremely high precision.
- Scintillometers measure atmospheric optical disturbances.
- Fiber optic sensors.

Category	Property	Typical
Imaging	Spatial resolution	250 nm (in plane) at $\lambda = 500$ nm
	Field of view	~50 μm (high magnification)
	Dynamic range	8–16 bits (256–65, 536 intensity levels)
	Lowest detectable signal	Shot-noise limited
Spectroscopy	Spectral resolution	1–20 nm (may depend on λ)
	Spectral range	400–900 nm

figure 1.5 typical characteristics of an image spectrometer

1.4 Imaging spectroscopy ⊂ *electromagnetic spectroscopy*

Imaging spectroscopy (also spectral imaging or chemical imaging) is the application of reflectance spectroscopy to every pixel in a spatial image. Spectroscopy can be used to detect individual absorption features due to specific chemical bonds in a solid, liquid, or gas. Solids can be either crystalline (i.e. minerals) or amorphous (like glasses). Every material is formed by chemical bonds, and has the potential for detection with spectroscopy. Actual detection is dependent on the spectral coverage, spectral resolution, and signal-to-noise of the spectrometer, the abundance of the material and the strength of absorption features for that material in the wavelength region measured.

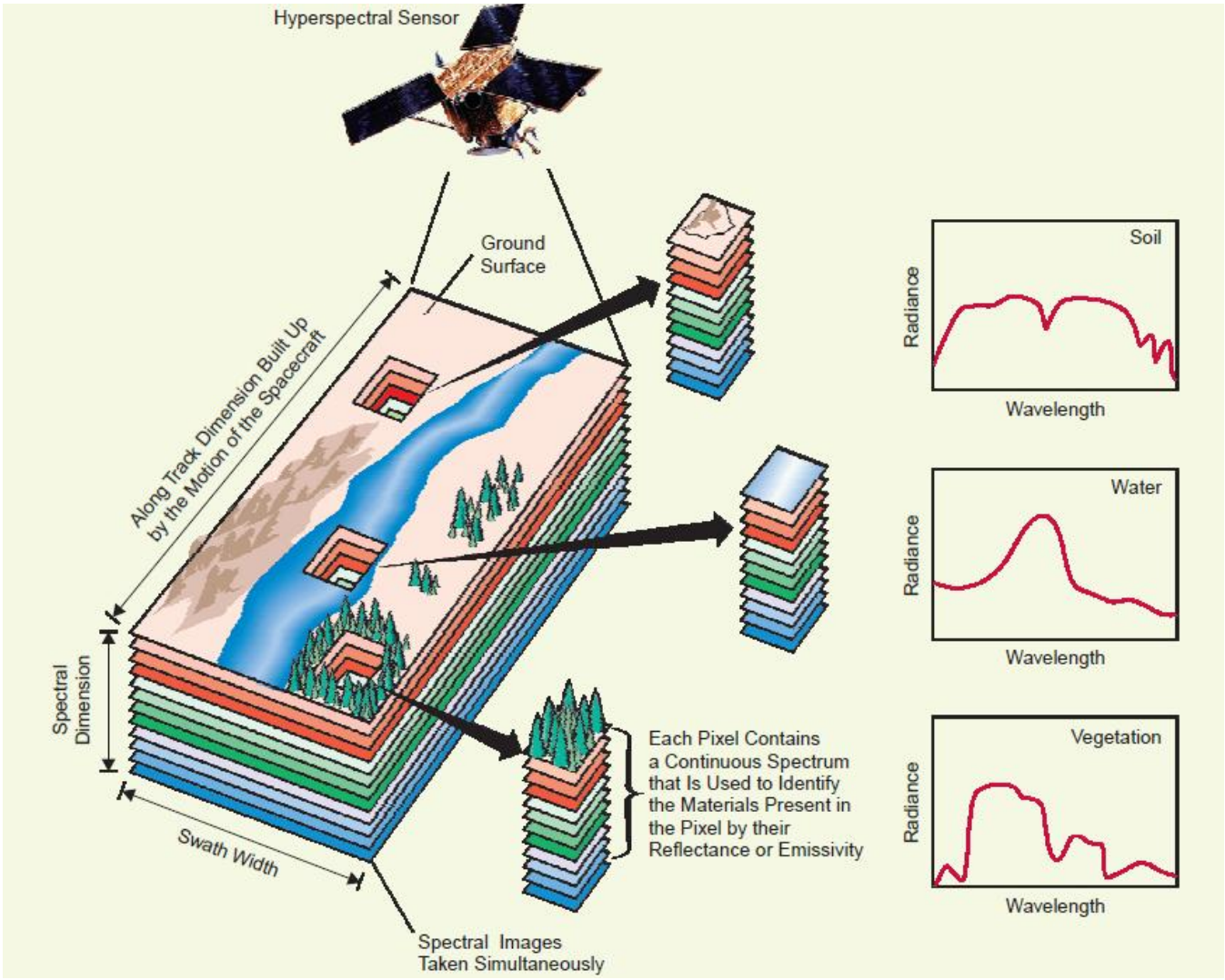


figure 1.6 Principle of (remote sensing) imaging spectroscopy

In remote sensing situations (figure 1.6), the surface materials mapped must be exposed in the optical surface (e.g., to map surface mineralogy it must not be covered with vegetation), and the diagnostic absorption

features must be in regions of the spectrum that are reasonably transparent to the atmosphere (the atmosphere can be corrected for all but the strongest absorptions). The optical surface is the same as what the geologist sees in the field with his or her eyes. Spectroscopy can be used in laboratories on hand samples, in the field with portable field spectrometers (spatial resolution in the millimeter to several meter range), from aircraft, and in the future from satellites. The aircraft systems now operational can image large areas in short time (~2 sq. km per second!), producing spectra for each pixel that can be analyzed for specific absorption bands and thus specific materials. These measurements can then be used for the unambiguous direct and indirect identification of surface materials and atmospheric trace gases, the measurement of their relative concentrations, subsequently the assignment of the proportional contribution of mixed pixel signals (e.g., the spectral unmixing problem), the derivation of their spatial distribution (mapping problem), and finally their study over time (multi-temporal analysis).

Imaging spectroscopy can be considered as the equivalent of color photography, but each pixel needs to acquire many bands of light intensity data from the spectrum, instead of just the three bands of the RGB color model. More precisely, it is the simultaneous acquisition of spatially coregistered images in many spectrally contiguous bands.

Some spectral images contain only a few image planes of spectral data, while others are better thought of as full spectra at every location in the image. For example, solar physicists use spectroheliograms, images of the Sun built up by scanning the slit of a spectrograph, to study the behavior of surface features on the Sun; such a spectroheliogram may have a spectral resolution of over 100,000 (λ / $\Delta\lambda$) and be used to measure local motion (via the Doppler shift) and even the magnetic field at each location in the image plane. The multispectral images collected by the Opportunity rover, in contrast, have only four wavelength bands and hence are only a little more than 3-color images. To be scientifically useful, such measurement should be done using an internationally recognized system of units.

1.5 Hyperspectral imaging

For the last one hundred years detectors have been developed for radiation of almost any region of the electromagnetic spectrum. Recent developments in detector technology incorporate point sensors into integrated detector arrays, which allows setting up imaging radiometers instead of point measuring devices. Quantitative measurements of the spatial distribution of radiometric properties are now available for (remote) sensing at almost any wavelength.

Hyperspectral imaging collects and processes information from across the electromagnetic spectrum. Unlike the human eye, which just sees visible light, hyperspectral imaging is more like the eyes of the mantis shrimp, which can see visible light as well as from the ultraviolet to infrared. Hyperspectral capabilities enable the mantis shrimp to recognize different types of coral, prey, or predators, all which may appear as the same color to the human eye (figure 1.7).

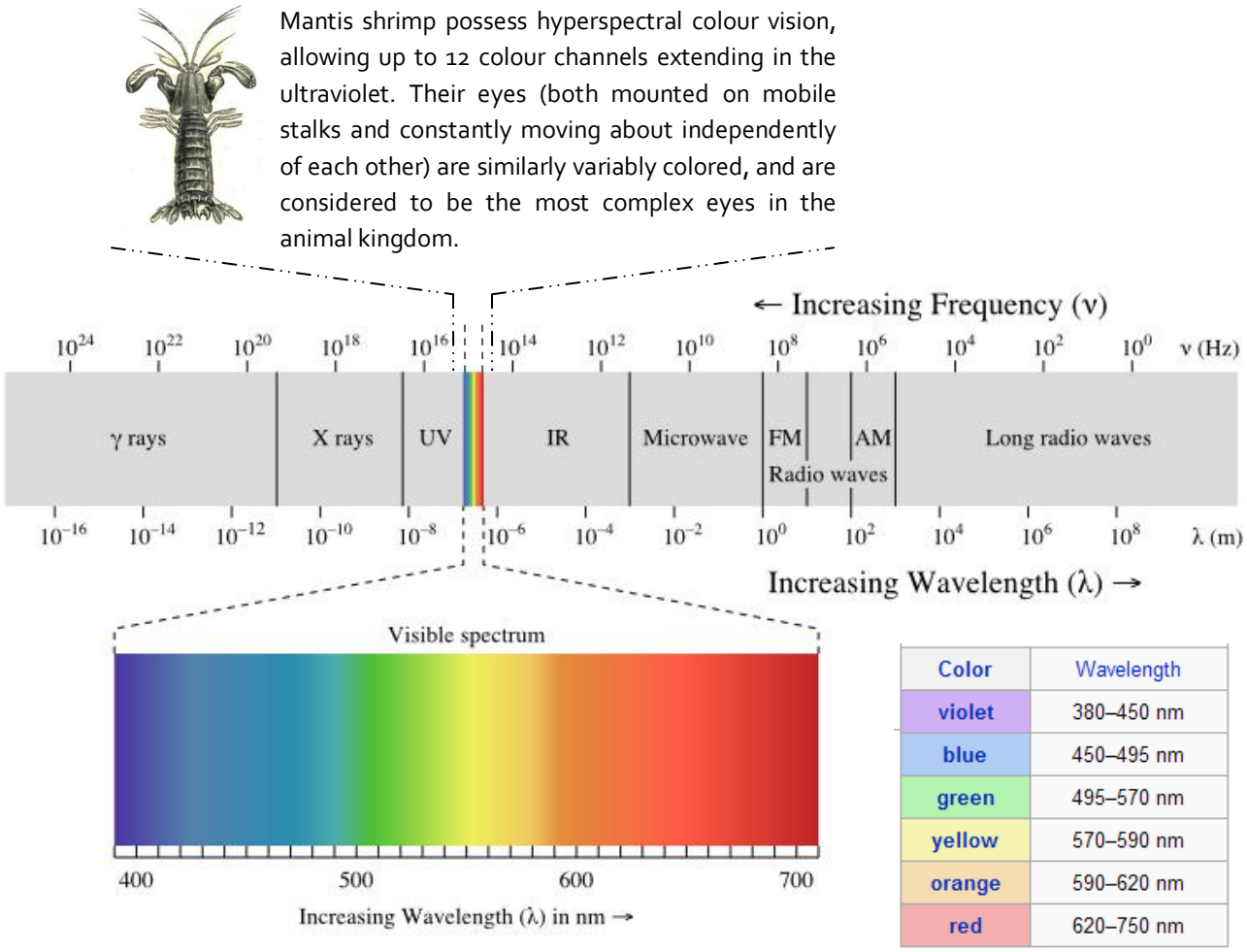


figure 1.7 mantis shrimp's spectacular eye vision spectral range

Humans build sensors and processing systems to provide the same type of capability for application in agriculture, mineralogy, physics, and surveillance and other fields of science. Hyperspectral sensors collect information as a set of 'images'. Each image represents a range of the electromagnetic spectrum and is also known as a spectral band. Hyperspectral sensors look at objects using a vast portion of the electromagnetic spectrum. Certain objects leave unique 'fingerprints' across the electromagnetic spectrum. These 'fingerprints' are known as

spectral signatures and enable identification of the materials that make up a scanned object. For example, having the spectral signature for oil helps mineralogists find new oil fields.

The precision of these sensors is typically measured in spectral resolution, which is the width of each band of the spectrum that is captured. If the scanner picks up on a large number of fairly small wavelengths, it is possible to identify objects even if said objects are only captured in a handful of pixels. However, spatial resolution is a factor in addition to spectral resolution. If the pixels are too large, then multiple objects are captured in the same pixel and become difficult to identify. If the pixels are too small, then the energy captured by each sensor-cell is low, and the decreased signal-to-noise ratio reduces the reliability of measured features.

Chapter 2
Hyperspectral Data Processing

The high spectral resolution characteristic of hyperspectral sensors preserves important aspects of the spectrum (e.g. shape of narrow absorption bands) and makes differentiation of different materials possible. The spatially and spectrally sampled information can be described as a datacube, whose face is a function of the spatial coordinates and depth is a function of spectral band (or wavelength).The data in each band corresponds to a narrow band image of the surface covered by the field of view of the sensor, where as along the wavelength dimension, each image pixel provides a spectrum characterizing the materials within the pixel. The nature and organization of the collected data is illustrated in figure 2.

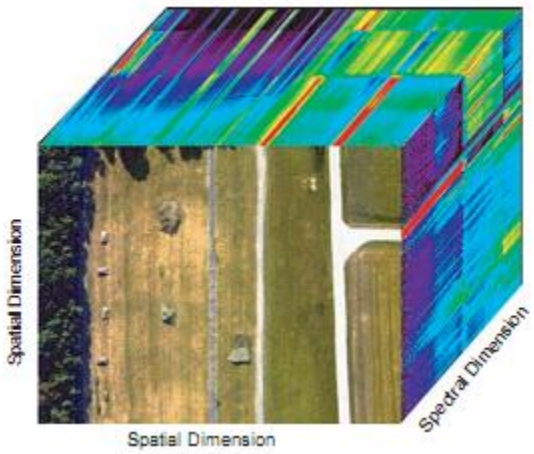


figure 2: Imaging spectrometry data cube illustrating

the 3-D spatial and spectral character of the data.

2.1 Signal representation

The data that is supplied by such systems is best represented in the form of an N-dimensional vector for each pixel where *N* is the number of spectral bands. This viewpoint of the data is referred to as a feature space representation, as compared to the image space and spectral space presentation in figure 2.1. Typically there are several hundred thousand pixels per data set. The spectral space graph of figure 2. 1(b) might lead one to believe that each ground cover material is appropriately represented by a single spectral curve; some use the term "spectral signature." To proceed from this assumption gives up a considerable amount of potential. The angle of the sun, and thus the time of day, season and latitude, the direction of view, the atmospheric condition, and a number of other such uncontrollable variables substantially affects the spectral response of any given material. From a scientific point of view, it has been of interest to try to make adjustments for these variables. However, this proves to be quite a daunting problem as it is difficult to accumulate the needed data for each pixel and each

column of atmosphere to enough precision to do more than have a cosmetic effect on the data in image space. Sound application of appropriate analysis algorithms are not usually much improved by such adjustments.

More significantly, beyond these observational variables, at least when considering remote sensing situations, the Earth's surface itself is a highly variable and dynamic place from a spectral point of view. Consider the grassy areas of figure 2.1(a). Even in terms of the three bands used to generate this image, it is apparent to the unaided eye that the spectral response of the class "grass" varies significantly over that scene. From a data analysis point of view, it is important to recognize that this variation in the ground scene response is not all "noise." Some (most) of this variation is information bearing. Thus, from a data analysis standpoint, a more effective and complete representation of diagnostic spectral responses is in terms of class-conditional probability d

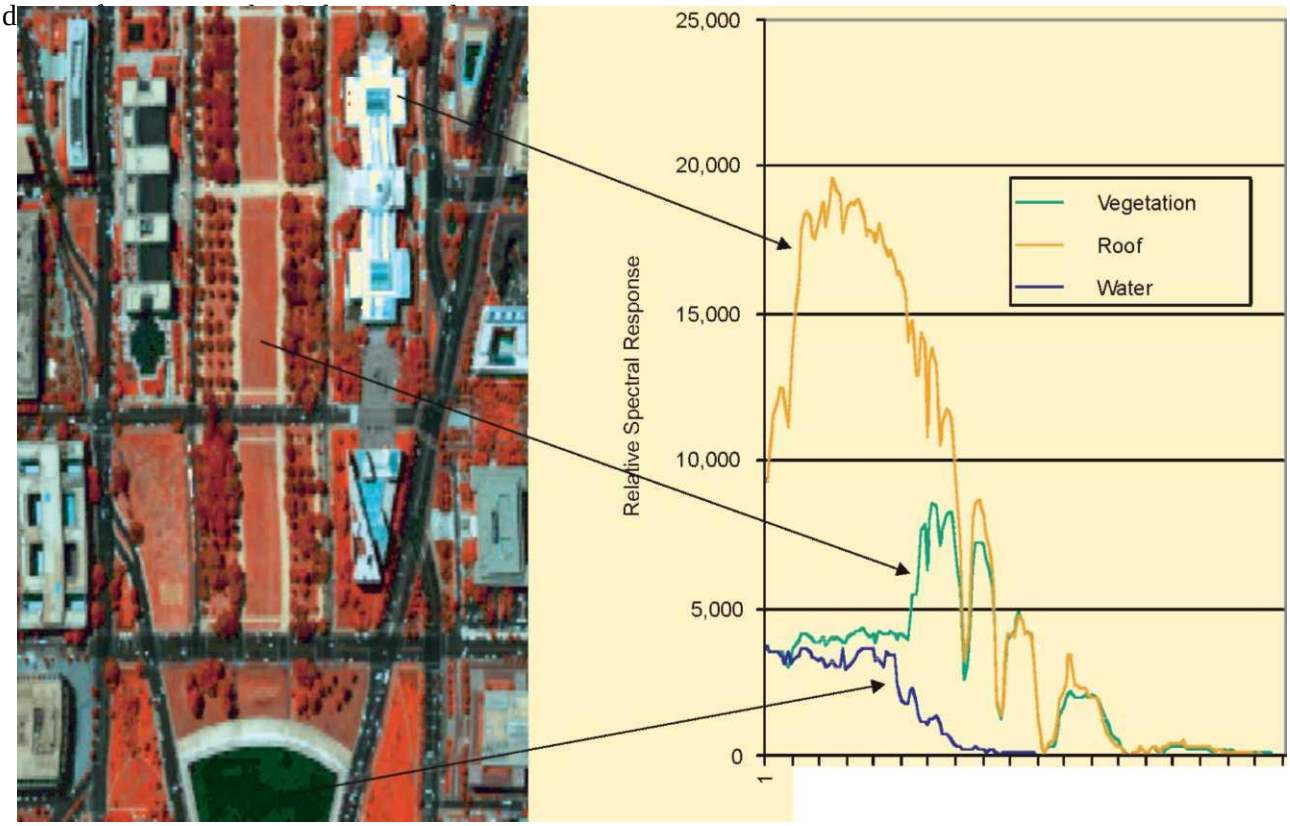


figure. 2.1. (a) A simulated color IR image of an urban area, the Washington, D.C., mall. This image is made using three bands of the 210 bands collected by the sensor system, one band from the visible green, one from the visible red, and one from the near infrared. Such displays are referred to as displays in image space. (b) A display of the data of pixels of three materials as a function of wavelength by spectral band number. The bands in this case are approximately 10 nm wide over the range of 0.4-2.4 pm. This type of data display is referred to as a display in spectral space.

2.2 Class Discrimination – Feature Selection

It is in such a representation, where not only the average spectral response but also the manner of variation of a material's response about its average exhibits, that is the most information bearing. To make clearer the value of this model in discriminating between two classes, one of the most common ways to predetermine the

separability of two classes of materials is by the use of a statistical distance measure [1]. A simple straightforward measure is the Euclidean distance, where only mean differences are used, neglecting the covariance of the classes. The Mahalanobis distance use same mean covariance matrix, whereas Bhattacharyya and Jeffreys-Matusita account for the different class covariances. Equations are in the respective distances in their Gaussian form:

$$Euclidean\ Distance:\ D_E = (m_i - m_j)^T (m_i - m_j)$$

$$Mahalanobis\ Distance:\ D_M = (m_i - m_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (m_i - m_j)$$

$$Bhattacharyya\ Distance:\ D_B = \frac{1}{8} D_M + \frac{1}{2} ln \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}}$$

$$Jeffries - Matusita\ Distance:\ D_{JM} = 2(1 - e^{D_B})$$

Note that the first term on the right measures the portion of the class (the two classes are referred as i and j) separation due to the difference in means, while the second term measures the separation of the classes due to the covariances. Thus, to use only a single spectral curve to model a class (a "spectral signature?"), even if it is the average of a number of actual spectral responses makes use of only the separability measured by the first term on the right of the above equation. Further, even from this partial modeling of the class densities, it is clear that, though two classes might have the same mean values, making that first term on the right zero, they may still be quite separable.

Modeling each class in terms of a probability density function allows one to capture the information about a class also by the "shape" of the class in feature space, as quantified by all higher order statistics. Then classification can conveniently be implemented via the discriminant function concept. Other measures for calculating the class separability are the Receivers Operating Characteristic curve(ROC curve,[2]).

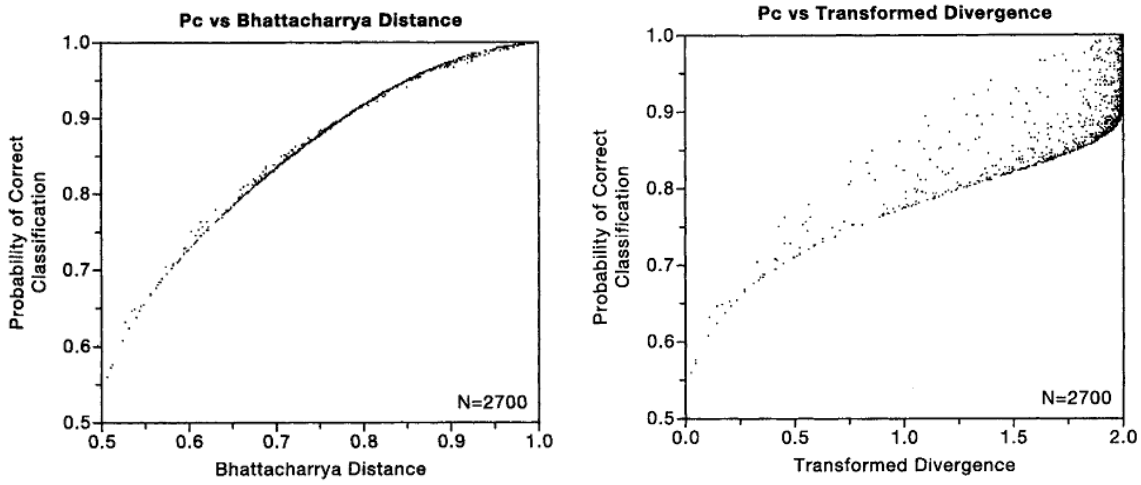


figure 2.2 : Results of a Monte Carlo study of the relationship between Probability of Correct Classification and two different statistical separability measures.

Unfortunately, the relationship between distance measures and classification accuracy is not precisely one-to-one for any of the measures above, meaning that a given value of a distance measure does not imply a specific value of probability of correct classification. Rather, the best that can be said is that (1) a given value of a distance measure implies a certain range of possible classification accuracies, and (2) usually (but not always), a larger value of a distance measure implies a larger value of classification accuracy. This is an important limitation, and has been studied in the literature extensively. Only in a few cases has it been possible to derive bounds on the probability of correct classification. In cases where it has not been possible to derive such error bounds analytically, it is useful to study the property of separability measures empirically. In the graphs of figure 2.2 are shown graphically the results of Monte Carlo studies of the relationship between several distance measures and classification accuracy. The results that fall in a narrower range imply a greater degree of unique one-to-one mapping between the distance measure and classification accuracy.

From these results for two-dimensional data, the Error Function Bhattacharrya Distance appears to most nearly provide the kind of performance desired, with direct Bhattacharrya Distance not far behind. Given the larger amount of computation required for the error function transformation, the direct Bhattacharrya Distance measure may be a good choice for many circumstances.

A naive search for a set of N features results in a univariate set of M best features. But ordering all the features by some class separability measure and selecting the *M* best, does not necessarily lead to the best multidimensional feature set. A more successful approach is a multivariate one, in which we try to find the optimal *combination* of available features. The best subset of M variables out of N may be found by evaluating the class separability measure (D) for all possible combinations of M variables. However, the number of all possible combinations, becomes huge even for modest values of N and M. Therefore, several approaches have been suggested to avoid the exhaustive search. We divide them into three groups: sequential, randomized and exponential algorithms. An example of a randomized algorithm is the *tabu search,* first presented by Glover [3]. Zang used it for feature selection [4]. The tabu search starts with an initial feature set of the desired length M. A random number i from 1 to M determines which feature is scanned. All unused features are tried in position i of the feature set to improve the performance of the feature set. The best performing feature replaces the old feature on that position. Notice that the old feature is always replaced, even if this means a loss in performance of the new feature set. If we would only replace the old feature in case of a performance gain, we were bound to be trapped in the first local minimum encountered.

Exponential algorithms evaluate a number of subsets that grow exponentially with the dimensionality of the search space, e.g., exhaustive search, branch and bound [5]. Sequential algorithms add and remove features sequentially, e.g. sequential forward and backward selection.

2.3 High dimensional Data

A two-channel feature space plot for the area marked by the dashed rectangle in the three-channel image space figure is shown in figure 2.3. From the image space presentation, which utilizes three of the 12 bands available in this data set, it appears that there are two fairly distinct classes of ground cover in the rectangular area, but this is not so apparent from a visual observation of the two-dimensional feature space presentation. For these two classes in these two bands, the data appears to be heavily overlapped, and the two classes do not appear to be spectrally distinct. However, an advantage of the feature space representation is that its dimensionality is easily expanded, while that of the image space is not. If one adds a third dimension to this feature space or a fourth, one might well be able to visualize that spreading these same data points over the larger volume of the higher dimensional space would allow for greater potential separability. Increasing the dimensionality further would spread the data over an even greater volume, thus reducing overlap and enhancing the potential for discrimination, so long as the fundamental assumption that different materials do have diagnostically different characteristics remains valid. (We note in passing that for multispectral data of Earth observational scenes, like the case illustrated above, classes of data in N-dimensional feature space usually do not occur in distinct clusters. Rather they occur in a sparse continuum, making the process of quantitatively specifying to considerable precision the classes to be discriminated a key to successful data analysis). As an extreme illustration of this, consider that one has 10-bit data in 100-dimensional space, a very feasible circumstance today. The 10-bit data implies 1024 possible discrete values in each of the 100 dimensions, or that there are approximately $(10^3)^{100} = 10^{300}$ discrete locations in this feature space. The volume of this space is so great that even for a data set of 10^6 pixels, the probability of any two pixels landing in the same digital cell or even fairly adjacent cells is vanishingly small. Thus there is no overlap, and in theory, anything is separable from anything. However, there are complexities that must be dealt effectively in such a space in order to approach this potential.

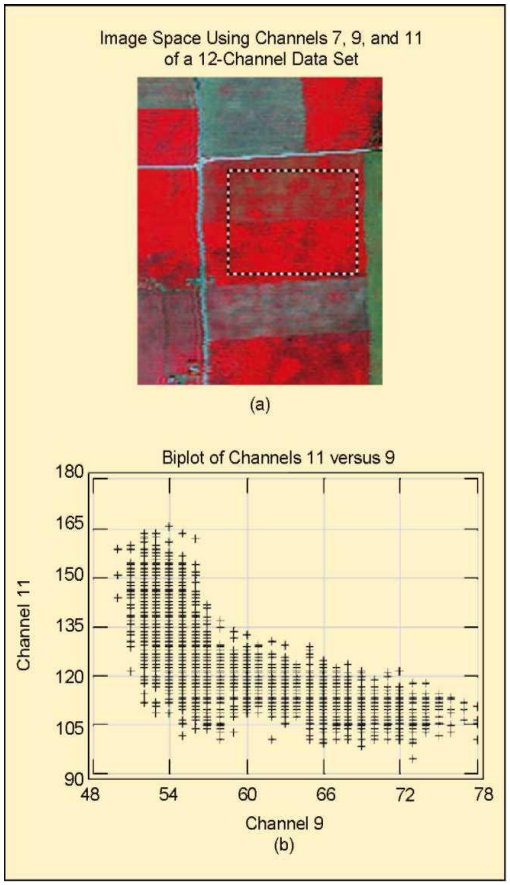


figure. 2.3: Two agricultural species in (a) three-dimensional image space and (b) two-dimensional feature space

High-dimensional vector spaces have been found by mathematicians to have some rather unusual and unintuitive characteristics [6]. It has been shown [7] that high-dimensional space is mostly empty, which implies that multivariate data in R^d is usually in a lower dimensional structure. As a consequence, for any given analysis task, high-dimensional data can be projected to a lower di-

mensional subspace without losing significant information in terms of separability among the different statistical classes. However, the specific subspace will surely be different for each different data set and analysis task. A second consequence of the foregoing, is that normally distributed data will have a tendency to concentrate in the tails; similarly, uniformly distributed data will be more likely to be collected in the corners, making density estimation more difficult. Local neighborhoods are almost surely empty, producing the effect of losing detailed density estimation.

It turns out that this difficulty in density estimation is one of the chief challenges facing the data analyst. Due to the large number of parameters of the scene and its observation, one must expect to have to train a classifier for each new data set that is to be analyzed. The labeling of training samples and accumulation of the information by which to do so nearly always means that there will be a paucity of training samples with which to model each of the class density functions. Thus, one must determine the parameters of a high dimensional density function with a relatively small number of samples.

In a very general context, Hughes was able to demonstrate the impact of this problem on a theoretical basis some years ago [20]. One of his results is displayed in figure: 2.4, which shows the mean recognition accuracy averaged over the ensemble of possible classifiers, versus the measurement complexity. Here, measurement complexity is related to the number of discrete cells in the feature space, and therefore the number of spectral bands and the bit precision in each. The parameter, *m*, of the individual graphs of the figure is the number of training samples available to define the classes.

It is seen that the expected accuracy starts at 50% for this two-class case, i.e., chance performance. For the case of an infinite number of training samples, the curve proceeds upward to the right as measurement complexity increases, rapidly at first but then more slowly, becoming asymptotic to its final value. However, for any finite number of training samples, the result has a maximum value. This is because there will then be estimation error in determining the values of the parameters of the classifier, and for a given number of training

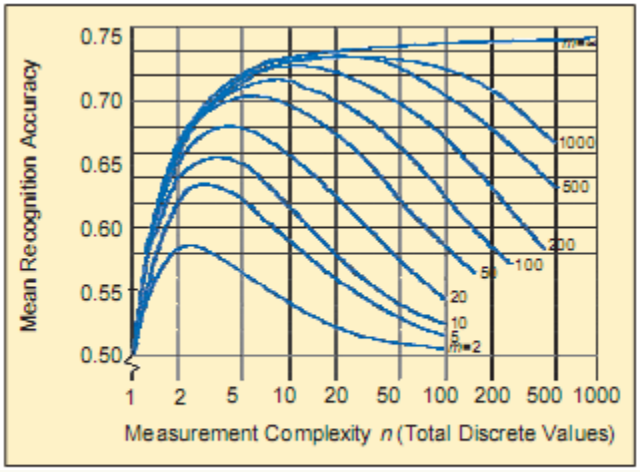


figure. 2.4: Mean recognition accuracy versus measurement complexity for the finite training case.

samples, the greater the measurement complexity the greater the estimation error and the poorer the performance. This may be the explanation for less complex classifiers sometimes outperforming more complex ones. The maximum value of each curve does increase with increasing numbers of training samples, and in this case, occurs at a higher measurement complexity. Thus on average, to achieve higher accuracy will require increased numbers of features and/or an increase in SNR reflected in the number of bits or discrete values per feature. Thus the number of spectral features and the SNR are interrelated with the number of training samples available per class.

2.4 Data Analysis Procedure [31]

The major question that the analyst must deal with is how to choose and train a suitable sequence of algorithms by which to accomplish the desired analysis, given the circumstances found in an experiment. The problem of optimally training a classifier comes down to how completely and precisely one models the data set and the specific classes one wishes to discriminate between. The classification process ordinarily involves assigning each pixel to one of a list of classes. Thus one must set up an exhaustive list of classes, so that there is a logical class to which to assign each pixel of the data set, even though one may be interested in only one or a small number of classes in the scene. The rule for establishing the list of classes then is that the classes must be:

- Of informational value. The list must contain all of the classes of interest to the information consumer.
- Exhaustive. In addition to those desired by the user, it must contain enough additional classes so that there is a logical class to which to assign each pixel in the data set.
- Separable. The classes must be separable in terms of available spectral features.

Further, each class must be modeled to adequate completeness and precision. As pointed out earlier, one must specify not only the mean response of a given class, but also how the response for that class varies about its mean, since this variation is often quite diagnostic of the class. Modeling the class response in terms of a multidimensional probability density function is perhaps the most effective way of doing this. However, as the measurement complexity, defined by the number of features and the bit precision of the data reflecting S/N, increases, this becomes more daunting. There will usually only be a limited number of samples that can be made available for defining a class density function model. The number of training samples needed varies greatly with the specific situation. A number many times as large as the number of features is highly desirable, although there are algorithms becoming available to mitigate this condition to some extent.

In addition, the analyst has the further challenge that the samples used for training the classifier must be truly representative of the class intended. If one wishes to define a class to be called "corn" in an agricultural problem, how thick must be the stand of corn in a pixel for it to be desired to call the pixel corn, how much weeds should be allowed, what range of varieties (include popcorn?) what range of planting dates and thus maturity

level, and many other variables must be considered. Clearly the process is not in any sense "automatic" as it must reflect the specific requirements of the user.

Data flow through a system to a final analysis generally requires the application of a sequence of algorithms. Figure 2.5 outlines such a sequence. The numbered paragraphs refer to the numbered boxes in the diagram:

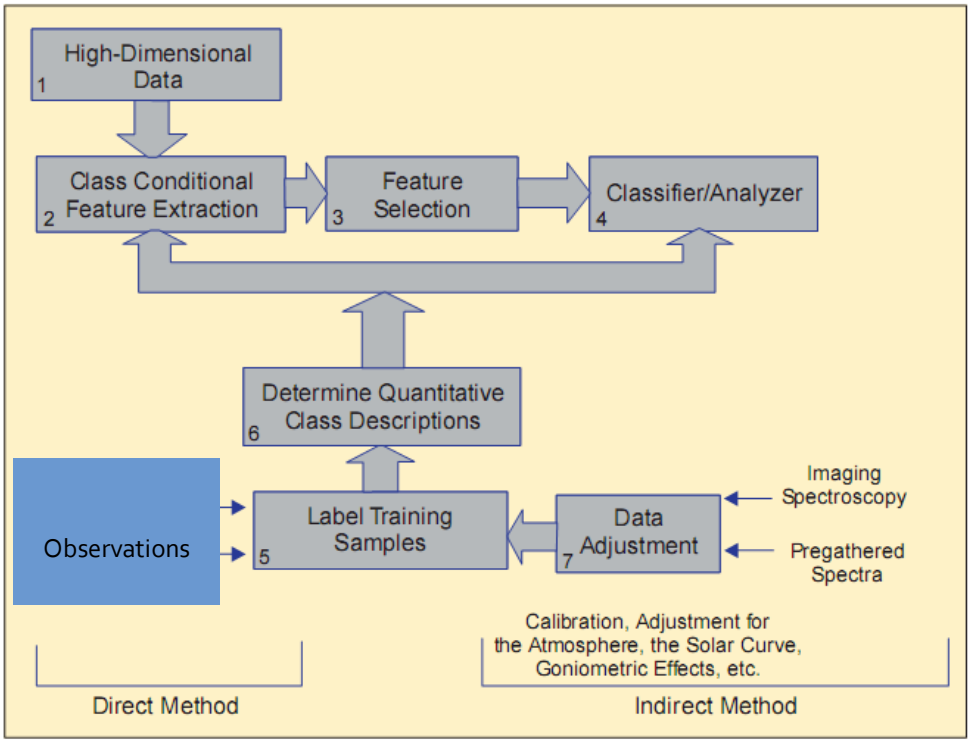


Figure. 2.5. A schematic diagram of the hyperspectral data analysis process.

1)Hyperspectral data consists of data gathered in more than one spectral band. There is no accepted definition for where the boundary is between data termed multispectral and hyperspectral. It is well established that the geometry of vector spaces changes continually as the dimensionality of the space increases, and indeed that it is materially different from the familiar three-dimensional geometry by the time dimensionality reaches seven to ten. Further, it usually requires a dimensionality of the order often or more to satisfactorily accomplish many practical analysis tasks. Thus it will be assumed that the data to be analyzed contains at least ten and perhaps as many as several hundred spectral bands.

2)Again assuming that the data were gathered in a larger number of bands than is necessary or desirable for the particular analysis at hand, an important early step is to form the feature subset that is to be used in the analysis. This should be done in a situation-specific way, that is, using the description of the specific classes desired. Thus, a feature extraction algorithm such as those described above is applied at this point.

3) Given box 2, there may still remain the decision as to how many of the generated features to utilize. The choice here and that in box 4 will depend to some extent upon the individual classes and the precision with which they have been modeled.

4) There remains, then, the application of the specific classification algorithm to be used. Again, the choice of algorithm depends upon the class model precision and the level of detail of the classes.

5) As has been detailed above, the labeling of adequate sets of training samples is a key step, perhaps the most important step of the entire process.

6)Having labeled a set of samples for each class that are assumed to be truly representative of an exhaustive list of classes that includes the desired classes, the task here is to use those samples to define as precise an N-dimensional model of the classes in the feature space as possible. Except in very simple cases where a single point in feature space is adequate, this will nearly always consist of modeling the entire distribution of each class. This may involve use of an iterative scheme, or it may simply consist of computing first- and second-order statistics. However classes may require modeling in terms of more than one mode, with the training samples divided between the various modes. There are also additional algorithms that can further assist in mitigating the small training sample problem [8].

7)Box 7 suggests one option for labeling training pixels being an attempt to adjust all or a part of the data for the various observational variables that were present, depending on the precise conditions of the scene and the sensor system at the time each pixel measurement was made. If one could do this adequately, this would make possible the use of some additional sources of reference data on which to base the labeling, as indicated on the diagram. The adjustment of the data for all of these variables is a very complex task and is problematic. It often cannot be done with as much precision as needed. Because of this, the overall scheme above is designed to not necessarily require calibrated data that has been so adjusted.

Chapter 3
Hyperspectral Classification
3.1 Introduction

Statistical classification is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (referred to as traits, variables, characters, etc) and based on a training set of previously labeled items. These characteristics are generally their response in different spectral ranges. Labeling is implemented through pattern classification procedures. The term "pattern" refers to the set of radiance/reflectance measurements obtained in the various wavebands for a given pixel, and spectral pattern classification refers to the family of classification procedures that utilizes this pixel-by-pixel spectral information as the basis for land cover classification, painting material classification, cancer tissue identification etc. In contrast, spatial pattern recognition involves the classification of image pixels on the basis of their spatial relationship with pixels surrounding them. Temporal pattern recognition uses change in spectral reflectance over time as the basis of feature identification.

The classification process has two main stages. In the first stage, the number and nature of the categories are determined, whilst in the second stage every unknown or unseen element is assigned to one of the categories according to its level of resemblance (or similarity) to the basic patterns. These stages are often called classification and identification, respectively. In the context of hyperspectral remote sensing, the categories could be land cover features or cloud types, and the assignment to one of the categories is carried out by allocating numerical labels, corresponding to the classes, to individual pixels. Hence, for a researcher working in the remote sensing field, classification basically means determining the class membership of each pixel in an image by comparing the characteristics of that pixel to those of categories known a priori.

3.2 The classification process

Image classification is the process of creating a meaningful digital thematic map from an image dataset. The classes shown on the map are derived either from known cover types or by algorithms that search the data for similar pixels. Once data values are known for the distinct cover types in the image, a computer algorithm can be used to divide or segment the image into regions that correspond to each cover type or class. Image classification can be done using a single image dataset, multiple images acquired at different times, or image data with additional information such as elevation measurements, or expert knowledge about the area. Traditionally, image classification involves several steps :

(i) "Feature extraction/selection : The term feature refers to a single element of a pattern. More generally, a feature can be thought of "...as a distillation of that information contained in the measurements which is

useful for deciding on the class to which the pattern belongs" (Swain and Davis, 1978). In addition data are often highly correlated between spectral bands, which may not be useful for land cover classification and even may reduce classification accuracy. Thus, feature extraction performs two functions:

- separation of useful information from noise or non-information and
- reduction of the dimensionality of the data in order to simplify the calculations performed by the classifier, and to increase the efficiency of statistical estimators in a statistical classifier.

These aims can be achieved by applying spatial or spectral transform to the image, such as selection of a subset of bands, or a principal component transformation to reduce the data dimensionality.

(ii) Training: The term "training" arose from the fact that many pattern recognition systems were "trainable"; i.e., they learned the discriminant functions in the feature space by adjusting their parameters when applied to a training pattern (pixel vector) whose true class is known. This process of training a classifier is either supervised by the analyst or unsupervised.

(iii) *Labeling:* The process of allocating individual pixels to their most likely class is known as labelling. This process of labeling can be approached in one of two ways. If the analyst knows the number of separable pixels that exist in the area covered by the image, and if it is possible to estimate the statistical properties of the values taken on by the features describing each of these pixels (in statistical classifiers), then individual pixels (test pixels) can be labeled as belonging to the classes based on these statistical properties. The other method is where the analyst has no clear idea of the number and character of the classes present in the images. A method of allocating and reallocating the individual pixels to one of an initial set of randomly-chosen pixels is used. At each stage, each pixel in turn is given the label of one of these randomly chosen pixels using some classifier. At the end of first iteration, when every pixel has been labeled, the randomly chosen pixels can be altered in character (either by combining, splitting, and removing some of the pixels) according to the nature of the pixels which have been associated with them. This process of pixel labeling is repeated until the process converges.

Generally, image classification techniques can be divided into supervised and unsupervised methods based on the involvement of the user during the classification process. Methods can be further sub-divided into parametric and non-parametric techniques, based on whether or not the classifier employs some distributional assumption about the data.

Supervised learning is the more useful technique when the data samples have known outcomes that the user wants to predict. On the other hand, unsupervised learning is more appropriate when the user does not know the subdivisions into which the data samples should be divided. Prior categorical division may not be obvious because the problem may be a new one, for which the user has little experience. In such a case, an unsupervised learning procedure can provide insight into groupings that may make physical sense and facilitate future analysis.

Supervised classification techniques require training areas to be defined by the analyst in order to determine the characteristics of each category. Each pixel in the image is, thus, assigned to one of the categories using the extracted discriminating information. Problems of diagnosis, pattern recognition, identification, assignment and allocation are essentially supervised classification problems, since in each case the aim is to classify an object into one of a pre-specified set of classes. Unsupervised classification, on the other hand, searches for natural groups of pixels, called clusters, present within the data by means of assessing the relative locations of the pixels in the feature space. In these classification systems, an algorithm is used to identify unique clusters of points in feature space, which are then assumed to represent unique categories. These are automated procedures and therefore require minimal user interaction.

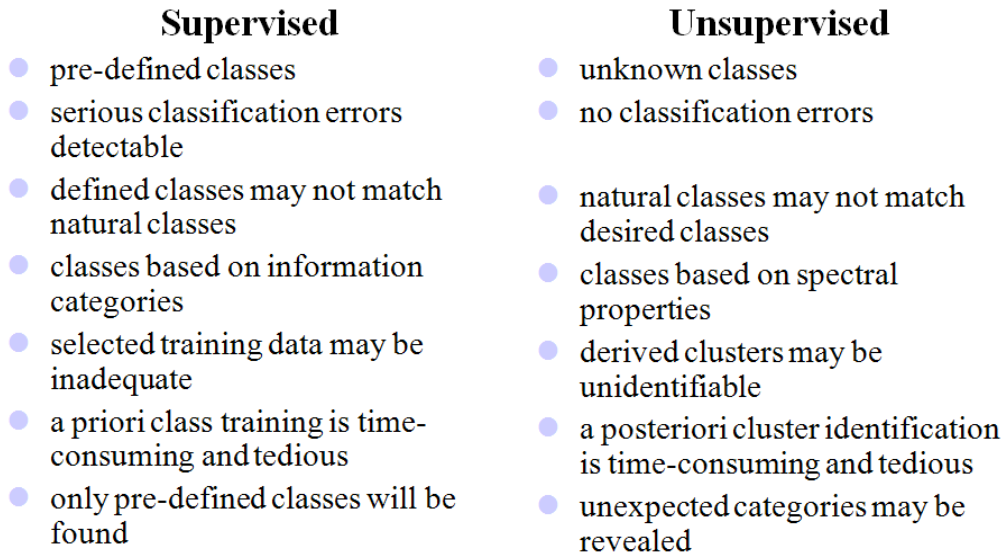


figure 3.1: Supervised VS. Unsupervised classification

Parametric classification procedures use some statistical measures to derive rules from the data, which leads to some assumptions. The most common assumption of this kind is that of the normal (Gaussian) frequency distribution of the data being used. However, non-parametric methods do not make any assumptions about the frequency distribution of the data used, and do not use statistical estimates. The minimum distance and maximum likelihood classifiers are examples of statistical classification methods, whilst the artificial neural network, support vector machine, and decision tree methods can be given as examples of non-parametric classification methods.

3.2.1 Data reduction

One of the problems with hyperspectral (high-dimensional) datasets is that, in many cases, not all the measured variables are important for understanding the underlying phenomena of interest. While certain computationally expensive novel methods can construct predictive models with high accuracy from high-

dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modeling of the data. In mathematical terms, the problem can be stated as follows: given the M-dimensional random variable $\boldsymbol{x}=[x_1\ x_2\ \dots\ x_M]$, *f*ind a lower dimensional representation of it, $\boldsymbol{y}=[y_1\ y_2\ \dots\ y_M]$ with k<M, that captures the content in the original data, according to some criterion. The components of $\boldsymbol{y}$ are sometimes called the hidden components.

Principal components analysis (PCA) [9] is a classical method that provides a sequence of best linear approximations to a given high-dimensional observation. It is one of the most popular techniques for dimensionality reduction. However, its effectiveness is limited by its global linearity. Multidimensional scaling (MDS) [10], which is closely related to PCA, suffers from the same drawback. Factor analysis [11] and independent component analysis (ICA) [12] also assume that the underling manifold is a linear subspace. However, they differ from PCA in the way they identify and model the subspace. The subspace modeled by PCA captures the maximum variability in the data, and can be viewed as modeling the covariance structure of the data, whereas factor analysis models the correlation structure. ICA starts from a factor analysis solution and searches for rotations that lead to independent components .The main drawback with all these classical dimensionality reduction approaches is that they only characterize linear subspaces (manifolds) in the data. In order to resolve the problem of dimensionality reduction in nonlinear cases, many recent techniques, including kernel PCA [14], Laplacian eigenmaps (LEM) [15], etc.

3.2.1.1 Principal Components Analysis (PCA)

Principal component analysis (PCA) is the best, in the mean-square error sense, linear dimension reduction technique. Being based on the covariance matrix of the variables, it is a second-order method. In various fields, it is also known as the singular value decomposition (SVD), the Karhunen-Loeve transform, the Hotelling transform, and the empirical orthogonal function (EOF) method. In essence, PCA seeks to reduce the dimension of the data by finding few orthogonal linear combinations (the PCs) of the original variables with the largest variance. In other words, the goal of principal component analysis is to compute the most meaningful (orthonormal) basis to re-express a noisy data set. The hope is that this new basis will filter out the noise and reveal hidden structure.

By assuming linearity PCA seeks a linear combination of the original basis that best re-express a given dataset. The first component, s_1 , is the linear combination with the largest variance. We have $s_1 = \boldsymbol{x}^T\boldsymbol{w_1}$, where the p-dimensional coefficient vector $\boldsymbol{w}=[w_1\ w_2\ \dots\ w_p]$ *solves*

$$w_1 = argmax_{||w=1||} Var\{\boldsymbol{x}^T\boldsymbol{w}\}$$

The second PC is the linear combination with the second largest variance and orthogonal to the first PC, and so on. There are as many PCs as the number of the original variables. For many datasets, the first several PCs explain most of the variance, so that the rest can be disregarded with minimal loss of information. Since the variance depends on the scale of the variables, it is customary to first standardize each variable to have mean zero and standard deviation one. After the standardization, the original variables with possibly different units of measurement are all in comparable units. Assuming a standardized data with the empirical covariance matrix

$$\Sigma_{pxp} = \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^T.$$

Note that $\boldsymbol{X}\boldsymbol{X}^T$ is a symmetric matrix so it can be diagonalized by its orthonormal eigenvectors. Thus

$$\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T,$$

Where $\boldsymbol{\Lambda}$=$diag[\lambda_1\ \lambda_2\ \ldots\ \lambda_p]$ is the diagonal matrix of the ordered eigenvalues $\lambda_1 < \lambda_2 < \ldots < \lambda_p$ and $\boldsymbol{U}$ is a p x p orthogonal matrix containing the eigenvectors. It can be shown [21] that the PCs are given by the p rows of the p x n matrix **S**, where **S** = $\boldsymbol{U}^T\boldsymbol{X}$ (the weight matrix **W** is given by $\boldsymbol{U}^T$). The subspace spanned by the first k eigenvectors has the smallest mean square deviation from **X** among all subspaces of dimension k. Another property of the eigenvalue decomposition is that the total variation is equal to the sum of the eigenvalues of the covariance matrix,

$$\sum_{i=1}^{p} Var(PC_i) = \sum_{i=1}^{p} \lambda_\iota = \sum_{i=1}^{p} trace(\boldsymbol{\Sigma})$$

and that the fraction $\sum_{i=1}^{\kappa} \lambda_\iota / trace(\boldsymbol{\Sigma})$ gives the cumulative proportion of the variance explained by the first κ PCs.

Performing PCA is quite simple in practice:

1 .Organize a data set as an mxn matrix, where m is the number of measurement types and n is the number of trials.

2. Subtract of the mean for each measurement type or row x_i .

3. Calculate the eigenvectors (or the SVD) of the covariance.

One benefit of *PCA i*s that we can examine the variances ($Var(PC_i)$) associated with the principle components. Often one finds that large variances associated with the first k<p principal components, and then a precipitous

drop-off. One can conclude that most interesting dynamics occur only in the first k dimensions. Both the strength and weakness of *PCA* is that it is a *non-parametric* analysis.

The interpretation of the PCs can be difficult at times. Although they are uncorrelated variables constructed as linear combinations of the original variables, and have some desirable properties, they do not necessarily correspond to meaningful physical quantities. In some cases, such loss of interpretability is not satisfactory to the domain scientists.

3.2.1.2 kernel Principal Components Analysis (k-PCA)

Kernel principal component analysis (kernel PCA) is an extension of principal component analysis (PCA) using techniques of kernel methods. Using a kernel, the originally linear operations of PCA are done in a reproducing kernel Hilbert space with a non-linear mapping. Kernel Methods approach the problem by mapping the data into a high dimensional feature space, where each co-ordinate corresponds to one feature of the data items, transforming the data into a set of points in a Euclidean space. In that space, a variety of methods can be used to find relations in the data. Since the mapping can be quite general (not necessarily linear, for example), the relations found in this way are accordingly very general. This approach is called the kernel trick.

Kernel trick is a method for using a linear classifier algorithm to solve a non-linear problem by mapping the original non-linear observations into a higher-dimensional space, where the linear classifier is subsequently used; this makes a linear classification in the new space equivalent to non-linear classification in the original space. This is done using Mercer's theorem, which states that any continuous, symmetric, positive semi-definite kernel function K(x, y) can be expressed as a dot product in a high-dimensional space. More specifically, if the arguments to the kernel are in a measurable space X, and if the kernel is positive semi-definite then there exists a function φ(x) whose range is in an inner product space of possibly high dimension, such that

$$K(x, y) = \varphi(x)\varphi(y)$$

The kernel trick transforms any algorithm that solely depends on the dot product between two vectors. Wherever a dot product is used, it is replaced with the kernel function. Thus, a linear algorithm can easily be transformed into a non-linear algorithm. This non-linear algorithm is equivalent to the linear algorithm operating in the range space of φ. However, because kernels are used, the φ function is never explicitly computed. This is desirable, because the high-dimensional space may be infinite-dimensional (as is the case when the kernel is a Gaussian).

By assuming non linearity, k-PCA seeks a non linear combination of the original basis that best re-express a given dataset. The first component, s_1 , is the component with the largest variance. We have $s_1 = \boldsymbol{x}^T \boldsymbol{w_1}$, where the p-dimensional coefficient vector $\boldsymbol{w}=[w_1\ w_2\ \ldots\ w_p]$ *solves*

$$w_1 = argmax_{||w=1||} Var\{\varphi(\boldsymbol{x})^T \varphi(\boldsymbol{w})\} = argmax_{||w=1||} Var\{K(\boldsymbol{x}, \boldsymbol{w})\}$$

Assuming a standardized data with the empirical covariance matrix

$$\Sigma_{pxp}{}' = \frac{1}{n-1} \varphi(\boldsymbol{X}) \varphi(\boldsymbol{X})^T,$$

we must - like in the linear case – diagonise the covariance matrix. If λ, **V** eigenvalues and eigenvectors of Σ'_{pxp} then :

$$\lambda \boldsymbol{V} = \Sigma' \boldsymbol{V} \overset{\boldsymbol{KM}}{\Longrightarrow}$$

$$\begin{cases} \lambda(\varphi(x_k)\boldsymbol{V}) = \varphi(x_k)\Sigma' \boldsymbol{V}, for\ k = 1, \dots p \\ \boldsymbol{V} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \alpha_i \varphi(x_i) = \boldsymbol{a\varphi} \end{cases}$$

$$\Longrightarrow \lambda \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \alpha_i \varphi(x_i)\, \varphi(x_k) = \frac{1}{n-1} \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \alpha_i (\varphi(x_k) \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \varphi(x_j))(\varphi(x_j)\varphi(x_i)) \overset{\boldsymbol{Kernel\ trick}}{\Longrightarrow}$$

$$\Rightarrow (n-1)\lambda \boldsymbol{K\alpha} = \boldsymbol{K^2 \alpha} \Rightarrow (n-1)\lambda \boldsymbol{\alpha} = \boldsymbol{K\alpha}\ (\boldsymbol{K}\ symmetric)$$

By definition of kernels ,K is positive semi definite. K's eigenvalues will be nonnegative and will give the solutions (n-1)λ of $(n-1)\lambda \boldsymbol{\alpha} = \boldsymbol{K\alpha}$**.** We therefore only need to diagonalized **K.** Let $\lambda_1 < \lambda_2 < \dots < \lambda_p$ denote the Eigenvalues and $\boldsymbol{a^1}, \boldsymbol{a^2}, \dots, \boldsymbol{a^p}$ the corresponding complete set of eigenvectors. We normalize $\boldsymbol{a^1}, \boldsymbol{a^2}, \dots, \boldsymbol{a^p}$ by requiring the corresponding vectors in our Hilbert space to be normalized: $\boldsymbol{V^k V^k} = 1 \Rightarrow \lambda_k(\boldsymbol{\alpha}^\kappa \boldsymbol{\alpha}^\kappa) = 1.$

For the purpose of principal component extraction, we need to compute projections on the eigenvectors $\boldsymbol{V^k}$**.** Let **x** be a test point:

$$kPC_k(\boldsymbol{x}) = \boldsymbol{V}^k \varphi(\boldsymbol{x}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} a_{\boldsymbol{i}}^{\boldsymbol{k}}\ \varphi(\boldsymbol{x_i})\varphi(\boldsymbol{x}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} a_{\boldsymbol{i}}^{\boldsymbol{k}}\ K(\boldsymbol{x_i}, \boldsymbol{x})$$

3.3 Unsupervised classification

An unsupervised classification method is used to determine the number of spectrally-separable groups or clusters in an image for which there is insufficient reference information available. These unsupervised methods can be viewed as techniques of identifying natural groups, or structures, within multispectral image data. While applying an unsupervised method, the analyst generally specifies only the number of classes (or the upper and lower bound on the number of classes) and some statistical measure, depending upon the type of clustering

algorithms used. These methods generate the specified number of clusters in feature space, and the user assigns these clusters (spectral classes) to information classes depending on his or her knowledge of the area. Determination of the clusters is performed by estimating the distances or comparison of the variance within and between the clusters. These automated classification methods are expected to delineate (or extract) those land cover features that are desired by the analyst. After the specified number of groups is determined, they are labeled by allocating pixels to land cover features present in the scene. However, some groups may be inappropriate since they represent either irrelevant categories for the purpose of the study or else they are mixed classes. Therefore, the spectral characteristics of the area of interest should be sufficiently well known to the analyst to allow him to correctly label the clusters representing actual features.

Unsupervised classification techniques generally require user interaction in specifying the number of groups to be recognized and in labeling the correctly identified areas with the individual feature (or class) label. Owing to the minimal amount of user involvement, they are usually considered as automated procedures. Clustering has been used for several decades in various fields for grouping data. There are numerous clustering algorithms that can be used to determine the natural spectral grouping present in the data set, each having its own characteristics. Some procedures iterate to a local minimum for the average distance from each pixel to the nearest cluster means. For example, the most popular clustering algorithms used in remote sensing image classification are k-means [15], fuzzy C-means [16] and isodata [17], statistical clustering methods, and the SOM (self organising feature maps), an unsupervised neural classification method.

3.3.1 *K-means*

The **k-means algorithm** is an algorithm to cluster n objects based on attributes into k partitions, k < n. 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. It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are k clusters S_i, i = 1, 2, ..., k, and μ_i is the centroid or mean point of all the points $x_j \in S_i$.

The algorithm is composed of the following steps:

1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*

2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

A drawback of the k-means algorithm is that the number of clusters *k* is an input parameter. An inappropriate choice of *k* may yield poor results. The algorithm also assumes that the variance is an appropriate measure of cluster scatter.

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading. Figure shown below illustrates this with an example of the width and height measurements of an object. Despite both measurements being taken in the same physical units, an informed decision has to be made as to the relative scaling. As the figure shows, different scalings can lead to different clusterings.

Suppose that we have n sample feature vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ all from the same class, and we know that they fall into k compact clusters, k < n. Let $\mathbf{m}_i$ be the mean of the vectors in cluster i. If the clusters are well separated, we can use a minimum-distance classifier to separate them. That is, we can say that **x** is in cluster i if $|| \mathbf{x} - \mathbf{m}_i ||$ is the minimum of all the k distances. This suggests the following procedure for finding the k means:

- Make initial guesses for the means $m_1, m_2, ..., m_k$
- Until there are no changes in any mean
    - Use the estimated means to classify the samples into clusters
    - For i from 1 to k
        - Replace $\mathbf{m}_i$ with the mean of all of the samples for cluster i
    - end_for
- end_until

Here is an example (figure 3.2) showing how the means $\mathbf{m}_1$ and $\mathbf{m}_2$ move into the centers of two clusters:

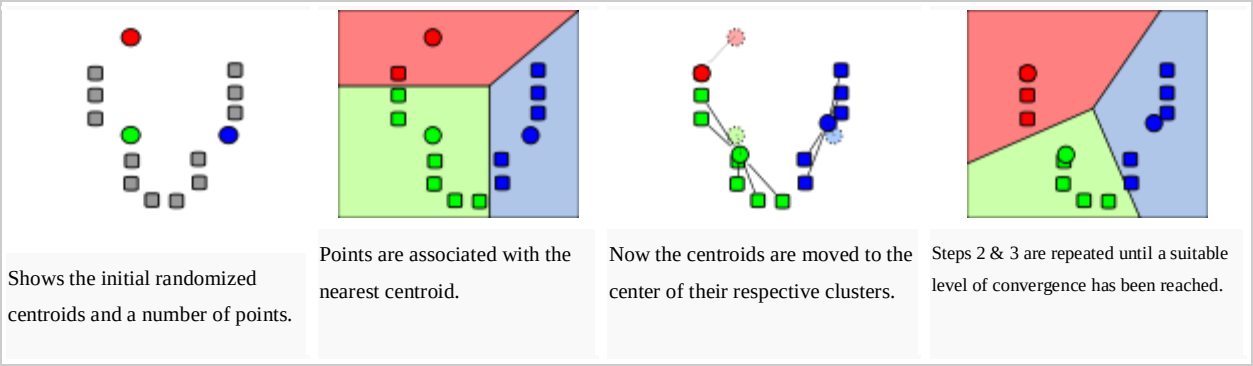


figure 3.2: k-means clustering process

3.3.2 Fuzzy *C*-Means

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m. The algorithm is composed of the following steps:

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*

2. *At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

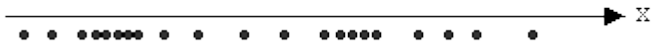
3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

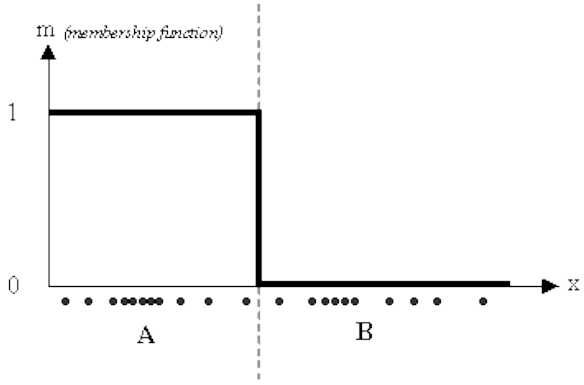
4. *If $|| U^{(k+1)} - U^{(k)}|| < \varepsilon$ then STOP; otherwise return to step 2.*

As already told, data are bound to each cluster by means of a Membership Function, which represents the fuzzy behavior of this algorithm. To do that, we simply have to build an appropriate matrix named U whose factors are numbers between 0 and 1, and represent the degree of membership between data and centers of clusters.

For a better understanding, we may consider this simple mono-dimensional example. Given a certain data set, suppose to represent it as distributed on an axis. The figure below shows this:

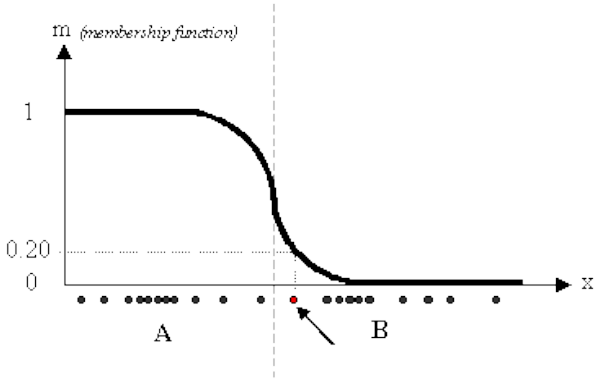


Looking at the picture, we may identify two clusters in proximity of the two data concentrations. We will refer to them using ‘A’ and ‘B’.



In the FCM approach, instead, the same given datum does not belong exclusively to a well defined cluster, but it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate that every

datum may belong to several clusters with different values of the membership coefficient.



In the figure on the left, the datum shown as a red marked spot belongs more to the B cluster rather than the A cluster. The value 0.2 of 'm' indicates the degree of membership to A for such datum. Now, instead of using a graphical representation, we introduce a matrix U whose factors are the ones taken from the membership functions:

$$U_{MC} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \end{bmatrix} \quad U_{MC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

The number of rows and columns depends on how many data and clusters we are considering. More exactly we have C = 2 columns (C = 2 clusters) and N rows, where C is the total number of clusters and N is the total number of data. We can notice that in the first case the coefficients are always unitary. It is so to indicate the fact that each datum can belong only to one cluster.

Other properties are shown below:

- $u_{ij} \in [0,1] \qquad \forall i, j$
- $\sum_{j=1}^{C} u_{ik} = 1 \qquad \forall i$
- $0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall N$

3.3.3 *Isodata*

Isodata stands for *Iterative Self-Organizing Data Analysis Techniques*. This is a more sophisticated algorithm which allows the number of clusters to be automatically adjusted during the iteration by merging similar clusters and splitting clusters with large standard deviations. In the migrating means (or ISODATA, or nearest mean) algorithm (Ball and Hall, 1965), the value of the function to be minimized is the average

Euclidean distance between each sample point and the corresponding cluster mean. Intuitively, this is equivalent to generating spherical clusters with small variances or scatter. There is no analytical method for generating clusters that minimizes the value of this function. There are a number of different forms of this algorithm, but in all of them at least two parameters must be specified by the user: the number of clusters and the maximum number of iterations. The latter parameter ensures the method will terminate if convergence is not achieved:

We first define the following parameters:

1. K = number of clusters desired;
2. I = maximum number of iterations allowed;
3. P = maximum number of pairs of cluster which can be merged;
4. Θ_N = a threshold value for minimum number of samples in each cluster can have (used for discarding clusters);
5. Θ_S = a threshold value for standard deviation (used for split operation);
6. Θ_C = a threshold value for pairwise distances (used for merge operation).

The algorithm:

- Step 1. Arbitrarily choose k (not necessarily equal to K) initial cluster centers: $m_1, m_2, \dots, m_k$ from the data set $\{x_i,\ i = 1,2\ , \dots , N\}$.
- Step 2. Assign each of the N samples to the closest cluster center:

$$\boldsymbol{x} \in \omega_j \ if\ D_L(\boldsymbol{x}, \boldsymbol{m_j}) = \max\{D_L(\boldsymbol{x}, \boldsymbol{m_i}), \qquad i = 1,2\ , \dots , k\}$$

- Step 3. Discard clusters with fewer than Θ_N members, i.e., if for any j, $N_j < \Theta_N$, then discard ω_j and κ <-- κ-1.
- Step 4. Update each cluster center: $\boldsymbol{m_j} = \frac{\mathbf{1}}{\boldsymbol{N_j}} \sum_{x \in \boldsymbol{\omega_j}} \boldsymbol{x}\ (j = 1,2\ , \dots , k)$
- Step 5. Compute the average distance D_j of samples in cluster $\boldsymbol{\omega_j}$ from their corresponding cluster center: $D_j = \frac{1}{N_j} \sum_{x \in \boldsymbol{\omega_j}} D_L(\boldsymbol{x}, \boldsymbol{m_j})\ (j = 1,2\ , \dots , k)$

- Step 6. Compute the overall average distance of the samples from their respective cluster centers: $D = \frac{1}{N}\sum_{j=1}^{k} N_j D_j$
- Step 7. If $k \leq \frac{k}{2}$ (too few clusters), go to Step 8; else if $k > 2k$ (too many clusters), go to Step 11; else go to Step 14.

(Steps 8 through 10 are for split operation, Steps 11 through 13 are for merge operation.)

- Step 8. First step to split. Find the standard deviation vector $\sigma_j = [\sigma_1{}^{(j)}, \dots, \sigma_N{}^{(j)}]^T$ for each cluster: $\sigma_i{}^{(j)} = \sqrt{\frac{1}{N_j}\sum_{x \in \omega_j}(x_i - m_i{}^{(j)})^2}$, $(i = 1,2,\dots,n,\ j = 1,2,\dots,k)$,

  where $m_i{}^{(j)}$ is the ith component of $\boldsymbol{m_j}$ and σ_i is the standard deviation of the samples in ω_j along the ith coordinate axis. N_j is the number of samples in ω_j.

- Step 9. Find the maximum component of each Σ_j and denote it by $\sigma_{max}{}^{(j)}$; Do this for $j = 1,2,\dots,k$
- Step 10. If for any $\sigma_{max}{}^{(j)}$, $j = 1,2,\dots,k$
  - $\sigma_{max}^{(j)} > \Theta_S$,
  - $D_j > D$,
  - $N_j > 2\Theta_N$
  - t**hen** *split* $\boldsymbol{m_j}$ into two new cluster centers $\boldsymbol{m_j}^+$ and $\boldsymbol{m_j}^-$ by adding +δ, -δ to the component of $\boldsymbol{m_j}$ corresponding to $\sigma_{max}{}^{(j)}$, where δ can be $\alpha\sigma_{max}{}^{(j)}$, for some $\alpha > 0$,. Then delete $\boldsymbol{m_j}$ and let κ ← κ+ 1. Goto Step 2

    else Go to Step 14.

- Step 11. First step to merge. Compute the pair wise distances D_{ij} between every two cluster centers: $D_{ij} = D_L(\boldsymbol{m_i}, \boldsymbol{m_j})\ for\ all\ i \neq j$ and arrange $\frac{k(k-1)}{2}$ of these distances in ascending order.
- Step 12. Find no more than P smallest D_{ij}'s which are also smaller than Θ_C and keep them in ascending order:

$$D_{i_1 j_1} \leq D_{i_2 j_2} \leq \cdots \leq D_{i_P j_p}$$

- Step 13. Perform *pair wise merge*: for $l = 1,2, \ldots P$ do the following:
- If neither of $\boldsymbol{m}_{jl}$ nor $\boldsymbol{m}_{il}$ has been used in this iteration, then merge them to form a new center: $\boldsymbol{m} = \frac{1}{N_{jl}+N_{il}}[N_{jl}\boldsymbol{m}_{jl} + N_{il}\boldsymbol{m}_{il}]$. Delete $\boldsymbol{m}_{il}$ and $\boldsymbol{m}_{jl}$, and let k ← k - 1 and go to Step 2.
- Step 14. Terminate if maximum number of iterations I is reached. Otherwise go to Step 2.

The Isodata algorithm is more flexible than the K-mean method. But the user has to choose empirically many more parameters listed previously.

3.4 Supervised classification

Supervised classification methods are most commonly used in hyperspectral sensing and based on the knowledge of the area to be classified. "These methods are often central to the image analysis process, since these concerns the direct transformation from pixel counts to thematic map" (Wilkinson, 2000). Supervised classification may be defined as the process of identifying unknown objects by using the spectral information derived from training data provided by the analyst. The result of the identification is the assignment of unknown pixels to pre-defined categories. The main difference between the unsupervised and supervised classification approaches is that supervised classification requires training data. The analyst locates specific sites in the remotely sensed image that represent homogeneous examples of known land cover types. These areas are commonly referred to as training sites because the spectral characteristics of these known areas are used to train the classifier. The training data thus extracted is used to find the properties of each individual class. The training data are generally derived from fieldwork, analysis of aerial photographs, from the study of appropriate maps, or from personal experience.

Supervised classification is performed in two stages; the first stage is the training of the classifier, and the second stage is testing the performance of the trained classifier on unknown pixels. In the training stage, the analyst defines the regions that will be used to extract training data, from which statistical estimates of the data properties are computed. At the classification stage, every unknown pixel in the test image is labeled in terms of its spectral similarity to specified land cover features. If a pixel is not spectrally similar to any of the classes, then it can be allocated to an unknown class. As a result, an output image, or thematic map is produced, showing every pixel with a class label. The characteristics of the training data selected by the analyst have a considerable effect on the reliability and the performance of a supervised classification process. The training data must be defined by the analyst in such a way that they accurately represent the characteristics of each individual feature and class used in the analysis. Two features of the training data are of key importance. One is that data must represent the range of variability within class and the other is that the size of the training data set should be sufficient. In order to have a representative set of data, the pixels should be so selected that they correctly

represent the spectral diversity of each class. Pixels should be selected from each of the fields to include all spectral classes. The best sampling strategy is to select training pixels randomly from the whole test image. Unfortunately, this is generally not possible in practice, as data for the whole area are generally not available.

The size of the training data set is also very important in supervised classification, if statistical estimates are to be reliable. Sample size is mainly related to the number of features whose statistical properties are to be estimated. Typically, it is recommended that the minimum training set size is some 10-30 times the number of wave bands per class being used for classification (Mather, 1999; Piper, 1992). Generally, a large training set is required for mapping from multispectral data sets. Supervised classification methods require more user interaction, especially in the collection of training data. The accuracy of supervised classification is determined partly by the quality of the ground truth data and partly by how well the set of ground truth pixels are representative of the full image. In order to measure the accuracy, it is common practice to use only part of the ground truth data for training the classifier and to use the remaining pixels for testing, that is to see if the classifier output corresponds to reality.

3.4.1 Parametric Classifiers

Parametric approaches to classification make use of a parameterized model of the classes in the spectral feature space. These are generally more powerful than non-parametric methods and lead to higher overall classification accuracy if the data used satisfy the requirements of the model. The maximum likelihood method is the most common parametric approach. This procedure models classes according to the frequency distributions of the training pixels. Most often classes are modeled by using the multivariate form of the normal probability density function. Pixels are then classified by assigning them to the class to which they have the highest statistical likelihood of belonging.

3.4.1.1 Bayesian Classifier

Bayesian Decision Theory shows us how to design an optimal classifier if we know the prior probabilities $P(w_i)$ and the class-conditional densities $p(x|w_i)$. Unfortunately, we rarely have complete knowledge of the probabilistic structure. However, we can often find design samples or training data that include particular representatives of the patterns we want to classify. To simplify the problem, we can assume some parametric form for the conditional densities and estimate these parameters using training data. Then, we can use the resulting estimates as if they were the true values and perform classification using the Bayesian decision rule. Bayesian classification and decision making is based on probability theory and the principle of choosing the most probable or the lowest risk (expected cost) option [18].

Assume that there is a classification task to classify feature vectors (samples) to K different classes. A feature vector is denoted as $\mathbf{x} = [x\backslash, x_2, \cdots, x_D]^T$ where D is the dimension of a vector. The probability that a feature vector x belongs to class w_k is $P(w_k|x)$, and it is often referred to as a posteriori probability. The classification of the vector is done according to posterior probabilities or decision risks calculated from the probabilities. The posterior probabilities can be computed with the Bayes formula

$$P(\omega_k|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_k)P(\omega_k)}{P(\boldsymbol{x})},$$

,where $P(\boldsymbol{x}|\omega_k)$ is the probability density function of class ω_k in the feature space and $P(\omega_k|\boldsymbol{x})$ is the a priori probability, which tells the probability of the class before measuring any features. If prior probabilities are not actually known, they can be estimated by the class proportions in the training set. The divisor:

$$P(\boldsymbol{x}) = \sum_{i=1}^{K} P(\boldsymbol{x}|\omega_i)P(\omega_i)$$

is merely a scaling factor to assure that posterior probabilities are really probabilities, i.e., their sum is one. It can be shown that choosing the class of the highest posterior probability produces the minimum error probability [18,19]. However, if the cost of making different kinds of errors is not uniform, the decision can be made with a risk function that computes the expected cost using the posterior probabilities, and choose the class with minimum risk.

The major problem in the Bayesian classifier is the class-conditional probability density function $P(\boldsymbol{x}|\omega_k)$. The function tells the distribution of feature vectors in the feature space inside a particular class, i.e., it describes the class model. In practice it is always unknown, except in some artificial classification tasks. The distribution can be estimated from the training set with a range of methods.

3.4.1.2 Maximum likelihood (*multivariate Gaussian case, Gaussian mixture model case)*

The maximum likelihood method is a well known supervised classification algorithm that is based on the assumption that the probability density function for each class is normal (Gaussian) (Tou and Gonzalez, 1974). The normal distribution describes the probability of a single feature and it is specified by two parameters, the mean and the variance. The mean of the distribution controls the location of the distribution and the variance controls the spread of the data. When more than one feature is involved, then the multivariate generalization of the normal distribution has to be used, i.e. the multivariate normal distribution. Instead of a single mean controlling the location of the distribution there is now one mean for each feature making up a mean vector. The multivariate equivalent of the variance is the variance-covariance matrix, representing the variability of pixel values for each feature within a particular class and the correlations between the features. These two parameters are computed for

each sample, and they are used to describe each class. The maximum likelihood classifier generates estimates of both the variance-covariance matrix and mean of the category spectral response patterns during the classifier training process. These estimates are derived by selecting samples that represent each class to be recognized from the total population to be classified.

The assumption of normality is generally reasonable for common spectral response distributions. Under this assumption, the distribution of a class response pattern can be completely described by the mean vector and the covariance matrix. With these parameters, it is possible to compute the statistical probability of a given pixel being a member of a particular land cover class. The pixel is assigned to the class for which the probability of membership is the highest. Although in practice the assumption of "normally distributed" data is not generally met, the classifier generally outputs an acceptable result.

Assume that there is a set of independent samples X = {x_1,..., $\mathbf{x_N}$} drawn from a single distribution described by a probability density function p*(X;* **θ***)* where **θ** is the PDF parameter list. The likelihood function

$$\mathcal{L}(\mathrm{X};\boldsymbol{\theta}) = \prod_{\mathbf{n=1}}^{\mathbf{N}} \mathrm{p}(\mathbf{x}_\mathrm{n};\boldsymbol{\theta})$$

tells the likelihood of the data X given the distribution or, more specifically, given the distribution parameters **θ**. The goal is to find **θ** that maximizes the likelihood:

$$\hat{\theta} = argmax_\theta \mathcal{L}(\mathrm{X};\boldsymbol{\theta})$$

Usually this function is not maximized directly but the logarithm

$$L(X;\boldsymbol{\theta}) = ln\mathcal{L}(\mathrm{X};\boldsymbol{\theta}) = \sum_{\mathbf{n}=1}^{\mathrm{N}} \ln \mathrm{p}(\mathbf{x}_\mathrm{n};\boldsymbol{\theta})$$

called the log-likelihood function which is analytically easier to handle. Because of the monotonicity of the logarithm function the solution to $argmax_\theta \mathcal{L}(\mathrm{X};\boldsymbol{\theta})$ is the same using $\mathcal{L}(\mathrm{X};\boldsymbol{\theta})$ or $L(X;\boldsymbol{\theta})$ [18].

- *Maximizing the log likelihood function for multivariate Gaussian case (figure 3.3)* :

The Gaussian probability density function in one dimension is a bell shaped curve defined by two parameters, mean μ and variance σ^2. In the D-dimensional space it is defined in a matrix form as

$$N(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma(\boldsymbol{x}-\boldsymbol{\mu})]$$

,where μ is the mean vector and Σ the covariance matrix.

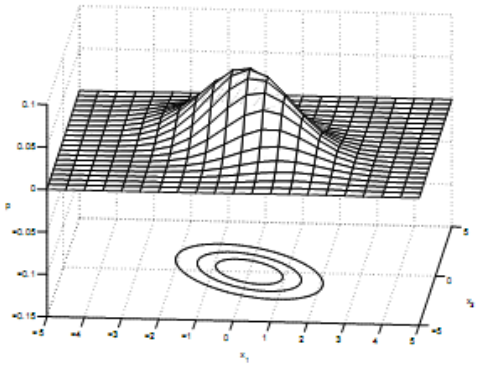


figure 3.3: An example surface of two-dimensional Gaussian PDF with $\mu = [0; 0]$ and $\Sigma = [1.56, -0.97; -0.97, 2.68]$ and contour plots (equiprobability surfaces) to emphasize the shape.

The Gaussian distribution is usually quite good approximation for a class model shape in a suitably selected feature space. It is a mathematically sound function and extends easily to multiple dimensions. In the Gaussian distribution lies an assumption that the class model is truly a model of one basic class. If the actual model ,the actual probability density function, is multimodal, it fails. For example, if we are searching for different face parts from a picture and there are several basic types of eyes , because of people from different races perhaps, the single Gaussian approximation would describe a wide mixture of all eye types, including patterns that might not look like an eye at all.

Depending on p(x; θ) it might be possible to find the maximum analytically by setting the derivatives of the log-likelihood function to zero and solving θ. It can be done for a Gaussian PDF, which leads to the well-known intuitive estimates for a mean and variance [18], but usually the analytical approach is intractable. In the case of a Gaussian pdf we can use the estimated μ and Σ to classify test vectors with log version of pdf: $-2\ln p(x) = +\frac{1}{2}\ln|\boldsymbol{\Sigma}_{\theta_i}| + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\theta_i})^T \Sigma_{\theta_i}{}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\theta_i})$. An observation vector will be assigned to the class i for which the value $-2\ln p(x)$ is the smallest. The reliability of the results obtained with this classifier declines when the frequency distribution of the data departs from normality, especially when the distribution is bimodal. In extreme cases, where the multivariate normal assumption does not properly describe the data distribution in feature space, the results can be misleading. The other drawback of this method is the computational cost required to classify each pixel. This is particularly important in circumstances where data to be classified are measured in a large number of spectral bands, or include many spectral classes to be discriminated. The reliability of the estimates of mean vector and variance-covariance matrix, which are fundamental to the calculation of the likelihood, is affected by the relationship between sample size and the number of features. It should also be noted that all features are used to discriminate between classes, rather than the minimum effective set. It is not possible

to use categorical data with this classifier as the classifier assumes that the data forming each class are normally distributed. The maximum likelihood classification method is available in almost all remote sensing and image processing software packages, and it is generally used as the standard supervised classification method.

- *Maximizing the log likelihood function for Gaussian mixture model case (figure 3.4):*

Gaussian mixture model (GMM) is a mixture of several Gaussian distributions and can therefore represent different subclasses inside one class. The probability density function is defined as a weighted sum of Gaussians

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{c=1}^{C} a_c N(\boldsymbol{x}, \boldsymbol{\mu_c}, \boldsymbol{\Sigma_c})$$

where a_c is the weight of the component c, $0 < a_c < 1$ for all components, and $\sum_{c=1}^{C} a_c = 1$. The parameter list $\boldsymbol{\theta} = \{a_1, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}, \ldots, a_c, \boldsymbol{\mu_c}, \boldsymbol{\Sigma_c}\}$ defines a particular Gaussian mixture probability density function.

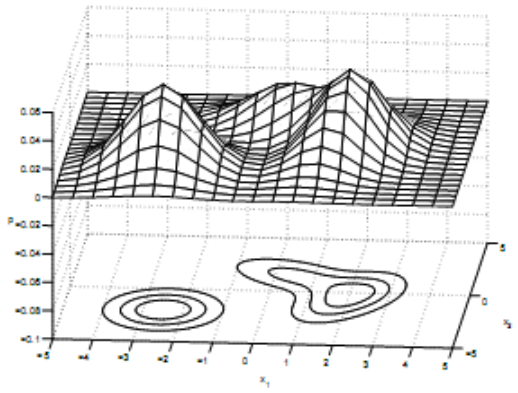


figure 3.4 An example surface of a two-dimensional Gaussian mixture PDF with three components: $\alpha_1 = 0.40$, $\boldsymbol{\mu}_1 = [-2.5; -2]$, $\Sigma_1 = [0.81, 0; 0, 1.44]$, $\alpha_2 = 0.25$, $\boldsymbol{\mu}_2 = [0.5; 1.5]$, $\Sigma_2 = [1.30, -0.66; -0.66, 1.30]$ and $\alpha_3 = 0.35$, $\boldsymbol{\mu}_3 = [2.0; -0.5]$, $\Sigma_3 = [0.69, 0.61; 0.61, 2.36]$.

Estimation of the Gaussian mixture parameters for one class can be considered as un-supervised learning of the case where samples are generated by individual components of the mixture distribution and without the knowledge of which sample was generated by which component. Clustering usually tries to identify the exact components, but Gaussian mixtures can also be used as an approximation of an arbitrary distribution. For Gaussian mixture PDF the analytical approach of estimating parameters is intractable. In practice an iterative method such as the expectation maximization (EM) algorithm is used. Maximizing the likelihood may in some cases lead to singular estimates, which is the fundamental problem of maximum likelihood methods with Gaussian mixture models[22].

The expectation maximization (EM) algorithm is an iterative method for calculating maximum likelihood distribution parameter estimates from incomplete data (some elements missing in some feature vectors). It can

also be used to handle cases where an analytical approach for maximum likelihood estimation is infeasible, such as Gaussian mixtures with unknown and unrestricted covariance matrices and means.

Assume that each training sample contains known features and missing or unknown features. Mark all good features of all samples with X and all unknown features of all samples with Y. The expectation step (E-step) for the EM algorithm is to form the function

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta^i}) \triangleq E_Y[ln\mathcal{L}(X, Y; \boldsymbol{\theta})|X; \boldsymbol{\theta^i}]$$

where $\boldsymbol{\theta^i}$ is the previous estimate for the distribution parameters and θ is the variable for a new estimate describing the (full) distribution. The function calculates the likelihood of the data, including the unknown feature Y marginalized with respect to the current estimate of the distribution described by $\boldsymbol{\theta^i}$. The maximization step (M-step) is to maximize $Q(\boldsymbol{\theta}; \boldsymbol{\theta^i})$ with respect to $\boldsymbol{\theta^i}$ and set

$$\boldsymbol{\theta^{i+1}} \underset{\mathbf{i+1}}{\leftarrow} \text{argmax}_\theta Q(\boldsymbol{\theta}; \boldsymbol{\theta^i})$$

The steps are repeated until a convergence criterion is met [19]. For the convergence criterion it is suggested in [18] that

$$Q(\boldsymbol{\theta^{i+1}}; \boldsymbol{\theta^i}) - Q(\boldsymbol{\theta^i}; \boldsymbol{\theta^{i-1}}) \leq T$$

with a suitably selected T and in [19] that

$$\|\boldsymbol{\theta^{i+1}} - \boldsymbol{\theta^i}\| \leq \varepsilon$$

for an appropriately chosen vector norm and ε . Common for both of these criteria is that iterations are stopped when the change in the values falls below a threshold. A more sophisticated criterion can be derived by using a relative rather than absolute rate of change.

The EM algorithm starts from an initial guess $\boldsymbol{\theta}^\circ$ for the distribution parameters and the log-likelihood is guaranteed to increase on each iteration until it converges. The convergence leads to a local or global maximum, but it can also lead to singular estimates, which is true particularly for Gaussian mixture distributions with arbitrary covariance matrices. The description of the general EM algorithm and also its application for the Gaussian mixture model can be found in [18,19,22].

The initialization is one of the problems of the EM algorithm. The selection of $\boldsymbol{\theta}^{\circ}$ (partly) determines where the algorithm converges or hits the boundary of the parameter space producing singular, meaningless results. Some solutions use multiple random starts or a clustering algorithm for initialization. [23]

The application of the EM algorithm to Gaussian mixtures according to [23] goes as follows. The known data X is interpreted as incomplete data. The missing part Y is the knowledge of which component produced each sample x_n. For each $\mathbf{x_n}$ there is a binary vector $\mathbf{y_n} = (y_{n,1}, \dots, y_{n,c}\}$, where $y_{n,c} = 1$, if the sample was produced by the component c, or zero otherwise. The complete data log-likelihood is

$$ln\mathcal{L}(\mathrm{X},\mathrm{Y};\boldsymbol{\theta}) = \sum_{\mathrm{n}=1}^{\mathbf{N}}\sum_{\mathbf{c}=\mathbf{1}}^{\mathbf{C}} \mathrm{y}_{\mathrm{n,c}} \ln(a_{\boldsymbol{c}} p(\boldsymbol{x}_n|c;\boldsymbol{\theta}))$$

The E-step is to compute the conditional expectation of the complete data log-likelihood, the Q-function, given X and the current estimate of the parameters. Since the complete data log-likelihood $ln\mathcal{L}(\mathrm{X},\mathrm{Y};\boldsymbol{\theta})$ is linear with respect to the missing Y, the conditional expectation $\mathrm{W}=\mathrm{E}[\mathrm{Y}|\mathrm{X},\boldsymbol{\theta}]$ has simply to be computed and put it into $ln\mathcal{L}(\mathrm{X},\mathrm{Y};\boldsymbol{\theta})$.

Therefore

$$\mathrm{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{\mathbf{i}}\right) \triangleq \mathrm{E}\left[ln\mathcal{L}(\mathrm{X},\mathrm{Y};\boldsymbol{\theta})\middle|\mathrm{X},\boldsymbol{\theta}^{\mathbf{i}}\right] = \ln(\mathrm{X},\mathrm{W};\boldsymbol{\theta})$$

where the elements of W are defined as

$$\mathrm{w}_{\mathrm{n,c}} \triangleq \mathrm{E}\left[\mathrm{y}_{\mathrm{n,c}}\middle|\mathrm{X},\boldsymbol{\theta}^{\mathbf{i}}\right] = \mathrm{P}[\mathrm{y}_{\mathrm{n,c}} = 1|\mathrm{x}_{\mathrm{n,c}},\boldsymbol{\theta}^{\mathbf{i}}]$$

The probability can be calculated with the Bayes law

$$\mathrm{w}_{\mathrm{n,c}} = \frac{a_{\boldsymbol{c}}{}^{\boldsymbol{i}} p(\boldsymbol{x}_n|c;\boldsymbol{\theta}^{\boldsymbol{i}})}{\sum_{\mathrm{j}=1}^{\mathrm{C}} a_{\boldsymbol{j}}{}^{\boldsymbol{i}} p(\boldsymbol{x}_n|j;\boldsymbol{\theta}^{\boldsymbol{i}})} \qquad (1)$$

where $a_{\boldsymbol{c}}{}^{\boldsymbol{i}}$ is the a priori probability (of estimate $\boldsymbol{\theta}^{\mathbf{i}}$) and $\mathrm{w}_{\mathrm{n,c}}$ is the a posteriori probability that $\mathrm{y}_{\mathrm{n,c}} = 1$ after observing $\mathbf{x_n}$. In other words , $\mathrm{w}_{\mathrm{n,c}}$ is the probability that $\mathbf{x_n}$ was produced by component c.

Applying the M-step to the problem of estimating the distribution parameters for C-component Gaussian mixture with arbitrary covariance matrices, the resulting iteration formulas are as follows:

$$a_{\boldsymbol{c}}{}^{\boldsymbol{i+1}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{w}_{\mathrm{n,c}} \qquad (2)$$

$$\mu_c^{i+1} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n \, \mathrm{w}_{\mathrm{n,c}}}{\sum_{n=1}^{N} \mathrm{w}_{\mathrm{n,c}}} \qquad (3)$$

$$\Sigma_c^{i+1} = \frac{\sum_{n=1}^{N} \mathrm{w}_{\mathrm{n,c}} (\boldsymbol{x}_n - \mu_c^{i+1})(\boldsymbol{x}_n - \mu_c^{i+1})^{\mathrm{T}}}{\sum_{n=1}^{N} \mathrm{w}_{\mathrm{n,c}}} \quad (4)$$

The new estimates are gathered to θ^{i+1}. If the convergence criterion is not satisfied, i ← i + 1 and Eqs. 1-4 are evaluated again with new estimates. [18]

The interpretation of the Eqs. 2-4 is actually quite intuitive. The weight a_c of a component is the portion of samples belonging to that component. It is computed by approximating the component-conditional PDF with the previous parameter estimates and taking the posterior probability of each sample point belonging to the component c (Eq. 1). The component mean $\boldsymbol{\mu_0}$ and covariance matrix E_c are estimated in the same way. The samples are weighted with their probabilities of belonging to the component, and then the sample mean and sample covariance matrix are computed.

It is worthwhile to note that so far the number of components C was assumed to be known. Clustering techniques try to find the true clusters and components from a training set, but our task of training a classifier only needs a good enough approximation of the distribution of each class. Therefore, C does not need to be guessed accurately, it is just a parameter defining the complexity of the approximating distribution. Too small C prevents the classifier from learning the sample distributions well enough and too large C may lead to an overfitted classifier. More importantly, too large C will definitely lead to singularities when the amount of training data becomes insufficient.

3.4.2 Non Parametric Classifiers

The simplest forms of classifier rely on non-parametric methods, because these algorithms make no assumptions about the probability distribution of the data, and are often considered robust because they may work well for a wide variety of class distributions, as long as the class signatures are reasonably distinct. A wide variety of non-parametric spectral classifiers is available. These consist of statistical methods such as the parallelepiped or box classifier, the minimum distance classifier, and non-statistical methods such as the neural network, support vector machines, and decision tree classifiers.

3.4.2.1 *The Minimum Distance classifier*

Minimum distance classifier is a simple non-parametric classification method, which uses the minimum distance between the pixel and the centroid or the most representative spectra of the training class. This

classification method uses different kind of distance metrics in multidimensional feature space to measure the degree of dissimilarity between pixels and class centroids computed from training data. The pixel is assigned to the least dissimilar class centroid. Like the parallelepiped classifier, this algorithm does not take all the training data into consideration. For example, in order to assign a pixel to a specified class, Euclidean distances are calculated for each reference vector (or mean), and then the minimum value, i.e. the shortest distance, is determined. As a result, the pixel is allocated to the class that is the closest in terms of the estimated multidimensional Euclidean distance.

This type of classifier is mathematically simple and computationally efficient, but has certain limitations. Most importantly, it is sensitive to different degrees of variance in the spectral response data. Due to these problems, this classifier is not widely used in applications where spectral classes are close to one another in measurement space and have high variance. However, it can give results that are comparable to other statistical classifiers, such as the maximum likelihood classifier in cases where the classes are well defined in feature space.

Each classifier uses different functions to calculate distance. Such a distance function is known as a metric. Together with the set, it makes up a metric space. In mathematics, a metric space is a set where a notion of distance (called a metric) between elements of the set is defined. The metric space which most closely corresponds to our intuitive understanding of space is the 3-dimensional Euclidean space. In fact, the notion of "metric" is a generalization of the Euclidean metric arising from the four long known properties of the Euclidean distance. The Euclidean metric defines the distance between two points as the length of the straight line connecting them. More generally:

A metric on a set Ω *is a function called the distance function or simply distance* $d: \Omega \times \Omega \rightarrow R$ *(where R is the set of real numbers).*

$\forall\, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in R$, *d(.,.) is required to satisfy the following conditions:*

- $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ *(equality holds if* $x = y$*)*
- $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$
- $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$

These conditions express intuitive notions about the concept of distance. For example distance function between distinct points is positive and the distance operator should yield the same value independent of the order of the operands. The triangle inequality means that the distance traversed directly between x and z, is not larger than the distance to traverse in going first from x to y, and then from y to z.

a) *Spectral Angle mapper & Euclidean distance*

a) *Spectral Angle mapper & Euclidean distance*

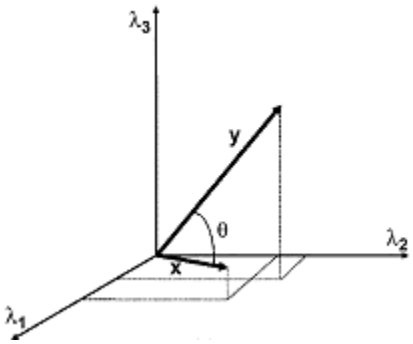


figure 3.5 . Vector representation of two spectra for three different λ.

Let $\boldsymbol{x}, \boldsymbol{y} \in R^M$ ($\boldsymbol{x}, \boldsymbol{y}$ non negative, M-dimensional data). Spectral angle mapper [24] calculates the angle (figure 3.5) between spectra in a M-dimensional space using the concept of inner products:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\|\|\boldsymbol{y}\| \cos(\boldsymbol{x}, \boldsymbol{y}) \rightarrow$$

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = SAM(\boldsymbol{x}, \boldsymbol{y}) = \cos^{-1}\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\right), \qquad 0 \leq \theta \leq \frac{\pi}{2}$$

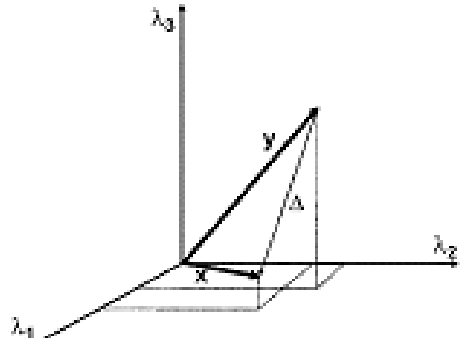


figure 3.6 . Vector representation of two spectra for three different λ.

Euclidean distance or Euclidean metric is the "ordinary" distance between two spectra that one would measure with a "ruler", which can be proven by repeated application of the Pythagorean theorem (figure 3.6). By using this formula as distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean norm.

Let $\boldsymbol{x}, \boldsymbol{y} \in R^M$ ($\boldsymbol{x}, \boldsymbol{y}$ non negative, M-dimensional data):

$$\Delta(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2}$$

1. Invariance to Multiplicative Scaling:

The angle measured by SAM is invariant to multiplication of and by scalars, because multiplication of a vector by a scalar simply increases its extent in a particular direction, but it does not alter the angle it creates with another vector. This invariance is useful for hyperspectral processing because atmospheric compensation algorithms are limited in their ability to estimate reflectances to within a multiplicative constant. Also certain light conditions in a lab driven experiment could favor bands (non uniform light source) by scaling overall spectra.

$$\theta(\alpha\boldsymbol{x}, \beta\boldsymbol{y}) = SAM(\alpha\boldsymbol{x}, \beta\boldsymbol{y})$$

$$= \cos^{-1}\left(\frac{\langle \alpha\boldsymbol{x}, \beta\boldsymbol{y} \rangle}{\|\alpha\boldsymbol{x}\|\|\beta\boldsymbol{y}\|}\right)$$

$$= \cos^{-1}\left(\frac{\alpha x_1 \beta y_1 + \alpha x_2 \beta y_2 + \ldots + \alpha x_M \beta y_M}{\sqrt{(\alpha x_1)^2 + (\alpha x_2)^2 + \ldots + (\alpha x_M)^2}\sqrt{(\beta y_1)^2 + (\beta y_2)^2 + \ldots + (\beta y_M)^2}}\right)$$

$$= \cos^{-1}\left(\frac{a\beta(x_1 y_1 + x_2 y_2 + \ldots + x_M y_M)}{|a||b|\sqrt{{x_1}^2 + {x_2}^2 + \ldots + {x_M}^2}\sqrt{{y_1}^2 + {y_2}^2 + \ldots + {y_M}^2}}\right) = \begin{cases} \cos^{-1}\left(\frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\right), & 0 \leq a\beta \\ \cos^{-1}\left(-\frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\right), & a\beta \leq 0 \end{cases}$$

$= SAM(\boldsymbol{x}, \boldsymbol{y})$

In contrast, distances measured by EMD depend on the multiplication of and by scalars, because multiplication of a vector by a scalar increases each of the $(x_i - y_i)^2$ terms :

$$\Delta(\alpha\boldsymbol{x}, \beta\boldsymbol{y}) = \|\alpha\boldsymbol{x} - \beta\boldsymbol{y}\| = \sqrt{\sum_{i=1}^{M}(\alpha x_i - \beta y_i)^2} \neq \Delta(\boldsymbol{x}, \boldsymbol{y})$$

Different variations of EMD can be used in hyperspectral analysis by normalizing the distance:

$$\Delta_*(\frac{\boldsymbol{x}}{\bar{\boldsymbol{x}}}, \frac{\boldsymbol{y}}{\bar{\boldsymbol{y}}}) = \left\|\frac{\boldsymbol{x}}{\bar{\boldsymbol{x}}} - \frac{\boldsymbol{y}}{\bar{\boldsymbol{y}}}\right\| = \sqrt{\sum_{i=1}^{M}(\frac{x_i}{\bar{\boldsymbol{x}}} - \frac{y_i}{\bar{\boldsymbol{y}}})^2}$$

$$\Delta_{**}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{M}\sqrt{\sum_{i=1}^{M}\frac{(x_i - y_i)^2}{(x_i^{max} - y_i^{min})^2}}$$

2. *Additivity*

Let $\boldsymbol{x}, \boldsymbol{y} \in R^M$ *expressed as partitions of vector elements :* $\boldsymbol{x} = [\boldsymbol{x_a}\ \boldsymbol{x_b}], \boldsymbol{y} = [\boldsymbol{y_a}\ \boldsymbol{y_b}]$, *where M = α + b and* $\boldsymbol{x_a}, \boldsymbol{y_a} \in R^{\alpha}$, $\boldsymbol{x_b}, \boldsymbol{y_b} \in R^{b}$. *A distance metric is* ***non additive*** *if*

$d(\boldsymbol{x}, \boldsymbol{y}) \neq d(\boldsymbol{x_a}, \boldsymbol{y_a}) + d(\boldsymbol{x_b}, \boldsymbol{y_b})$

SAM is a non additive distance function [24]:

$$cos\theta(\boldsymbol{x}, \boldsymbol{y}) = \cos\big(SAM(\boldsymbol{x}, \boldsymbol{y})\big)$$

$$= \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{\langle \boldsymbol{x_a}, \boldsymbol{y_a}\rangle + \langle \boldsymbol{x_b}, \boldsymbol{y_b}\rangle}{\sqrt{\|\boldsymbol{x}_a\|^2 + \|\boldsymbol{x}_b\|^2}\sqrt{\|\boldsymbol{y}_a\|^2 + \|\boldsymbol{y}_b\|^2}} = = \frac{1 + \frac{\langle \boldsymbol{x_b}, \boldsymbol{y_b}\rangle}{\langle \boldsymbol{x_a}, \boldsymbol{y_a}\rangle}}{\sqrt{1 + \frac{\|\boldsymbol{x}_b\|^2}{\|\boldsymbol{x}_a\|^2}}\sqrt{1 + \frac{\|\boldsymbol{y}_b\|^2}{\|\boldsymbol{y}_a\|^2}}} \frac{\langle \boldsymbol{x_a}, \boldsymbol{y_a}\rangle}{\|\boldsymbol{x}_a\|\|\boldsymbol{y}_a\|}$$

$$= \frac{1 + \frac{\langle \boldsymbol{x_b}, \boldsymbol{y_b}\rangle}{\langle \boldsymbol{x_a}, \boldsymbol{y_a}\rangle}}{\sqrt{1 + \frac{\|\boldsymbol{x}_b\|^2}{\|\boldsymbol{x}_a\|^2}}\sqrt{1 + \frac{\|\boldsymbol{y}_b\|^2}{\|\boldsymbol{y}_a\|^2}}}\ cos\theta_{\alpha} = \beta cos\theta_{\alpha}\big(= \beta' cos\theta_b\big),$$

where β,β' depend only in $\boldsymbol{x_a}, \boldsymbol{y_a}, \boldsymbol{x_b}, \boldsymbol{y_b}$.

$cos\theta(\boldsymbol{x},\boldsymbol{y}) = \beta cos\theta_\alpha \neq cos\theta_\alpha + cos\theta_b \rightarrow \theta(\boldsymbol{x},\boldsymbol{y}) \neq \theta_\alpha + \theta_b, (\theta = \theta(\boldsymbol{x},\boldsymbol{y}),\ \theta_\alpha = \theta_\alpha(\boldsymbol{x_a},\boldsymbol{y_a}),\ \theta_b = \theta_b(\boldsymbol{x_b},\boldsymbol{y_b})$

By observing $\beta = \frac{1+\frac{\langle \boldsymbol{x_b},\boldsymbol{y_b}\rangle}{\langle \boldsymbol{x_a},\boldsymbol{y_a}\rangle}}{\sqrt{1+\frac{\|x_b\|^2}{\|x_a\|^2}}\sqrt{1+\frac{\|y_b\|^2}{\|y_a\|^2}}} = cos\theta(\boldsymbol{x_a},\boldsymbol{x_b})cos\theta(\boldsymbol{y_a},\boldsymbol{y_b})(1+\frac{\langle \boldsymbol{x_b},\boldsymbol{y_b}\rangle}{\langle \boldsymbol{x_a},\boldsymbol{y_a}\rangle})$, the first two terms are always less than one and the third term is unconstrained (although for reflectance signals from the reflective regime, which must be positive, this term is necessarily greater than one). Given two (subset) spectra $\boldsymbol{x_a},\boldsymbol{y_a}$, β is calculated with one or more bands $(\boldsymbol{x_b},\boldsymbol{y_b})$ that are selected from the unused bands and appended to the starting set [24]. After selecting the initial subset of bands, we would like β to be as small as possible to broaden the initial angle. Any unused band may be ranked by its associated value of β, and the band having the lowest value of β is added to the initial subset of bands. Then, is reevaluated with the new band, and is updated for the remaining unused bands. The process may be repeated iteratively, until a stopping condition is met. One logical criterion is when no remaining bands exist that yield β<1. This is equivalent to adding bands to increase the angle until no bands exist having β<1 [24].

The property of additivity for Euclidean distance can only be satisfied for $(EMD)^2$ because of the square root:

$$\Delta^2(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{M}(x_i - y_i)^2 = \sum_{i\,\in\,\alpha}(x_i - y_i)^2 + \sum_{i\,\in\,\beta}(x_i - y_i)^2$$

3. Monotonicity

A distance metric is monotonic if its value increase monotonically as the dimension of its operands, x and y, increase. By examining $(1+\frac{\langle \boldsymbol{x_b},\boldsymbol{y_b}\rangle}{\langle \boldsymbol{x_a},\boldsymbol{y_a}\rangle})$, it is clear that the term may be greater or less than one, depending on the values in $\boldsymbol{x_a},\boldsymbol{y_a},\boldsymbol{x_b},\boldsymbol{y_b}$. Thus, the addition of more spectral bands does not always guarantee an increase in angle (non monotonic) . On the other hand for EMD it is evident that an addition of bands to **x,y** cannot decrease distances because of additional terms $(x_i - y_i)^2$. Thus, nonzero spectral bands necessarily lead to an increase in the distance metric (monotonic). We can conclude that the contrast between two signals measured by EMD increases with the number of bands.

Various distance functions have been proposed for hyperspectral analysis but they have not gained much acceptance and credibility as SAM and EMD. This may be because the metric is not as intuitive, or because it does not have any physically meaningful properties, as SAM does, or even because they do not meet the criteria for a metric. However, this fact alone does not disqualify it from being useful. The three properties of a distance metric are desirable, but not absolutely necessary. There exist several different approaches for what a distance function could "measure". In most cases a direct transformation on the spectra's values reveal underlying properties such as correlation and independence, surface reflectance attributes or even probabilistic behaviors .

b) *Spectral Correlation Mapper (SCM) [25]*

Let $\boldsymbol{x}, \boldsymbol{y} \in R^M$ and $\bar{x}, \bar{y}$ are the sample means of $\boldsymbol{x}, \boldsymbol{y}$ respectively. The distance that SCM calculates is:

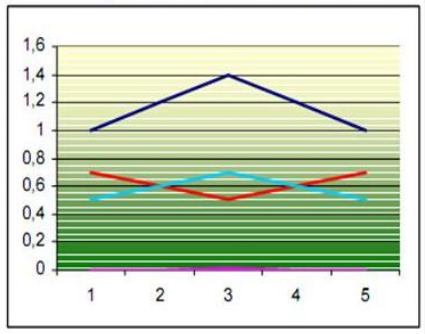
$$SCM(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

This distance function effectively calculates a statistical measure of independence known as *Pearson correlation coefficient* . In probability theory and statistics, correlation, (often measured as a correlation coefficient), indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data. A number of different coefficients are used for different situations. The best known is the Pearson product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations. From a geometric point of view a correlation angle is defined as:

$$SCA(\boldsymbol{x}, \boldsymbol{y}) = arccos(\frac{SCM(\boldsymbol{x}, \boldsymbol{y}) + 1}{2})$$

Table 1. Comparison between SAM Estimate and Pearsonian Correlation Coefficient

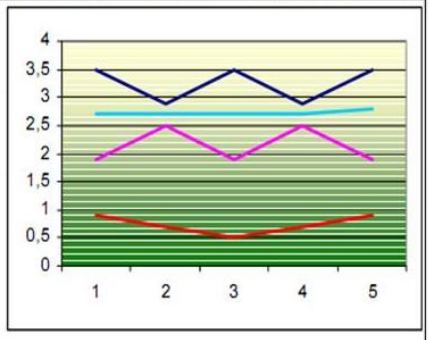
Band	Ref.	Target A	Target B	Target C
1	0,7	0,5	1	0,005
2	0,6	0,6	1,2	0,006
3	0,5	0,7	1,4	0,007
4	0,6	0,6	1,2	0,006
5	0,7	0,5	1	0,005
Total	3,1	2,9	5,8	0,029
Mean	0,62	0,58	1,16	0,0058



	Value of the cos (SAM)	Pearson Correlation
Reference x Target A	0.969299	-1
Reference x Target B	0.969299	-1
Reference x Target C	0.969299	-1

Table 2. Comparison between SAM Estimate and Pearsonian Correlation Coefficient

Band	Ref.	Target A	Target B	Target C
1	0,9	1,9	2,7	3,5
2	0,7	2,5	2,7	2,9
3	0,5	1,9	2,7	3,5
4	0,7	2,5	2,7	2,9
5	0,9	1,9	2,8	3,5
Total	3,7	10,7	13,6	16,3
Mean	0,74	2,14	2,72	3,26



	Value of cos (SAM)	Pearson Correlation
Reference x Target A	0,965150	-0,218220
Reference x Target B	0,981606	0,534522
Reference x Target C	0,980079	0,218218

Table 3.7 : SAM & SCM comparison

The term $SCM(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$ is the cosine of $\boldsymbol{x}, \boldsymbol{y},$ shifted by their sample means. In other words SCM resemblances SAM as a normalized inner product shifted by mean: $cos(SAM(\boldsymbol{x} - \bar{x}, \boldsymbol{y} - \bar{y})) = SCM(\boldsymbol{x}, \boldsymbol{y})$ SCM is an improved version of SAM, because it can recognize both positive and negative correlations. The above table summarizes SCM, SAN differences.

In the left of table 3.7 two spectra that we want to classify (blue + light blue) present different monotony from reference spectrum . SAM's cosine measures differences without recognizing the negative correlation between them. On the other hand SCM recognizes the negative cross-correlation of data "measuring" -1, that is to say data change values at opposite directions. In the second table where spectra values have more fluctuations SAM measures almost identical distances (target B, target C) for vectors that differ considerably as it appears with a simple observation.

c) *Spectral Information Divergence (SID) [26]*

Spectral Information Divergence (SID) is an information theoretic spectral metric which is derived from the concept of divergence in information theory. In order to describe the probabilistic behavior of spectra signatures we must define an appropriate probability space (Ω, Σ, P) associated with it.

Let $\mathbf{x}, \mathbf{y} \in \mathrm{R}^{\mathrm{M}}$ hyperspectral pixel vectors: $\boldsymbol{x}=[x_1\ x_2\ \dots\ x_M]$, $\boldsymbol{y}=[y_1\ y_2\ \dots\ y_M]$.

Then $\mathbf{x}, \mathbf{y}$ can be modeled as random variables by defining appropriate probability distributions. All component in $\mathbf{x}, \mathbf{y}$ are nonnegative due to the nature of radiance or reflectance so we can normalize to the range [0,1] by defining

$p_j = \frac{x_j}{\sum_{i=1}^{M} x_i}$, $q_j = \frac{y_j}{\sum_{i=1}^{M} y_i}$ so that $\boldsymbol{p} = \{p_m\}_{m=1}^{M}$, $\boldsymbol{q} = \{q_m\}_{m=1}^{M}$ are the desired probability vectors resulting from pixel vectors $\mathbf{x}, \mathbf{y}$ respectively.

Using $\boldsymbol{p}, \boldsymbol{q}$ we define Spectral Information Divergence by

$SID(\mathbf{x}, \mathbf{y}) = \mathrm{D}(\mathbf{x} \,||\, \mathbf{y}) + (\mathbf{y} \,||\, \mathbf{x})$, where

$\mathrm{D}(\mathbf{x} \,||\, \mathbf{y}) = \sum_{\mathrm{i}=1}^{\mathrm{M}} p_i \log(\frac{p_i}{q_i})$ and $\mathrm{D}(\mathbf{y} \,||\, \mathbf{x}) = \sum_{\mathrm{i}=1}^{\mathrm{M}} q_i \log(\frac{q_i}{p_i})$.

SID considers each pixel as a random variable and uses its spectral histogram to define a probability distribution. In other words this distance function views each pixel spectrum as a random variable and then measures the discrepancy of probabilistic behaviors between their spectra. The term $\mathrm{D}(\mathbf{x} \,||\, \mathbf{y})$ is called the relative entropy of Y with respect to $\mathbf{x}$ which is also known as Kullack-Leihler information function, directed divergence or cross entropy. Spectral Information Divergence can be used to measure the spectral similarity between two pixel vectors $\mathbf{x}$ and reference pixel vector $\mathbf{y}$.

d) Spectral Gradient Angle (SGA)

Spectral gradient [27] is a surface reflectance descriptor which is invariant to scene geometry and incident illumination for smooth diffuse surfaces. The invariant properties of the spectral gradients by examining the rate of change in reflected intensity with respect to wavelength make them a particularly appealing tool in many diverse areas of computer vision such as color constancy, tracking, scene classification, material classification, stereo correspondence, even re-illumination of a scene.

The measurement of the energy of radiation, an objective quantity that can be measured in W, is called radiometry. When the spectral sensitivity of the eye is taken into account, the measurement is called photometry, where light is measured in lumens. Photometry is semi-objective, intermediate between the physical stimulus of energy and the psychophysical response of brightness.

Photometric data is a readily available dense source of information in intensity images, but is not widely used in computer vision because of its dependence on viewpoint and incident illumination. Computer vision algorithms main objective is to identify objects from an imaged scene. Although reflected light is a primary source of information for object identification, slight variations in viewing conditions often cause large changes in an object's appearance. Different approaches have been developed to eliminate variations:

- Color/cue constancy algorithms,
- identifying reflectance-based object properties that are invariant to illumination

Most color techniques assume that the spectral reflectance functions have the same degrees of freedom as the number of photoreceptor classes (typically three.) Thus, none of these methods can be used in grayscale images for extracting illumination invariant color information.

Spectral Gradient is an invariant to illumination distance metric by identifying materials with the same reflectance under variable viewing conditions and discriminating materials with distinct reflectance functions.

Reflected light I from each point $p = (x, y, \zeta)$ at wavelength λ in a imaged scene depends on the light source , E and the surface reflectance S of the materials composing the scene:

$$I(\rho, \lambda) = E(\rho, \lambda)S(\rho, \lambda)$$

The reflectance function $S(p, \lambda)$ may depend on the surface material, the geometry of the scene and the viewing and incidence angles. When the spectral distribution of the incident light does not vary with the position of the light, the geometric and spectral components of the incident illumination are separable:

$$E(\rho, \theta, \varphi, \lambda) = e(\rho, \lambda)E(\rho, \theta, \varphi)$$

where $e(\lambda)$ is the illumination spectrum.

$$I(\rho,\lambda) = e(\rho,\lambda)E(\rho, \theta, \varphi,)S(\rho, \lambda) \rightarrow$$

$$ln[I(\rho,\lambda)] = ln[e(\rho,\lambda)] + \ln[E(\rho, \theta, \varphi,)] + \ln[S(\rho, \lambda)] \rightarrow$$

$$\frac{\partial\, ln[I(\rho,\lambda)]}{\partial\lambda} = \frac{1}{e(\rho,\lambda)}\frac{\partial e(\rho,\lambda)}{\partial\lambda} + \frac{1}{S(\rho,\lambda)}\frac{\partial S(\rho,\lambda)}{\partial\lambda}$$

1. *Invariance to Incident Illumination*

   Although the spectral distribution of the most commonly used indoor-scene illuminations sources (i.e., tungsten and fluorescent light) is not constant, one can assume that *e* changes very slowly over small increments of λ. This means that its derivative with respect to wavelength is approximately zero.

$$\frac{\partial\, ln[I(\rho,\lambda)]}{\partial\lambda} = \frac{1}{S(\rho,\lambda)}\frac{\partial S(\rho,\lambda)}{\partial\lambda}$$

2. *Invariance to Geometry and Viewpoint*

   - *Lambertian Model*

     $S(\rho,\lambda) = cos\theta(\rho)p(\rho,\lambda)$, where $p(\rho, \lambda)$ is the albedo or diffuse reflection coefficient at point p.

   - Smooth Diffuse Reflectance Model olff[30]):

     $S(\rho,\lambda) = cos\theta(\rho)p(\rho,\lambda)\left(1 - F\left(\theta(\rho), n(\rho)\right)\right)\left(1 - F\left(sin^{-1}\left(\frac{sin\varphi\,(\rho)}{n(\rho)}, \frac{1}{n(\rho)}\right)\right)\right)$, where $\theta(p)$ and $\varphi(p)$ are the incidence and viewing angles respectively, $p(p, \lambda)$ is the surface albedo, F() is the Fresnel reflection coefficient, and *n* is the index of refraction.

*T*he partial derivative of the incident illumination with respect to wavelength for both models is always a function of albedo because all the other terms depend only in a specific point ρ:

$$\frac{\partial\, ln[I(\rho,\lambda)]}{\partial\lambda} = \frac{1}{p(\rho,\lambda)}\frac{\partial p(\rho,\lambda)}{\partial\lambda}$$

Notice from the last derivative that spectral gradient encodes information at discrete spectral locations about how fast the surface albedo changes as the spectrum changes. It is a profile of the rate of change of albedo with respect to wavelength over a range of wavelengths.

In order to use spectral gradient as a pixel distance metric we can built the spectral gradient angle [29]:

$$\theta_{SG}(\rho_1, \rho_2) = \cos^{-1}\left(\frac{\langle \frac{\partial\, ln[I(\rho_1,\lambda)]}{\partial\lambda}, \frac{\partial\, ln[I(\rho_2,\lambda)]}{\partial\lambda} \rangle}{\left\|\frac{\partial\, ln[I(\rho_1,\lambda)]}{\partial\lambda}\right\| \left\|\frac{\partial\, ln[I(\rho_2,\lambda)]}{\partial\lambda}\right\|}\right), \qquad 0 \le \theta \le \frac{\pi}{2}$$

e) *SID – SAM Mixed Measure [26]*

Combining SID and SAM into a new measure which is the product of them, a new hyperspectral distance function can be defined as it increases spectral discriminability because it makes two similar spectral signatures even more similar and two dissimilar spectral signatures more distinct:

$$SIDSAM_1(\boldsymbol{x}, \boldsymbol{y}) = SID(\boldsymbol{x}, \boldsymbol{y}) * \tan(SAM(\boldsymbol{x}, \boldsymbol{y}))$$

$$SIDSAM_2(\boldsymbol{x}, \boldsymbol{y}) = SID(\boldsymbol{x}, \boldsymbol{y}) * \sin(SAM(\boldsymbol{x}, \boldsymbol{y}))$$

f) Spectral similarity scale (SSS) [28]

The mathematic definition of the SSS is founded on the definition of vector identity (two identical vectors have the same magnitude and direction) and the assumption that vector magnitude is independent of direction. Extrapolating from the definition of vector identity to similarity leads to the statement that two

similar vectors must have both similar magnitude and direction. Thus, the similarity of two reflectance spectra is completely defined by a "similarity vector" consisting of two elements that separately describe magnitude and direction similarity. The magnitude of the similarity vector is the scalar representation of spectral similarity while the direction of the similarity vector indicates the relative influence of the two components. This concept can be applied to high dimensional reflectance data and thus defines the Spectral Similarity Scale; numbers on the scale are termed Spectral Similarity Values (SSV). The elements of magnitude and direction are described by a traditional distance function like EMD and pearson coefficient respectively. It is possible to use different components that capture the notion of magnitude and direction to define "similarity vectors". Another combination could be SID or GSA for magnitude and SCM for direction.

$$SSV_1 = \sqrt{EMD^2 + SCM^2}$$
$$SSV_2 = \sqrt{SID^2 + SCM^2}$$

g) *Minimum Distance classifiers with statistical parameters*

- Minimum Distance to Means:

Let $\mathbf{x}, \mathbf{y} \in \mathrm{R}^{\mathrm{M}}$ *hyperspectral pixel vectors:* $\boldsymbol{x}=[x_1\ x_2\ \dots\ x_M]$, $\boldsymbol{\mu_y}=[\mu_y\ \mu_y\ \dots\ \mu_y]$, *where* μ_y = *sample mean of reference vector y* →

$$EMD^2(\boldsymbol{x}, \boldsymbol{\mu_y}) = \|\boldsymbol{x} - \boldsymbol{\mu_y}\|^2 = \sum_{i=1}^{M} (x_i - \mu_y)^2 = (\boldsymbol{x} - \boldsymbol{\mu_y})^T (\boldsymbol{x} - \boldsymbol{\mu_y})$$

In this case, pixels are assigned to whichever class has the smallest Euclidean distance to its mean. The classes are, by default, assumed to have probability density functions with common covariances that are equal to the identity matrix. This is equivalent to assuming that the classes all have unit variance in all features and the features are all uncorrelated to one another. Geometrically, for the two-class, two-feature case, figure shows the decision boundaries for this classifier would appear for a given set of classes distributed as the two elongated oval areas. It is seen in this case that the decision boundary for the Minimum Euclidean Distance classifier is linear and is in fact the perpendicular bisector of the line between the two class mean values. Its location is uninfluenced by the shapes of the two class distributions.

- Fisher's Linear Discriminant: $d_{FLD} = (\boldsymbol{x} - \boldsymbol{\mu_y})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu_y})$

In this case, the classes are assumed to have density functions with common covariance specified by Σ^{-1}. This is equivalent to assuming that classes do not have a common covariance all features, the features are not necessarily uncorrelated, but all classes have the same variance and correlation structure. In this case the decision boundary in feature space will be linear, but its location and orientation between the class mean values will depend on the combined covariance for all the classes in addition to the class means.

Geometrically, for the two-class, two-feature case, Figure shows how the decision boundary for this classifier would appear for a given set of classes indicated by the two elongated oval areas. It is seen in this case that the decision boundary for the Fisher Linear Discriminant classifier is also linear but its orientation and location of the combined distribution are influenced by the overall shape.

- Quadratic (Gaussian) Classifier: $d_{QG} = -\frac{1}{2}\ln|\Sigma_y| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_y})^T \Sigma_y{}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_y})$

In this case the classes are not assumed to have the same covariance, each being specified by Σ_y. The decision boundary in feature space will be a second-order hyper surface (or several segments of second order

hyper surfaces if more than one subclass per class is assumed), and its form and location between the class mean values will depend on the Σ_y 's.

Geometrically, for the two-class, two-feature case, Figure shows how the decision boundaries for this classifier would appear for a given set of classes indicated by the two elongated oval areas. Interestingly, if the data for the two classes had the mean values and class and feature variances as indicated, but with features uncorrelated with one another, the data would be spread over the regions indicated by the circles. Under these circumstances the error rate, as indicated by the overlapped area would be substantially greater. This is another illustration of the fact that correlation does not always imply redundancy and features being uncorrelated is not necessarily desirable. Note that as described, the minimum distance to means, the Fisher linear discriminant, and the quadratic classifier are all maximum likelihood classifiers. They form a hierarchically related set and differ only in the assumptions about the details of the class covariance functions.

3.5 Accuracy assessment

The results of any classification process applied to hyperspectral data classification must be quantitatively assessed in order to determine their accuracy. As suggested by Lillesand and Kiefer (1994), a classification process is not complete until its accuracy is assessed. There may be different ways to assess the accuracy of a classification process. Accuracy assessment can be qualitative or quantitative, expensive or inexpensive, quick or time consuming, well-designed and efficient. The purpose of quantitative accuracy assessment is the identification and measurement of map errors. Quantitative accuracy assessment involves comparison of an area on a map against reference information of the same area, assuming reference data to correct. There are number of ways to determine the degree of error in the end-product, which is typically a thematic map or image, by measuring overall classification accuracy, and calculating the Kappa statistics for a given number of test data.

3.5.1 Confusion Matrix

The accuracy of classification has traditionally been measured by the overall accuracy by generating a confusion matrix and determining accuracy levels by dividing the total number of correctly classified pixels (sum of major diagonal of confusion matrix, also called actual agreement) by the total number of reference pixels. However as a single measure of accuracy, the overall accuracy gives no insight into how well the classifier is performing for each of the different classes (Fitzgerald and Lees, 1994). In particular, a classifier might perform well for a single class that accounts for a large proportion of the test data and this will create a bias in overall accuracy, despite low class accuracies for other classes. To avoid such a bias when assessing the accuracy of a classifier, it is important to consider the individual class accuracies. Individual class accuracy can be obtained by dividing the total number of correctly classified pixels in that category by the total number of pixels of that category. Individual class accuracy can be determined by using the reference data (called producer's accuracy).

The resulting percentage accuracy indicates the probability that a reference pixel will be correctly classified. Story and Congalton (1986) suggested that producer's accuracy is a measure of error of omission. However, a misclassification error is not only an omission from the correct class but also a commission into another class. Individual class accuracy obtained from the classified data in that category (user's accuracy) is a measure of error of commission (Story and Congalton, 1986).

3.5.2 *Kappa* statistics

Generally, the confusion matrix is an appropriate tool for assessing the accuracy of hyperspectral classifications. However, Congalton (1991) suggested the use of the Kappa coefficient as a suitable measure of the accuracy of a thematic classification. It is a measure of the randomness of the classification results. It measures the difference between the actual agreement in the confusion matrix (i.e., the agreement between the classification and the reference data as indicated by the major diagonal) and the chance agreement which is indicated by row and column totals. It provides a better measure of the accuracy of a classifier than the overall accuracy, and it takes into account the whole confusion matrix rather than the diagonal elements alone. Cohen's kappa coefficient is a statistical measure of inter-rater agreement. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The equation for κ is:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

,where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

The Kappa statistic is calculated from the confusion matrix by using the following formula:

$$K = \frac{n \sum_{i=1}^{p} x_{ii} - \sum_{i=1}^{p} x_{io}\, x_{oi}}{n^2 - \sum_{i=1}^{p} x_{io}\, x_{oi}}$$

,where n = total number of pixels used for testing the accuracy of a classifier ,P=number of classes, $\sum_{i=1}^{p} x_{ii}$ = sum of diagonal elements of confusion matrix, $\sum_{i=1}^{p} x_{io}$ = sum of row I, $\sum_{i=1}^{p} x_{oi}$ = sum of column i.

3.5.3 *Statistical separability measures*

In many applications we are required to identify a target pixel extracted from an unknown image scene using an existing spectral library or database Δ. In this case, it is of interest to know what is the likelihood of the pixel in question being identified as one of the signature spectra in Δ. Let $\{s_j\}_{j=1}^{J}$ be J spectral signatures in Δ and t be a target pixel to be identified using . Let m(.,.) be any given hyper spectral measure. We define the spectral discriminatory probabilities of all s_j 's in Δ with respect to t as follows:

$$p_{t,\Delta}^{m}(i) = \frac{m(t, s_i)}{\sum_{j=1}^{J} m(t, s_j)}, \qquad \text{for } i = 1,2, \dots, J$$

where $\sum_{j=1}^{J} m(t, s_j)$ is a normalization constant determined by ***t*** and Δ. The resulting probability vector

$$p_{t,\Delta}^{m} = (p_{t,\Delta}^{m}(1), p_{t,\Delta}^{m}(2), \dots, p_{t,\Delta}^{m}(J))^{T}$$

is called the spectral discriminatory probability vector of Δ with respect to t[26]. The target pixel t can be identified by selecting the one with the smallest spectral discriminatory probability because t and the selected one have the minimum spectral discrimination.

If we are given two spectral similarity measures, how do we evaluate which of the two is more effective? To meet this need ,a criterion, spectral discriminatory power [26], is developed. It is designed based on the power of discriminating one pixel from another relative to a reference pixel d. Assume that m(.,.) is any given hyperspectral measure. Let d be the spectral signature of a reference pixel and s_i, s_j be the spectral signatures of two pixels. We define spectral discriminatory power of m(.,.) by

$$PW^{m}(s_i, s_j; d) = max\{\frac{m(s_i,d)}{m(s_j,d)}, \frac{m(s_j,d)}{m(s_i,d)}\}$$

The PW provides an index of spectral discrimination capability of a specific hyperspectral measure m(.,.) between any two spectral signatures relative to d. Obviously, the higher the PW is,the better discriminatory power m(.,.) has.

In addition, PW is symmetric and bounded below by one, i.e., PW >= 1 with equality if and only if $s_i = s_j$.

Since $p_{t,\Delta}^{m} = (p_{t,\Delta}^{m}(1), p_{t,\Delta}^{m}(2), \dots, p_{t,\Delta}^{m}(J))^{T}$ is the spectral discriminatory probability vector of t using a spectral library Δ, we can further define τhe spectral discriminatory entropy [26] of Δ with respect to t by

$$H^{m}(t; \Delta) = \sum_{j=1}^{J} p_{t,\Delta}^{m}(j) \log p_{t,\Delta}^{m}(j)$$

which provides the uncertainty measure of identifying t using the spectral signatures in Δ. A smaller $H^{m}(t; \Delta)$ indicates a better chance to identify t.

Chapter 4
Experimental Methods and Results (Case study: El Greco's pigment identification)

As in forensics, analyzing a piece of artwork typically begins with the most fundamental visualization tool of all: the human eye. A microscopically small sample of pigment is taken from a painting and placed under a standard microscope for examination. A scanning electron microscope is then typically used to look at the same sample with higher magnification. X-ray fluorescence and radiography are also routinely used to gather elemental data. Various spectroscopic techniques are then employed, depending on the desired application for the identification of pigments.

Application of physical/chemical destructive methods for pigment analysis, as mentioned above, involves the extraction of a specimen in order to study its structure and chemical composition. Although the dimensions of specimens must not exceed a few square millimeters in area and a few cubic millimeters in volume, researchers have to identify surfaces characterized by high spatial heterogeneity. Also, repetitive sampling is a prohibited procedure due to the historical value of the artwork, its small dimensions or even its maintenance condition. Microchemistry, Selective coloration with Electronic Microanalysis, X-ray Diffractometry, Mass Spectrometry, Gas Chromatography, Thin layer Chromatography, Neutron/Proton Activation Analysis [32] , laser induced breakdown spectroscopy are some of the established methods for exploring pigment materials.

In an attempt to cancel the invasive nature of these methods, optical spectroscopy has been extensively applied and evaluated in the scientific analysis and documentation of artworks. Application of modern spectral processing systems refers only to the study of various colors' spectral responses composing an artwork without physically extracting pigment materials. X-ray Fluorescence spectroscopy, Infrared Absorption spectroscopy, Ultraviolet Fluorescence spectroscopy, Near-Infrared Reflectroscopy, Gamma Spectroscopy, (surface-enhanced) Raman spectroscopy, Fourier transform infrared etc. have provided a unique insight into the material composition, technique of construction and deterioration effects which are essential for the analysis and helpful in determining the optimum preservation scheme. However, the above techniques suffer from the major drawback that they are capable of acquiring spectral information from only one -visually selected- spatial point.

These limitations highlight the need for the development of spatially resolved spectral acquisition methods and technologies, capable of performing spectral mapping of the area under study. Imaging spectroscopy as the application of reflectance spectroscopy to every pixel in a spatial image of an artwork can be used to detect individual absorption features of a color pigment due to specific chemical bonds. Hyperspectral processing systems implemented by high spatial and spectral resolution imaging monochromators and devices (HySI) can be

employed for the analysis of artworks as they provide a non-destructive information extraction framework, enabling an in-depth, in-situ and real time analysis without compromising the artwork's integrity.

Besides the non-destructive analysis advantage of imaging spectroscopy, HySI systems offer a fast, high quality and real time provision of information data for comparative analysis available even after the completion of conservation works. Moreover, their capabilities of capturing images in a range of hundreds narrow spectral bands combined with versatile software, provide an attractive technology for a close examination of subtle variances in color pigments' materials and binding media.

Other non-destructive methods are: Reflected UV photography, Ultraviolet Fluorescence Photography, Infrared Reflection Photography, Photography with Color infrared film, Optical coherence tomography, Radiography.

4.1 Hyperspectral acquisition system

Typical HySI configurations involve coupling an imaging monochromator with an imaging detector, while both are interfaced with personal computer elements and controlling units. Imaging monochromators although implemented by different technologies operate as band pass filters (acousto-optic, liquid crystal tunable filters, and Fourier transformed interferometers) tuning the full spectral range to the specific application requirements. Then an imaging detector captures the reflected image and special developed software is used to analyze the acquired spectral images. However these configurations fail to provide a sufficient amount of diagnostic information due to its technological limitations. In particular they suffer from narrow spectral range, since different modules are required to cover the visible or the (N)IR spectral range, low throughput and image shifting. Low throughput of these systems necessitates the use of a high power light source for the illumination of the object in order to obtain acceptable image brightness. But this could be harmful for the object, since high power illumination can provoke photo thermal and/or photochemical damage.

In an attempt to overcome the above-mentioned limitations and provide a HySI system capable of real time spectral imaging (both reflectance and fluorescence) with high spectral resolution and high throughput ratio, Dr. Balas et al [34,35] developed an all-optical imaging monochromator functioning as an electronically tunable narrow band pass optical filter. Displacement of the optical elements of the latter, results in the tuning of the imaging wavelength, which is performed with the aid of electromechanical manipulators controlled from the PC via microcontroller. Mu.SIS HS' (figure 4.1) technical features are:

- Spectral imaging acquisition of 5nm full width half maximum (FWHM), performing in 34 spectral bands of about 20nm each, in the spectral range 360nm (Ultraviolet)–1550nm (Near Infrared).
- Real time capturing & displaying images with an analysis of 1600x1200 pixels.
- Minimum transmittance is 40% across its operational spectral range, which determines the high throughput of the developed monochromator.
- Tuning spectral range of the filtering system is matched with the responsivity spectral range of the charge coupled device(CCD) image sensor, with the capability of extending to longer wavelengths, up to the mid-infrared range (photocathode).
- A megapixel CCD camera, for feeding back the monochromator signal, based on the IEEE-1394 data transferring protocol, capable of acquiring images at a rate of 15 frames/s at full resolution and of more than 30 frames/sat VGA resolution.
- A special calibration procedure [36] is executed before any imaging procedures, compensating for the wavelength dependence of the response of the electro optical parts of the system, such as CCD, illuminators, etc, thus ensuring the full exploitation of the CCD's dynamic range.
- Operating in imaging mode, an image at each wavelength band is acquired while, in spectroscopy mode, a fully resolved diffuse reflectance and/or fluorescence spectrum per image pixel can be recorded (image spectral cube). The combination of spectral and color imaging with calibration enables the system to operate as either Imaging Spectrometer or Imaging Colorimeter.

figure 4.1: MuSIS HS

4.2 Data Description

Evaluating Mu.SIS HS performance in the non-destructive analysis of objects of artistic and historic value, spectral imaging was performed for several cases of interest with different data sets such as: (1) recovery of overwritten script, (2) pigment identification and mapping (3) assessment of laser cleaning effects [33], (4) pigment identification in El greco's Concert of Angels (1608-1614) [37].

By observing the reflectance spectra at different wavelengths it was possible to determine the ideal band (600nm) in which the underlying previously erased inscription became non-transparent, for the case of overwritten script. Moreover, pigment identification was possible by determining the reflectance behavior of color pigments in a series of spectral images and matching them with spectral profiles of reference samples developed for comparison. For example, it was clearly seen that the variation of reflectance as a function of the wavelength of a red paint under study, matched with a vermilion reference sample, while it was in clear contrast with a mars red reference sample. In the case of cleaning effects, a set of spectral images of a detail of an old manuscript cleaned with a 532nm-second harmonic Q-switched Nd:YAG and an excimer laser at 248 nm were used. The Nd:YAG induced surface alterations which, while were not seen in the visible, they were well depicted in the 380nm images. The captured spectral images also showed an in-depth damage provoked by the excimer laser.

A similar methodology for identifying El greco's Concert of Angels (figure 4.2) pigment materials was used:

Figure 4.2: Spectral imaging on El greco's Concert of Angels (1608-1614).

1. The first step was to develop a manufacturing process for replicating both color pigments and the underlying layer of canvas or wooden panel with the ones used in painter's era, acquiring materials of very similar chemical composition. In particular, two series of pigment samples were composed, one for oil-painting on prepared canvas and another one for egg painting (egg tempera pigments) on prepared wooden plate.
    - Wooden plates for egg pigments were constructed from sea plywood, coated with a mixture of animal bone glues and covered with cotton material of cellular texture (the mixture and cotton combination was known as a "size"). Once the size dried, layer upon layer of gesso applied, each layer sanded down before the next applied (2-3 layers totally) before a smooth hard surface emerged.
    - For oil painting sampling, canvas was stretched across 5 wooden frames made of linen. Early canvas as in the case of El Greco's Concert of Angels was made of linen, a sturdy brownish fabric of considerable strength. Linen is particularly suitable for the use of oil paint. Gesso of 10-12 layers totally was applied.
    - 34 different pigments of various colors (red, blue, green, ochre, black & white for egg-painting and red, blue, green, ochre and yellow for oil-painting, table 4.3) from KREMER PIGMENTE's catalogue [40] were selected and applied on the wooden panels or canvas with different preparation procedures according to oil (linseed oil as binding media) or egg (egg yolk as binding media) sampling scheme.
    - Gradually darker tones of each pigment were applied, by painting consecutive layers in different areas of plates or canvas, acquiring a sufficient amount of 108 representative samples (figure 4.4-4.11)
    - Carbon black lines were drawn on the substrate for simulating the presence of under drawings.In order to assess the selected color pigments as suitable for building up an El Greco color reference database, in terms of similarity with the original pigments, X-ray Diffractometry technology was used. The crystalline structure of pigments was compared with crystalline profiles of pigments known to have been used by El Greco. The latter profiles have been obtained from other works using destructive methods [37].

Pigment Name	Pigment code	Pigment's Concentration per egg yolk ml	Pigment's Concentration per linseed oil ml
Μαύρες Χρωστικές			
Bone black	47100	C=0,15 gr/ml	
Plant black (vine black)	47000	C=0,10 gr/ml	
Pyrolusite	47500	C=0,10 gr/ml	
Ασπρες χρωστικές			
Lead White	46000	C=0,75 gr/ml	
Calsite	58500	C=0,50 gr/ml	
Κόκκινες χρωστικές			
Realgar	10800	C=0,50 gr/ml	C_1=0,84 gr/ml C_2=0,93 gr/ml C_3=1,05 gr/ml C_4=1,20 gr/ml
Minio	42500	C=0,40 gr/ml	C_1=0,66 gr/ml C_2=0,73 gr/ml C_3=0,83 gr/ml C_4=0,95 gr/ml
Hematite	48600	C=0,20 gr/ml	C_1=0,27 gr/ml C_2=0,28 gr/ml C_3=0,285 gr/ml C_4=0,29 gr/ml
Cinnabar	10620		C_1=1,07 gr/ml C_2=1,39 gr/ml C_3=1,65 gr/ml C_4=2,50 gr/ml
Cinnabar (manufactured)	42000	C=0,40 gr/ml	
Carmine lake	42100		C_1=0,27 gr/ml C_2=0,28 gr/ml C_3=0,285 gr/ml C_4=0,29 gr/ml
Alizarin Crimson lake (manufactured)	23610		C_1=0,36 gr/ml C_2=0,38 gr/ml C_3=0,41 gr/ml C_4=0,45 gr/ml
Red lake dark		C=0,10 gr/ml	
Red lake light		C=0,10 gr/ml	
Μπλε χρωστικές			
Lapis Lazuli (nat. Ultramarine)	10510	$3Na_2O$.$3Al_2O_3$.$6SiO_2$ .Na_2S	π.Χ.
Azurite	10200	2Cu CO_3 .$Cu(OH)_2$	π.Χ.
Egyptian Blue	10060	$CaO.CuO.4SiO_2$ ή $CaCuSi_4O_{10}$	π.Χ.
Indigo	36000.-A	Προέρχεται από φυτά του γένους Indigofera tinctoria	Σε μικρή ποσότητα στην Ευρώπη μέχρι το 1516. Σε αφθονία μετά το 1602.
Smalt	10010	65-71%SiO_2 + 16-21%K_2O + 6-7%CoO + 0-8%Al_2O_3	16ος αιώνας
Κίτρινες χρωστικές			
Lead-tin yellow	10110	2PbO. SnO_2	14ος-15ος αιώνας
Massicot	43010	PbO	π.Χ.
Naples yellow	43130	$Pb_3(SbO_4)_2$	17ος-18ος αιώνας

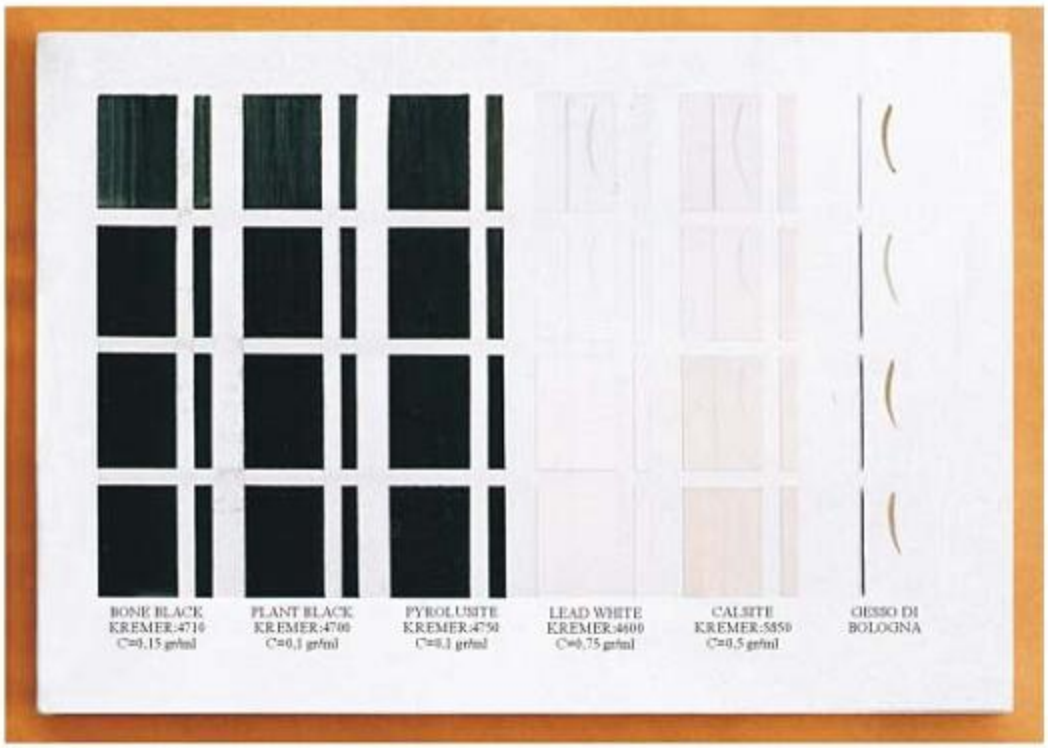
Γαιώδης χρωστικές			
Gold Ochre Spain	11500		π.Χ.
Gold Ochre Italy	40220		π.Χ.
Yellow Ochre (manufactured)	48000	Αργιλλοπυριτικά οξείδια (SiO_2 ή $Al_2O_3 .2SiO_2 .H_2O$,	π.Χ.
Raw Umber	40610	$Fe_2O_3.H_2O$+$MnO_2$8-16%	π.Χ.
Burnt Umber	40720	Fe_2O_3 + MnO_2 8-16%	π.Χ.
Raw Sienna Monte Amiata	17050	$Fe_2O_3.H_2O$+$MnO_2$0,6-1,5%	π.Χ.
Burnt Sienna	17100	Fe_2O_3 - MnO_2 0,6-1,5%	π.Χ.
Πράσινες χρωστικές			
Malachite	10300	$Cu\ CO_3 .Cu(OH)_2$	π.Χ.
Verdigris (manufactured)	44450	$Cu(C_2H_3O_2)_2.2Cu(OH)_2$	π.Χ.
Green Earth (Celadonite) Cote D'Azur	11250	(Υδροπυριτική ένωση Fe, Al, Mg και K)	π.Χ.
Green Earth Boemia	40810	(Υδροπυριτική ένωση Fe, Al, Mg και K)	π.Χ.
Green Earth Italy	40820	(Υδροπυριτική ένωση Fe, Al, Mg και K)	π.Χ.

Smalt	10010	C=1,00 gr/ml	C_1=0,20 gr/ml C_2=0,40 gr/ml C_3=0,65 gr/ml C_4=0,80 gr/ml
Κίτρινες χρωστικές			
Lead-tin yellow	10110		C_1=1,13 gr/ml C_2=1,31 gr/ml C_3=1,47 gr/ml C_4=1,67 gr/ml
Massicot	43010		C_1=1,10 gr/ml C_2=1,30 gr/ml C_3=1,50 gr/ml C_4=1,70 gr/ml
Naples yellow	43130		C_1=0,36 gr/ml C_2=0,53 gr/ml C_3=0,75 gr/ml C_4=2,00 gr/ml

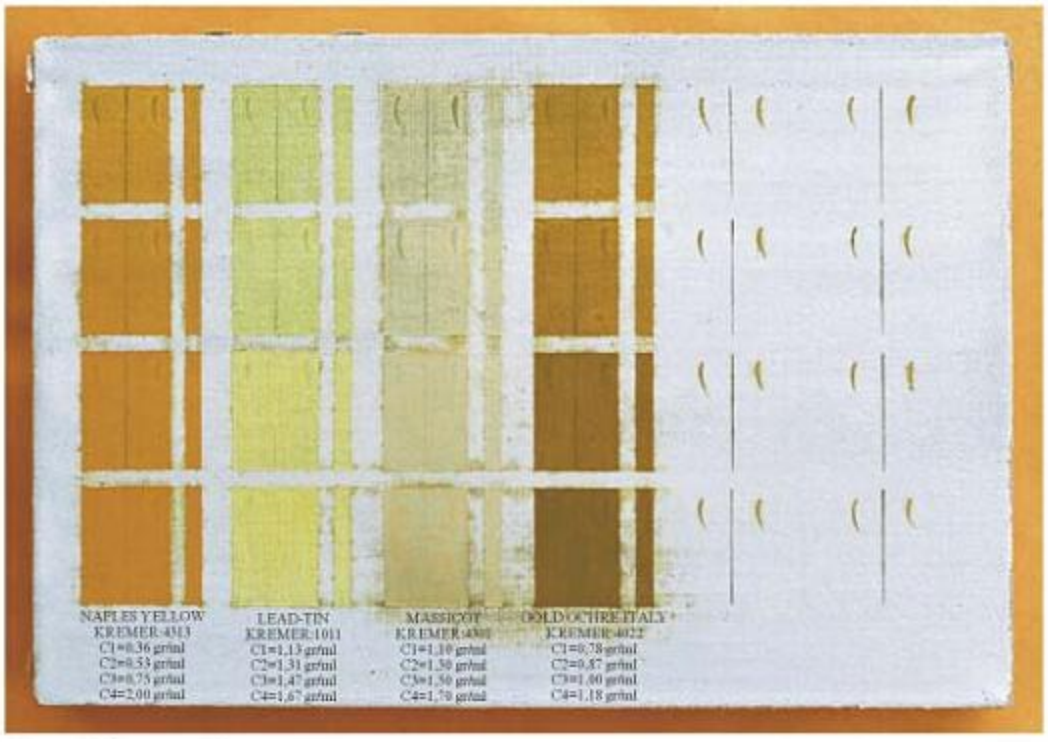
Γαιώδης χρωστικές			
Gold Ochre Spain	11500	C=0,60 gr/ml	
Gold Ochre Italy	40220		C_1=0,33 gr/ml C_2=0,38 gr/ml C_3=0,45 gr/ml C_4=0,60 gr/ml
Yellow Ochre (manufactured)	48000	C=0,225 gr/ml	C_1=0,15 gr/ml C_2=0,19 gr/ml C_3=0,26 gr/ml C_4=0,40 gr/ml
Raw Umber	40610	C=0,20 gr/ml	C_1=0,26 gr/ml C_2=0,29 gr/ml C_3=0,33 gr/ml C_4=0,40 gr/ml
Burnt Umber	40720	C=0,20 gr/ml	C_1=0,26 gr/ml C_2=0,29 gr/ml C_3=0,33 gr/ml C_4=0,40 gr/ml
Raw Sienna Monte Amiata	17050	C=0,175 gr/ml	C_1=0,29 gr/ml C_2=0,32 gr/ml C_3=0,34 gr/ml C_4=0,40 gr/ml
Burnt Sienna Monte Amiata	17100	C=0,20 gr/ml	C_1=0,15 gr/ml C_2=0,21 gr/ml C_3=0,24 gr/ml C_4=0,35 gr/ml
Πράσινες χρωστικές			
Malachite	10300	C=0,625 gr/ml	C_1=1,32 gr/ml C_2=1,47 gr/ml C_3=2,00 gr/ml C_4=2,40 gr/ml
Verdigris (manufactured)	44450	C=0,50 gr/ml	C_1=0,68 gr/ml C_2=0,80 gr/ml C_3=1,00 gr/ml C_4=1,25 gr/ml
Green Earth (Celadonite) Cote D'Azur	11250	C=0,50 gr/ml	C_1=0,65 gr/ml C_2=0,80 gr/ml C_3=1,10 gr/ml C_4=1,30 gr/ml
Green Earth Boemia	40810	C=0,40 gr/ml	C_1=0.50 gr/ml C_2=0,60 gr/ml C_3=0,80 gr/ml C_4=1,10 gr/ml
Green Earth Italy	40820	C=0,75 gr/ml	C_1=0,82 gr/ml C_2=1,00 gr/ml C_3=1,25 gr/ml C_4=1,55 gr/ml

Table 4.3 : Selected pigments from KREMER PIGMENTES's catalog

figures 4.4,4.5: reference samples of black, white and yellow pigments.



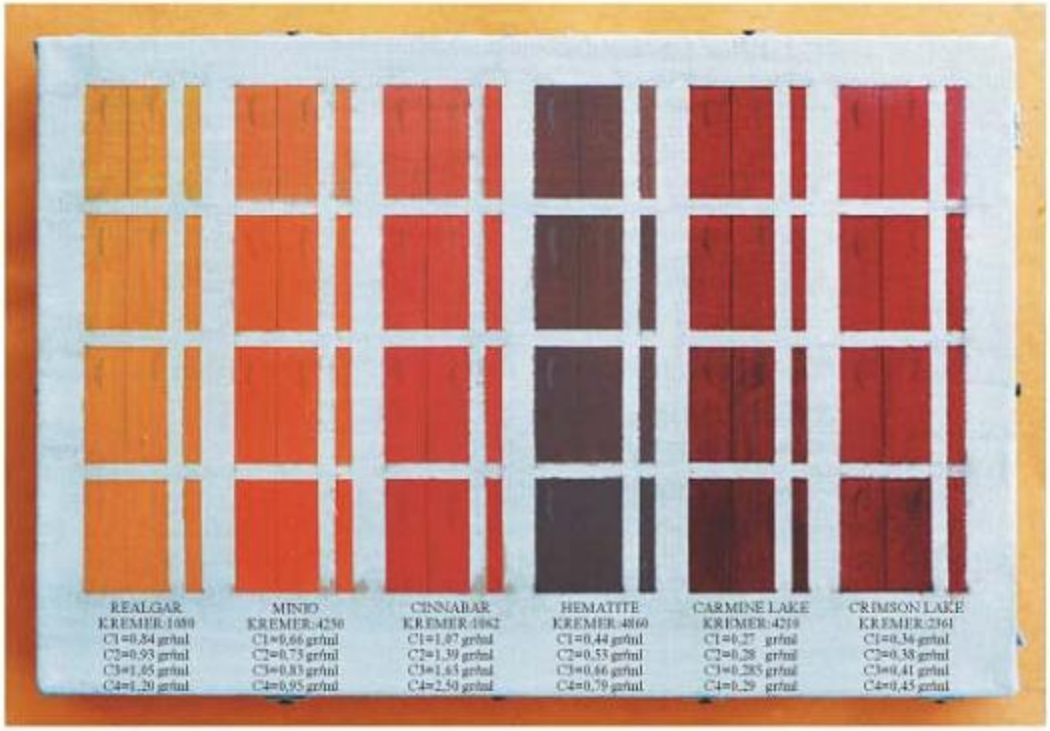
Wood plate



Canvas

figures 4.4,4.5: reference samples of red pigments.

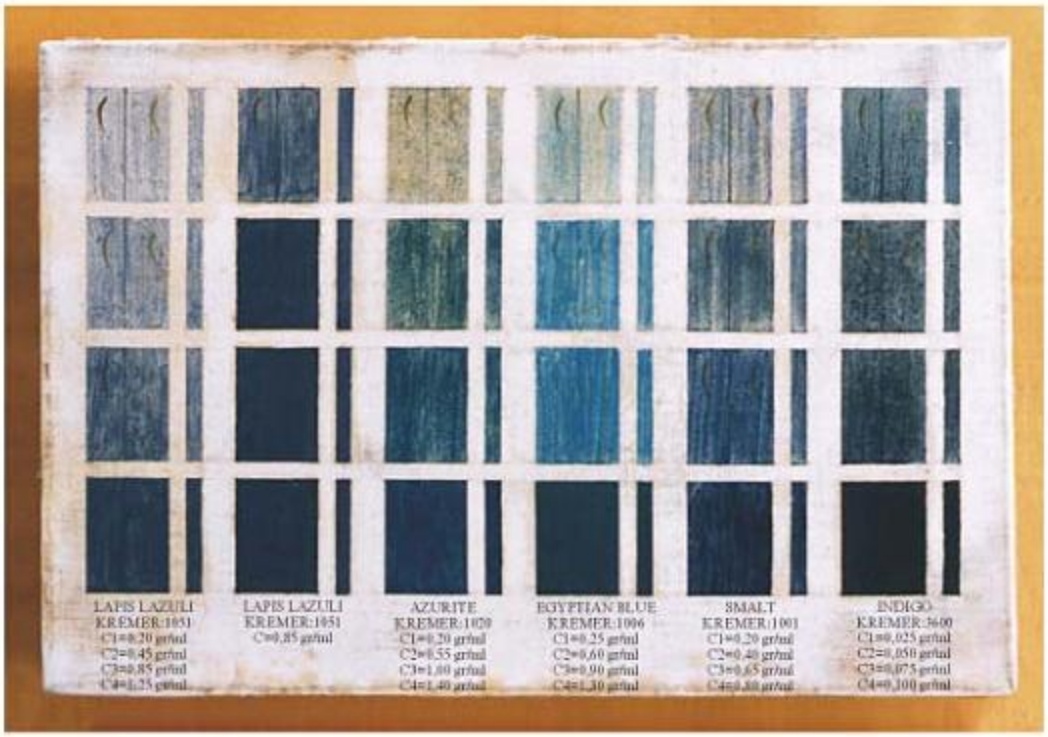
Wood plate



Canvas

figures 4.6,4.7: reference samples of blue pigments.

Wood plate



Canvas

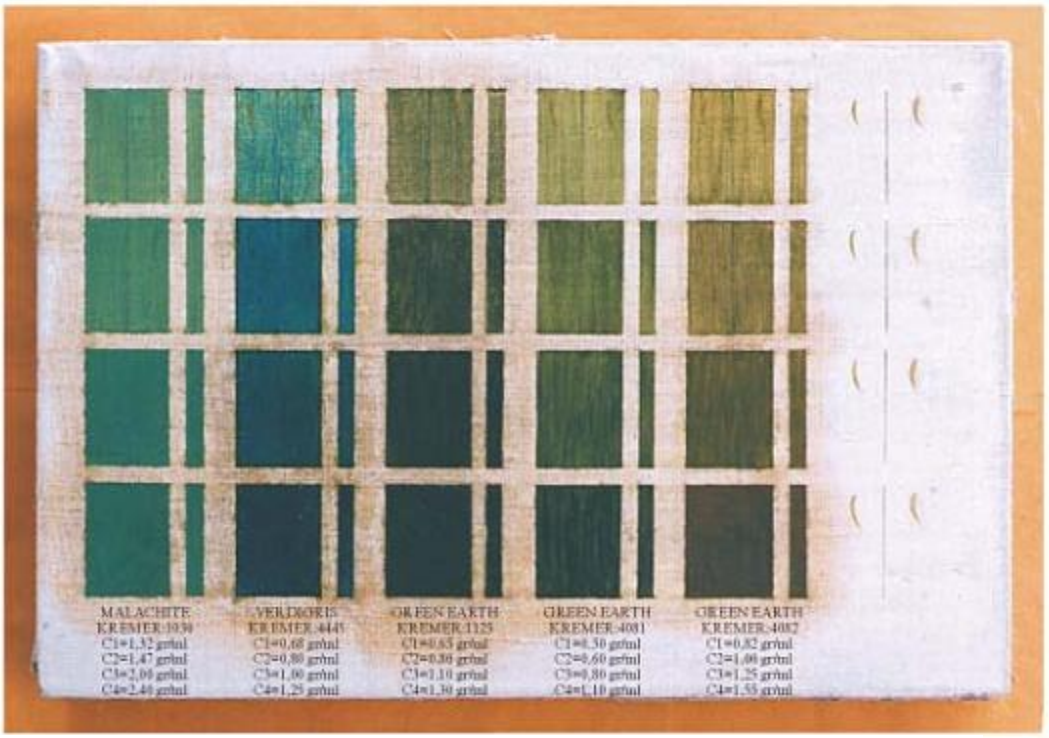
figures 4.8,4.9: reference samples of ochre pigments.

Wood plate

Canvas

figures 4.10,4.11: reference samples of green pigments.

Wood plate



Canvas

2. Spectral imaging with Mu.SIS was performed at 320-1550 nm both for pigments (figures 4.12) and for several areas of El Greco's Concert of Angels (figure 4.13).

figure 4.12: example of spectral imaging of red egg pigments at various wavelengths.

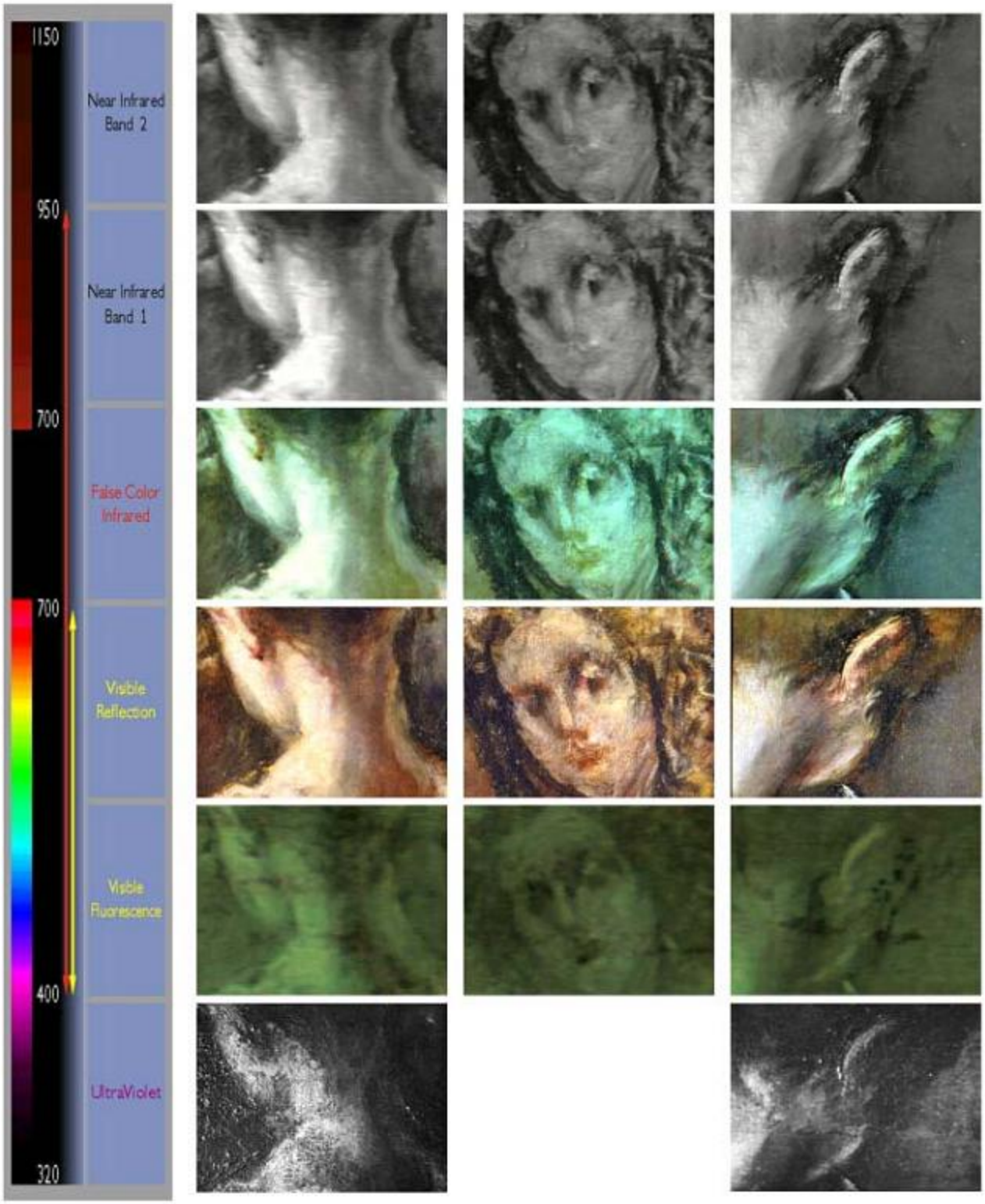
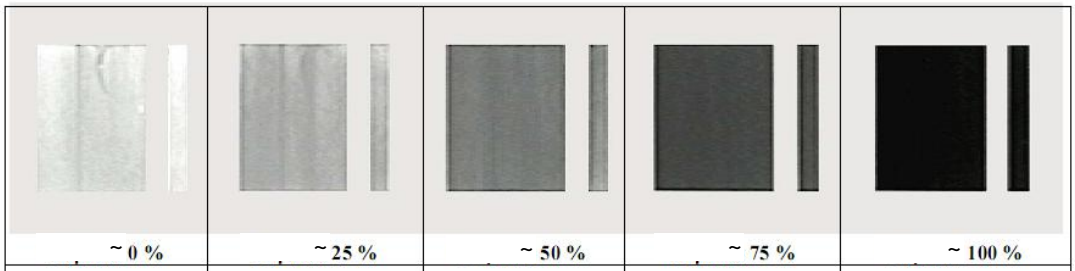


figure 4.13: example of spectral imaging of an upper left portion from El Greco's painting.

3. After processing reference color pigments with an spectrometer, an absorption profile (table 4.14) based on near infrared reflectance was developed, as at these wavelengths, pigments exhibited the most discrimination. Each color pigment presented a specific absorption behavior at different spectral ranges and with pseudocolor information from selected (N)IR bands, enabled the identification of unknown pigments on Angels Concert [37]. For example, the green cloth from the second angel from the left, exposed definite spectral similarities with Malachite's reference color pigment. In particular, spectral images sampled from different areas of the green cloth, displayed blue-green pseudocolor, 100% absorption at IR bands and low percentage of fluorescence at Ultraviolet, which matched quite well with malachite's spectral profile. Further identification of the rest sample areas across the painting was aided by appropriate software.



~ 0 %	~ 25 %	~ 50 %	~ 75 %	~ 100 %
Lead White Calsite Realgar Minio Red lake light Lead-tin yellow Naples yellow	Cinnabar (manufactured) Cinnabar Red lake dark Carmine lake Alizarin Crimson lake Smalt Indigo Massicot Gold Ochre Spain Gold Ochre Italy Yellow Ochre Raw Sienna Monte Amiata Burnt Sienna Monte Amiata	Hematite	Lapis Lazuli Egyptian Blue Green Earth Boemia Green Earth Italy Raw Umber	Bone black Plant black (vine black) Pyrolusite Azurite Burnt Umber Malachite Verdigris Green Earth

Figure 4.14: % absorption at near infrared of reference samples

The results of the above non-destructive analysis for all cases clearly shown that Mu.SIS could detect, identify and map the distribution of A&M materials—both original and added, based on their spectral characteristics, in a strictly non-destructive way. It was possible to differentiate and identify A&H materials with similar coloration but of different chemical nature, by simply tuning the imaging wavelength, inspecting the narrow band images of both A&H object and material models and comparing their reflectance (or fluorescence) characteristics. Also, pigment identification of El Greco's Concert of Angels was possible by comparing sample

areas with reference absorption profiles based on reflectance (or fluorescence) responses at selected spectral ranges.

4.3 Experimental methods and results

Previous works on using multispectral imaging for pigment identification [38],[39] involved applying spectroscopic methods (Vidicon system, CCD camera combined with a liquid crystal tunable filter) on a limited set of sample materials and then perform a classification or segmentation task. Based on observations and classification results, both studies proposed general guidelines for the application to an actual work of art. In particular, PCI score plots of spectral responses, used for discriminating a limited set of egg tempera pigments [38]. Preliminary results on Luca Signorelli's "Predella della Trinita" showed that imaging spectroscopy in combination with PCA is a useful methodology for detecting zones of paintings characterized by different chemical composition or different physical properties. LDA and FCM clustering was used [39] for discriminating four substances with differing near-IR spectra (graphite, blue pen ink, indocyanine green and dysprosium chloride). Moreover, LDA classification of the spectroscopic imaging data from a 16th century drawing (Winnipeg Art Gallery archival photograph of Untitled Hamlet with Bridge, Mountains in Background) revealed regions of the drawing where small amounts of the ink or its decomposition products remained after an unsuccessful cleaning attempt.

Facilitating further the development of a spectral database for color pigments and to improve MuSIS HS' diagnostic capabilities in non-destructive analysis of artworks, several classification and segmentation algorithms have been tested with El Greco's oil & egg reference pigments. Isodata, Maximum Likelihood, Expectation Maximization, Spectral Angle Mapper, Spectral Correlation Mapper, Spectral Information Divergence, Normalized Euclidean and Spectral Gradient Angle 's performance evaluated with the sample library of El Greco's oil and egg pigments.

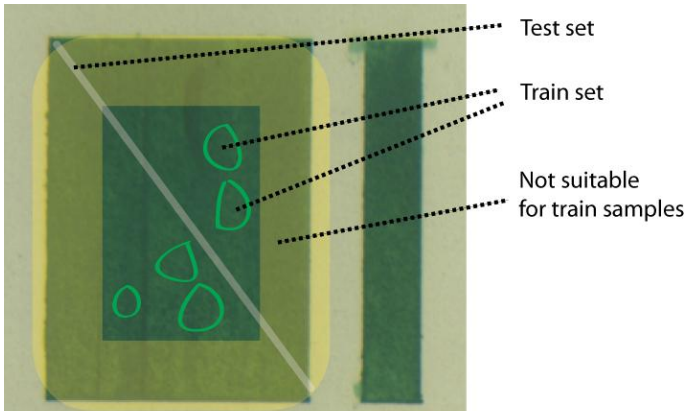


Figure 4.15: possible pigments pools for assembling a suitable train set

Spectral imaging acquisition was performed with MuSIS HS, in the spectral range 400–1000 nm with 20 nm tuning step. Spectral cubes with 1600x1200 resolution obtained after calibration for the 10 reference wooden plates and canvas. For each cube, spectral images co-registered, saved and transformed from reflectance to intensity values with MuSIS software for further classification processing. The classification process involved assigning each pixel to one of a list of classes. Thus, one must set up an exhaustive list of classes, so that there is a logical class to which

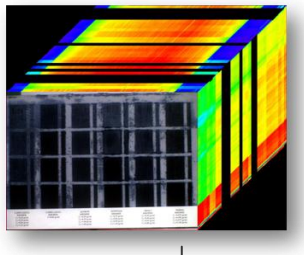
to assign each pixel of the data set, even though one may be interested in only one or a small number of classes in the scene. All of the 34x4 different sample pigments have been assigned as classes, plus 4x2 classes for canvas and wooden plate. For the maximum likelihood classifier, a key requirement is to have the training data size for each class equal to from 10-30 times the number of features (Mather, 1999). The required training set size may therefore be large, and acquiring such training sets may be difficult where a large number of classes is involved or utilizing data acquired in many wavebands. In our case, a training sample was composed as a vector of 30 intensity values in the spectral range [400, 420, 440, … ,980, 1000] nm. Each training sample represents a pixel from the reference images. A systematic sampling approach was developed with the aid of a gui application (figure 4.16), rather than random sampling of pigments, in order to deliver sufficient classification accuracy. Specifically, samples near the centre of the painted area were most suitable to built train sets for each pigment (figure 4.15). Also test sets were extracted from a diagonal area (3/7 ratio with train set approximately). Validation of the algorithms is demonstrated with pseudocolor maps for every color and every tone separately. EM algorithm tends to converge to a singular solution, when training data is (nearly) insufficient as in the case of oil pigments, where it was not possible to collect enough samples to train the classifier. For Spectral Angle Mapper and the rest distance metrics, 4 samples were selected for each color (one for each tone) as reference vectors for computing distances with all the other pixels. After thresholding the minimum distance (or angle in most cases) a pseudocolor map displayed the classification results. For every algorithm, confusion matrixes and Kappa statistics were obtained. Establishing an appropriate train set and therefore a list of classes of informational value, exhaustive and separable, several discriminality measures were utilized such as Bhattacharyya distance, Relative Spectral Discriminality Power, Relative Spectral Discriminality rate and Relative Spectral Discriminality Entropy.

The Classification process is described in the following diagram (Classification Process for Spectral Database):

Spectral aquitition

Intensity transform

Image registration



K - means

Isodata

Thematic Maps

Data extraction – Feature Selection

Pool of 30- dimensional vectors

Test set

Train set

ML

EM

SAM

SID

SCM

NEUC

SGA

Confusion Matrix
Kappa Statistics

Class Discrimination tests (Bhattacharyya distance, Relative Spectral Discriminality rate and Relative Spectral Discriminality Entropy)

Build Classifier

thresholding

Thematic Maps

Spectral aquitition

Intensity transform

Image registration

PCI

Classification Process for Spectral Database

1 *Spectral acquisition of reference pigments (MuSIS)*
2 *Image registration / Intensity transform (MuSIS software) (unsupervised classification)*
3a *K-means/Isodata*
4a *Thematic Maps (supervised classification)*
3b *Data extraction*
4b *train set (Bh. Dist., RSDR, RSDE) / test set (7/3)*
5b *ML, EM and distance metrics classification*
6b *Confusion matrix / Kappa*
7b *Thematic (pseudocolor) maps*

Classification Process for El Greco's paintings

1 *Spectral acquisition of reference pigments (MuSIS)*
2 *Image registration / Intensity transform (MuSIS software)*
3a *PCI*
4a *False color images with principal components (supervised classification)*
3b *train set from Spectral Database*
4b *ML, SAM, SID, SCM classification*
5b *Thematic (pseudocolor) maps*

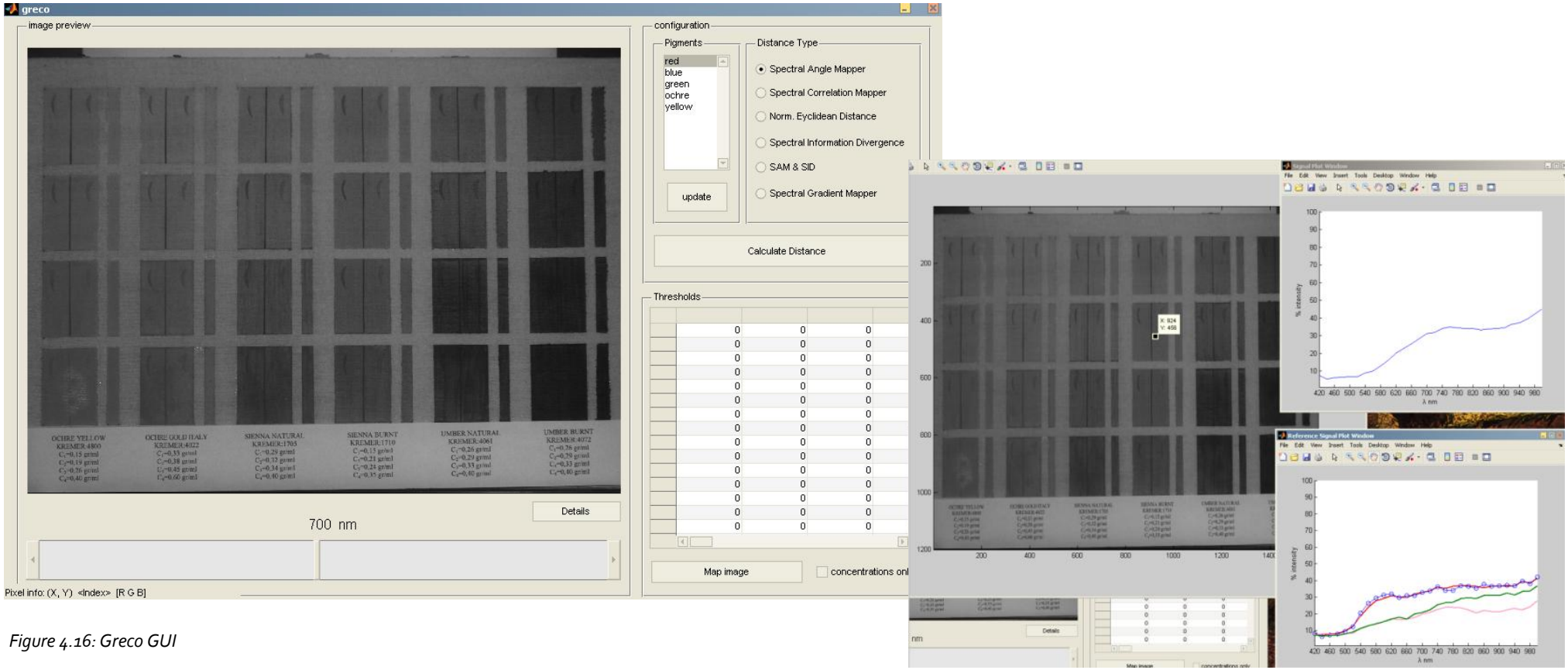


Figure 4.16: Greco GUI

4.3.1 k-means/Isodata clustering

Several parameters with isodata algorithm have been tested both for oil and egg (appendix A). Best results acquired with the following parameters:

- Number of classes: 20(min)-32(max)
- Maximum iterations: 10
- Change Threshold %: 5
- Minimum # Pixel in Class: 1
- Maximum Class (std) deviation: 0.5
- Minimum Class Distance: 2

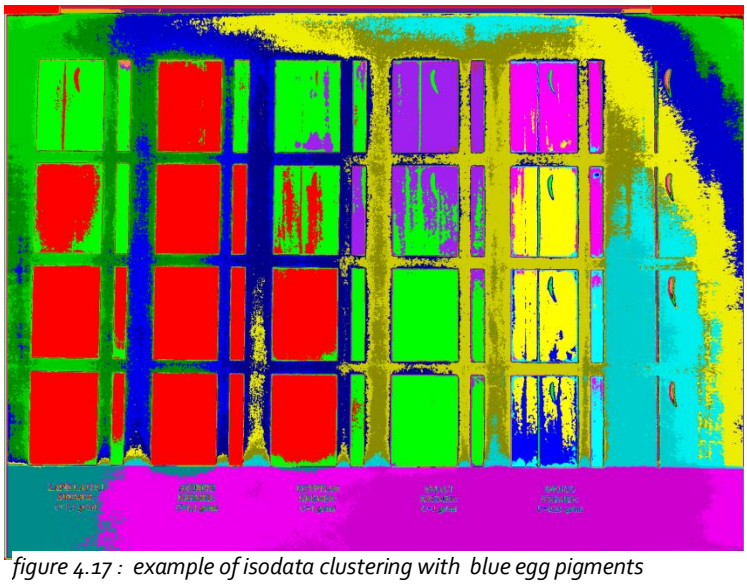


figure 4.17 : example of isodata clustering with blue egg pigments

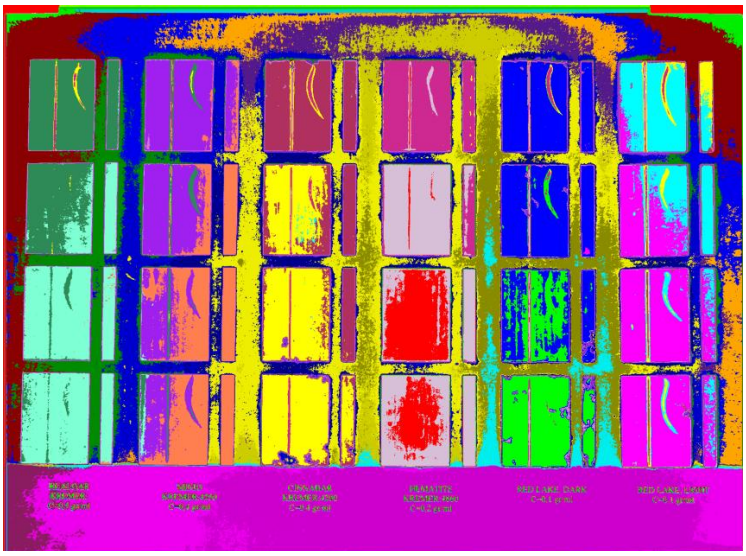


figure 4.18: example of isodata clustering with red egg pigments

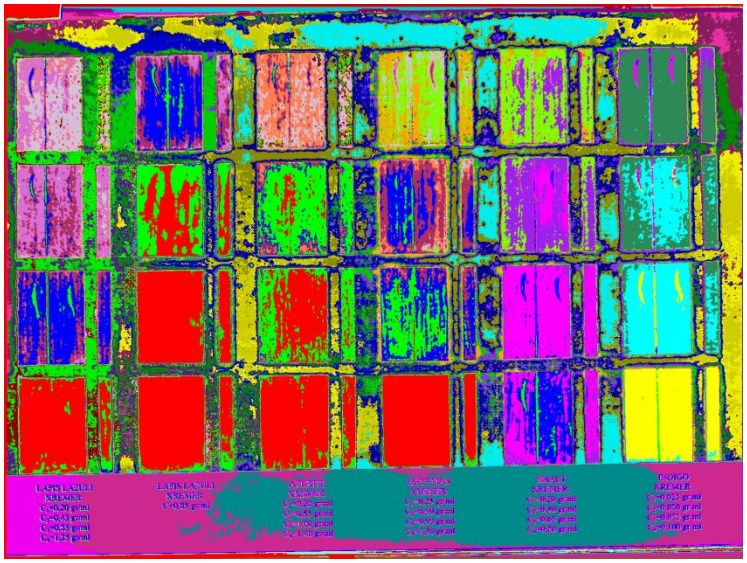


figure 4.19: example of isodata clustering with blue oil pigments

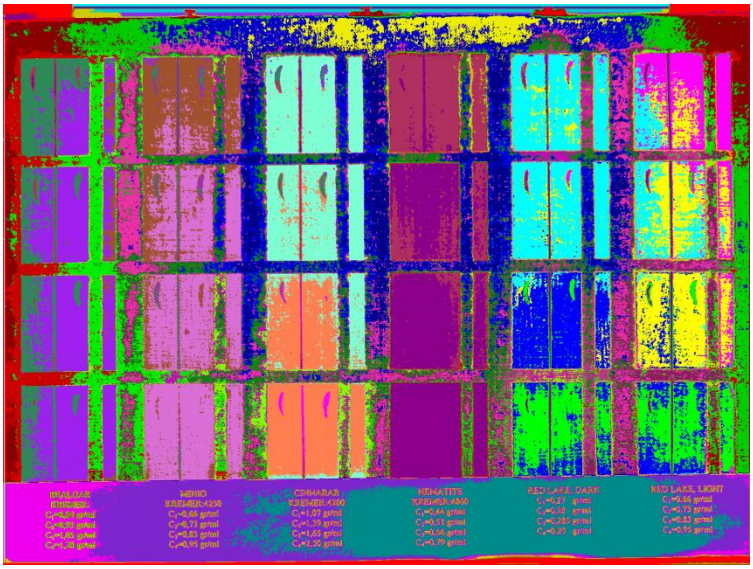


figure 4.20: example of isodata clustering with red oil pigments

4.3.2 Maximum Likelihood results

Each reference canvas or wooden plate was tested with ML classifier. As the number of egg and oil pigments, including different concentrations, corresponds to >90 classes, 4 (egg) or 5 (oil) pseudocolor maps (thematic maps) used to display classification accuracy [Appendix E].Also in Appendix D are listed all the confusion matrixes for every color set separately. Bhattacharyya distance computed for testing class separability [Appendix B]. Sampling of classes for gathering different train sets, repeated as many times required to obtain adequate classification accuracy.

Classification example for red egg pigments (one color for every 4 concentrations of each class):

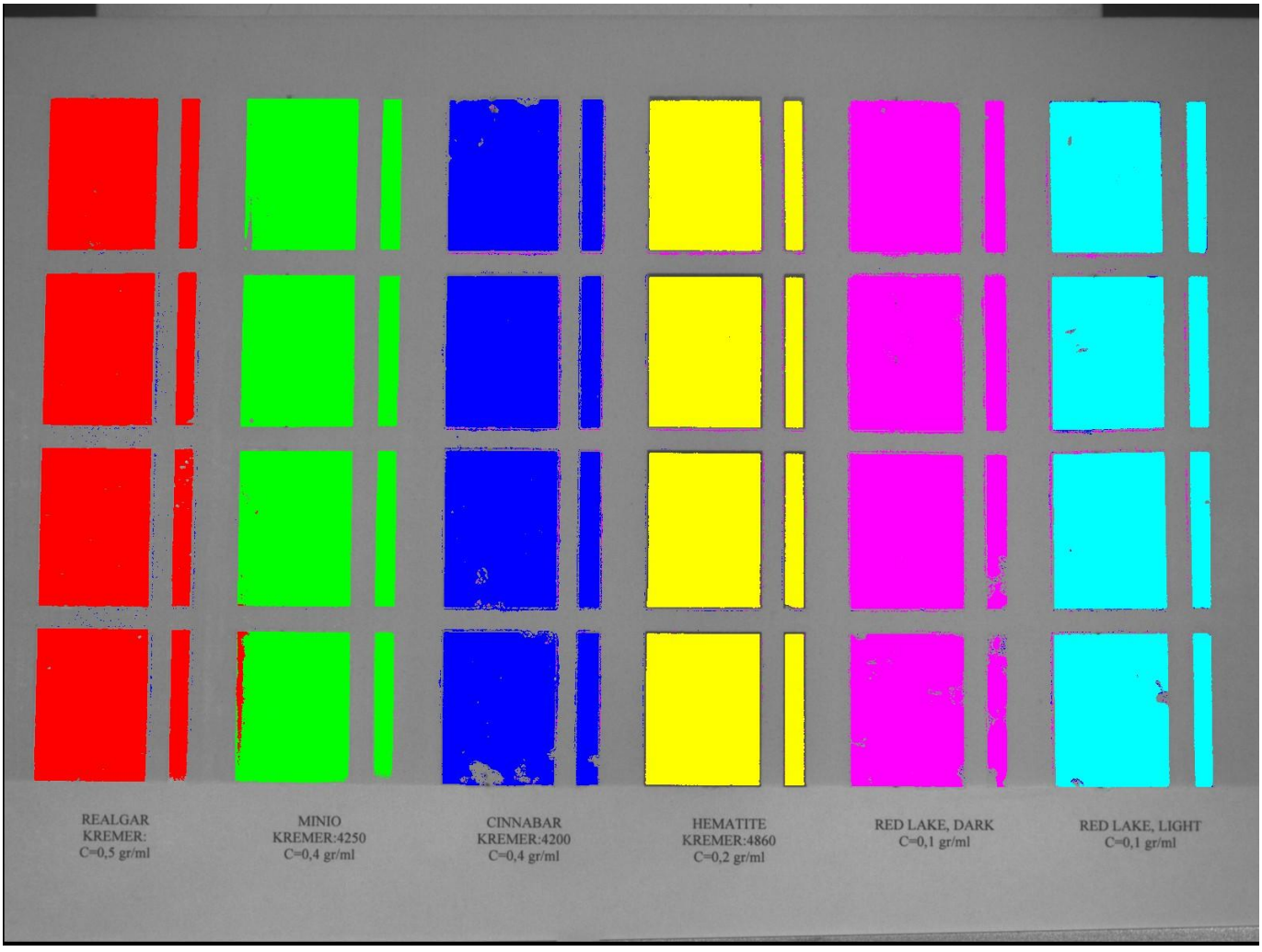


figure 4.21*: (ML) Pseudocolor map of red pigments reference wooden plate with red color classes.*

red	<->	Realgar Kremer:4250
green	<->	Minio Kremer:4200
blue	<->	Cinnabar Kremer:4860
yellow	<->	Hematite Kremer:4860
magenta	<->	Red Lake, Dark
cyan	<->	Red Lake, Light

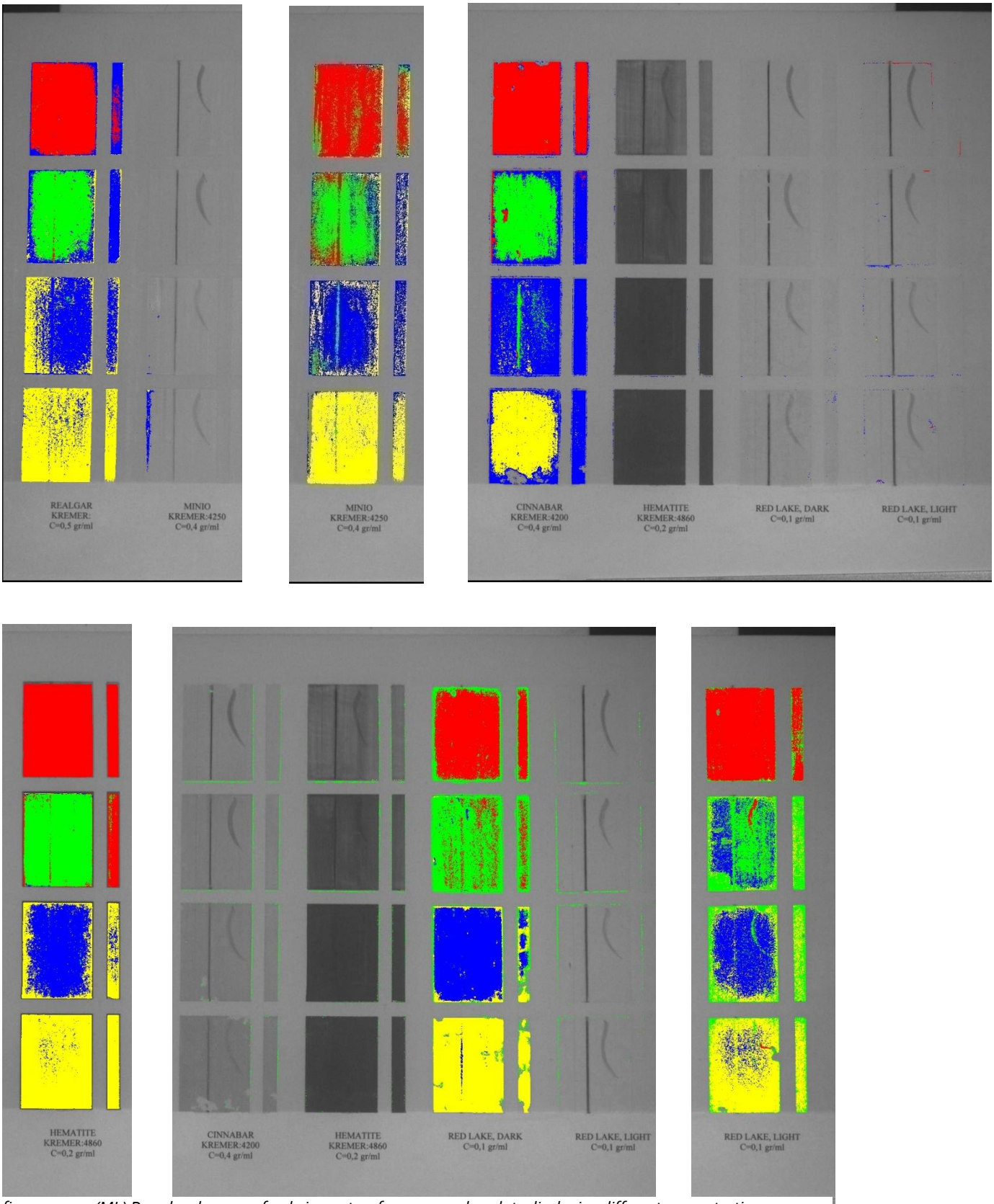


figure 4.22: (ML) Pseudocolor map of red pigments reference wooden plate displaying different concentrations

lighter < ... < ... < darker

red < green < blue < yellow

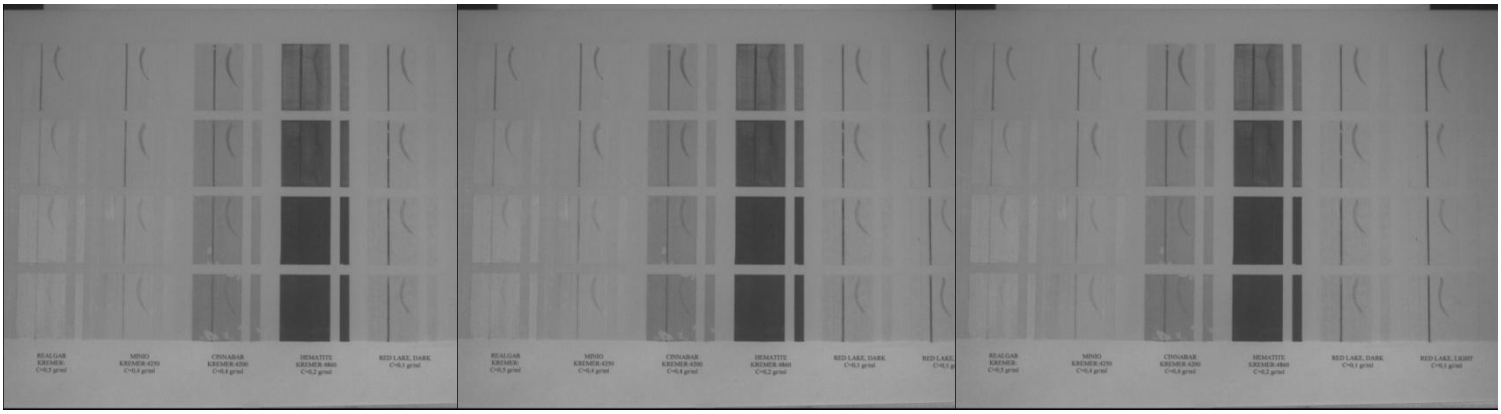


figure 4.23 (ML) Pseudocolor map of red pigments reference wooden plate with blue/green/ochre color classes. Maximum Likelihood classifies only red pigments on red reference plate without mixing with other color pigments.

(norm.) Bhattacharyya distance-reds	REALGAR KREMER	MINIO KREMER	CINNABAR KREMER	HEMATITE KREMER	RED LAKE,DARK	RED LAKE,LIGHT
REALGAR KREMER	0	0,004242	0,022344	0,710386	0,024293	0,001534
MINIO KREMER		0	0,051674	0,88392	0,056983	0,00486
CINNABAR KREMER			0	0,471107	0,02476	0,036856
HEMATITE KREMER				0	0,77888	1
RED LAKE,DARK					0	0,026796
RED LAKE,LIGHT						0

Table 4.24: (norm.) Bhattacharyya distance of red pigments

Red (egg) pigments	REALGAR KREMER	MINIO KREMER	CINNABAR KREMER	HEMATITE KREMER	RED LAKE,DARK	RED LAKE,LIGHT	*Users (%)*
REALGAR KREMER	6944	0	0	0	0	0	**100**
MINIO KREMER	0	4924	0	0	0	0	**100**
CINNABAR KREMER	0	0	5246	0	0	0	**100**
HEMATITE KREMER	0	0	0	7175	0	0	**100**
RED LAKE,DARK	0	0	0	0	4587	0	**100**
RED LAKE,LIGHT	0	0	0	0	12	4508	**99,73451**
Producers (%)	**100**	**100**	**100**	**100**	**99,73907**	**100**	**98,0815**
COHEN KAPPA	kappa value : 0,999566		Variance : 6,16E-06		z (k/sqrt(var)) : 402,5909		

Classification example for oil egg pigments:

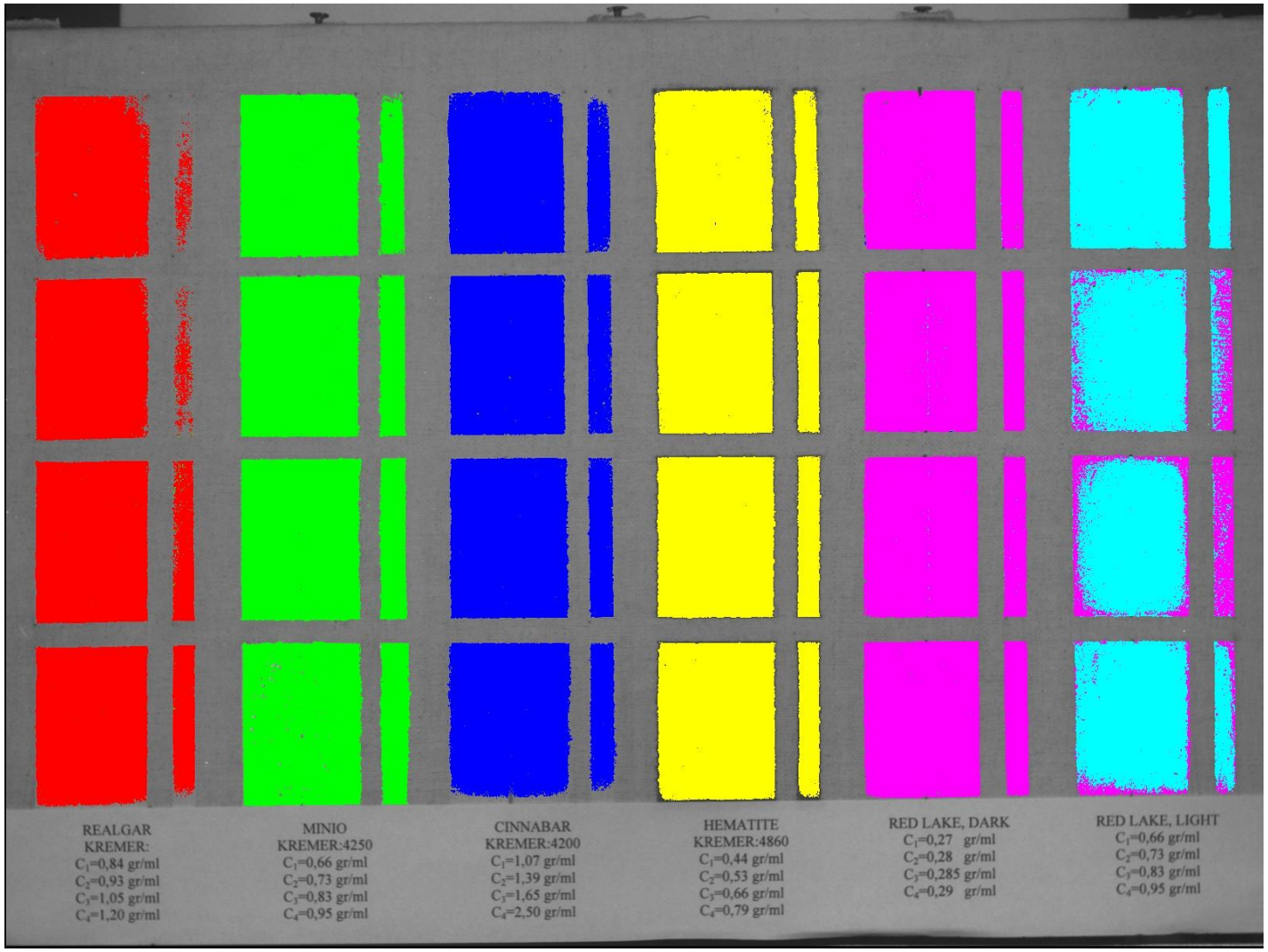


figure 4.25 *(ML) Pseudocolor map of red pigments reference canvas with red color classes.*

red <-> Realgar Kremer

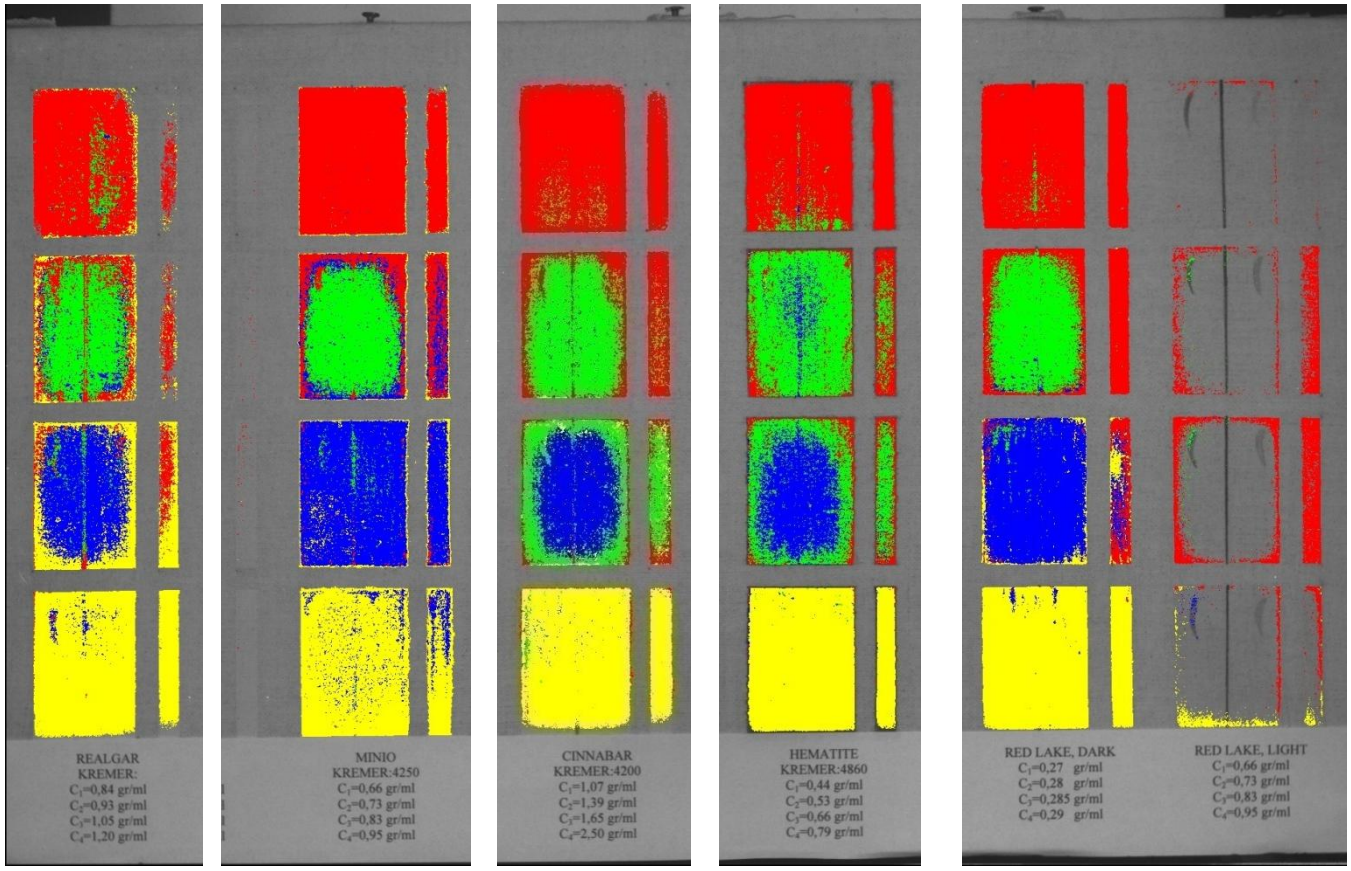
green <-> Minio Kremer:4250

blue <-> Cinnabar Kremer:4200

yellow <-> Hematite Kremer:4860

magenta <-> Red Lake, Dark

cyan <-> Red Lake, Light



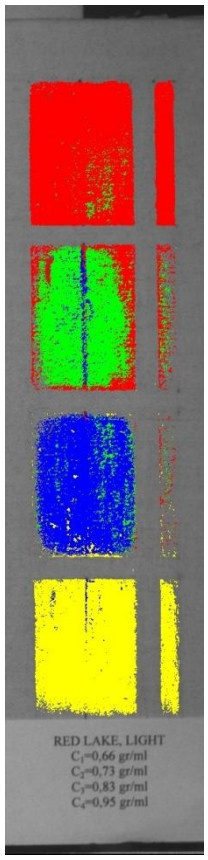


figure 4.26: (ML) Pseudocolor map of red pigments reference canvas displaying different concentrations

lighter < … < … < darker

red < green < blue < yellow

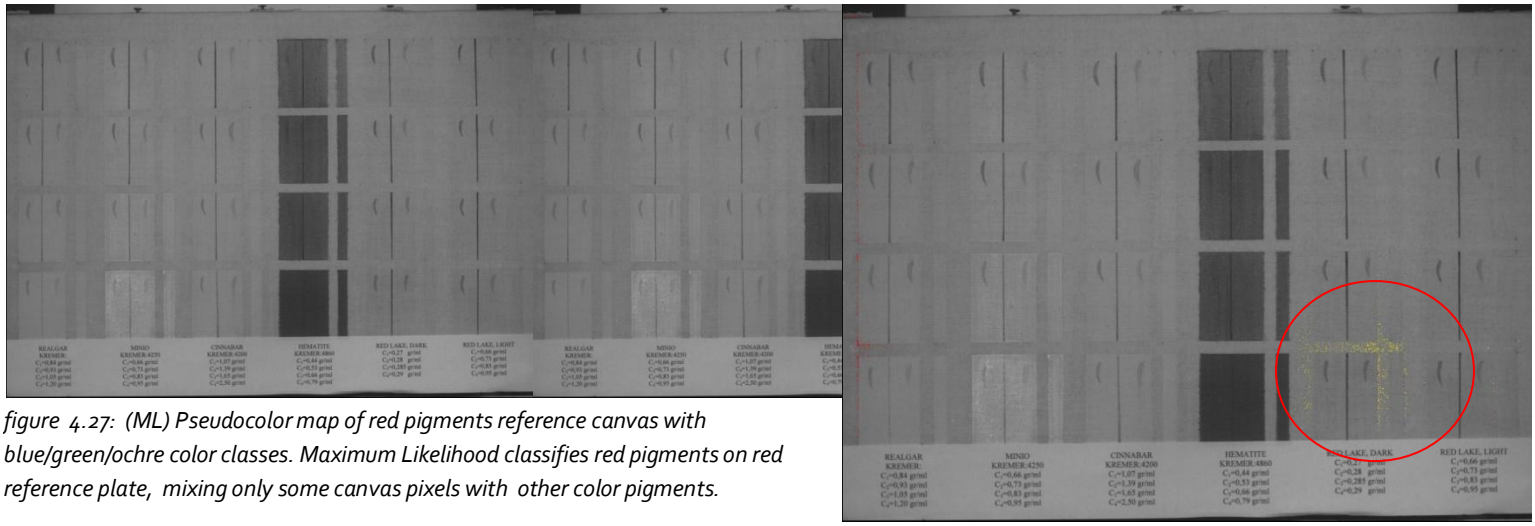


figure 4.27: (ML) Pseudocolor map of red pigments reference canvas with blue/green/ochre color classes. Maximum Likelihood classifies red pigments on red reference plate, mixing only some canvas pixels with other color pigments.

(norm.) Bhattacharyya distance-reds	REALGAR KREMER	MINIO KREMER(4250)	CINNABAR KREMER(4200)	HEMATITE KREMER(4860)	RED LAKE,DARK	RED LAKE,LIGHT
REALGAR KREMER	0	0,051686	0,019017	0,357114	0,024157	0,029058
MINIO KREMER(4250)		0	0,012609	1	0,212152	0,228739
CINNABAR KREMER(4200)			0	0,736931	0,10731	0,116339
HEMATITE KREMER(4860)				0	0,44611	0,347163
RED LAKE,DARK					0	0,002689
RED LAKE,LIGHT						0

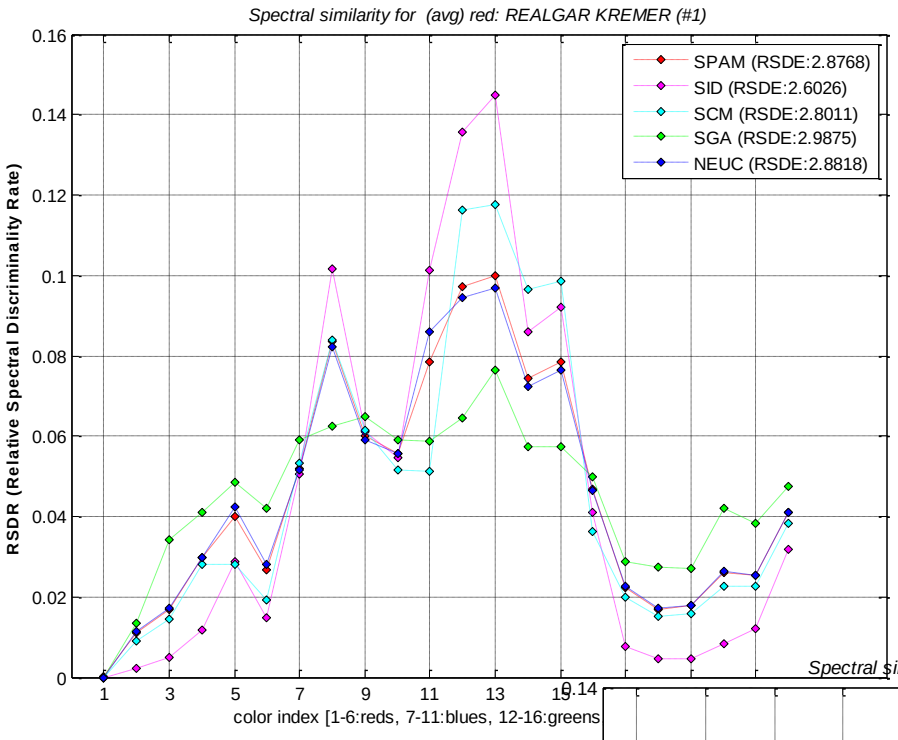
Table 4.28: (norm.) Bhattacharyya distance of red oil pigments

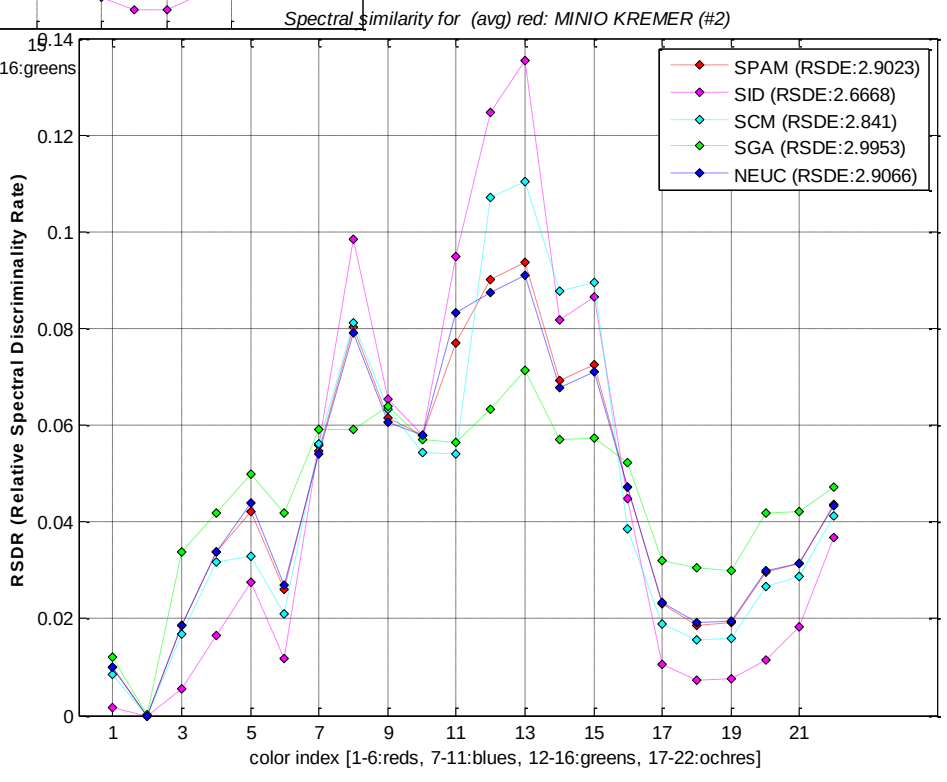
Red (egg) pigments	REALGAR KREMER	MINIO KREMER(4250)	CINNABAR KREMER(4200)	HEMATITE KREMER(4860)	RED LAKE,DARK	RED LAKE,LIGHT	*Users (%)*
REALGAR KREMER	5865	2	0	0	0	0	**99,96591**
MINIO KREMER(4250)	0	7086	0	0	0	0	**100**
CINNABAR KREMER(4200)	0	0	4371	0	0	0	**100**
HEMATITE KREMER(4860)	0	0	0	5868	0	0	**100**
RED LAKE,DARK	0	0	2	0	2148	3	**99,76777**
RED LAKE,LIGHT	0	0	0	0	153	4103	**96,40508**
Producers (%)	**100**	**99,97178**	**99,95426**	**100**	**93,35072**	**99,92694**	**97,19894**
COHEN KAPPA	kappa value :	0,993382	Variance :	7,36E-06	z (k/sqrt(var)) :	366,0446	

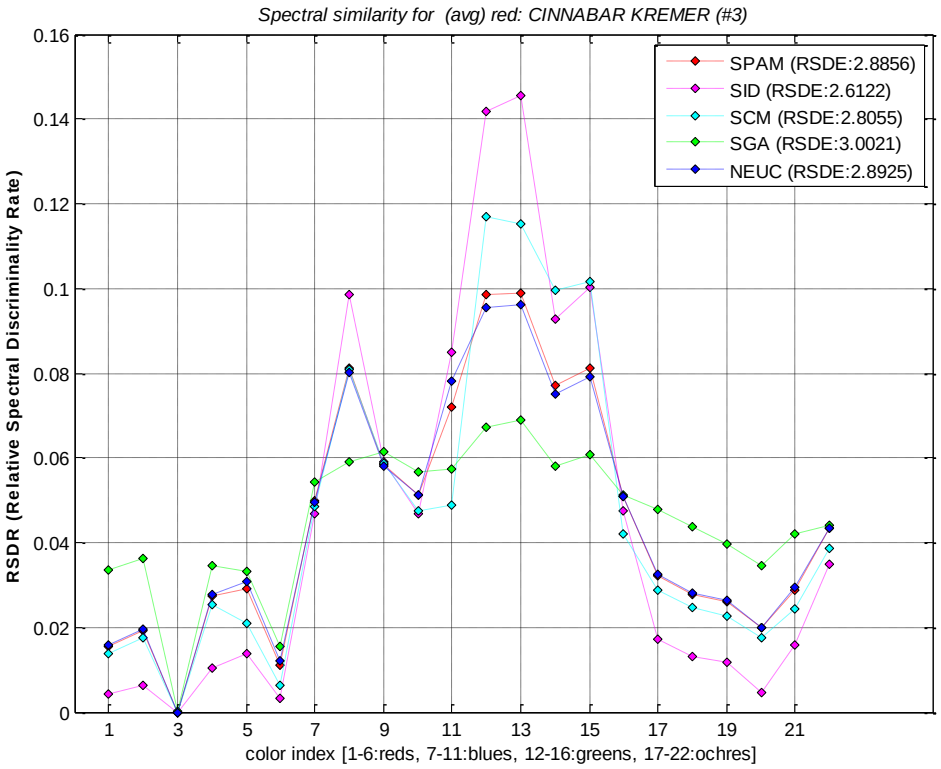
4.3.3 Distance metrics results

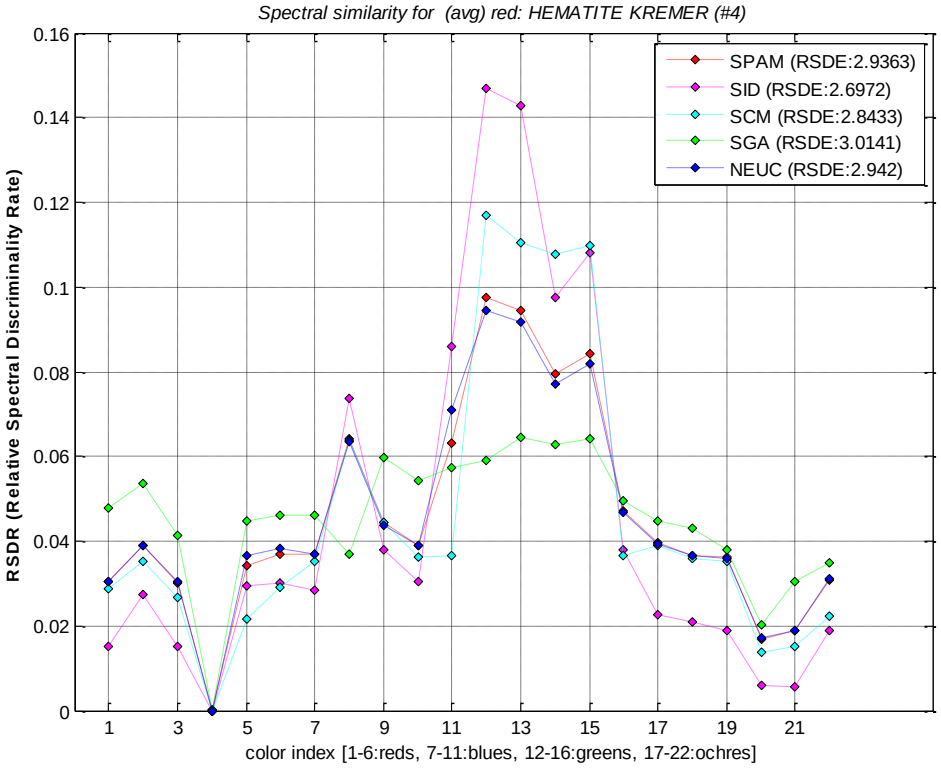
Each reference canvas or wooden plate was tested with SPAM, SCM, NEUC, SGA, SID distance metrics and pseudocolor maps used to display classification accuracy for different color sets separately [Appendix E] and confusion matrixes [Appendix D] . Also, RSDR and RSDE computed for testing classes separability [Appendix C]. The goal was to achieve classification accuracy close to maximum likelihood results with the fewest possible reference vectors. For representing each pigment, 4 vectors were selected for a color class (one for each concentration) [Appendix F].

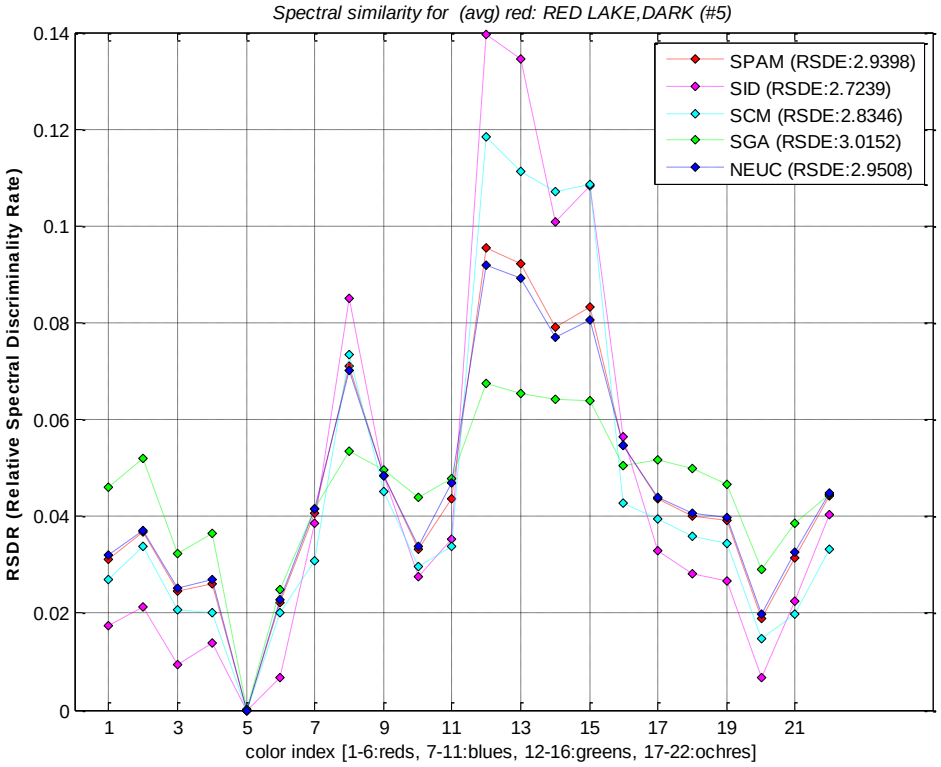
- Classification example for red egg pigments:






Spectral similarity for (avg) red: CINNABAR KREMER (#3)
SPAM (RSDE:2.8856)
SID (RSDE:2.6122)
SCM (RSDE:2.8055)
SGA (RSDE:3.0021)
NEUC (RSDE:2.8925)
RSDR (Relative Spectral Discriminality Rate)
color index [1-6:reds, 7-11:blues, 12-16:greens, 17-22:ochres]


Spectral similarity for (avg) red: HEMATITE KREMER (#4)
SPAM (RSDE:2.9363)
SID (RSDE:2.6972)
SCM (RSDE:2.8433)
SGA (RSDE:3.0141)
NEUC (RSDE:2.942)
RSDR (Relative Spectral Discriminality Rate)
color index [1-6:reds, 7-11:blues, 12-16:greens, 17-22:ochres]



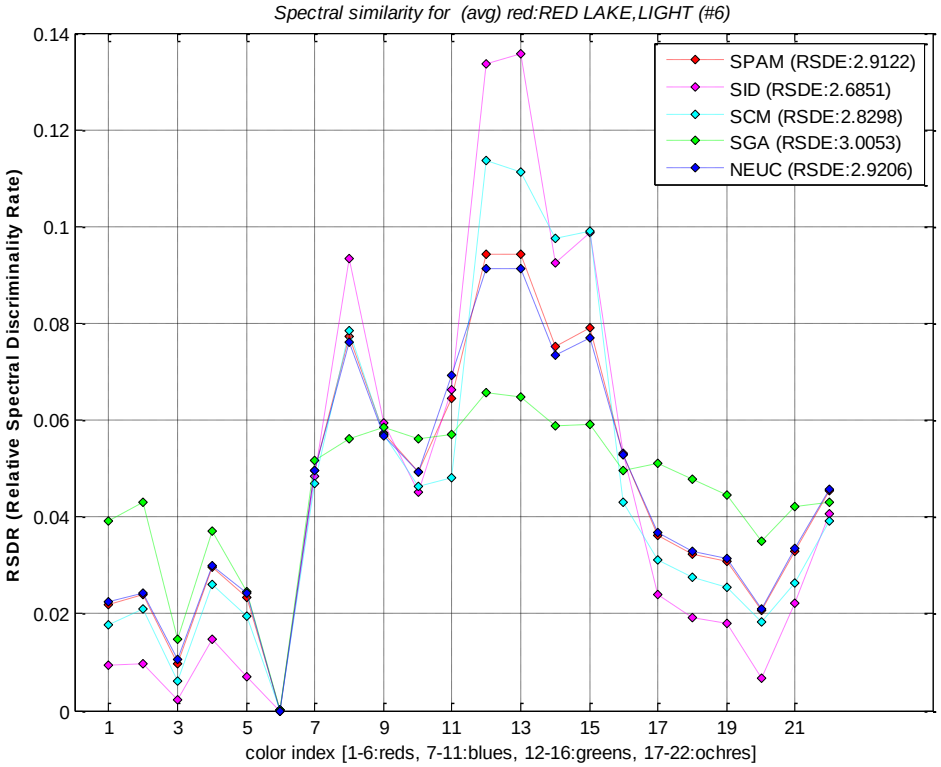


figure 29-34 : Spectral similarity of red reference pigments with the other reference pigments in the train set.

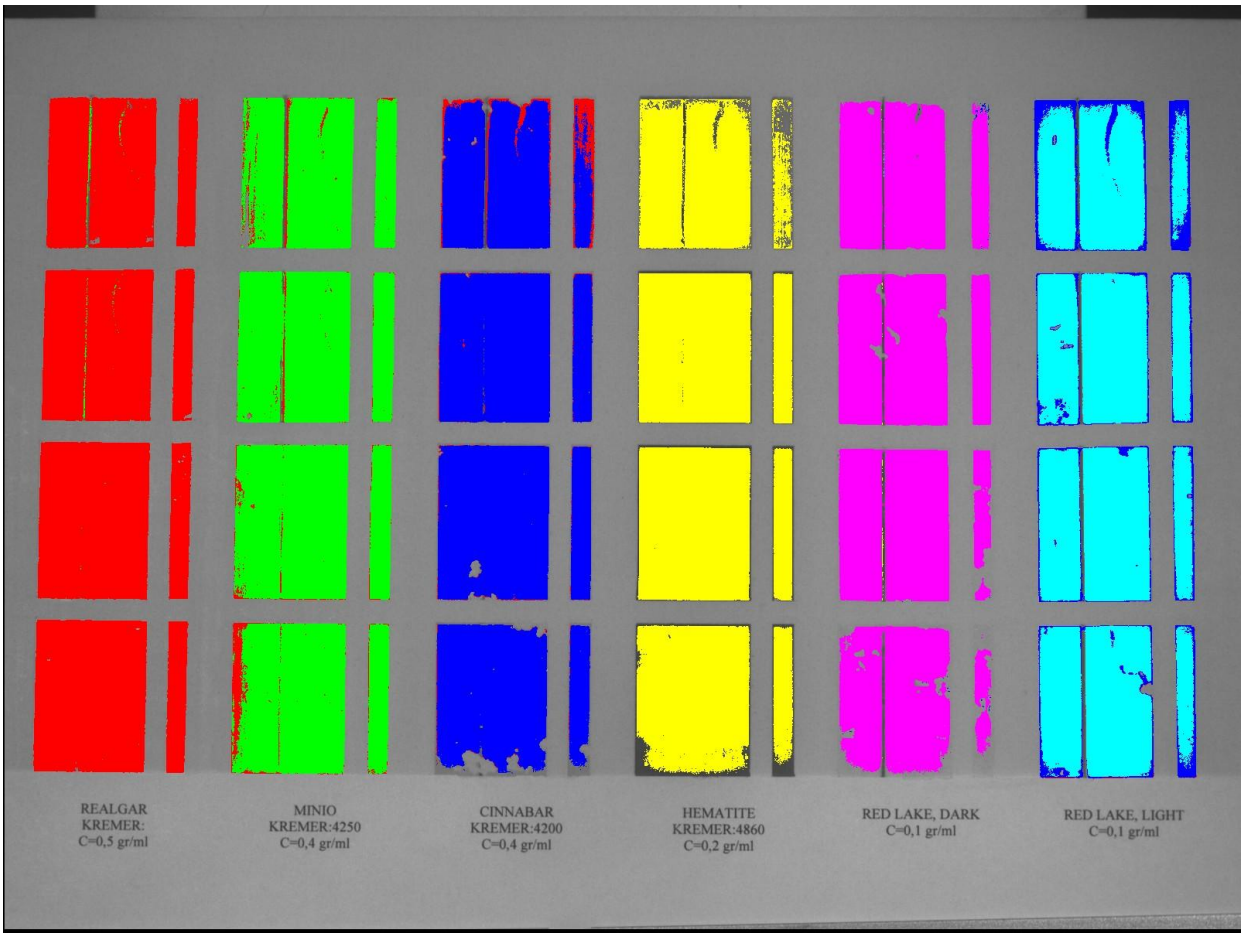


figure 4.35: (NEUC pseudocolor map of red egg pigments)

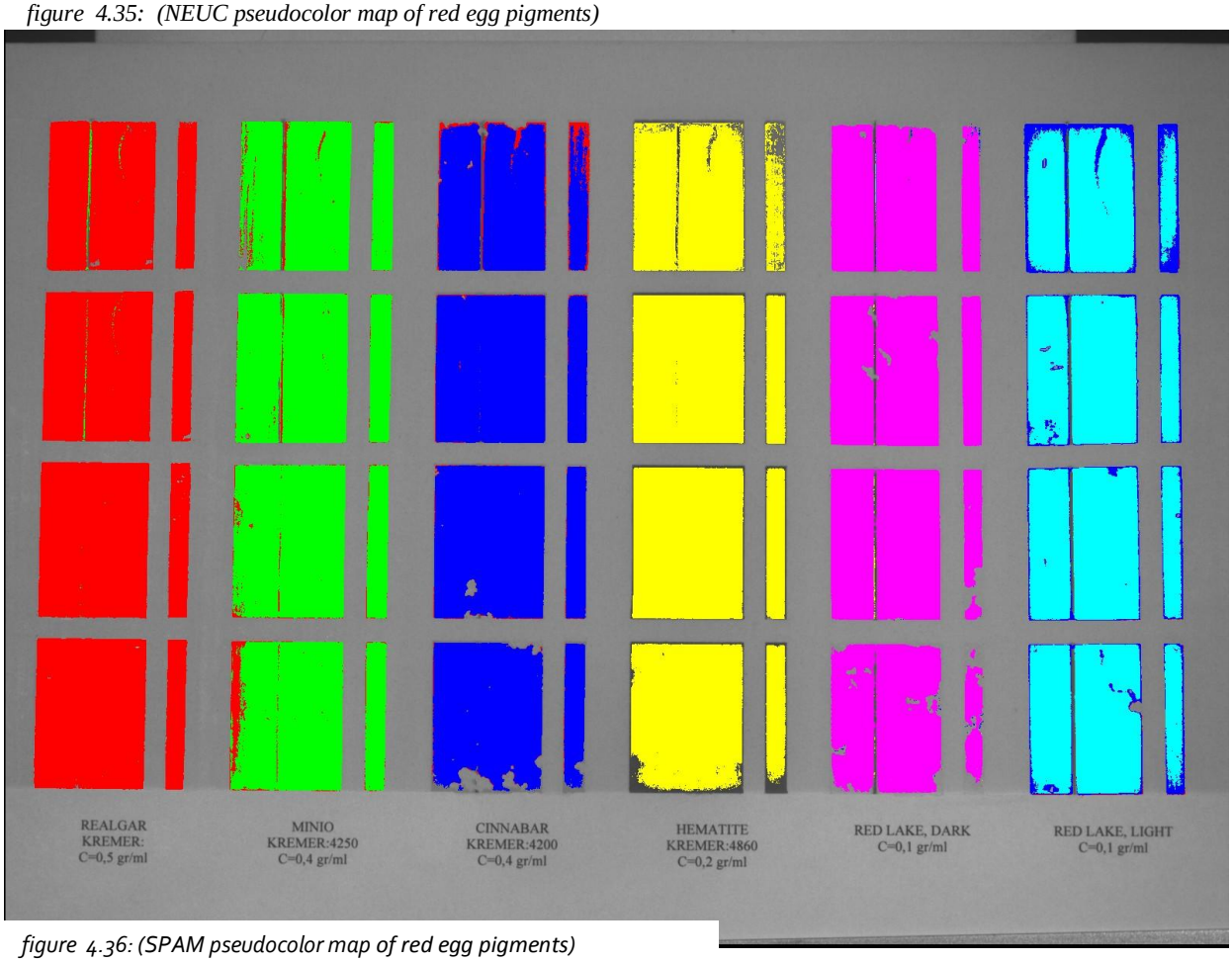


figure 4.36: (SPAM pseudocolor map of red egg pigments)

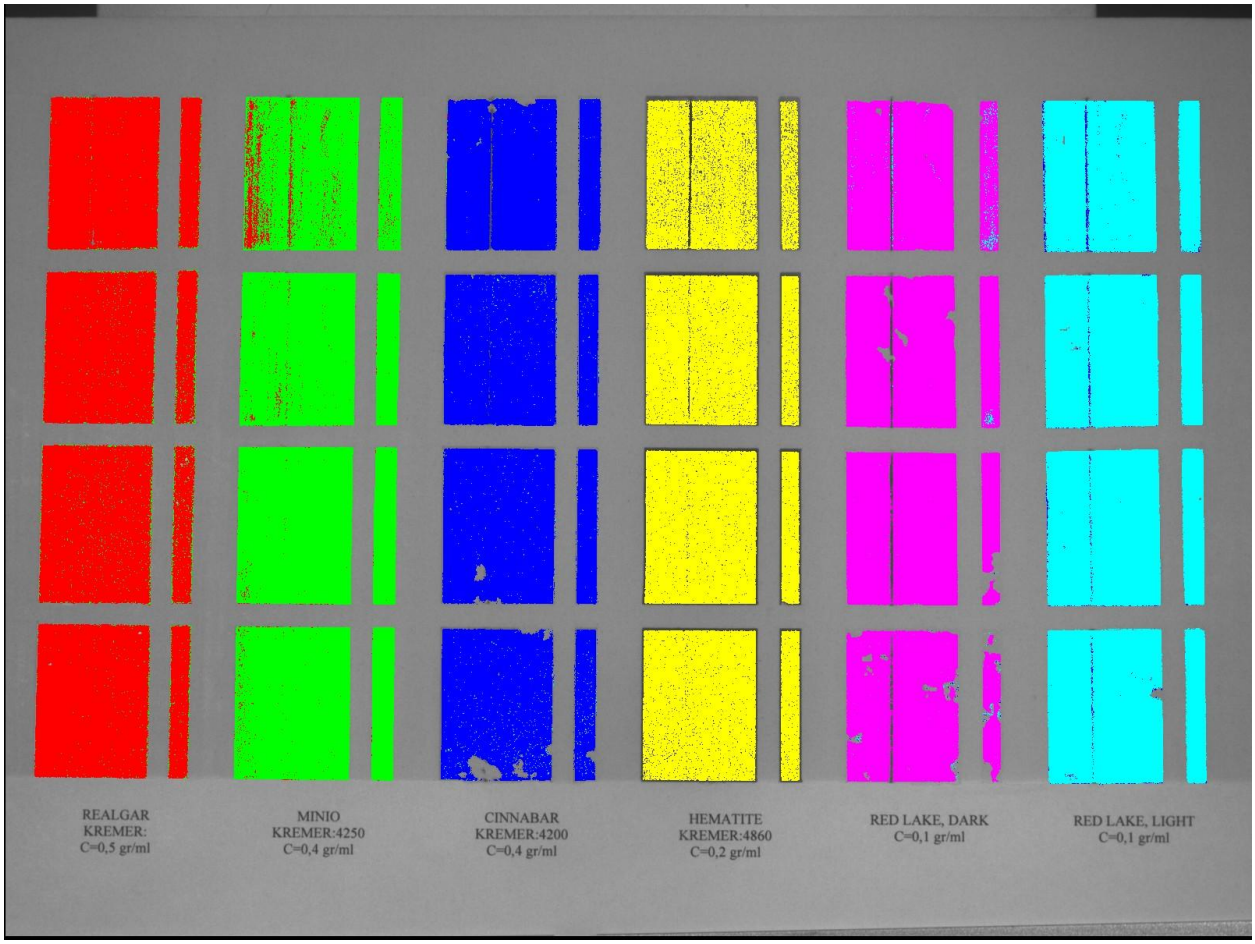


figure 4.37 (SGA pseudocolor map of red egg pigments)

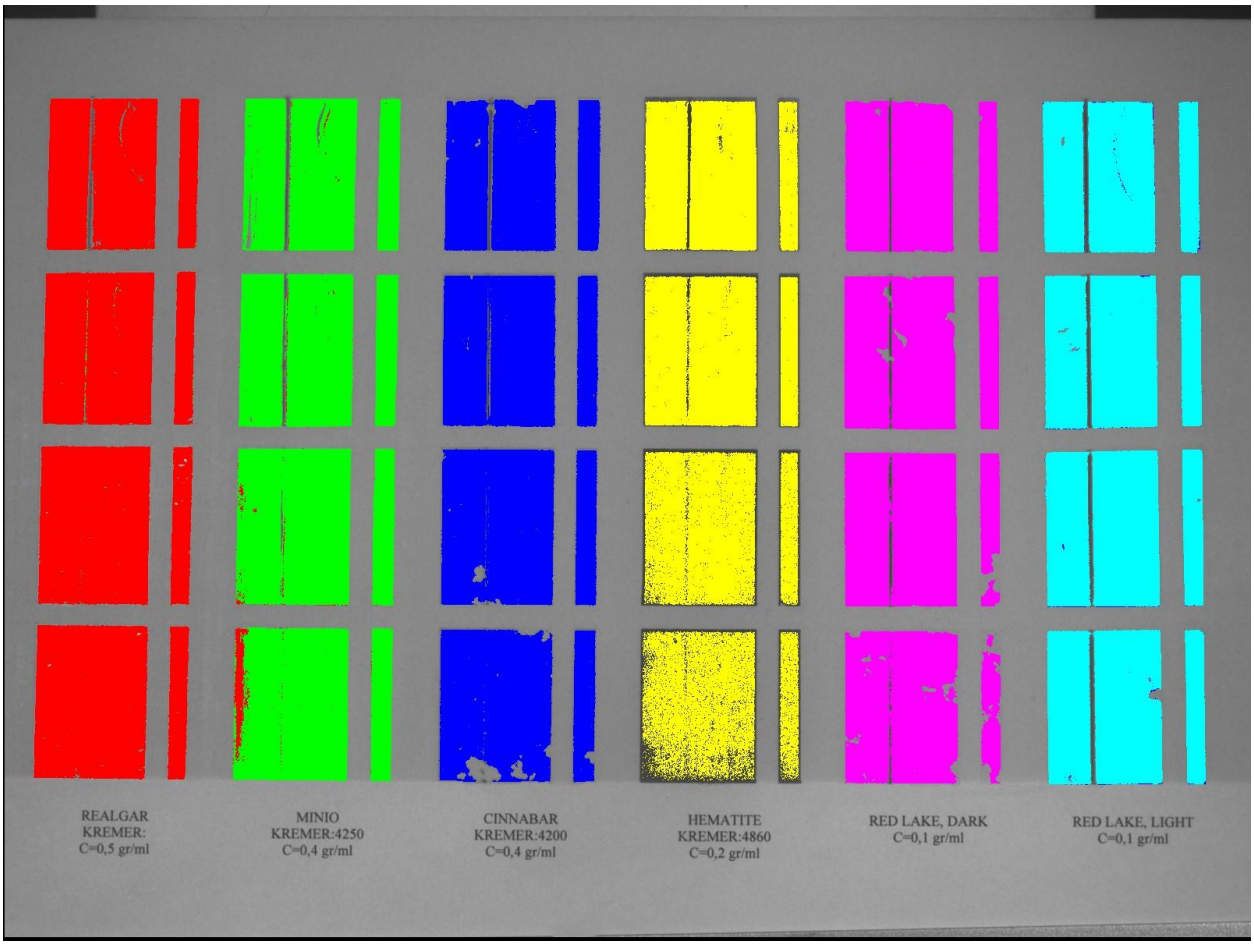


figure 4.38 (SCM pseudocolor map of red egg pigments)

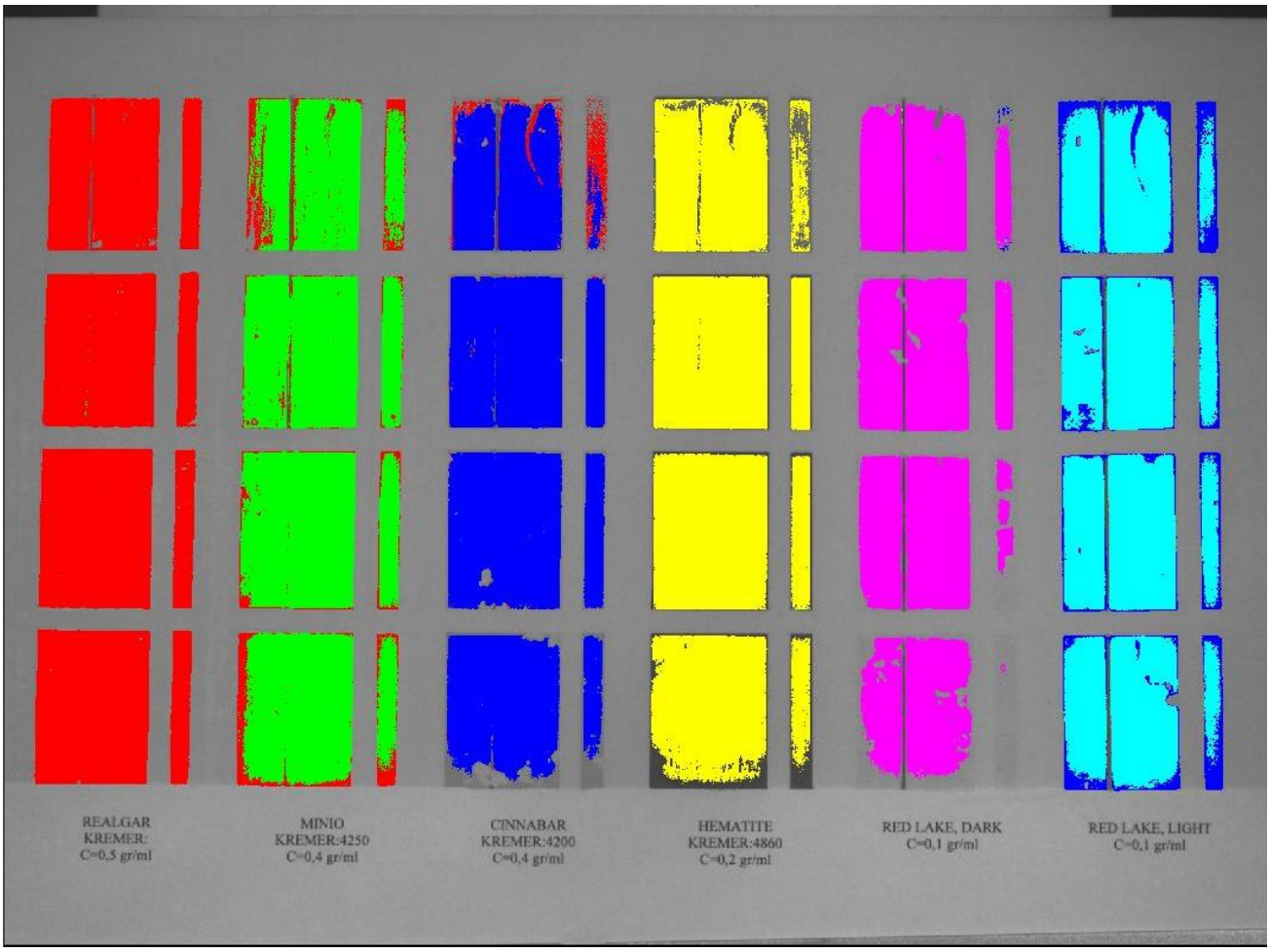


figure 4.39 (SID pseudocolor map of red egg pigments)

Red (egg) pigments SPAM	REALGAR KREMER	MINIO KREMER	CINNABAR KREMER	HEMATITE KREMER	RED LAKE,DARK	RED LAKE,LIGHT	*Users (%)*
REALGAR KREMER	7000	38	0	0	0	0	**99,46007**
MINIO KREMER	144	4847	0	0	0	0	**97,11481**
CINNABAR KREMER	59	0	5213	0	0	0	**98,88088**
HEMATITE KREMER	0	0	0	7080	0	0	**100**
RED LAKE,DARK	0	0	0	9	4593	0	**99,80443**
RED LAKE,LIGHT	5	0	160	0	0	4428	**96,40758**
Producers (%)	**97,11432**	**99,22211**	**97,02215**	**99,87304**	**100**	**100**	**97,42633**
COHEN KAPPA	kappa value : 0,985053		Variance : 6,14E-06		z (k/sqrt(var)) : 397,6817		

Red (egg) pigments SCM	REALGAR KREMER	MINIO KREMER	CINNABAR KREMER	HEMATITE KREMER	RED LAKE,DARK	RED LAKE,LIGHT	*Users (%)*
REALGAR KREMER	7016	62	9	0	0	0	**98,99817**
MINIO KREMER	99	4929	0	0	0	0	**98,03103**
CINNABAR KREMER	1	0	5275	0	0	0	**99,98105**
HEMATITE KREMER	0	0	0	7215	0	0	**100**
RED LAKE,DARK	0	0	0	0	4631	0	**100**
RED LAKE,LIGHT	0	0	9	0	0	4593	**99,80443**
Producers (%)	**98,59472**	**98,75776**	**99,65993**	**100**	**100**	**100**	**98,88944**

COHEN KAPPA kappa value : 0,989764 Variance : 6,11E-06 z (k/sqrt(var)) : 400,5672

Red (egg) pigments - SGA	REALGAR KREMER	MINIO KREMER	CINNABAR KREMER	HEMATITE KREMER	RED LAKE,DARK	RED LAKE,LIGHT	*Users (%)*
REALGAR KREMER	6972	93	0	0	0	0	**98,68365**
MINIO KREMER	121	4893	0	0	0	0	**97,58676**
CINNABAR KREMER	0	0	5247	0	0	29	**99,45034**
HEMATITE KREMER	0	1	0	7036	0	0	**99,98579**
RED LAKE,DARK	0	0	5	7	4628	14	**99,44134**
RED LAKE,LIGHT	0	0	15	0	0	4586	**99,67398**
Producers (%)	**98,29409**	**98,1151**	**99,62028**	**99,90061**	**100**	**99,07107**	**98,01686**

COHEN KAPPA	kappa value : 0,985053	Variance : 6,14E-06	z (k/sqrt(var)) : 397,6817

Red (egg) pigments - SID	REALGAR KREMER	MINIO KREMER	CINNABAR KREMER	HEMATITE KREMER	RED LAKE,DARK	RED LAKE,LIGHT	*Users (%)*
REALGAR KREMER	6979	7	0	35	0	0	**99,40179**
MINIO KREMER	240	4714	0	7	0	0	**95,02117**
CINNABAR KREMER	232	0	5033	3	0	0	**95,5391**
HEMATITE KREMER	0	0	0	6902	0	0	**100**
RED LAKE,DARK	0	0	0	40	4569	0	**99,13213**
RED LAKE,LIGHT	0	0	174	23	0	4397	**95,7118**
Producers (%)	**93,66528**	**99,85173**	**96,65834**	**98,45934**	**100**	**100**	**95,7605**

COHEN KAPPA	kappa value : 0,972402	Variance : 6,19E-06	z (k/sqrt(var)) : 390,9475

Red (egg) pigments NEUC	REALGAR KREMER	MINIO KREMER	CINNABAR KREMER	HEMATITE KREMER	RED LAKE,DARK	RED LAKE,LIGHT	*Users (%)*
REALGAR KREMER	6999	36	0	0	0	0	**99,48827**
MINIO KREMER	156	4830	0	0	0	0	**96,87124**
CINNABAR KREMER	77	0	5195	0	0	0	**98,53945**
HEMATITE KREMER	0	0	0	6996	0	0	**100**
RED LAKE,DARK	0	0	0	9	4592	0	**99,80439**
RED LAKE,LIGHT	8	0	165	1	0	4419	**96,21163**
Producers (%)	**96,67127**	**99,26017**	**96,92164**	**99,85727**	**100**	**100**	**97,04439**

COHEN KAPPA	kappa value : 0,983678	Variance : 6,15E-06	z (k/sqrt(var)) : 396,6698

(Confusion Matrixes for red egg pigments with SPAM, NEUC, SID, SGA, SCM distance metrics)

accuracy	Maximum likelihood		SAM		NEUC		SID		SCM		SGA	
	%	kappa	%	kappa	%	kappa	%	kappa	%	kappa	%	kappa
(egg) red	98,0815	0,99566	97,42633	0,985053	97,4439	0,973678	95,7605	0,972402	98,88944	0,985053	98,01686	0,989764
(egg) blue	99,75482	1	99,95773	0,999464	99,95773	0,999464	99,9535	0,999411	99,92814	0,999089	99,89855	0,998715
(egg) green	99,65648	0,99599	96,42338	0,955708	96,31022	0,954291	95,72433	0,947062	91,71516	0,898689	82,07646	0,783188
(egg) ochre	91,80865	0,905788	79,668	0,788225	78,34062	0,75453	72,95825	0,710483	97,25021	0,967478	74,52322	0,733453
(oil) red	84,19894	0,993382	72,34722	0,882595	70,08237	0,866539	65,05462	0,732788	78,36184	0,907312	78,93096	0,915065
(oil) blue	86,78327	0,86835	76,30148	0,75628	75,38559	0,744692	74,72094	0,742798	81,51693	0,839506	81,57736	0,847937
(oil) green	84,84561	0,877941	75,51325	0,79461	63,11801	0,609101	64,25459	0,616915	68,30366	0,671958	54,7689	0,505582
(oil) ochre	94,389	0,980824	71,20871	0,830218	69,19566	0,819875	63,17699	0,703517	84,12968	0,885817	61,15516	0,633121
(oil) yellow	80,83664	0,837055	79,3176	0,8061	78,98107	0,80451	71,31573	0,791338	79,3102	0,796952	60,43001	0,765396

accuracy	Exp.Max.(2 Comp.)		Exp.Max.(3 Comp.)		Exp.Max.(4 Comp.)	
	%	kappa	%	kappa	%	kappa
(egg) red	97,94399	0,9637	99,28292	0,990618	98,19902	0,979131
(egg) blue	99,9535	1	99,94505	1	99,9535	1
(egg) green	98,39961	0,97991	99,1311	0,989392	99,05432	0,98429
(egg) ochre	97,80143	0,974953	98,47813	0,984055	97,81727	0,976107

Table 4.40: accuracy scores for EM,ML and distance metrics.

The above table summarizes the classification accuracy of all the algorithms used for pigment identification. Maximum Likelihood and Spectral correlation Mapper seem to be efficient enough for classifying pigments as they scored higher than the other algorithms. Probabilistic approach of ML and the ability of SCM to recognize negative correlation –especially near the ends of each sample area- affect classification accuracy both for canvas and wooden plates. However, linseed oil as binding media for oil pigments used in canvas samples increased the difficulty of acquiring train/test sets and thus evaluating algorithms performance, as it expanded beyond the (distinct) boundaries of each color's sample area. On the other hand, egg pigments are more easily classified, because egg yolks compose color materials of higher density that doesn't bleed on wooden plates. Classification rate of distance metrics for egg pigments range from 72% to 99% (kappa: 0.71-0.99), while for oil pigments range from 54% to 83%(kappa: 0.50-0.88). ML appears to be less sensitive to binding media as it scores >80% for both cases. Generally, SAM and NEUC perform similar, while SID and SGA don't appear to improve classification rate, especially for oil pigments. Expectation Maximization provides a minor performance improvement over ML for most pigments, in the expense of time needed to complete a classification task ((average classification time on a AthlonXP 3000+, 2GB memory), EM(3 components):1380.5567 seconds, ML: 955.624581 seconds, SAM: 266.515542 seconds, SCM: 276.694712 seconds, NEUC: 203.890509 seconds, SID: 627.508936 seconds). Thematic maps support these observations, providing an overall aspect of classification results for comparative evaluation of ML, EM and distance metrics. For example, comparing pseudocolor maps of distance metrics of red egg pigments (figures 4.35 – 4.39) with the pseudocolor map of ML (figure 4.21), it is clearly seen that SCM (with only 4 reference vectors per color) approximates ML's performance.

4.3.4 El Greco's paintings

During fall-winter of 2008 the Museum of Cycladic Art, in collaboration with SEACEX (Sociedad Estatal Para la Acción Cultural Exterior), presented an exhibition of El Greco and his workshop. The exhibition focused on El Greco's workshop in Spain and its impact on 17th century Spanish painting, including paintings made by the master himself, his students Jorge Manuel Theotocopoulos (1578-1631) and Luis Tristan (ca. 1585-1624) and other artists who were affiliated with or influenced by his workshop. Two or more versions of each iconographic theme were presented: the original, made by El Greco himself, and copies, produced by his students. The exhibition addressed also the issue of authenticity as El Greco has been extensively forged in the 20th century. Modern technology allows for the detection of fakes, but as some collectors and museums refuse to test their Greco paintings, controversy over the authenticity of a considerable number of works is sustained.

However, employing modern diagnostic methods of non invasive nature could address authenticity issues without destructive tests. Dr. Balas and his scientific team performed spectral imaging with Mu.SIS HS on several paintings of El Greco and his workshop. The acquired spectral cubes combined with the train sets of El

Greco's pigments were tested with several classification algorithms (figure 4.41). However, realistically, aging of the media often leads to sample spectra which deviate considerably from the ideal reference target spectra. Thus, strict spectral library searches may not adequately match regions with the target spectra of a painting. Generally, a supervised classifier containing a set of training spectra that span the variability encountered for the type and age of work will outperform a simple spectral library search. Although our spectral library is not exhausting or complete, classification thematic maps showed definite similarities with the materials used in El Greco's paintings. Also, using spectral discrimination measures on sample areas, it is possible to detect similarities between paintings.

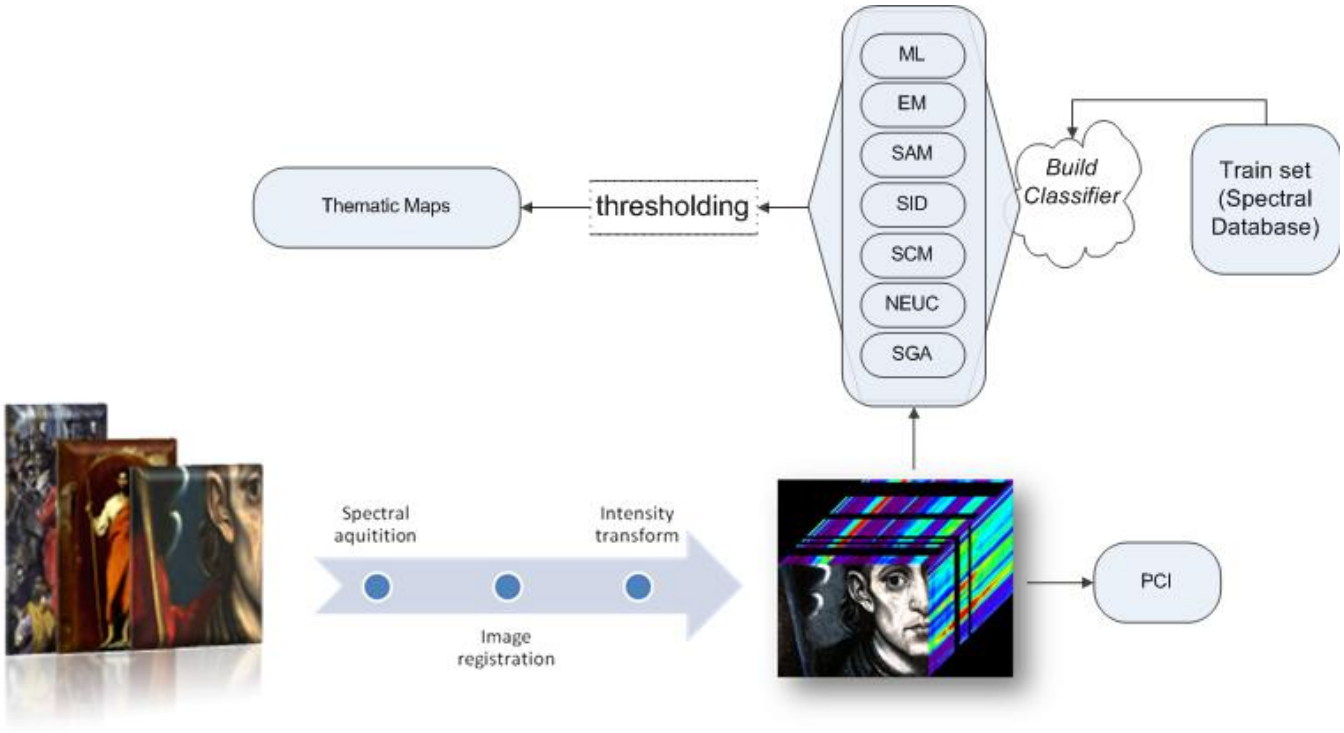


figure 4.41: classification process for El Greco's paintings

Classification Process for El Greco's paintings

1	Spectral acquisition of reference pigments (MuSIS)
2	Image registration / Intensity transform (MuSIS software)
3a	PCI
4a	False color images with principal components (supervised classification)
3b	train set from Spectral Database
4b	ML, SAM, SID, SCM classification
5b	Thematic (pseudocolor) maps

The following classification examples of paintings are demonstrated:

1. *"Saint James, Apostle and Pilgrim", El Greco*
    - *red kirtle detail*
2. *"Espolio", J.M Theotokopoulos*
    - *red kirtle detail*
    - *yellow kirtle detail*
3. *"Saint Louis, King of France", J.M. Theotokopoulos*
    - *face detail*
    - *red kirtle detail*
    - *blue background detail*

Analysis of the spectroscopic imaging data from the three paintings (details) was carried out with Maximum Likelihood classifier and SPAM, SCM, SID distance metrics. Also, PCI falsecolor maps (red for first component, green for second component, blue for third or fourth component) were used either for distinguishing zones which, in the rgb image, appear to be painted with the same pigment or for enhancing image color discrimination. Because of the age of the works, large variations in spectra of the paintings were observed, but in most cases thematic maps agreed on pigment's identity.

In El Greco's Saint James, Apostle and Pilgrim (figure 4.42), the kirtle, according to ML classifier, appears to have been painted with red pigment Minio Kremer (4250) (figure 4.44), while distance metrics recognize two different areas, as PCI image depicts (figure 4.43), with Minio Kremer (4250) and Cinnabar Kremer (4200) pigments.

In J.M. Theotokopoulos' Espolio (figure 4.46) and Saint Louis, King of France (figure 4.50) the darkest areas of the red kirtles appear to have been painted with Red Lake, while the lighter area with Cinnabar Kremer (figure 4.49, figure 4.52). Also, the yellow kirtle on the down right of Espolio with Naples Yellow Kremer (4313) or Massicot Kremer (4301) or most probably a mix of the two. For the details on face of Saint Louis a lead-Tin Kremer (1011) or a Massicot Kremer (4301) was used, while for the blue background both ML and distance metrics agree on Azurite Kremer.

figure 4.42: Saint James, Apostle and Pilgrim, El Greco

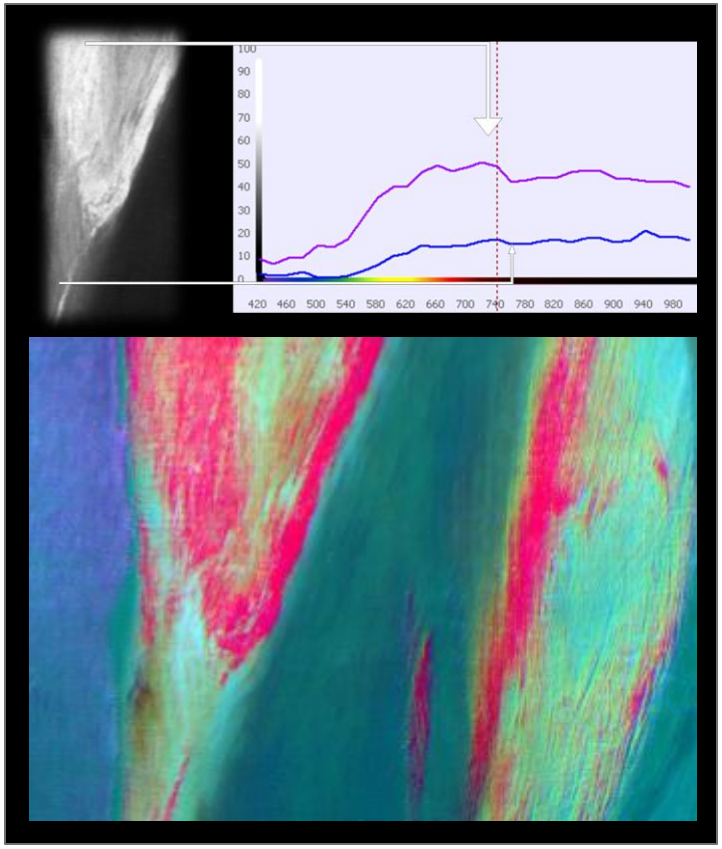


figure 4.43: PCI falsecolor image of kirtle (detail)

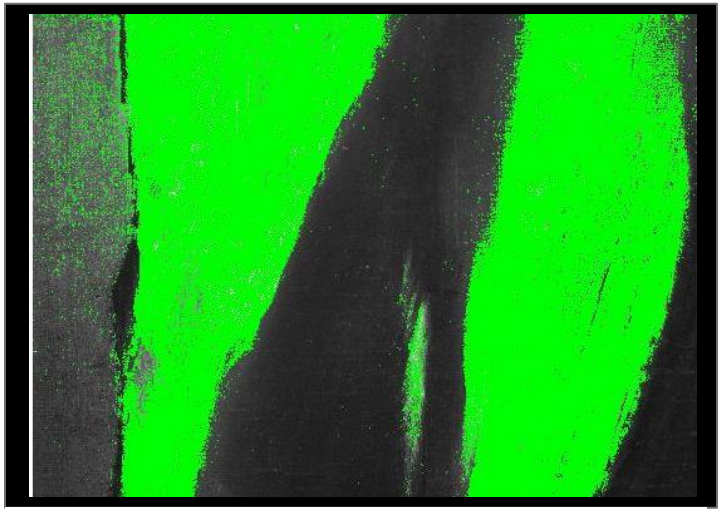


figure 4.44: thematic map of red pigments (ML classifier)

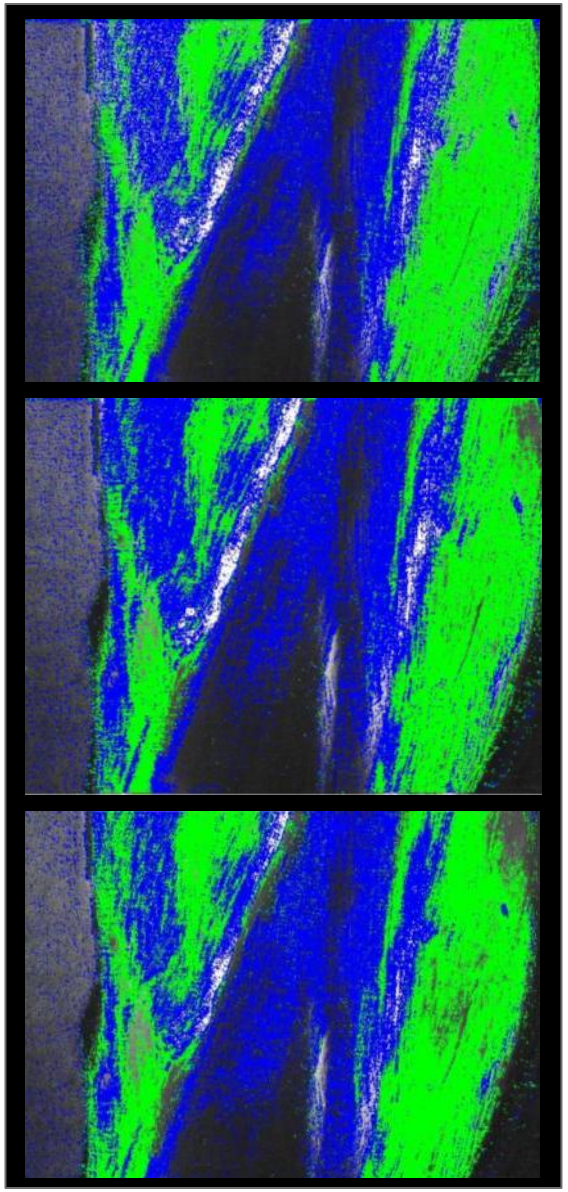


figure 4.45: thematic map of red pigments (SCM/SID/SPAM distance metric)

Color legend

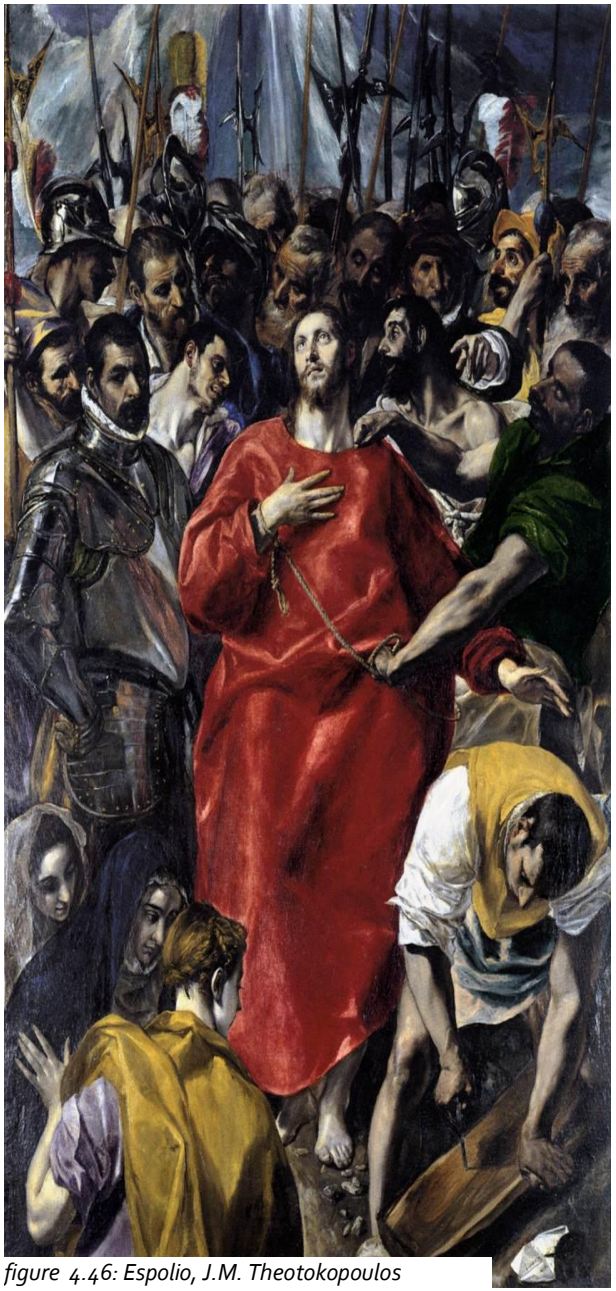


figure 4.46: Espolio, J.M. Theotokopoulos

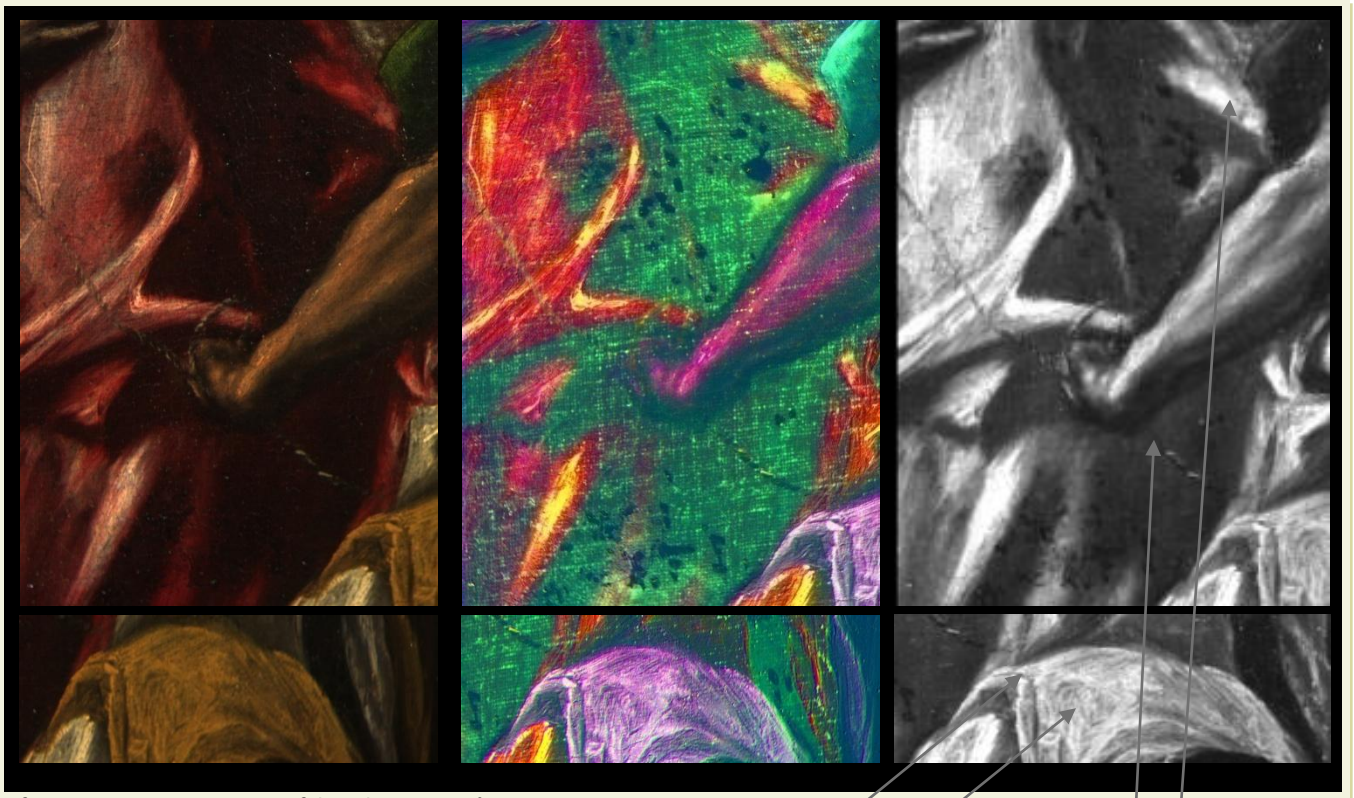


figure 4.47: RGB image / PCI falsecolor image/ (λ=780,760nm) image (details)

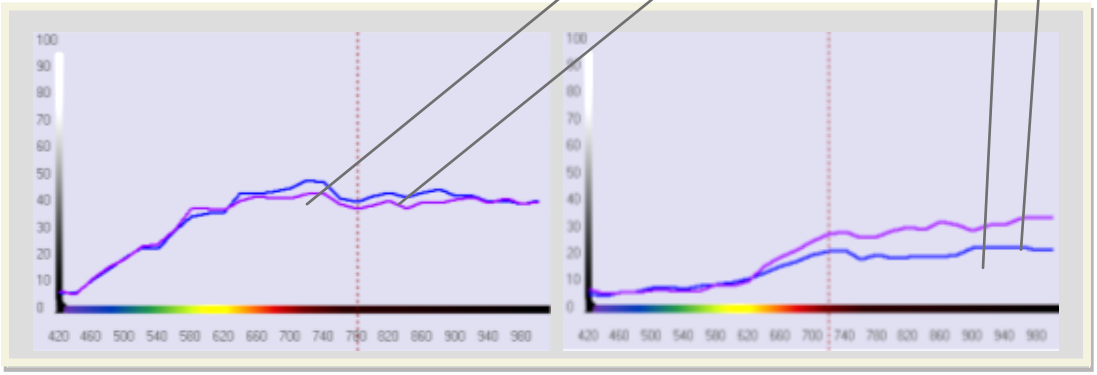
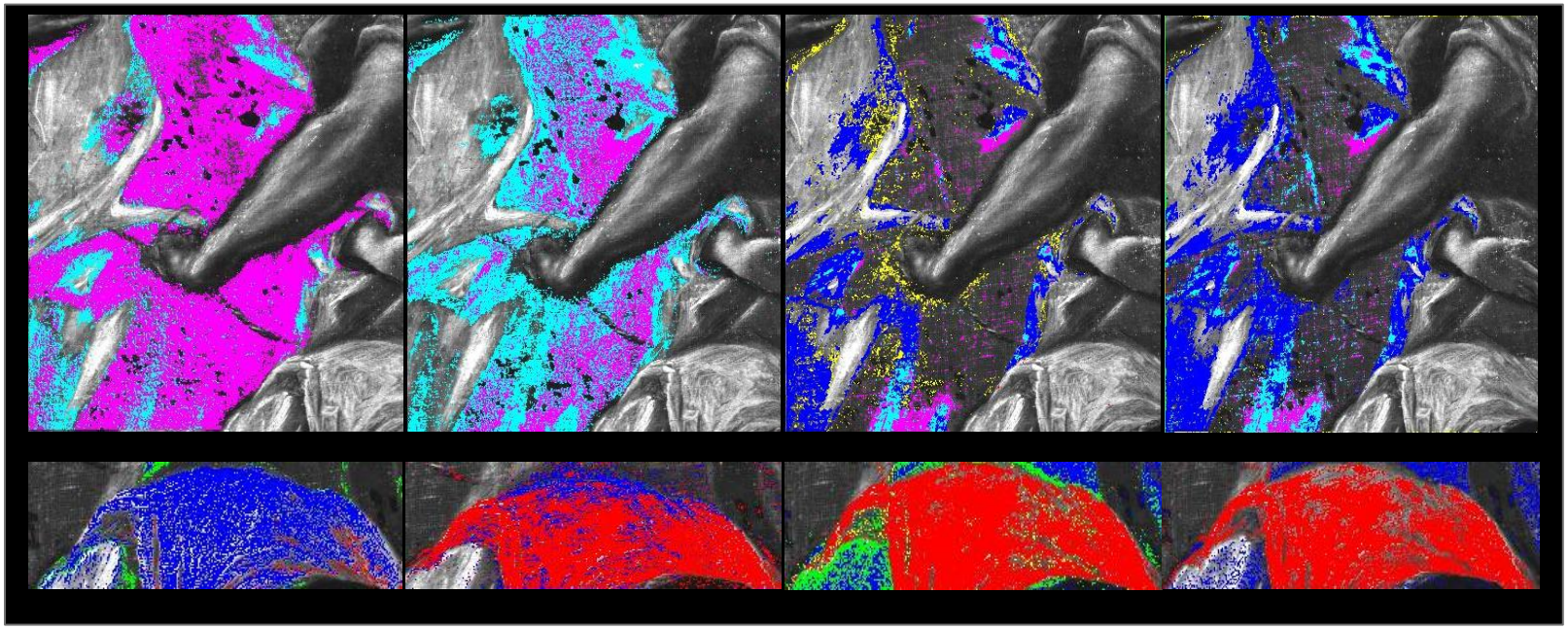


figure 4.48: spectral response of four points from Espolio (details)



Thematic map of red pigments

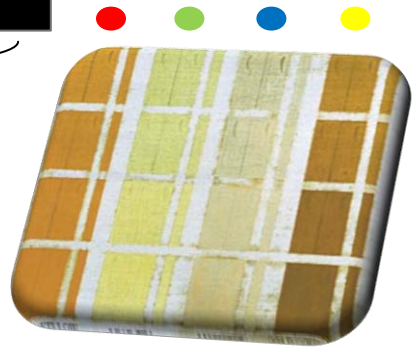
(ML / SCM / SID / SPAM)

figure 4.49: Thematic maps of red & yellow pigments.

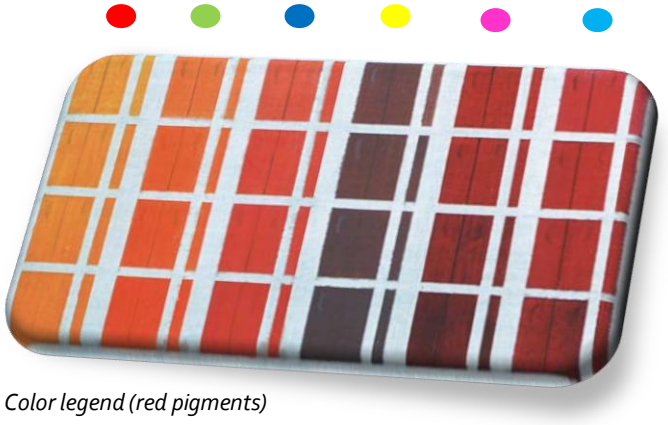
Thematic map of yellow pigments

(ML / SCM / SID / SPAM)

class/color	*(oil) red*	*(oil) yellow*
red	REALGAR KREMER	NAPLES YELLOW KREMER(4313)
green	MINIO KREMER(4250)	LEAD-TIN KREMER(1011)
blue	CINNABAR KREMER(4200)	MASSICOT KREMER(4301)
yellow	HEMATITE KREMER(4860)	OCHRE GOLD ITALY KREMER(4800)
magenta	RED LAKE,DARK	
cyan	RED LAKE,LIGHT	



Color legend (yellow pigments)



Color legend (red pigments)

figure 4.50 Saint Louis, King of France,

J.M. Theotokopoulos

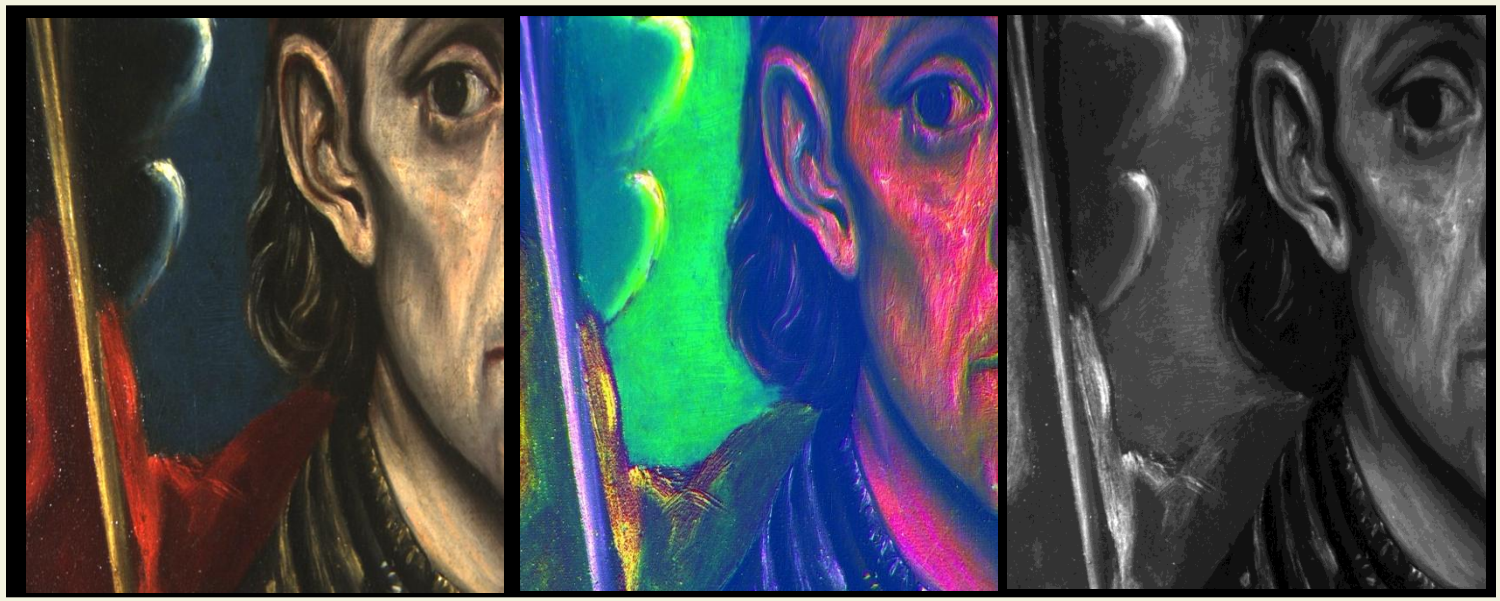


figure 4.51: RGB image / PCI falsecolor image/ NIR (λ=1000 nm) image (detail)

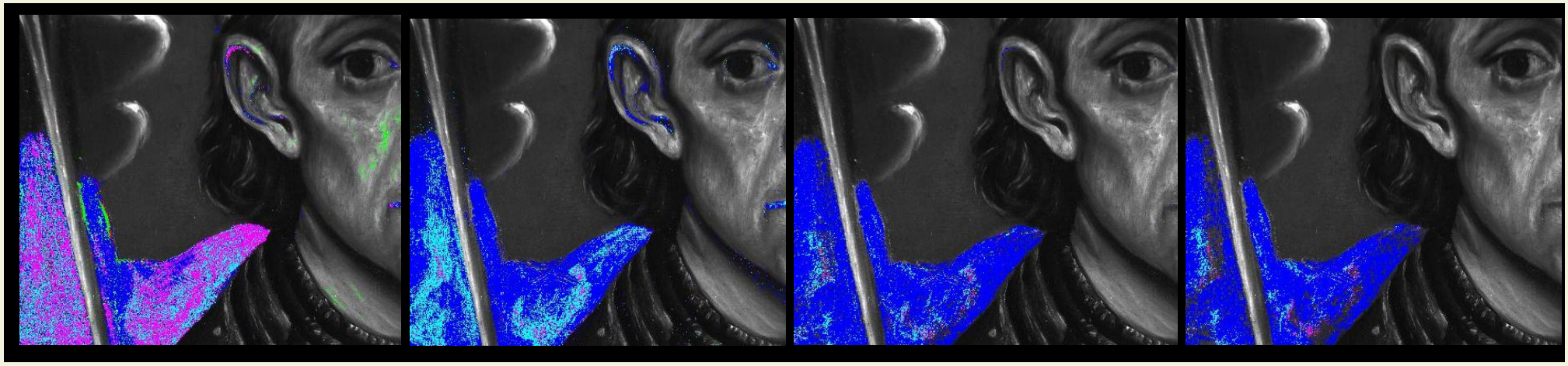


figure 4.52: Thematic map of red pigments (ML / SCM / SID / SPAM)

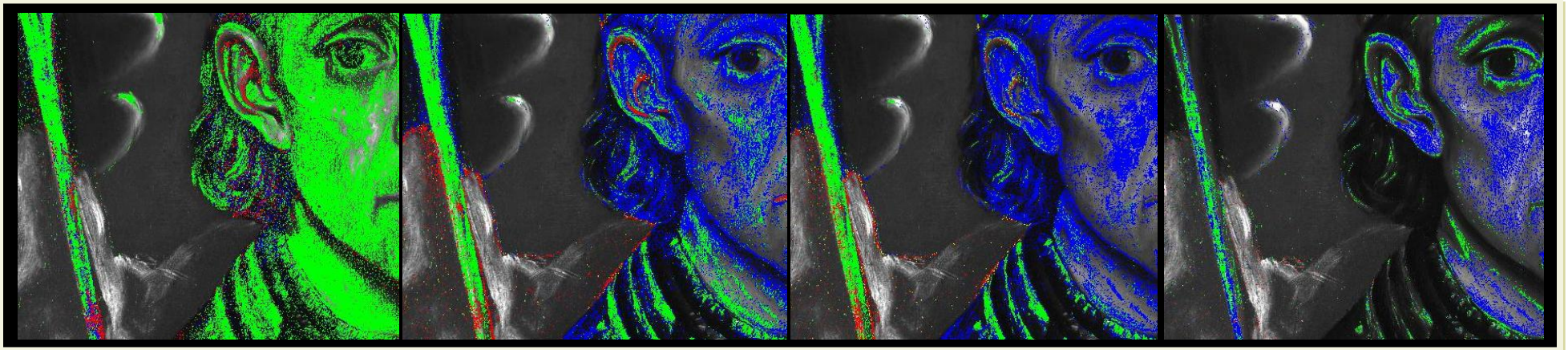


figure 4.53: Thematic map of yellow pigments (ML / SCM / SID / SPAM)

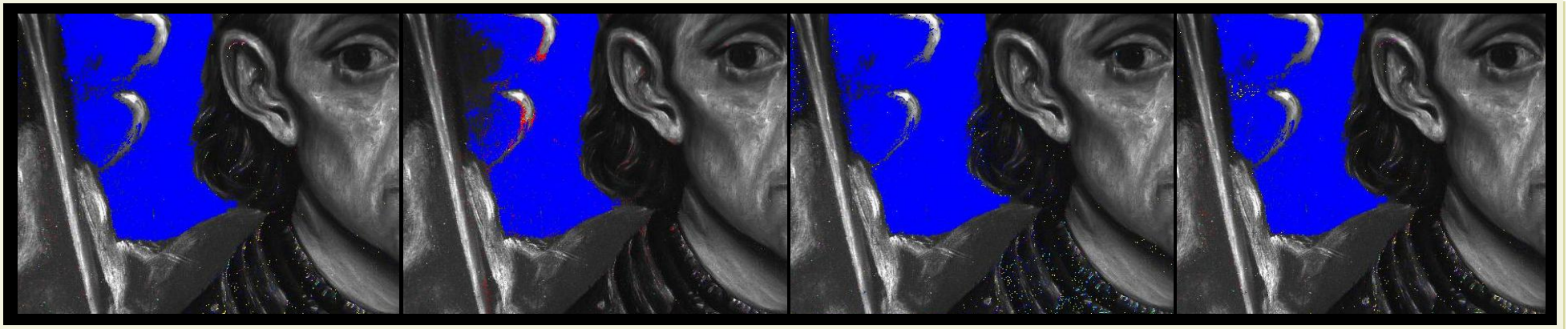
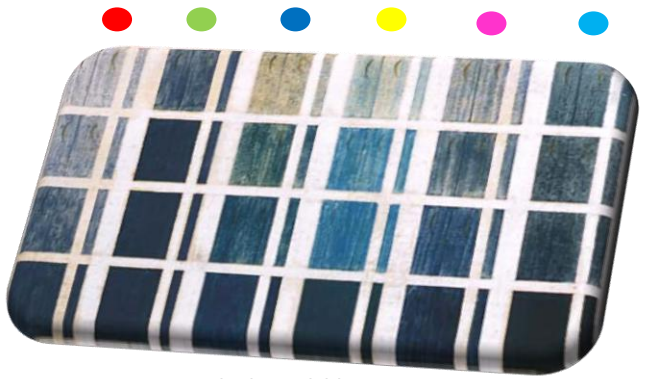


figure 4.54: Thematic map of blue pigments (ML / SCM / SID / SPAM)

class/color	(oil) red	(oil) blue	(oil) yellow
red	REALGAR KREMER	LAPIS LAZULI KREMER	NAPLES YELLOW KREMER(4313)
green	MINIO KREMER(4250)	LAPIS LAZULI KREMER(C=0.85)	LEAD-TIN KREMER(1011)
blue	CINNABAR KREMER(4200)	AZURITE KREMER	MASSICOT KREMER(4301)
yellow	HEMATITE KREMER(4860)	EGYPTIAN KREMER	OCHRE GOOLD ITALY KREMER(4800)
magenta	RED LAKE,DARK	SMALT KREMER	
cyan	RED LAKE,LIGHT	INDIGO KREMER	



Color legend (blue pigments)

Chapter 5
Conclusion and future work

An important step towards our understanding of works of art and their history is the composition of the pigments used by the artists. Knowledge of their main elements, such as the pigments and bonding agents used, is vital for the preservation of paintings. Examinations of such compounds are usually done with destructive techniques in which a sample or microsample is required. Due to the uniqueness of the artwork, samples for external studies can only be taken with the utmost caution and only when they are vital for gathering additional information for the preservation of the piece. As a result, the need for the development of non-invasive techniques to identify pigments is growing.

MuSIS HS constitutes a hyper-spectral imaging apparatus, capable of performing spectral imaging in a wide spectral range of 34 bands from UV to NIR. The combination of an imaging monochromator with a CCD camera resolves spatial resolution issues that similar optical spectroscopy technologies fail to address. Thus, it can be used as a non-invasive technique for extracting spectral information from pigments and paintings. A spectral library of El Greco's pigments composed by materials with known chemical and structural characteristics, following their original development process, was an ideal candidate for testing MuSIS diagnostic capabilities. Isodata unsupervised classifier was successful at discriminating pigments areas from canvas or wooden plate and was able to discriminate some (a few) colors. Employing different supervised classification algorithms, pigment identification was possible with adequate accuracy. In particular, Maximum Likelihood, a well- known classification tool used in remote sensing for the past 20 years, was able to discriminate successfully several color pigments with >80% accuracy for oil colors and >91% for tempera colors. A systematic sampling process was used instead of random sampling to achieve the above results. Expectation Maximization offered a marginal improvement over ML's classification accuracy in the expense of classification time. Also a more strict sampling scheme from reference pigments was employed, as EM was very sensitive producing singular results when train sample area was not homogenous, i.e pigments separated from oil or egg bonding agents. In the case of oil colors on canvas with EM it was not possible to acquire sufficient classification accuracy due to sensitivity to sample size and sampling plan.

The concept of distance as a classification technique was also explored in this project. Referencing a class only using a few training samples normally leads to a decrease of accuracy comparing with a parametric classifier. However classifiers incorporating distance metrics have been proven useful in applications where a training set of adequate size is not easy to acquire (oil pigments) or due to the high dimensionality of the data. A series of classification algorithms for discriminating spectra using generalized ideas of distance involving statistical correlation, spectral gradients, information theory etc. was employed for discriminating pigments. In the case of

egg pigments, Spectral Correlation Mapper, with 4 samples per color, was able to perform similar with ML, whereas the other distance metrics achieved a classification accuracy of >78% . On the other hand, oil painting was a more difficult classification task in order to maintain a high accuracy score for distance metrics (ML appears to be less sensitive to binding media). SCM was able, for most color sets, to perform similar with ML, while other distance metrics for some color pigments appear to perform well, for other color sets classification accuracy declines. In particular, with the red color set, both oil and egg, all the classifiers performed adequate (>95% for egg, >70% for oil) , while for distance metrics ,except SCM, the ochre & green color sets were the most problematic. Thematic maps supported these observations, providing an overall aspect of classification results for comparative evaluation of ML, EM and distance metrics, especially in canvas samples where the difficulty of acquiring train/test sets increased , as the linseed oil expanded beyond the (distinct) boundaries of each color's sample area. In these cases, RSDR and RSDE were useful for selecting appropriate reference samples.

Testing MuSIS HS and the ML, SID, SCM, SPAM classifiers with oil train sets on El Greco's paintings provided by the Museum of Cycladic Art, it was able to demonstrate spectral similarity and close relation with color pigments used in paintings and the spectral library. In most cases, classifiers agreed on identifying the same reference pigments with the ones used in the paintings. Also, PCA on spectral cubes can be used for aiding the classification analysis, providing an enhanced image of the painting or revealing details not visible in the 400-700 spectral range. A pseudocolor color image of 3 PCA components was utilized for this purpose.

The classification methodology discussed in this diploma thesis aided by a simple-to-use GUI application could facilitate the work of an art work preserver, help pigment identification if an adequate spectral library is provided, while implementation of other classification algorithms is possible. Preliminary results with SVMs and spectral unmixing confirm that MuSIS HS is suitable for pigment identification tasks under several supervised classification schemes.

References

[1] J.A. Richards and X.Jia, "Remote Sensing Digital Image Analysis" , 3rd ed. New York: Springer, 1999.

[2] T. Fawcett (2004). "ROC Graphs: Notes and Practical Considerations for Researchers"

[3] H.Zhang and G.Sun. "Feature selection using tabu search method. Pattern Recognition", 35:701711,2002.

[4] F.Glover. "Future paths for integer programming and links to artificial intelligence. Computers and Operations Research", 5:533549,1986.

[5] P.M. Narendra and K.Fukunaga. "A branch and bound algorithm for feature subset selection". IEEE Transactions on Computers, C-31:917922,1977.

[6] L.Jimenez and D.Landgrebe, "Supervised classification in high dimensional space :geometrical, statistical and asymptotical properties of multivariate data," IEEE Trans.Syst., Man,Cybernet.,C, vol. 28, pp. 39-54,Feb. 1998.

[7] M.G. Kendall, "A Course in the Geometry of n-Dimensions", New York: Hafner, 1961.

[8] J.P.Hoffbeckand D.A .Landgrebe, "Covariance matrix estimation and classification with limited training data," IEEE Trans. Pattern Anal. Machine Intell., vol. 18, pp. 763-767, July 1996.

[9] Jolliffe I.T. "Principal Component Analysis, Series: Springer Series in Statistics", 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28

[10] Borg, I. and Groenen, P.: "Modern Multidimensional Scaling: theory and applications" (2nd ed.), Springer-Verlag New York, 2005

[11] "Exploratory Factor Analysis" - A Book Manuscript by Tucker, L. & MacCallum R. (1993)

[12] T.-W. Lee (1998): "Independent component analysis: Theory and applications", Boston, Mass: Kluwer Academic Publishers

[13] B. Schölkopf, A. Smola, K.-R. Muller, "Kernel Principal Component Analysis", Bernhard, 1999, MIT Press Cambridge, MA, USA, 327-352

[14] Mikhail Belkin, Partha Niyogi , "Laplacian eigenmaps and spectral techniques for embedding and clustering" (2002)

[15] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

[16] Romesburg, H. Clarles, "Cluster Analysis for Researchers", 2004, 340 pp. ISBN 1-4116-0617-5

[17] N. Memarsadeghi, D. M. Mount, N. S. Netanyahu, and J. Le Moigne,"A Fast Implementation of the ISODATA Clustering" Algorithm,Internat. J. Comput. Geom. Appl., 17 (2007)

[18] S. Theodoridis and K. Koutroumbas. "Pattern Recognition". Academic Press, 1999. ISBN 0-12-686140-4.

[19] R.O. Duda, P.E. Hart, and D.G. Stork. "Pattern Classification". John Wiley & Sons,Inc., 2ndedition, 2001

[20] G.F. Hughes, "On the mean accuracy of statistical patternrecognizers," EEE Trans. Inform. Theory, vol. IT-14, pp. 55-63, Jan. 1968.

[21] Benzécri, J.-P. (1973). "L'Analyse des Données. Volume II. L'Analyse des Correspondences". Paris, France: Dunod

[22] B.S. Everitt and D.J. Hand. "Finite Mixture Distributions. Monographs on Applied Probability and Statistics" .Chapman and Hall, 1981.

[23] M.A.T. Figueiredo and A.K. Jain. "Unsupervised learning of fnite mixture models". IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):381–396, Mar 2002

[24] Kruse, F.A., J.W. Boardman, A.B. Lefkoff, K.B. Heidebrecht, A.T. Shapiro, P.J. Barloon, and A.F.H. Goetz, 1993." The Spectral Image Processing System (SIPS): Interactive visualization and analysis of imaging spectrometer data". Remote Sensing of Environment 44: 45-163.

[24] Nirmal Keshava, "Distance Metrics and Band Selection in Hyperspectral Processing With Applications to Material identification and Spectra" Libraries, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOW .42 ,NO. 7, JUL Y2004

[25] Carvalho Junior, O.A. and Menezes, P.R. 2000, "Spectral Correlation Mapper (SCM): an improving Spectral Angle Mapper (SAM)". Proceedings of the Nineth JPL Airborne Earth Science Workshop, JPL Publication 00-18, pp 65-74

[26] H. Du, C.-I. Chang, H. Ren, C-C Chang, J. O. Jensen, and F. M.D'Amico, "New hyperspectral discrimination measure for spectral characterization", Optical Engineering, 43, no 8, 2004, 1777–1786

[27] E. Angelopoulou, S. W. Lee, R. Bajcsy, "Spectral gradients: A material descriptor invariant to geometry and incident illumination", Proc IEEE Int Conf on Comp Vision. IEEE Computer Society Press, 1999, 861-867.

[28] Saeid Homayouni, Michel Roux, "HYPERSPECTRAL IMAGE ANALYSIS FOR MATERIAL MAPPING USING SPECTRAL MATCHING"

[29] S.A. Robila, "An Investigation of Spectral Metrics in HyperspectraImage Preprocessing for Classification", Proc ASPRS Annual Conf 2005

[30] Wolff,L.B.Diffuse, "Refectance Model for Smooth Dielectric Surfaces", Journal of the Optical Society of America A, Vol. 11, No. 11, November 1994, pp. 2956-2968.

[31] D.A.Landgrebe, "Hyperspectral Image Analysis", JANUARY 2002 IEEE SIGNAL PROCESSING MAGAZINE 1053-5888

[32] Α. Αλεξόπουλος, Αγορανός, Γ. Χρυσουλάκης, «θετικές επιστήμες και έργα Τέχνης», Αθήνα, 1993, σ. 126

[33] Costas Balas ,Vassilis Papadakis ,Nicolas Papadakis ,Antonis Papadakis,Eleftheria Vazgiouraki, ,George Themelis, "A novel hyper-spectral imaging apparatus for the non-destructive analysis of objects of artistic and historic value. Journal of Cultural Heritage 4 (2003)330s–337s

[34] D.Anglos, C.Balas, C.Fotakis, "Laser spectroscopic and optical imaging techniques in chemical and structural diagnosis of painted artworks" ,Am.Lab.(October)(1999)60–67.

[35] C.Balas, "An imaging method and apparatus for the non-destructive analysis of paintings and monuments", International PatentApp. PCT/GR00/00039.

[36] C.Balas, "A novel optical imaging method for the early detection, quantitative grading and mapping of cancerous and precancerous lesions of cervix", IEEE Trans. Biomed. Eng. 48(2001)96–104.

[37] Γ. Τσαϊρης, «Η χρήση ψηφιακού πολυφασματικού απεικονιστικού συστήματος στη μελέτη των χρωστικών του έργου Η συναυλία των Αγγέλων του Δομήνικου Θεοτοκόπουλου. Η συμβολή ψηφιακού υπερφασματικού απεικονιστικού συστήματος στην ταυτοποίηση χρωστικής ενός ζωγραφικού έργου», Πανεπιστήμιο Κρήτης, Τμήμα Ιστορίας-Αρχαιολογίας & Επιστήμης Υπολογιστών

[38] S. Baronti, A. Casini, F. Lotti, Simone Porcinai, "Principal component analysis of visible and near-infrared multispectral images of works of art", ELSEVIER Chemometrics and Intelligent Laboratory Systems, 39 (1997), 103-114

[39] James R. Mansfield ,Michael G . Sowa , Claudine Majzels ,Cathy Collins ,Edward Cloutis, ,Henry H. Mantsc, "Near infrared spectroscopic reflectance imaging: supervised vs .Unsupervised analysis using an art conservation application", ELSEVIER Vibrational Spectroscopy 19 (1999),33-45

[40] Technical bulletin, " Kremer Pigments, Raw materials for fine arts, Conservation, Wood finishing, Design", New York , 1998