Hyperspectral texture metrology

Rui Jian Chu

THÈSE

POUR L'OBTENTION DU GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE POITIERS

FACULTÉ DES SCIENCES FONDAMENTALES ET APPLIQUÉES

DIPLÔME NATIONAL - ARRÊTÉ DU 25 MAI 2016

Ecole Doctorale : Sciences et Ingénierie des Systèmes, Mathématiques, Informatique - SISMI

Secteur de Recherche : Traitement du Signal et des Images

Présentée par:
Rui Jian CHU

Hyperspectral Texture Metrology

Directrice de Thèse:
Christine FERNANDEZ-MALOIGNE

Co-directeur de Thèse:
Noël RICHARD

Soutenue le 07 juillet 2022
Devant la Commission d'Examen

JURY

Yannick BERTHOUMIEU ..... Rapporteur
Nicolas DOBIGEON ..... Rapporteur
Christine ANDRAUD ..... Examinateur
Faouzi GHORBEL ..... Examinateur
Alamin MANSOURI ..... Examinateur
Christine FERNANDEZ-MALOIGNE ..... Directrice
Noël RICHARD ..... Co-directeur
Jon Yngve HARDEBERG ..... Co-directeur

To the apple of my eye, Gao Yan.

Acknowledgements

First and foremost, I want to thank Noël for guiding me throughout these years. It has been a long and tough journey, and I have not been the easiest student to deal with. I want to thank him for having taught me so much, both in science and in life. He has been truly an inspiring figure in my academic endeavour, and I shall benefit from this for a very long time. His unconventional, yet revolutionary and visionary thoughts influence me a lot. His words still linger in my head and I shall slowly but surely understand most of his wisdom one day.

On the other hand, I am also indebted to Christine and Jon who have been supervising me. They have been instrumental in showing me the larger picture, sometimes beyond research. Apart from their expertise in image processing, I also benefit from their networks which prove tantamount to scientific achievements. I am equally indebted to the jury members especially the rapporteurs who have provided me precious feedbacks. I benefit so much from their instantaneous questioning during the defence as we exchanged insights and perspectives.

Last but not least, I want to thank my muse, Gao Yan who I dedicate this thesis to. Though she is not always present physically, I depend on her mentally all the time. I admire her strength and resourcefulness from which I derive my energy for daily tasks. She has been with me through thick and thin, sometimes I wonder if one deserves such perfection. This journey is impossible without her consistent companion, understanding, and love, which make me a better, braver, and more aspiring person than I was. Obviously, my family members also have been extremely supportive of my research which I am forever grateful for. It is not without their blessings that I can venture so freely, knowing that I am loved.

Abstract

Sixty years since the proposal of the first Julesz conjecture, many texture feature extraction methods have been published. However, few are developed for metrological purposes. According to metrology i.e. science of measurement, a feature must be developed in conjunction with distance or similarity measurement. In the context of our work on hyperspectral texture metrology, this requires the feature to be relatively independent of the sensor technology and thus from the number of spectral bands, spectral range, etc. In other words, the developed feature must be general with results strictly dependent only on the texture content.

In this Ph.D. manuscript, we began by defining texture as the joint distribution of spectral and spatial properties. Whereas the former dictates the spectral diversity or distribution of pixels, the latter refers to the spatial arrangement or interaction of pixels. Accordingly, we developed fullband spectral analysis adapted to the continuous nature of spectrum. By working in metric space, the need for dimensionality reduction e.g. principal component analysis (PCA) is eliminated. Besides, the selection of spectral reference is fully automated thanks to entropy maximization. An analogy between our approach and PCA is drawn, with the former shown to be more robust and faithful to the physics of spectrum.

Taking inspiration from graylevel cooccurrence matrix (GLCM) and local binary pattern (LBP), we next developed fullband spatial analysis adapted to images of any number of spectral bands and spectral range. We took into account both the accuracy of GLCM features and the efficiency of LBP similarity measure to propose the neighboring difference histogram. We then related neighboring difference with the notion of image gradient to pursue

a parameterless approach. Subsequently, we demonstrated a mathematical decomposition of spectral gradient into chromatic and achromatic parts. Considering the scale space theory, we also developed automatic scale selection to reflect the local structure size.

To validate our approach, we subjected our features to texture classification, retrieval, and segmentation tasks. The results show that our features perform at comparable level with state of the art including machine and deep learning approaches. More importantly, we demonstrate the physical interpretability of our features facing a multitude of spectral-spatial variations. Thanks to metrology, we testified to the efficiency of adapted pair of feature and similarity measure without necessarily the need for sophisticated learning techniques.

Table of contents

1 Introduction **1**

1.1 Motivation 2

1.2 A critical analysis on the state of the art 3

1.2.1 Feature extraction 3

1.2.2 Similarity measurement 5

1.2.3 Adaptation in hyperspectral domain 7

1.3 Towards texture metrology 11

2 Spectral Characterization of Texture **15**

2.1 Background 16

2.2 The proposed formulation 21

2.2.1 Spectral decomposition into shape and intensity 21

2.2.2 Spectral difference histogram 24

2.2.3 Assessment of spectral diversity 30

2.2.4 Automatic selection of spectral reference 35

2.3 Summary 54

3 Spectral-Spatial Characterization of Texture: Differences-Based **55**

3.1 Background 56

3.2 The proposed formulation 58

3.2.1 Ideas from sum and difference histograms 58
3.2.2 Edge sharpness, scale, and direction 62
3.2.3 Relative spectral difference occurrence matrix (RSDOM) 74
3.2.4 Similarity measurement and statistical modelling 76
3.3 Summary 78

4 Spectral-Spatial Characterization of Texture: Gradients-Based 81
4.1 Background 82
4.2 The proposed formulation 85
4.2.1 Automatic scale selection 85
4.2.2 Spectral gradient magnitude, direction, and coherence 102
4.2.3 Spectral shape and intensity gradients 126
4.2.4 Gradient histogram of spectral texture (GHOST) 136
4.3 Summary 138

5 Performance Assessment of The Proposed Texture Features 139
5.1 Texture classification on *HyTexiLa* 140
5.1.1 Dataset and classification scheme 140
5.1.2 Feature computation 144
5.1.3 Result analysis 146
5.2 Texture retrieval on *HSI-ICONES* 148
5.2.1 Dataset and retrieval scheme 148
5.2.2 Feature computation 149
5.2.3 Result analysis 149

6 Conclusion 151

References 157

Notation	Signification
$\mathbb{R}^D$	Real number in D dimensions.
$(e_1, e_2, \ldots)$	Tuple of which the ordering of its elements e_i matters.
$\{e_1, e_2, \ldots\}$	Set of which the ordering of its elements e_i does not matter.
I	Image defined in two spatial dimensions i.e. $I : \mathbb{R}^2 \to \mathbb{R}^L$ with L spectral channels. $I(x)$ denotes the pixel value at the given spatial location x. $\bar{I}$ refers to normalized image such that each spectrum integrates to unity.
x	Spatial location in $\mathbb{R}^2$ i.e. $x = (x_1, x_2)$.
s	Spectrum, defined as a continuous function. Capitalized as $S = \{s_1, \ldots, s_n\}$ if a set of n spectra is referred to.
$\bar{s}$	Normalized spectrum such that $\bar{s} = s/\|s\|$. Capitalized as $\bar{S} = \{\bar{s}_1, \ldots, \bar{s}_n\}$ if a set of n normalized spectra is referred to.
$\hat{s}$	Spectral reference. Capitalized as $\hat{S} = \{\hat{s}_1, \ldots, \hat{s}_R\}$ if a set of R spectral references is referred to.
$\widetilde{S}$	An approximation of convex hull $Conv(\bar{S})$ of the normalized spectral set $\bar{S}$ with Z spectra such that $\widetilde{S} = \{\widetilde{s}_1, \ldots, \widetilde{s}_Z\} \approx Conv(\bar{S})$.
G	Spectral shape. ΔG referes to spectral shape difference.
W	Spectral intensity. ΔW referes to spectral intensity difference.
λ	Wavelength; λ_l denotes the l^{th} wavelength. $\Delta\lambda$ refers to spectral resolution.
p	Probability density function. Capitalized as P if referring to a probability mass function. $\mathscr{U}$ denotes uniform probability distribution.
$\mathrm{KLD}(p, p')$	Kullback-Leibler (KL) divergence; $\mathrm{KLD}(p, p') = \mathrm{KL}(p\|p') + \mathrm{KL}(p'\|p)$, where $\mathrm{KL}(p\|p') = \int p(y) \log\left(\frac{p(y)}{p'(y)}\right) \mathrm{d}y$ is the KL measure of information.
$\mathscr{O}$	Computational complexity.
s.t.	Such that.
$\stackrel{\text{def}}{=}$	Equal to by definition.
$H(\cdot)$	Shannon entropy, defined in discrete domain.
$h(\cdot)$	Differential entropy, defined in continuous domain.
$\mathscr{N}(\mu, \sigma)$	Normal distribution of mean μ and standard deviation σ. $\mathscr{N}(\mu, \Sigma)$ refers to multivariate normal distribution of mean μ and covariance matrix Σ.
$y \sim p$	y is distributed as p.
$\lfloor y \rfloor$	Highest integer lower or equal than y.
$Conv(S)$	Convex hull of set S.
$\mathscr{J}_1^{(\hat{S})}(I)$	Spectral difference histogram of the given image I based on the set of spectral references $\hat{S}$.
$\mathscr{J}_2^{(r,\theta)}(I)$	Neigbouring difference histogram of the given image I such that the pixel pairs are defined with distance r and direction θ.
$\mathscr{D}(I)$	Spectral gradient histogram of the given image I.
$\bigcap$	Logical AND operation.
$\mathscr{T}$	Texture feature, defined by either RSDOM or GHOST in this work.
$\|\cdot\|_p$	p-norm e.g. $\|\cdot\|_2 = \sqrt{\int \|\cdot\|^2}$.

Chapter 1

Introduction

Contents

1.1 Motivation 2

1.2 A critical analysis on the state of the art 3

1.2.1 Feature extraction 3

1.2.2 Similarity measurement 5

1.2.3 Adaptation in hyperspectral domain 7

1.3 Towards texture metrology 11

1.1 Motivation

There are a plethora of texture feature extraction methods in the literature. Few, however, define what texture is. This results in numerous ad hoc approaches which correlate little with metrology, despite attaining supreme accuracy in some specific tasks e.g. segmentation, classification and image retrieval. Consequently, one often finds difficulty in applying them for the metrology of industrial quality control or realtime remote sensing due to the lack of training dataset, complications in pre-processing and intricacy of parameter tuning.

In contrast, a metrological approach seeks to first define the measurand before performing the measurement. Under metrology, measurements are to be solely dependent on measurand and dissociated from sensor. Correspondingly, the similarity between textures is to be preserved regardless of the imaging sensor with discrepancy only in accuracy. Indeed, the varying number of bands ($L = 1$ for grayscale, $L = 3$ for trichromatic, $L \gg 1$ for spectral) indicates only differences in the bandwidth or sampling frequency in spectral dimension *but not in the image content*. Consequently, texture as measured from hyperspectral image is more accurate than that from grayscale image though within *similar* range. Feature interpretability is also of utmost importance for the establishment of a direct relationship between the measurement (feature) and the measurand (texture). In contrast to data-driven approach, metrological solutions allow the quantification of error, uncertainty, and bias for traceability.

In this thesis, we aim to develop texture features which are metrological and fully interpretable. We note that images are inherently continuous despite the post-acquisition digital representation. As such, we choose to develop our assessment in hyperspectral context with straightforward application in grayscale and colour domains. Hyperspectral imaging (HSI) is closely related to the measurement of physical surface properties and material composition. Thanks to the dense spectral sampling up to hundreds of bands, HSI is able to provide rich information that allows rapid and nondestructive assessment. However, highly

accurate acquisition does not warrant highly accurate results. The full exploitation of HSI necessitates proper data interpretation and enforcement of metrological processing.

1.2 A critical analysis on the state of the art

1.2.1 Feature extraction

Existent texture feature extraction techniques can be mainly divided into four categories [36, 79, 81]: statistical-, structural-, transform-, and model-based. In the following, some major techniques will be reviewed to provide a general perspective of the state of the art.

Statistical-based features

Statistical-based features are formulated under the pretext of texture being generated from stochastic processes. As the oldest attempt to characterise texture, autocorrelation function tells about the size of the tonal primitives which serves as a measure of coarseness (or fineness) and regularity of the texture [64]. Mathematically simple, the autocorrelation features are nevertheless inefficient due to their lack of comprehensiveness. The graylevel cooccurrence matrix (GLCM) captures the second-order statistic for a given spatial relationship [30], whose rotational invariance is achieved by correlogram [33]. The graylevel run length matrix (GRLM) records the "run" or connected length of a particular pixel in a given orientation [26, 23]. However, both GLCM and GRLM suffer from matrix sparsity problems which demand heavy image quantization. The six Tamura features are formulated in correspondence to human visual perception: coarseness, contrast, directionality, line-likeness, regularity, and roughness [76]. Although psychophysically meaningful, mathematically the features are defined non-uniquely with an ad hoc sense.

Structural-based features

The structural-based features are developed in the context of texture being a repetition of its primitives or textons. The classic method involves the determination of the primitives as well as its placement rules [85]. However, such approach proves too restrictive as it is only applicable for regular texture. A more versatile approach can be found by identifying microstructures instead. Local binary pattern (LBP) records the distribution of microtextons as defined by thresholding the neighbourhood of each pixel [56]. Evidently, LBP loses out on the first-order statistical information. The Leung-Malik (LM) [48], Schmid (S) [67], and Maximum Response (MR) [84] filterbank approaches define texton dictionary by the cluster centres in the filter response space. Whereas LM and MR filterbanks consist of multiscale and multidirectional Gaussian derivatives of order up to two, S filterbank is rotationally invariant with isotropic, "Gabor-like" filters. As with all optimization problems, the required clustering operation presents a computational overhead and algorithm-dependent efficiency.

Transform-based features

The transform-based features are the ones generated by various mathematical transform operations. The most straightforward operation is the Fourier transform which analyzes the frequency content of texture [77]. However, it is a global operation defined on the entire image with a complete loss of spatial information. To achieve joint spatial-frequency localization, the Gabor transform has been designed to mimic the operation of simple cells [14]. However, its efficiency is highly dependent on the filterbank design. With the advent of multiresolution analysis, wavelet analysis has been proposed [53]. The basis consists of dilations and translations of a single function called the mother wavelet whose optimal choice e.g. Haar, Daubechies, and biorthogonal dependent on texture. Apart from its lack of shift-invariance, the size of a given region determines its representative frequencies within the transform. This latter property seems particularly restrictive given that there is no reason,

in general, to assume that the frequency content of an image region should be related to its size.

Model-based features

The model-based features are constructed based on a given model. The fractal model assumes self-similarity of texture [59], but the feature i.e. the fractal dimension alone is not sufficient to capture all textural properties. The Markov random field (MRF) model relies on the assumption that the value of each pixel value is only dependent on that of neighbouring pixels [16]. However, the calculation of MRF parameters is subjected to computationally intensive optimization due to the nonlinearity of the likelihood function. The Weber local descriptor (WLD) is inspired by a psychological law [12], but its construction is ad hoc as it lacks physical or mathematical justification in relation to the texture (why consider differential excitation and gradient orientation but not others?). More recently, a gravitational-based model named pattern of local gravitational force (PLGF) has been developed [6]. Though conceptually interesting, its improvised concatenation of histograms (of force magnitude and of angle) is mathematically questionable (as opposed to joint histogram).

1.2.2 Similarity measurement

Facing the need for similarity measurement, some feature extraction techniques resort to simplification or summarization of the originally sound formulation. This is sometimes done without careful consideration of the underlying data distribution. Some features are also developed without regarding the need for similarity measurement, hence the ad-hocness. In the following, three common approaches for adapting the features for similarity measurement are presented.

p-norm measures with feature vectors

The first is expression by feature vector. For example, GLCM and GRLM are reduced into moments e.g. entropy, homogeneity, and correlation for summarizing the matrix [30, 26]. These moments are then concatenated to form the feature vector before the application of p-norm measures for similarity assessment. Although commonly employed in the literature, such treatment lacks mathematical plausibility. For example, the usage of the Euclidean norm assumes independence between the elements in the vector. However, it is well known that the moments as extracted from GLCM and GLRM are highly correlated which clearly violates the independence assumption. On the other hand, statistical moments e.g. mean and standard deviation are also commonly used to summarize filter responses of Gabor [3] and wavelet approaches. We remark that the usage of these statistical moments implicity assumes the normal distribution of the data, which is a strong statement.

Statistical distances with histograms

The second is expression by distribution. For example, correlogram is expressed as a flattened two-dimensional histogram; LBP is expressed as one-dimensional histogram of binary codes; whereas the wavelet coefficients are modelled by generalized Gaussian mixture model. Mathematically, such approach fares better than that of moments because it avoids making assumptions about the data distribution. However, it also possesses several problems due to the need for density estimation. Firstly, the feature efficiency is highly dependent on the efficiency of the statistical modelling which itself also presents a computational burden. A soundly formulated feature may perform poorly due to inaccurate statistical modelling. Secondly, density estimation becomes tricky in the case of multidimensional space thanks to the exponential increase of memory consumption and the curse of dimensionality. For example, a one-dimensional histogram of 2^8 bins requires 256 bits, but a two-dimensional and three-dimensional histogram of the same number of bins in each axis requires over 200

and 60000 times respectively the memory and storage requirements of the one-dimensional version. Furthermore, statistical modelling also tends to break down in high dimensional space. For example, it is known that in the case of kNN most of the mass is concentrated at the unit ellipsoid. These difficulties themselves lead to the reduced efficiency of otherwise well-intended feature extraction.

Machine or deep learning

Facing the difficulty in expressing the feature in an efficient manner, some authors skip the devising of similarity measurement altogether and opt for a classifier-based solution. For example, sophisticated machine learning algorithms with the like of support vector machine (SVM), random forest (RF), and autoencoders have been devised to tackle the high dimensionality problem. Recently, deep learning approaches like convolutional neural network (CNN), recurrent neural network (RNN), and graph convolutional network (GCN) has been optimized to learn the relationship between the inputs and intended outputs. While these algorithms can perform in a superman-like capacity, they represent a generation of "black box" approaches whereby there exists no explanation of the functionality of the code. This possesses difficulty in analyzing and interpreting the result which is of vital interest in fields such as medical diagnosis and military application. Although society has since realized such problem and shifted towards "explainable artificial intelligence" (XAI), such approach still suffers from the need for a wealth of annotated data for the learning, not to mention the expense of time and energy to perform the computation.

1.2.3 Adaptation in hyperspectral domain

In retrospect, most texture feature extraction is devised in the context of grayscale images. Indeed, Indeed, it exists a school of thought that considers *color* and *texture* as separate entities. Whereas colour is measured globally according to the image histogram ignoring local

neighbouring pixels, texture is typically characterized by the relationship of the intensities of neighbouring pixels ignoring their colour. Thanks to the well-established methods for accessing monovariate texture, the remaining work seemed to be on *adapting* the existing methods on multivariate images. In the following, three common approaches for processing hyperspectral texture are presented. Texture assessment in the hyperspectral domain represents a unique case due to the high dimensionality. Even if colour texture assessment is also performed in a multivariate fashion, the number of spectral channels is only three compared to possible hundreds and even thousands of spectral bands in the case of hyperspectral image.

Marginal processing

The first approach is to extract a representative image(s) from the hyperspectral image from which texture features can be extracted. The common argument is that most of the hyperspectral is highly redundant and can be compressed using dimensionality techniques. For example, principal component analysis (PCA) [27, 28, 92, 22, 90, 22, 27, 28, 35, 91, 82, 45] and minimum noise fraction (MNF) [87, 13] have been widely used in the remote sensing community which commonly deal with AVIRIS sensor images up to hundreds of spectral bands. Following the dimensionality reduction, only the first few principal components (PCs) are considered for the texture assessment. However, it exists no principled but empirical approach to selecting the optimum set of PCs. The other problem arising from such approach is the possible loss, alteration, or destruction of the geometrical properties in the image. Furthermore, these techniques are also applied without considering the physics. For example, PCA is defined for vectors, but it is well known that spectral values are highly correlated and are by definition, functions. On the other hand, PCA is appropriate for capturing the structure of data that are normally distributed, or in which the linear pairwise correlations are the most important form of statistical dependence in the data. But natural scenes contain many higher-order forms of statistical structure, and there is good reason to believe they form an

extremely nongaussian distribution that is not at all well captured by orthogonal components The approach of using the first few PCs as the representative image so that the texture features can be extracted also is a misconception; it has been reported that this is not necessarily the case. Another solution is to process the texture features on the panchromatic image. However, this misses the opportunity provided by hyperspectral data which has provided a wealth of spectral and spatial information for more accurate texture characterization.

Cross-channel processing

The second approach is cross-channel processing which is the direct extension of opponent process in the hyperspectral domain. In particular, opponent features are obtained by the differences between features derived from one spectral band and that from the other spectral band. This has been applied in the context of hyperspectral texture feature extraction using Gabor filter [70] and LBP [43]. Due to the manner in which the features are generated (cross-channel processing), a final feature size of $L^2 - L$ is obtained which can be gigantic depending on the number of spectral bands L. As such, dimensionality reduction is usually performed before opponent feature extraction. Well-founded for processing of colour or trichromatic texture (thanks to the opponent process in the human visual system), cross-channel processing is, however, hardly justifiable in the case of hyperspectral texture. In fact, it is a forced adaptation of the opponent process in the hyperspectral domain. Furthermore, it exists no proper mathematical or physical justification for cross-channel processing. The argument that such approach is devised for taking into account spatial correlation between spectral bands is deemed weak due to the ad hoc nature.

Vectorial processing

The third approach is vectorial processing. Apart from the two-dimensional spatial dimensions, the spectral bands are treated as the third dimension. Having established that, existing

two-dimensional features defined for grayscale images can thus be adapted in three dimensions. For example, GLCM is extended to graylevel cooccurrence tensor field (GLCTF) [80] with the spatial relationship defined in a three-dimensional (3D) manner. Expressed as a 3D matrix, GLCTF records the cooccurrence of three graylevels within a 3D moving box. In the case of LBP, it is extended to three-dimensional LBP (3D-LBP) [38] considering an ellipsoidal neighbourhood instead of a circular one. To simplify the calculation, only $P = 6$ vertices are considered with "radius" of the octahedron fix at $r = 1$. Similarly, 3D-Gabor [3], 3D-wavelet [29], and 3D-MRF [61] have been defined by performing joint spectral-spatial assessment. Although being mathematically plausible, the textural inferences derived from vectorial processing are questionable. Indeed, the physical meanings of GLCM, LBP, Gabor, wavelet, and MRF are originally defined in two (spatial) dimensions and may not be directly extendable to higher dimensions. Physically it is insufficient to merely claiming spectral information has been added on top of the originally defined spatial properties, for example by claiming the LBP codes as of "spatiochromatic" textons, or the Gabor energies as of "spatiochromatic" frequencies. Without proper interpretation of the third-order measures, vectorial processing can only be seen as some sort of mathematical convenience without any physical meaning in relation to the underlying texture complexity. On the other hand, the choice of parameters also becomes less defined in the context of the hyperspectral application. For example, what is the physical meaning of r in the spectral dimension in the case of GLCTF (3D version of GLCM)? In the case of GLCM, the optimal choice of r is proportional to texton size. Applying the same selection criterion in the spectral dimension, the optimal r thus must be corresponding to the spectral variation. Accordingly, smaller r is required for spectrally distinct pixel pairs whereas larger r is permissible for spectrally similar pairs. Another complexity can be foreseen if one were to pursue in such direction: do the texture features defined from these 3D measures biased towards the spatial or spectral variation?

Indeed, many questions remained unanswered in vectorial processing and care must be undertaken before applying them.

1.3 Towards texture metrology

A spatiochromatic definition

Julesz conjectured that preattentive texture discrimination is limited to differences up to second-order statistics [41]. In other words, humans can only perceive texture differences up to the degree to which a pair of points (separated at a distance) varies. Also directly implied is that difference in first-order statistics e.g. intensity and colour also contributes to the discrimination. Later, Julesz perfected his theory that texture discrimination is based on first-order statistics of textons[1] like line terminators, line crossings, and elongated blobs of a certain colour, orientation, size etc. [42] These psychophysical findings point to an important fact: spectral and spatial properties are different facets of texture, and that a complete characterization requires *joint description of both*. On the other hand, scientific studies into human vision have revealed that we are not equally sensitive to spatial variations in brightness and colour [21]. In fact, while we are very good at picking up very fine details in brightness (intensity), we are less sensitive to fine spatial changes of colour. As contrast is highly related to texture perception, these facts further support a spatiochromatic definition of texture.

A statistical treatment

Having said that, we propose to define texture as the *joint distribution of spectral and spatial properties*. Intended to be as generic as possible, the said properties can include intensity, colour, directionality etc. which can be defined later. Of particular interest should be the

[1] "The putative units of preattentive human texture perception".

used notion of *distribution* (specifically, a continuous one) which speaks aloud its intended information-theoretic characterization. Indeed, information in the real world is anything but deterministic. In the presence of noise as well as metrological constraints, uncertainty is the only constant[2]. Furthermore, spatiochromatic properties associated with natural (non-man-made) textures are clearly stochastic. Spectrally, this is obvious as the texture is a spatially varying phenomenon. The same can also be spoken about its spatial properties, for example, there can never be a texture with precisely defined directionality (directionalities) unless it is artificially generated. At best, a natural texture can only exhibit *main* or major directionality (directionalities) in the presence of other minor orientations. Therefore, the texture must be modelled as a distribution which allows quantization of the minor variations. That said, distribution can also efficiently encode major variations. For instance, a bichromatic (or trichromatic, quadchromatic etc.) texture can be intuitively modelled using bimodal distribution in the spectral dimension. Indeed, a distributional formulation also facilitates feature interpretability which is one of the pillars of metrology.

Full-band processing or dimensionality reduction?

Evidently, the large number of spectral bands renders any attempt to construct an image histogram out of a hyperspectral image impractical. Facing this problem, two approaches were proposed to reduce the number of bands. The first is **feature extraction**. In particular, principal component analysis (PCA) [58] has been extensively used for projecting spectral data into lower subspace. PCA uses a linear transformation to translate and rotate multiband data into a new coordinate system that maximizes the variance. The principal components (PCs) associated with large eigenvalues contain most of the information, while the PCs associated with small eigenvalues are dominated by noise. Thus, only the first few PCs can be retained to obtain lower-dimensional data while preserving as much of the data's variation as possible. Other techniques which have been used for dimensionality reduction include

[2]In resonance with the saying: the only constant in life is change.

independent component analysis (ICA) [15] and nonnegative matrix factorization (NMF) [47]. A review of the dimensionality reduction techniques for hyperspectral image processing may be found in [49].

However, the feature extraction techniques can be hardly justified in the context of metrology as they are data dependent. For example, in the context of PCA, the PCs are calculated based on the covariance matrix, which is in turn uniquely defined for each given image. As a result, spectral information extracted this way could not be compared across different images. In addition, there is no physical sense to the weighted sum of measurements coming from different spectral ranges and physics e.g. visible vs. infrared region. Moreover, PCA requires considering the spectrum as a Euclidean vector in L-dimensional space. This is manifested in the calculation of the covariance matrix which uses 2-norm as distance measures. The same requirement can also be found in ICA and NMF. This violates the physics which requires to consider spectrum as a continuous function, rendering these dimensionality reduction techniques unsuitable for metrological purposes from the beginning.

Another approach for reducing the number of spectral bands is **feature selection**, which is also known as band selection. It aims to select a small subset of hyperspectral bands to remove spectral redundancy and reduce computational costs while preserving the significant spectral information of ground objects. A review of various band selection techniques for hyperspectral images may be found in [74]. However, such approach seems ironic because, why acquire images in tens to hundreds of spectral bands for spectral accuracy, only to finally reduce them to a few bands for computational simplicity? Why not then just work with colour or multispectral images in the first place? Indeed, the interest in working with hyperspectral images is to obtain a more accurate spectral representation thanks to its dense spectral sampling. Not taking advantage of such rich information may constitute a missed opportunity while defying the whole purpose of hyperspectral imaging.

Chapter 2

Spectral Characterization of Texture

Contents

2.1 Background 16

2.2 The proposed formulation 21

2.2.1 Spectral decomposition into shape and intensity 21

2.2.2 Spectral difference histogram 24

2.2.3 Assessment of spectral diversity 30

2.2.4 Automatic selection of spectral reference 35

2.3 Summary 54

In this chapter, we developed a metrological approach to characterize the spectral properties of hyperspectral image(s). In particular, we proposed methods to quantify spectral shapes (chromatic parts) and spectral intensities (achromatic parts) based on fullband processing. We also proposed an automatic spectral reference selection for operation in metric space.

2.1 Background

We begin by briefly recalling the physics of spectrum. Then, we review some common measures of spectral difference in the hyperspectral community. Finally, we consider the problem of constructing a histogram for hyperspectraimagesge facing the high number of spectral channels.

Physics of spectrum

Physically, a spectrum s is a continuous function f of wavelength λ such that $s = f(\lambda)$. As such, the (spectral) "space" of spectrum is infinite dimensional. However, the "space" of a sampled (discretized) spectrum $s = \big(s(\lambda_1), s(\lambda_2), \ldots, s(\lambda_L)\big)$ is finite dimensional. In this case, the dimensionality of spectral space is given by the number of (spectral) samples $s(\lambda_l)$ acquired with L wavelengths. That said, not every point in the spectral space constitute the valid spectra. A spectrum is strictly positive and thus only "lives" in the postive hyperoctant[1].

In the literature, it is common to treat a spectrum acquired with L samples as a Euclidean vector in L-dimensional space which then allows its manipulation using linear algebra. However, such approach actually lacks physical plausibility. For the Euclidean vector representation to be physically valid, the spectral samples must be independent (orthogonal) and thus uncorrelated. Accordingly, they are considered to be of no particular order such that the distance between two spectra remains unchanged with varying order of samples. Obviously, these are not the case as neighbouring spectral samples are indeed highly correlated [63], one

[1] Analogous to first quadrant in the two-dimensional space

reason being overlap of the spectral sensitivity functions. In this work, **we shall adhere to the fact that spectrum is a continuous function**; not as a mere collection or list of samples.

Assessment of spectral difference

Treating spectra as Euclidean vectors, it is straightforward to employ Euclidean distance or root mean square error (RMSE) for calculating the difference between two spectra s and s':

$$\mathrm{RMSE}(s,s') = \sqrt{(s-s')\cdot(s-s')}. \tag{2.1}$$

RMSE is adapted for assessing spectral intensity difference but not spectral shape difference. This is because RMSE does not take into account the ordering of spectral samples.

On the other hand, spectral angle mapper (SAM) [88] defines spectral difference by calculating the angle (in Euclidean vector space) between two spectra s and s':

$$\mathrm{SAM}(s,s') = \frac{s\cdot s'}{\|s\|_2\,\|s'\|_2}. \tag{2.2}$$

Clearly, SAM is invariant to spectral intensity difference because the angle is independent of the vectors' length. Besides, SAM also does not take into account the ordering of samples.

To conform to physics, the spectral difference should be calculated based on the fact the hat spectrum is a continuous function. A first step is taken in the work of spectral information divergence (SID) [9] which treats spectra as subject of randomness and hence as probability distributions. However, spectral normalization is required for transformation into probability before calculating the spectral difference using Kullback-Leibler (KL) divergence:

$$\begin{aligned} \mathrm{SID}(s,s') &= \mathrm{KL}(\bar{s}\|\bar{s}') + \mathrm{KL}(\bar{s}'\|\bar{s}, \\ \text{s.t.} \quad \mathrm{KL}(\bar{s}\|\bar{s}') &= \sum_{l=1}^{L} \bar{s}(\lambda_l) \log \frac{\bar{s}(\lambda_l)}{\bar{s}'(\lambda_l)}, \\ \bar{s} &= s \Big/ \sum_{l=1}^{L} s(\lambda_l), \end{aligned} \tag{2.3}$$

where $\bar{s}$ and $\bar{s}'$ are the normalized spectra. Taking into account the order of samples, the efficiency of SID in measuring spectral shape difference is superior to that of RMSE and SAM [8, 9]. However, SID too is invariant to spectral intensity difference.

Contrary to SID, spectral normalization is not required for calculating Kullback-Leibler pseudodivergence (KLPD) [62] which otherwise removes spectral intensity information. Expressing spectrum $s = \|s\|_1 \bar{s}$ in terms of its spectral intensity $\|s\|_1$ and its normalized version $\bar{s}$, KLPD defines KL divergence between two (unnormalized) spectra s and s' as:

$$\begin{aligned} \mathrm{KLPD}(s,s') &= \Delta G + \Delta W, \\ \text{s.t.} \quad \Delta G &= \|s\|_1 \, \mathrm{KL}(\bar{s}\|\bar{s}' + \left\|s'\right\|_1 \mathrm{KL}(\bar{s}'\|\bar{s}, \\ \Delta W &= (\|s\|_1 - \left\|s'\right\|_1) \log \left(\frac{\|s\|_1}{\|s'\|_1} \right), \end{aligned} \tag{2.4}$$

where ΔG and ΔW are termed spectral shape and intensity differences respectively.

Histogram of hyperspectral image

An image histogram depicts the distribution of pixel values (spectral[2] distribution) in the given image. In the context of monovariate image $I : \mathbb{R}^2 \to \mathbb{R}$, the image histogram is given by the histogram of graylevels. Considering a typical 8-bits image quantization, the

[2]Not to be confused with spectral power distribution. In the context of our work, spectral distribution refers to the generalization of graylevel and colour distribution in a given image with any number of channels.

histogram then consists of $2^8 = 256$ bins. If it is normalized such that its bins sum to unity, the histogram dictates the probability of finding a particular graylevel at a given spatial location.

In the context of multivariate image $I : \mathbb{R}^2 \to \mathbb{R}^L$ with L spectral channels, two approaches exist for constructing the image histogram. The first is to construct a L-dimensional histogram with each axis corresponding to a particular spectral channel. However, such multidimensional histogram would be extremely sparse with large L as in the case of hyperspectral image. Considering a typical 8 bits quantization for each of the spectral channels, the L-dimensional histogram would then consist of $(2^8)^L$ bins which are computationally infeasible. The second approach is to assume independence among the spectral channels, which then allows one to consider just the marginal distribution of spectral samples. Accordingly, it is then sufficient to just use L histograms of graylevels (one for each channel) to characterize the entire spectral distribution. However, such assumption of independence actually never holds according to the physics of the spectrum. To recall, this is because adjacent spectral channels are highly correlated as they are contiguously sampled in the spectral dimension.

Meanwhile, a proposition has been made in [19] to employ spectral difference space as the feature space. Accordingly, spectral variability is assessed by the distribution of spectral differences with respect to a chosen spectral reference $\hat{s}$. Based on KLPD, the bidimensional histogram of spectral difference (BHSD) dictates the probability of finding a particular pair of spectral shape and intensity differences $(\Delta G, \Delta W)$ at a given spatial location x [62]:

$$\mathrm{BHSD}^{(\hat{s})}(I) = p\left(\mathrm{KLPD}\left(I(x), \hat{s}\right) = (\Delta G, \Delta W)\right). \tag{2.5}$$

Evidently, the distribution of spectral differences (and hence spectral discrimination) depends on the chosen $\hat{s}$. For maximum discrimination, $\hat{s}$ should be selected such that it lies on the convex hull $Conv(S)$ of the spectral set $S = \{s_1, \ldots, s_n\}$ from the given image with n pixels and L spectral channels [19]. As demonstration, consider an example in Fig. 2.1 concerning

three choices of $\hat{s}$ for an image with $L = 2$. Suppose that the pixels can be clustered into three groups: A, B, and C. In Fig. 2.1 (a), $\hat{s}$ is very far from $Conv(S)$. Consequently, the clusters are hardly distinguishable in the spectral difference space as KLPD saturates. In Fig. 2.1 (b), $\hat{s}$ is located inside $Conv(S)$. As a result, some of the clusters (A and B) cannot be completely discriminated as they are "equidistant" from $\hat{s}$. In Fig. 2.1 (c) whereby $\hat{s}$ is selected such that it lies on $Conv(S)$, the clusters are clearly differentiated in the spectral difference space. That said, such approach necessitates a prior computation of $Conv(S)$ of which the complexity is given by e.g. $\mathcal{O}(n^{\lfloor L/2 \rfloor + 1})$ based on the gift wrapping algorithm [18]. Clearly, this is infeasible for hyperspectral images with L typically ranging from tens to hundreds or even thousands.

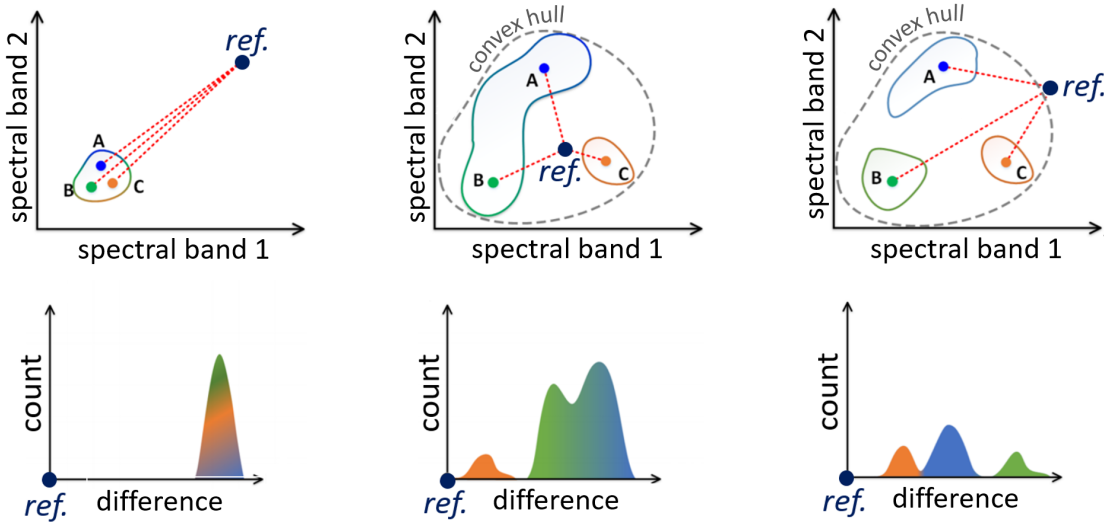


Fig. 2.1 Three possible choices of spectral references (top row) concerning an image with $L = 2$ spectral bands (channels), and their resulted representation of spectral clusters A, B, and C in spectral difference space (bottom row). For maximum discrimination, the spectral reference should be selected such that it lies on the convex hull of the spectral set as in (c). Image adapted from [19].

2.2 The proposed formulation

Having presented the problem of constructing histogram for hyperspectral image(s), in the following we provide a solution by first devising spectral decomposition into shape and intensity. Then, we address the representation of spectral shape in metric space. Finally, we propose the spectral difference histogram complete with an assessment of spectral diversity.

2.2.1 Spectral decomposition into shape and intensity

In colour image processing, the description of a given colour can be divided into two parts: luminance (alternatively termed brightness, lightness, value, etc.) and chromaticity (e.g. hue and saturation in HSV colour space) [69]. For example, *white* is a brighter colour than *gray* which is a darker version of that same white. Thus, the chromaticity of *white* and *gray* are the same as their luminance differs. Similarly, two colours can have different chromaticities but the same luminance. Indeed, luminance and chromaticity are independent of each other.

That said, we seek the same decomposition in the context of spectrum. Referring to Fig. 2.2, we introduce the notion of spectral intensity W and spectral shape G which are analogous to luminance and chromaticity respectively in colour science. Specifically, we **define spectral intensity W of a given spectrum s by its continuous integral regardless of its shape** [62]:

$$W \stackrel{\text{def}}{=} \int_{\mathbb{R}} s(\lambda) d\lambda, \tag{2.6}$$

whereby a discrete approximation can be obtained using e.g. trapezoidal rule:

$$\int_{\lambda_1}^{\lambda_L} s d\lambda \approx \sum_{l=1}^{L} \frac{s(\lambda_{l-1} + s(\lambda_l)}{2} \Delta\lambda_l, \tag{2.7}$$

where L refers to the number of spectral samples. Thanks to this metrological construction, spectral resolution $\Delta\lambda_l$ only affects the accuracy of the discrete approximation but not the

range of it. An alternative, but deemed less metrological approach is to define spectral intensity using discrete sum as in the case of SID for spectral normalization [8]:

$$W = \sum_{l=1}^{L} s(\lambda_l), \tag{2.8}$$

which is clearly not invariant to spectral resolution. Evidently, spectral intensity defined this way increases with the number of spectral bands. As such, spectra acquired from different spectral sensors are not directly comparable and thus not adapted for metrology.

On the other hand, we **define spectral shape G of a given spectrum s by its normalized version $\bar{s}$ regardless of its intensity**. Integrating to unity, it is mathematically expressed as:

$$G \overset{\text{def}}{=} \bar{s} = \frac{s}{W} \qquad \text{s.t.} \qquad \int_{\mathbb{R}} \bar{s}(\lambda) d\lambda = 1. \tag{2.9}$$

The handling and manipulation of spectral shape is, however, not as easy and straightforward as spectral intensity e.g. in constructing the histogram of spectral shape. This is because the spectral shape is by definition, a function in contrast to spectral intensity which is a scalar.

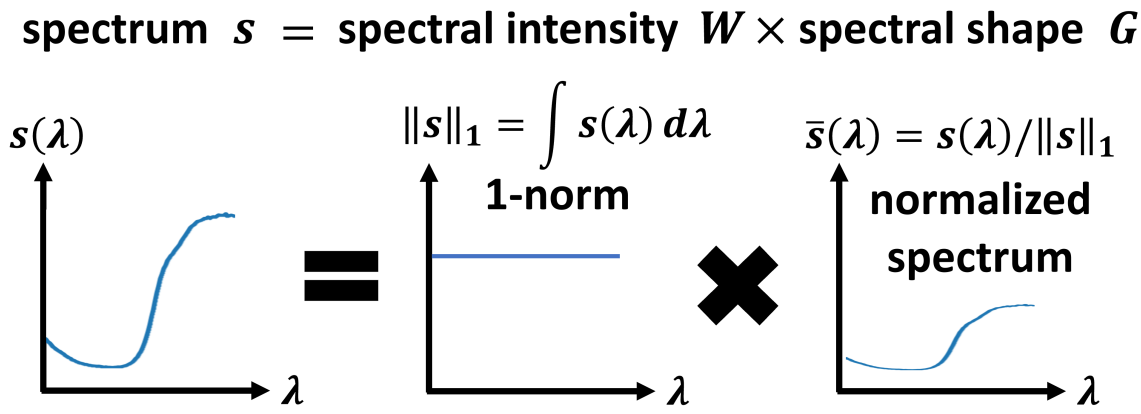


Fig. 2.2 The proposed decomposition of a given spectrum s in terms of spectral intensity W (as defined using 1-norm $\|s\|_1$) and spectral shape G (expressed as normalized spectrum $\bar{s}$).

Inspired by BHSD [62], we propose to characterize spectral shape G by working in metric space. Specifically, we **represent spectral shape G of a given spectrum s by the spectral shape difference ΔG between its normalized version $\bar{s}$ and a chosen spectral reference $\hat{s}$**:

$$G \xrightarrow{\text{metric}} \text{KLD}\,(\bar{s},\hat{s}) = \Delta G, \tag{2.10}$$

noting the discrepancy between ΔG defined here and the spectral shape difference as defined for KLPD (see eqn. 2.4). To be precise, the spectral shape difference space is actually a pseudometric space as KL divergence does not respect the triangle inequality [44].

To shed some insight into our approach, let us recall the operation of principal component analysis (PCA). The goal of PCA is to project the original data $Y = [y_1, y_2, \ldots, y_n]^T$ in L dimensions with n data points $y_i \in \mathbb{R}^L$ into a subspace which gives a more compact representation of the L variables [58]. To do so, PCA finds the projection planes which preserve, as much as possible, the distances between y_n. These projection planes called the principal components (PCs), are defined by the eigenvectors $v_l \in \mathbb{R}^L$ for $l \in [1, L]$ of the covariance matrix[3] $Y^T Y$, and ranked according to its variance which is defined by the corresponding eigenvalue. To project the original data onto the PCs, one calculates $Y_v = YV$ where $V = [v_1, v_2, \ldots, v_L]$. In other words, each of the original n data points can be transformed into the new axes by the dot products $y_n \cdot v_l$. For dimensionality reduction, one calculates $\widetilde{Y} = Y\widetilde{V}$ for $\widetilde{L} < L$, where $\widetilde{V} = [v_1, v_2, \ldots, v_{\widetilde{L}}]$ is the set of the top $\widetilde{L}$ ranked PCs.

In retrospect, **the spectral (shape) difference in eqn. 2.10 can be thought as a spectral projection onto the spectral reference**. Indeed, the calculation of spectral difference is analagous to the dot product $y_n \cdot v_l$ in PCA which projects a given data point y_n (analagous to the given spectrum) onto the PC v_l (analogous to spectral reference). We note that the dot product can be written as $y_n \cdot v_l = \|y_n\|_2 \, \|v_l\|_2 \cos\phi$, where ϕ is the angle between y_n and v_l. In the case of $\|y_n\|_2 = 1$ (analogous to normalized spectrum $\bar{s}$) and $\|v_l\|_2 = 1$, the dot

[3] Y should have zero mean to ensure that the first PC describes the direction of maximum variance, or otherwise the first PC might instead correspond more or less to the mean of the data

product is then given by the cosine similarity $\cos\phi$ which is analogous to spectral difference. Recalling the manner in which the PCs are defined (i.e. to maximize the variance of projected data), it is also straighforward to devise a strategy to select spectral reference such that it maximizes the spread[4] of spectral differences in the metric space (see sec. 2.2.4).

2.2.2 Spectral difference histogram

Following our proposed spectral decomposition, we intend to construct a histogram that jointly considers spectral shapes and spectral intensities of a given hyperspectral image $I : \mathbb{R}^2 \rightarrow \mathbb{R}^L$ with L spectral channels. Preliminarily, we define spectral difference histogram $\mathscr{J}_1^{(\hat{s})}(I)$ as the joint probability density function (PDF) of spectral shape difference ΔG (with respect to spectral reference $\hat{s}$) and spectral intensity W at a given spatial location x:

$$\mathscr{J}_1^{(\hat{s})}(I) = p\left(\begin{array}{c} \mathrm{KLD}\left(\bar{I}(x), \hat{s}\right) = \Delta G, \\ \\ \left\|I(x)\right\|_1 = W \end{array}\right) \qquad (2.11)$$

$$\text{s.t.} \qquad \left\|\bar{I}(x)\right\|_1 = \|\hat{s}\|_1 = 1.$$

Having established the conceptual equivalence between PCA and our spectral shape G representation in metric space i.e. through spectral shape difference ΔG, it is straightforward to represent G using multiple spectral references $\hat{s}_r$ for $r \in [1, R]$ where R is the number of spectral references. Indeed, a single spectral reference might be insufficient to account for all the spectral shape variability. This is because dissimilar spectral shapes may result in the same spectral difference with respect to a single spectral reference. With multiple spectral references (analogous to the projection of data by PCA in a subspace as defined using several

[4]The notion of variance is only defined in Euclidean vector space. As the metric (spectral difference) space is not Euclidean, we shall not quantify the spread of spectral differences using variance (see ssec. **??**).

principal components), it is expected that the spectral shapes being more thoroughly defined. That said, we expand the spectral shape representation in metric space (see eqn. 2.10) by:

$$G \xrightarrow{\text{metric}} \begin{cases} \text{KLD}(\bar{s}, \hat{s}_1) = \Delta G_1 \\ \text{KLD}(\bar{s}, \hat{s}_2) = \Delta G_2 \\ \vdots \end{cases} \tag{2.12}$$

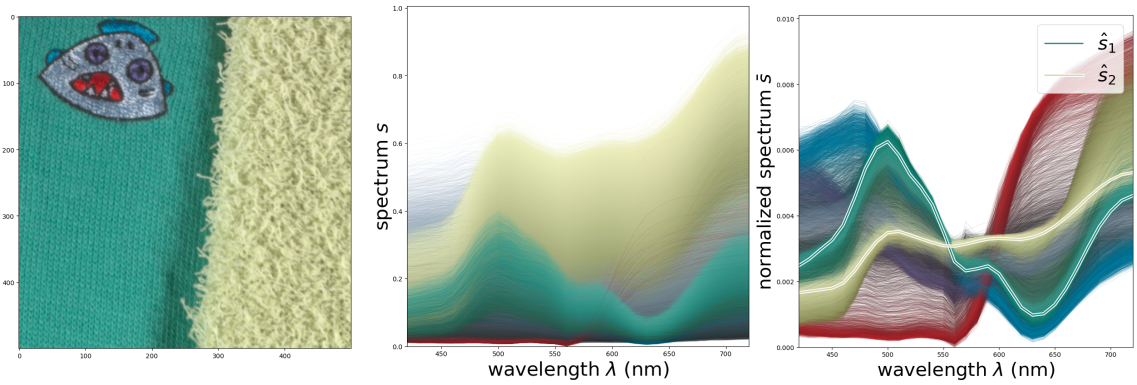
Considering R spectral references $\hat{S} = \{\hat{s}_1, \ldots, \hat{s}_R\}$, we define the spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ as the joint PDF of $R+1$ random variables i.e. the spectral shape differences $\{\Delta G_1, \ldots, \Delta G_R\}$ and spectral intensity W at a given spatial location x:

$$\mathscr{J}_1^{(\hat{S})}(I) = p\left(\begin{array}{c} \bigcap_{r=1}^{R} \left(\text{KLD}\left(\bar{I}(x), \hat{s}_r\right) = \Delta G_r \right), \\ \left\| I(x) \right\|_1 = W \end{array} \right) \tag{2.13}$$

$$\text{s.t.} \qquad \left\| \bar{I}(x) \right\|_1 = \|\hat{s}_r\|_1 = 1 \;\forall\; r \in [1, R].$$

We demonstrate the calculation of spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ based on the hyperspectral image *Cloth4* (500 $\times$ 500 pixels) I with wavelengths ranging from 420 nm to 720 nm at 10 nm interval [54] as shown in Fig. 2.3 (a). Referring to the spectra (reflectances) s as depicted in Fig. 2.3 (b), we can observe mainly five spectral groups of green, white, light violet, blue, and red colours. Specifically, a large majority of the spectra are either green or white, whereas much fewer are either light violet, blue, or red spectra. To characterize the spectral shapes G, we first normalize the spectra as illustrated in Fig. 2.3 (c). Clearly, the spectral shapes can now be observed effortlessly with the intensity information removed. Then, we randomly choose a spectrum from the green and from the white spectra as the spectral references $\hat{S} = \{\hat{s}_1, \hat{s}_2\}$. Note that our colour description here is not exact and only serves as reference (to be referred to easily in the following text). For example, the

green spectra are not genuinely green per se as they actually have more bluish and reddish components compared to true greens which theoretically peak around 550 nm. On the other hand, the white spectra are also not genuinely white as they actually have less bluish and more reddish components, thus leading to a warmer[5] white appearance. A genuine white spectrum is however, theoretically equinergy i.e. having constant energy throughout the visible spectrum. Such is the interest of hyperspectral imaging as it allows us to see beyond trichromatic colour and into the spectral power distribution of each pixel at each wavelength.



(a) (b) (c)

Fig. 2.3 (a) The hyperspectral image *Cloth4* (500 $\times$ 500 pixels) with wavelengths λ ranging from 420 nm to 720 nm at 10 nm interval, rendered in RGB for illustration; (b) all the (500 $\times$ 500 = 250000) spectra s of *Cloth4*; and (c) all the normalized spectra $\bar{s}$ of *Cloth4*.

We illustrate the distribution of spectral shape differences ΔG_1 and ΔG_2 in Fig. 2.4 (a) with a zoomed-in version in Fig. 2.4 (b). It can be seen that in general, the green spectra have lower ΔG_1 whereas the white spectra have lower ΔG_2. This is expected because $\hat{s}_1$ (associated with ΔG_1) and $\hat{s}_2$ (associated with ΔG_2) are one of the green and white spectra respectively. On the other hand, ΔG_1 of the blue spectra are generally lower than their ΔG_2. This suggests the higher spectral shape similarity between the blue spectra and $\hat{s}_1$ (green spectrum) compared to $\hat{s}_2$ (white spectrum) which is indeed the case as observable from Fig. 2.3 (c). As for the red spectra, their ΔG_1 is generally higher than their ΔG_2 which is

[5]A warm white is yellowish, reminiscent of light coming from tungsten light bulbs; a cool white leans towards bluish, resembling the white LED light.

expected considering their higher spectral shape similarity with $\hat{s}_2$. Concerning the light violet spectra, their ΔG_1 is comparable to ΔG_2, indicating the same level of spectral shape similarity (or dissimilarity) with both $\hat{s}_1$ and $\hat{s}_2$. Clearly, the spectral shape representation in metric (spectral difference) space is compact and fully interpretable. In comparison, we perform principal component analysis (PCA) of the normalized spectra $\bar{s}$. We illustrate the top two principal components (PCs) in Fig. 2.4 (c) which account for 96.5 % of explained variance. In contrast to $\hat{s}_1$ and $\hat{s}_2$, the PCs carry little physical interpretation as they are not real spectra but merely abstract algebraic constructs. Referring to the spectral distribution in the transformed subspace corresponding to the top two PCs as depicted in Fig. 2.4 (d), no physical interpretation too can be given except that the spectra are clustered according to their spectral shape as in the case of our approach. Indeed, PCA can be a solution if clustering is the only goal. However, our approach can be useful if physical interpretability is also required as in the case of data analysis concerning remote sensing and health applications.

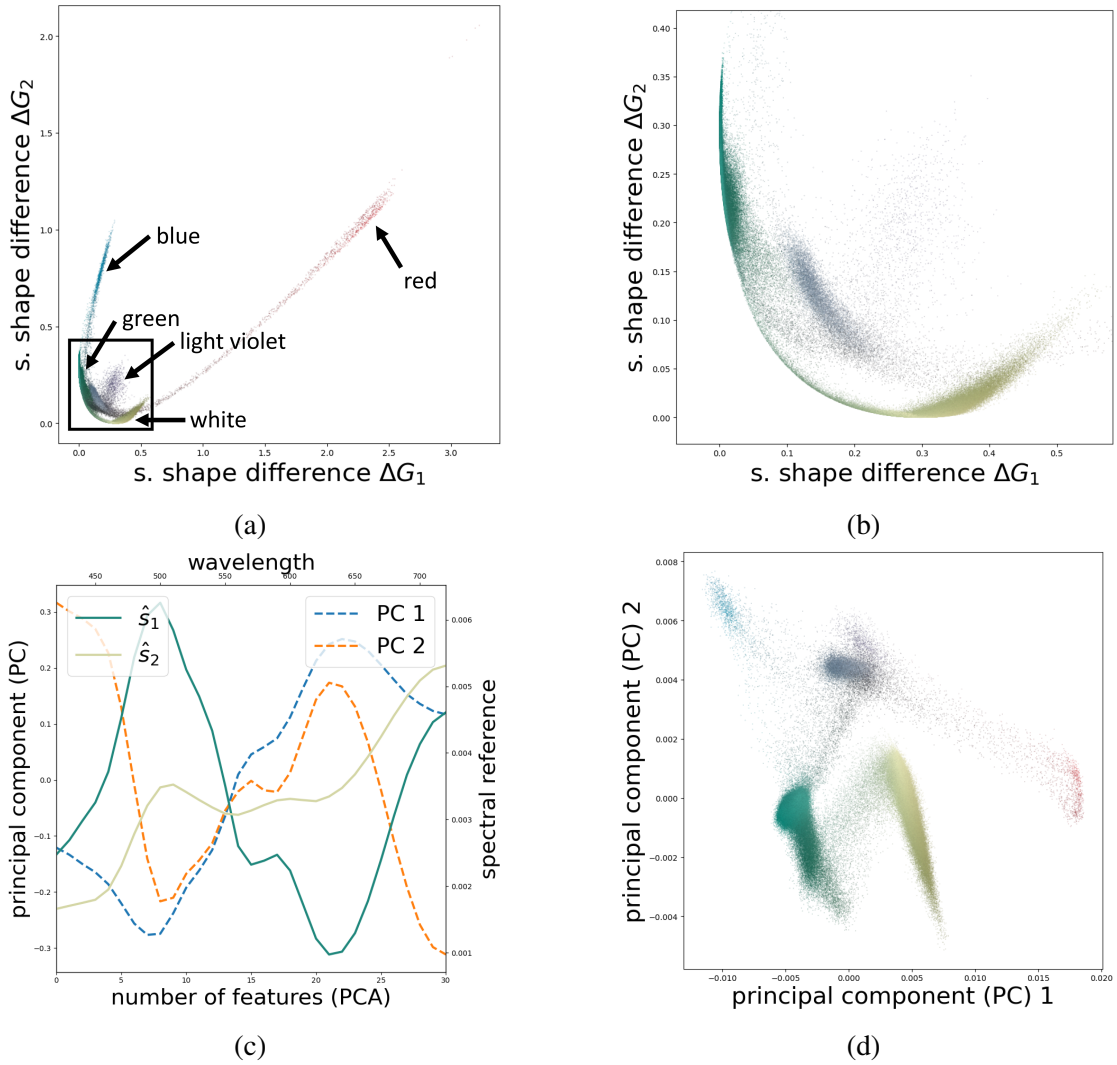
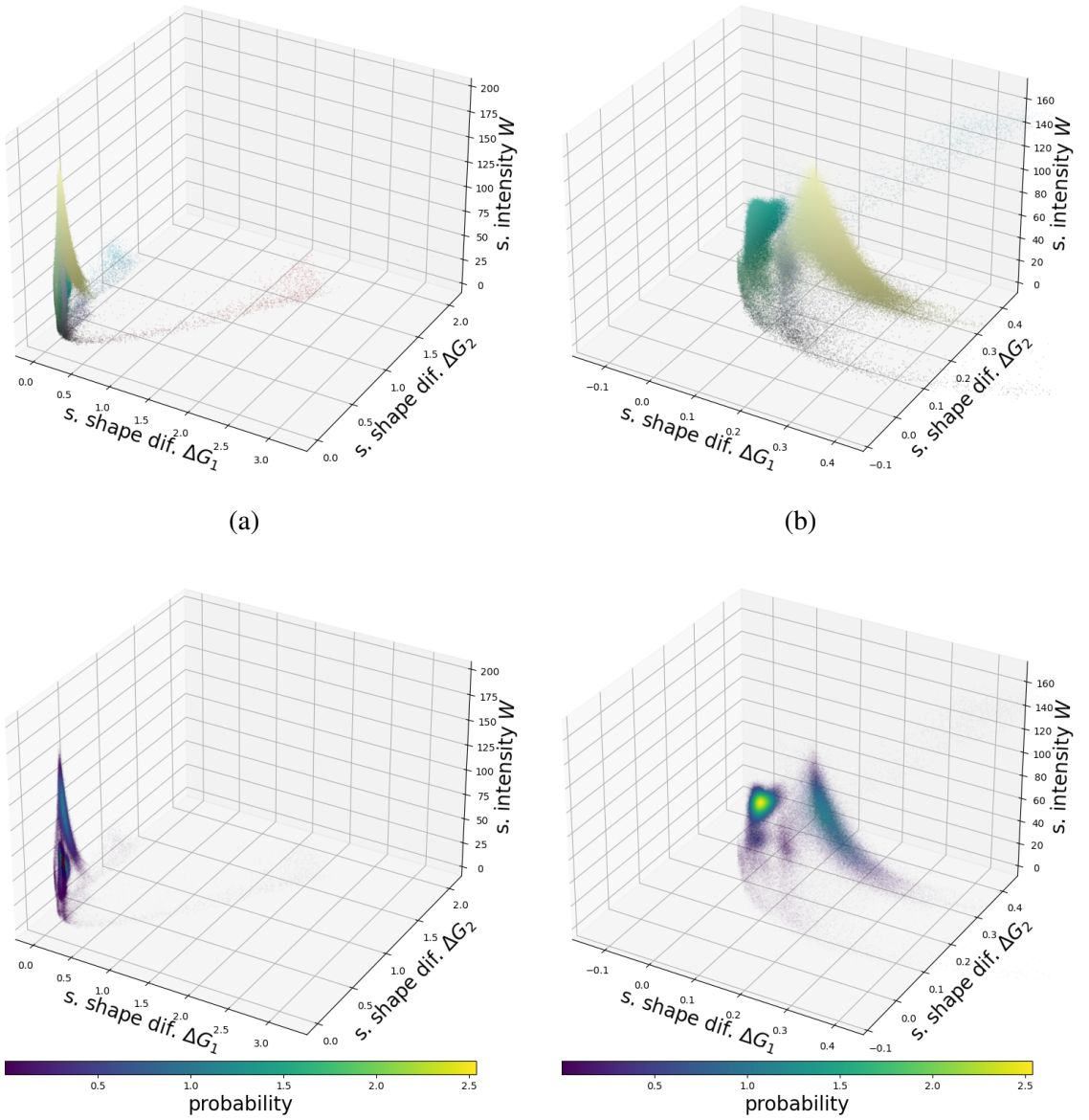


Fig. 2.4 (a) A bidimensional scatter plot of spectral shape differences ΔG_1 and ΔG_2 of *Cloth4* corresponding to the spectral references $\hat{s}_1$ and $\hat{s}_2$ respectively; (b) a zoomed in version of (a); (c) the top two principal components (PCs) as obtained from principal component analysis (PCA) of the normalized spectra $\bar{s}$, ovelaid with $\hat{s}_1$ and $\hat{s}_2$; and (d) a bidimensional scatter plot of $\bar{s}$ in the subspace spanned by the top two PCs.

For comprehensive representation, we illustrate the joint distribution of spectral shape differences ΔG_1 and ΔG_2 as well as spectral intensity W in Fig. 2.5 (a) with a zoomed-in version in Fig. 2.5 (b). It can be seen that in general, the white spectra have the highest W, in contrast to the red and blue spectra which have some of the lowest W. Such result agrees with our visual observation that the white pixels are generally brighter than others as seen

from Fig. 2.3 (a). Finally, we display the spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ in Fig. 2.5 (c) with a zoomed-in version in Fig. 2.5 (d). It can be seen that the green spectra are more compact compared to the white spectra. This indicates that the variability of the green spectra is lower compared to that of the white spectra. It is clear, then, information about spectral shapes and intensities can be communicated clearly and compactly thanks to $\mathscr{J}_1^{(\hat{S})}(I)$.



(a) (b)

(c) (d)

Fig. 2.5 (a) A three-dimensional scatter plot of spectral shape differences ΔG_1 and ΔG_2 as well as spectral intensity W of *Cloth4*; (b) a zoomed in version of (a); (c) the spectral difference histogram of *Cloth4*; and (d) a zoomed in version of (c).

2.2.3 Assessment of spectral diversity

Having defined the spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ to capture the spectral shape (differences) and intensity information of a given hyperspectral image I based on the selected

spectral references $\hat{S}$, we are now interested in devising a measure to quantify spectral diversity. As $\mathscr{J}_1^{(\hat{S})}(I)$ is a probability distribution, we are naturally directed towards the notion of entropy. In the field of information theory, the Shannon entropy $H(Y)$ of a discrete random variable Y reflects the average amount of information or uncertainty it carries [68]:

$$H(Y) = -\sum_{y \in Y} P(y) \log P(y), \tag{2.14}$$

where $P(y)$ is the probability mass function of the discrete random variable Y. To shed some insight on the measure of information by $H(Y)$, consider the alternative definition $H(Y) = \mathbb{E}\left[\mathbb{I}(y)\right]$ where $\mathbb{E}[\cdot]$ denotes expectation and $\mathbb{I}(y) = -\log P(y)$ refers to self-information or "surprisal" of measuring Y as y. Clearly, an event that has 100 % probability i.e. $P(y) = 1$ is unsurprising and carries no information as $\mathbb{I}(y) = 0$. Conversely, an event that has a lower probability is more suprising and carries more information. Accordingly, if there is only one probable event, $H(Y)$ is zero regardless of what the event is; if all the events are equally probable such that $P(y)$ is a uniform distribution, $H(Y)$ is maximum. Analogous to variance $\mathbb{E}\left[(Y-\mu)^2\right]$ which describes the spread of values around the mean μ in the Euclidean vector space, $H(Y)$ can thus be understood as the spread of values in the probabilistic space.

That said, we are obliged to consider an evaluation in continuous domain because the spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ is a probability density function (PDF) of spectral shape differences and spectral intensity. A seemingly direct measure is the differential entropy $h(Y)$ which replaces the discrete sum in eqn. 2.14 with continuous integral [68]:

$$h(Y) = -\int p(y) \log p(y) dy, \tag{2.15}$$

where $p(y)$ is the PDF of the continuous random variable Y. However, $h(Y)$ is not a good measure of information e.g. $h(Y)$ can be negative since PDFs can be greater than 1. For example, $h(Y) = \int_0^{1/2} (-2\log 2)\, dy = -\log 2$ for the uniform distribution $\mathscr{U}(0, 1/2)$, which

is unintelligible as information cannot be negative. Instead, the correct formula for assessing entropy in continuous domain is given by the limiting density of discrete points (LDDP) [37]:

$$\begin{aligned} H_n(Y) &= \log n - \mathrm{KL}(p(y)\|\mathscr{U}) \\ &= \log n - \int p(y) \log \frac{p(y)}{\mathscr{U}} dy \\ &= \log n - \left(\int p(y) \log p(y) dy - \int p(y) \log \mathscr{U}\, dy \right) \\ &= \log n + h(y) + \int p(y) \log \mathscr{U}\, dy, \end{aligned} \tag{2.16}$$

where n is the number of data points and $\mathrm{KL}(p(y)\|\mathscr{U})$ is the KL measure of information from uniform distribution $\mathscr{U}$ to $p(y)$. Suppose that $y \in (a,b)$, then $\mathscr{U} = \frac{1}{b-a}$, giving:

$$\begin{aligned} H_n(Y) &= \log n + h(y) + \log \frac{1}{b-a} \int p(y) dy \\ &= \log n + h(y) - \log(b-a) \\ &\propto h(y). \end{aligned} \tag{2.17}$$

Clearly, $H_n(Y)$ only differs from $h(y)$ by a constant. To avoid dependency on n and (a,b), which in the context of $\mathscr{J}_1^{(\hat{s})}(I)$ refer to the size of image and dynamic range of pixel values respectively, we **define spectral diversity as being proportional to the joint differential entropy of the spectral shape differences** $\{\Delta G_1, \ldots, \Delta G_R\}$ **and spectral intensity** W:

$$\begin{aligned}\text{spectral diversity} &\propto h(\Delta G_1,\ldots,\Delta G_R,W)\\ &= -\int\ldots\int\int\left(\mathscr{J}_1^{(\hat{S})}(I)\right)\log\left(\mathscr{J}_1^{(\hat{S})}(I)\right)d(\Delta G_1)\ldots d(\Delta G_R)dW\end{aligned} \tag{2.18}$$

$$\text{s.t.}\qquad \mathscr{J}_1^{(\hat{S})}(I) = p(\Delta G_1,\ldots,\Delta G_R,W)\,.$$

Depending on the chosen spectral references $\hat{S}$, the distribution of $\{\Delta G_1,\ldots,\Delta G_R\}$ would be different. Thus, the proposed measure is only comparable within the context of the same $\hat{S}$.

The interest in using entropy as the measure of spectral diversity is twofold. Thanks to its consideration of probabilities instead of raw data values, it is metrologically valid to consider the joint distribution of random variables of different physical properties as in the case of spectral shape differences and spectral intensity. Secondly, unlike variance which is only defined for normal distributions, entropy is a general measure applicable for any distributions. This is particularly crucial in the case of the spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ which can be multimodal, asymmetric, leptokurtic, etc. facing any form of spectral distribution.

Nevertheless, the true spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ which is a probability density function (PDF) cannot be known a priori unless the image is artificially generated with known parameters. Indeed, a digital image is just a realisation of the true scene which is subjected to various noise, sensor limitations, etc. As a result, the PDF can only be estimated based on the acquired spectra which can be thought of as the random samples from the real population. Hence, we are faced with the need for density estimation of $\mathscr{J}_1^{(\hat{S})}(I)$. A normalized histogram is perhaps the most basic method. For simplicity, let us revert to the continuous random variable Y. The histogram is constructed by first dividing the data points into a series of intervals called "bins", before counting how many data points fall into each

bin. Then, the histogram is normalized by simply dividing each bin by the total number of data points so that the bins sum to unity. Then, the differential entropy $h(y)$ is calculated by:

$$h(Y) = -\sum_{y \in Y} P\left(Y \in [y, y+\Delta y)\right) \log P\left(Y \in [y, y+\Delta y)\right), \tag{2.19}$$

where Δy here refers to the bin width. One of the problems with such approach is, however, the manner in which the bin width is chosen. Common methods such as Sturges' formula [72] and Freedman–Diaconis rule [25] are developed based on the assumption that the values are normally distributed. This proves too restrictive for $\mathscr{J}_1^{(\hat{S})}(I)$ which can be multimodal, asymmetric, leptokurtic, etc. On the other hand, the multidimensionality of $\mathscr{J}_1^{(\hat{S})}(I)$ also poses problem for the histogram construction. Consider a typical 8-bits quantization, the number of bins required would be $(2^8)^{R+1}$ recalling that R is the number of spectral references. Obviously, this renders the histogram computationally infeasible. Furthermore, the multidimensional histogram would be extremely sparse which hinders any statistical significance. The quantization of data due to binning also causes information loss.

In the literature, it exists another density estimation method which is based on k-nearest neighbours (kNN) [46]. For each of the n data points in D dimensions, their distance ω to their k^{th} nearest neighbor is calculated. The kNN density estimation is then given by $p(y) = k/nV_D\omega_i$ for $y \in Y$, where $V_D = \pi^{D/2}/\Gamma(D/2+1)$ is the volume of a unit D-dimensional ball and Γ is the Gamma function. The differential entropy $h(y)$ is given by:

$$h(y) = \Gamma(n) - \Gamma(k) + \log V_D + \frac{D}{n}\sum_{i=1}^{n} \log \omega_i. \tag{2.20}$$

The main interest of such approach is that it is operated directly on the data without any binnings required. Hence, it does not suffer from any quantization, sparsity, and memory issues as in the case of histogram. Thanks to its simplicity, it can be implemented very efficiently with a complexity of $\mathscr{O}(Dn\log n)$ based on k-d trees [5]. On the other hand, the kNN based density estimation can be thought of as employing variable "bin" size adapted to

the local density of data as exemplified in Fig. 2.6 concerning $D = 2$. Hence, it suffers less from data heterogeneity which leads to a better balance between bias and variance. Although k needs to be optimally chosen such that $k \to \infty$ when $n \to \infty$, in practice we note that its selection does not affect much of its efficiency. That said, we shall employ $k = 3$ throughout this work.

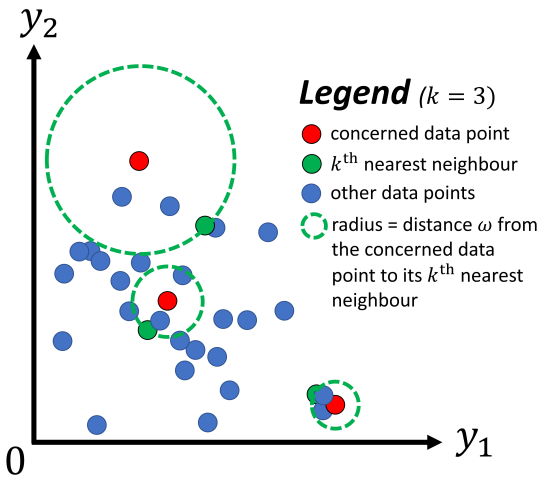


Fig. 2.6 An illustration of density estimation using k-NN, which can be thought of employing variable "bin" size adapted to the local density of data.

2.2.4 Automatic selection of spectral reference

Having defined spectral diversity by the joint differential entropy $h(\Delta G_1, \ldots, \Delta G_R, W)$ of the spectral shape differences $\{\Delta G_1, \ldots, \Delta G_R\}$ and spectral intensity W, it is straightforward to select the spectral references $\hat{S}$ which maximizes $h(\Delta G_1, \ldots, \Delta G_R, W)$. This is in accordance with the principle of maximum entropy which states that the probability density function (PDF) with the highest entropy is the one that is least biased; the one that best represents our current knowledge [37]. Perhaps a more direct justification is to recall the fact that entropy is the measure of information. By maximizing the entropy, the spectral information contained in the metric space is also maximized, which is obviously the desired scenario.

To explain further, let us recall the analogies we drew in sec. 2.2.1 between our spectral shape representation in metric space and principal component analysis (PCA). To begin, both project the original data in a subspace aiming to preserve the distance between the data points as much as possible. Whereas PCA does so through several top-ranked principal components (PCs), our approach requires to optimally defining a set of spectral references $\hat{S}$ that dictate the spectral projections. We summarize their correspondence in Table 2.1. It is clear, then, whereas PCA operates by maximizing data variance in the subspace spanned by the top PCs, we should, on the other hand, select $\hat{S}$ such that the (joint differential) entropy in the metric space (of the spectral shape differences $\{\Delta G_1, \ldots, \Delta G_R\}$) is maximized.

Our approach	**Comparison**	**PCA**
probability	Operation in … space.	vector
spectral reference	… as direction of projection.	principal component
entropy	… as measure of spread.	variance
spectral (shape) difference	Projection as defined by ….	dot product

Table 2.1 Analogies between our spectral shape representation in metric (spectral difference) space and principal component analysis (PCA).

Suppose that R spectral references $\hat{S} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_R\}$ are to be chosen among the normalized spectra of the image, we **perform the automatic spectral reference selection by maximizing the joint differential entropy of the spectral shape differences** $\{\Delta G_1, \ldots, \Delta G_R\}$:

$$\begin{aligned} \hat{S} &= \underset{\vec{s}'_r \in \bar{S}}{\arg\max} \left\{ h\left(\Delta G_1, \ldots, \Delta G_R\right) \right\} \\ \text{s.t.} \qquad & \Delta G_r = \mathrm{KLD}(\bar{s}, \vec{s}'_r) \;\forall\; r \in [1, R], \end{aligned} \tag{2.21}$$

where $\bar{S} = \{\bar{s}_1, \ldots, \bar{s}_n\}$ is the normalized spectral set from the image I with n pixels, and $\bar{s}'_r \; \forall \; r \in [1, R]$ refers to any set of R normalized spectra from $\bar{S}$. Note that here we have dropped the spectral intensity W from the entropy calculation because it is independent of $\hat{S}$.

To determine the optimal number R of spectral references, consider the solution to a similar problem in PCA. While keeping all the PCs allows to retain 100 % of the data variability, it is often the case that the few top-ranked PCs alone can already account for a very large percentage of the variability that keeping extra PCs is of little interest. We note that the same can also be envisioned in our approach which thus requires a way of ranking spectral references. We propose to **rank a spectral reference ahead of others when it results in the highest increase of the joint differential entropy of the spectral shape differences**:

$$\begin{aligned} \hat{s}_r &= \underset{\bar{s}'_r \in \bar{S}}{\arg\max} \left\{ h\left(\Delta G_1, \ldots, \Delta G_r\right) - h\left(\Delta G_1, \ldots, \Delta G_{r-1}\right) \right\} \quad \forall r > 1, \\ \text{s.t.} \qquad & \Delta G_r = \mathrm{KLD}(\bar{s}, \bar{s}'_r) \; \forall \; r \in [1, R], \end{aligned} \tag{2.22}$$

where $\hat{s}_r$ refers to the r^{th}-ranked spectral reference. Obviously, eqn. 2.22 can only be worked out after determining the **first-ranked spectral reference $\hat{s}_1$, which is the normalized spectrum which results in the largest differential entropy of spectral shape difference**:

$$\begin{aligned} \hat{s}_1 &= \underset{\bar{s}' \in \bar{S}}{\arg\max} \left\{ h\left(\Delta G\right) \right\} \\ \text{s.t.} \qquad & \Delta G = \mathrm{KLD}(\bar{s}, \bar{s}'). \end{aligned} \tag{2.23}$$

One may wonder the reason for not applying a similar approach as in the case of $\hat{s}_1$ for the subsequent ranking, that is, by ranking spectral references based on the differential entropy $h(\Delta G_r)$ of their associated spectral shape difference ΔG_r (analogous to PCA's PC ranking

based on its variance). This is because whereas PCs are uncorrelated by construction, it is rarely the case for spectral references. As a result, the spectral shape information as projected by each spectral reference (in the form of spectral shape difference) is also not independent from each other i.e. $h(\Delta G_1, \ldots, \Delta G_R) \leq \sum_{r=1}^{R} h(\Delta G_r)$. Only if the spectral references are independent i.e. $h(\Delta G_1, \ldots, \Delta G_R) = \sum_{r=1}^{R} h(\Delta G_r)$, they can be directly ranked using $h(\Delta G_r)$.

With more spectral references being employed, we can expect the increase (but at a decreasing rate) of the joint differential entropy of the resulted spectral shape differences. Visually, the joint differential entropy is an increasing function which stabilizes or saturates after the knee point, which we propose to take as the optimum number of spectral references. In other words, **the number of spectral references is selected such that the joint differential entropy of the resulted spectral shape differences is no longer increasing rapidly, and is thus no longer worth the cost i.e. the further addition of spectral references**.

In the continuous domain, it is straightforward to identify the knee point as the one having the largest curvature [66]. However, the identification is not so direct in the discrete domain as in our case as the number of spectral references is an integer. Indeed, discrete approximation of curvature using finite differences can be used, but it is subjected to approximation errors especially when there is no abundance of data points as in our case (we do not expect the number of spectral references to be too high). One of the alternatives is the Menger curvature [78] which defines the local curvature at a given point as the curvature of the circle circumscribed about the given point and the two points immediately before and after it (so three points in total). However, such approach is less robust as it is calculated locally while ignoring the overall behaviour of the data points. In this work, we propose to identify the knee point which corresponds to the optimum number of spectral references R_{opt} by:

$$
\begin{aligned}
R_{opt} &= \underset{R\in[1,R_{max}]}{\arg\max} \left\{ \frac{\left|h(\Delta G_1,\ldots,\Delta G_R) - \alpha R - \beta\right|}{\sqrt{\alpha^2+1}} \right\} \\
\text{s.t.} \qquad \alpha &= \frac{h(\Delta G_1,\ldots,\Delta G_{R_{\max}}) - h(\Delta G_1)}{R_{max}-1} \\
\beta &= h(\Delta G_1) - \alpha,
\end{aligned}
\tag{2.24}
$$

where R_{max} is the maximum number of spectral references selected to be "large enough" such that the joint differential entropy of the resulted spectral shape differences stops varying much. Note that eqn. 2.24 is adapted from the formula for calculating the distance from a point to a line, whereas α and β refer to the gradient and y-intercept of the line respectively. In this case, $h(\Delta G_1,\ldots,\Delta G_R) = \alpha R + \beta$ is the straight line joining $\left(1, h(\Delta G_1)\right)$ and $\left(R_{max}, h(\Delta G_1,\ldots,\Delta G_{R_{max}}\right)$ as illustrated in Fig. 2.7. Accordingly, the knee point $\left(R_{opt}, h(\Delta G_1,\ldots,\Delta G_{R_{opt}})\right)$ is simply given by the furthest point from the line.

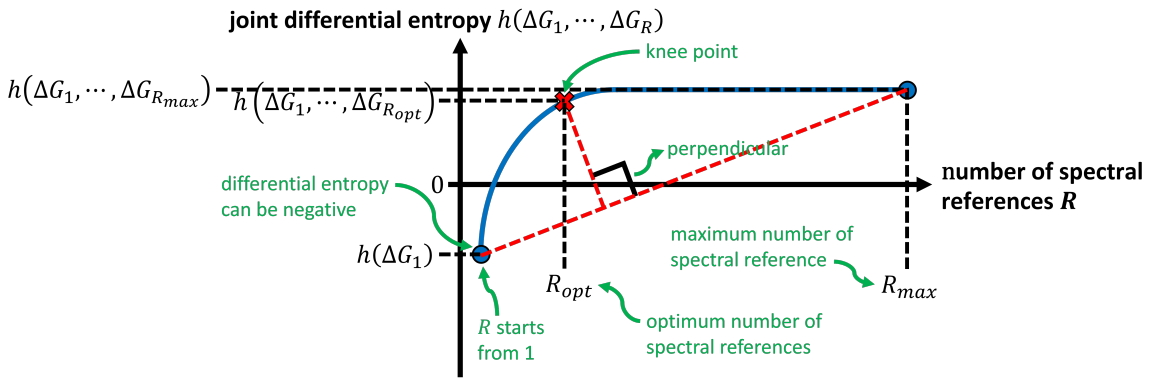


Fig. 2.7 The joint differential entropy $h(\Delta G_1,\ldots,\Delta G_R)$ of spectral shape differences corresponding to the top R spectral references is an increasing function which saturates after the knee point, which we propose to take as the optimum number of spectral references R_{opt}.

Up to now, the spectral references $\hat{S} = \{\hat{s}_1,\ldots,\hat{s}_R\}$ are to be chosen among the normalized spectra set $\bar{S} = \{\bar{s}_1,\ldots,\bar{s}_n\}$ from the image I with n pixels to maximize $h(\Delta G_1,\ldots,\Delta G_R)$:

$$\hat{S} = \underset{\vec{s}'_r \in \bar{S}}{\arg\max} \left\{ h\left(\Delta G_1, \ldots, \Delta G_R\right) \right\}$$
$$\text{s.t.} \qquad \Delta G_r = \mathrm{KLD}(\bar{s}, \vec{s}'_r) \; \forall \; r \in [1, R], \tag{2.25}$$

as recalled from eqn. 2.21. This requires to loop through all possible sets of R normalized spectra from $\bar{S}$, thus resulting in a complexity of at least $\mathscr{O}\left(n(n-1)(n-2)\ldots(n-R+1)\right)$. Suppose that $R = 3$ spectral references are to be chosen from an image (100×100 pixels) with $n = 10000$, the sets of R normalized spectra to be considered is 999700020000 or about 10^{12} which is computationally infeasible. However, we note that in fact, not all sets of the R normalized spectra need to be examined. Practically, we could **limit the spectral reference selection among spectra forming the convex hull of the normalized[6] spectral set *Conv*$(\bar{S})$** [19]. By the popular rubber band analogy, the convex hull of a set of points is the smallest convex set that contains it as illustrated in Fig. 2.11 concerning a two-dimensional case.

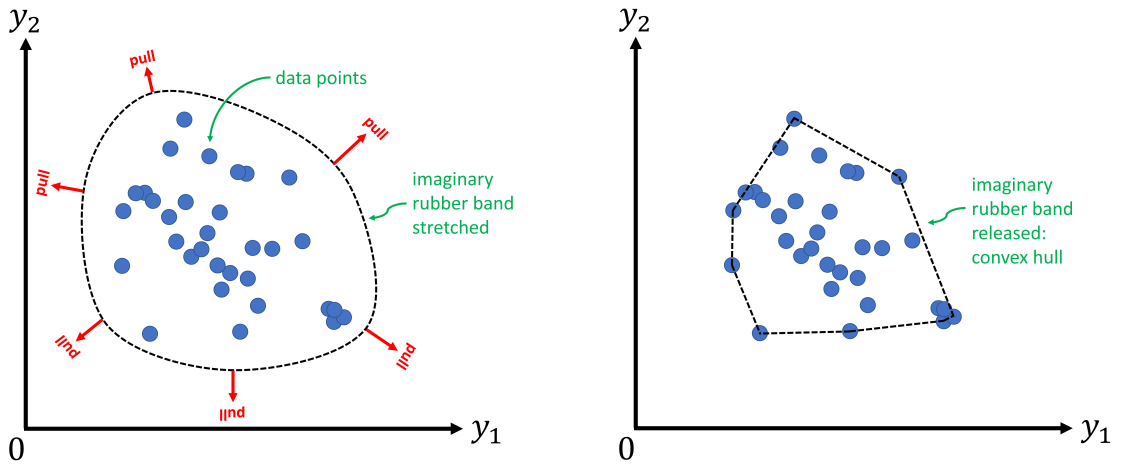


Fig. 2.8 An illustration of convex hull concerning a set of data points $y = (y_1, y_2)$ in two dimensions using the rubber band analogy.

[6] Our proposition here is adapted from [19] concerning spectral reference selection in the spectral difference space which also includes information from spectral intensity (in contrast to our spectral shape difference space).

We justify our proposition by first considering a set of normally distributed points $y \sim \mathcal{N}(0,1)$ as illustrated in Fig. 2.9 (a). Suppose that we want to represent each point $y \in \mathbb{R}$ in the metric space by its absolute[7] distance $\Delta y \in \mathbb{R}$ to a given reference $\hat{y} \in \mathbb{R}$ such that $y \xrightarrow{\text{metric}} |y - \hat{y}| = \Delta y$, what will be the optimal $\hat{y}$ for maximum discriminability of y in the metric space? Clearly, the extremum $\hat{y} = \min(y)$ or $\hat{y} = \max(y)$ makes the best choice because then, each distinct point y is represented by an unique Δy in the metric space as depicted in Fig. 2.9 (b) and (h). As the distributions of Δy too resembles $\mathcal{N}(0,1)$, also notice that the differential entropy of Δy is the same as that of y, reflecting the fact that the information of y in well preserved by Δy in the metric space. Meanwhile, the further the reference is from $\max(y)$ or from $\min(y)$, the lower the discriminability of y as Δy in the metric space as shown in Fig. 2.9 (c) - (g) as the distributions of Δy deviate from $\mathcal{N}(0,1)$. At the same time, also notice the decrease of $h(\Delta y)$ which reflects the loss of information in the metric space. That said, $\hat{y} = \text{median}(y)$ would be the worst choice[8] because then, y which are equidistant from $\text{median}(y)$, regardless of them being larger than or smaller than $\text{median}(y)$, would be indiscriminate as Δy in the metric space which is strictly positive (as induced by the absolute distance $|y - \hat{y}|$). Referring to Fig. 2.9 (e), it can be seen that approximately half of the originally normally distributed y is indistinguishable as Δy in the metric space with $\hat{y} = \text{median}(y)$. As the distribution of Δy resembles a half-normal distribution, its $h(\Delta y)$ is also the lowest i.e. about the half of $h(y)$, indicating that almost half of the information has been lost in the metric space representation. The same logic can also be applied in higher dimensions by regarding each point of the convex hull as some kind of "extremum" with the (geometric) median[9] as the "origin" as illustrated in Fig. 2.10. By the same arguments, the points of the convex hull are therefore the better choices of reference for maximum discriminability in the metric space. Naturally, this suggests spectral reference selection

[7] Analagous to spectral difference e.g. SAM [88] and KLPD [62] which are positive by definition.

[8] Here we make use of the median because we want the reference to be one of the given points, as we constrain our spectral reference to be, to avoid selecting a false spectrum. That said, in this example, the same argument also holds in the case of using the mean i.e. the arithmetic average of the points as the reference.

[9] Defined as the point which has the lowest cumulative distance to all the other points [1].

among normalized spectra forming the convex hull $Conv(\bar{S})$. As $Conv(\bar{S})$ is a much smaller set than the entire normalized spectral set $\bar{S}$, the process can thus be accelerated significantly.

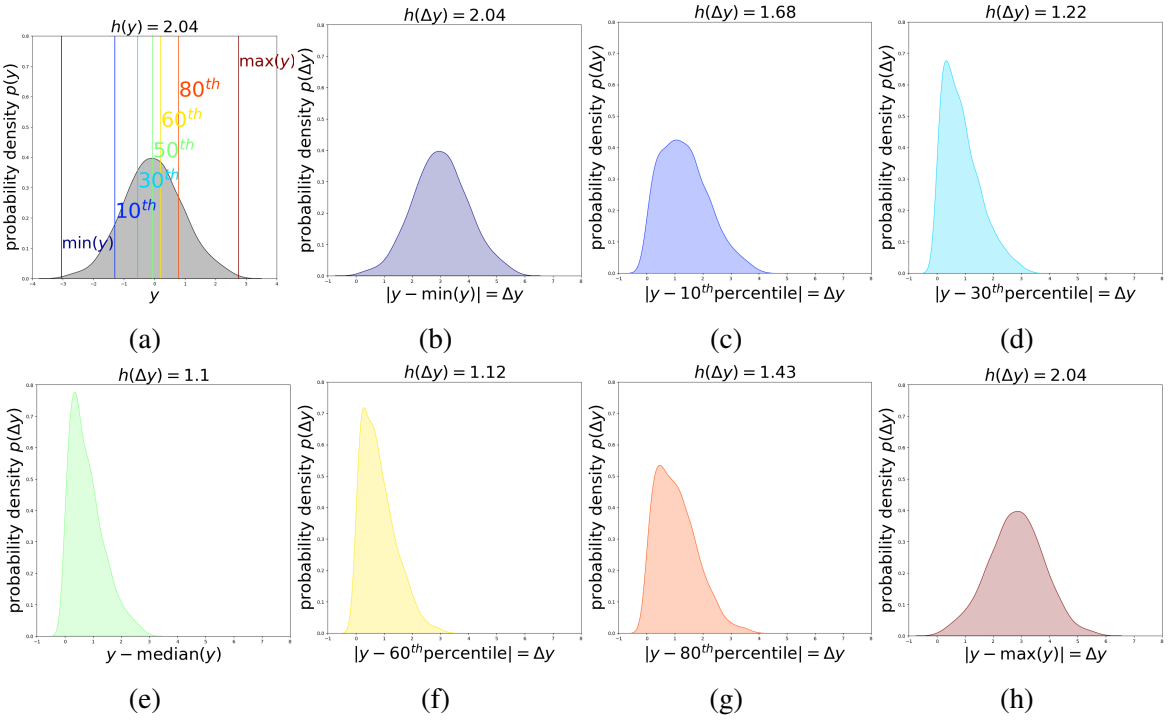


Fig. 2.9 The probability density of (a) a set of normally distributed points y; (b) the same set of points in metric (absolute distance) space as Δy with the reference being either $\min(y)$; (c) 10^{th} percentile; (d) 30^{th} percentile; (e) median; (f) 60^{th} percentile; (g) 80^{th} percentile; or (h) $\max(y)$. As written at the top of each plot is differential entropy as a measure of information.

We supplement our previous argument that the points of the convex hull can be thought of as some kind of "extrema" with the median as the "origin" by considering a similar problem i.e. endmember extraction. In linear spectral unmixing, endmembers are defined as the pure spectra from which other (mixed) spectra are formed by a linear combination [7]. Geometrically, the endmembers occupy the vertices of a simplex which contains all the other spectra [55]. Under this context, the points of the convex hull are thus the "extrema" such that spectra nearing them are "purer", whereas spectra nearing neither of them are "neutral" like an "origin". Recalling the analogy between PCA and our spectral shape representation in metric space as summarized in Table 2.1, we reassert our view that spectral shape differences

can be thought as a spectral projection onto the corresponding spectral reference. In other words, our approach can also be seen as expressing a given normalized spectrum in terms of spectral references. Indeed, our spectral reference selection is reminiscent of endmember extraction but with a key difference. Whereas linear spectral unmixing is formulated in the context of L-dimensional Euclidean vector space (with L spectral channels), our approach is conceived by respecting the physics of spectrum as a continuous function.

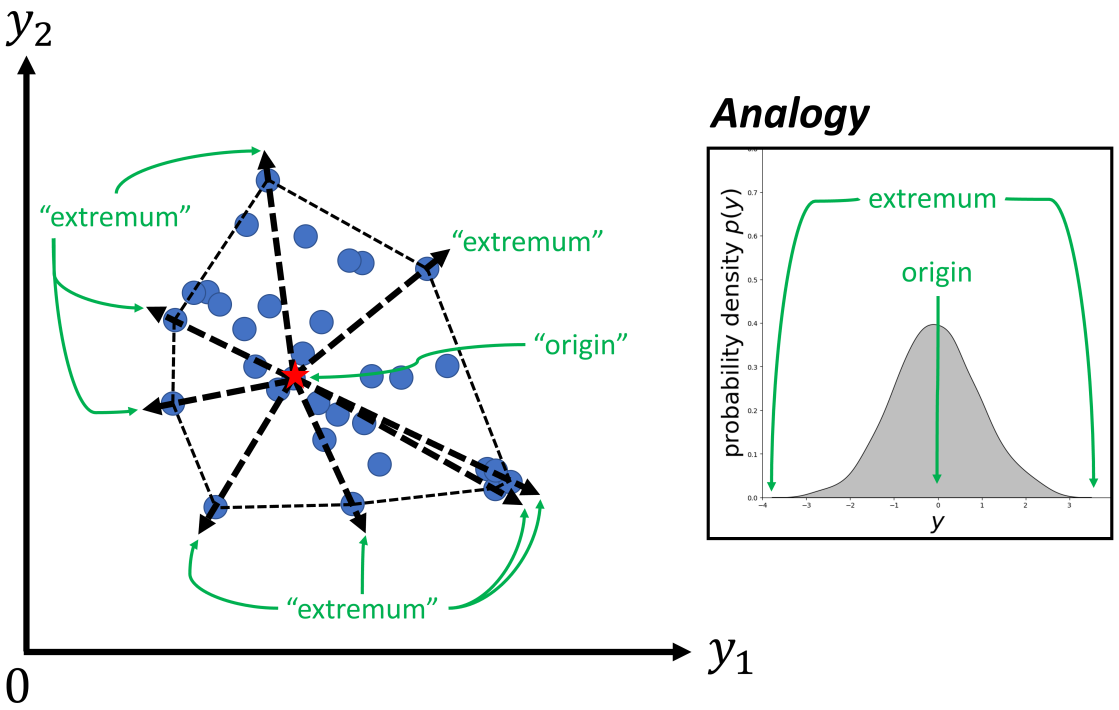


Fig. 2.10 An illustration of convex hull concerning a set of points $y = (y_1, y_2)$ in two dimensions. The points of the convex hull can be thought as some kind of "extrema" with the median as the "origin".

Having presented the case of accelerating spectral reference selection among spectra forming the convex hull $Conv(\bar{S})$ of the normalized spectral set $\bar{S}$, we are nevertheless, faced with another computational intractability as posed by the calculation of $Conv(\bar{S})$. Considering $n = 10000$ spectra (e.g. image of 100×100 pixels) with L wavelengths, the complexity of calculating $Conv(\bar{S})$ is given by e.g. $\mathcal{O}(n^{\lfloor L/2 \rfloor + 1})$ based on the gift wrapping algorithm

[18]. In light of this, we propose a distance-based method to approximate $Conv(\bar{S})$ by $\widetilde{S} = \{\widetilde{s}_1, \ldots, \widetilde{s}_Z\}$ supposing that there are Z spectra forming $Conv(\bar{S})$. We start by recognizing the fact that the two points furthest apart are part of $Conv(\bar{S})$. Accordingly, we identify the normalized spectral pair which has the largest spectral (shape) difference as parts of $\widetilde{S}$:

$$\widetilde{s}_1, \widetilde{s}_2 = \underset{\bar{s}, \bar{s}' \in \bar{S}}{\arg\max}\, \mathrm{KLD}(\bar{s}, \bar{s}'). \tag{2.26}$$

Then, we assume that the next point on $Conv(\bar{S})$ is collectively far from the other previously identified points. Accordingly, we identify the following part of $\widetilde{S}$ by:

$$\widetilde{s}_z = \underset{\bar{s} \in \bar{S}}{\arg\max} \bigcap_{z' < z} \mathrm{KLD}(\bar{s}, \widetilde{s}_{z'}) \quad \forall z > 2. \tag{2.27}$$

For simplicity, we choose to define the collective difference using the additive operation:

$$\widetilde{s}_z = \underset{\bar{s} \in \bar{S}}{\arg\max} \sum_{z' < z} \mathrm{KLD}(\bar{s}, \widetilde{s}_{z'}) \quad \forall z > 2. \tag{2.28}$$

An iterative procedure, the calculation will go on until all Z spectra have been identified.

Clearly, our proposition requires a prior definition of Z i.e. the number of spectra forming $Conv(\bar{S})$ which nevertheless, cannot be known a priori. In this work, we resort to selecting Z such that it is "large enough". We demonstrate the approximation of convex hull $Conv(y)$ by $\widetilde{Y} = \{\widetilde{y}_1, \ldots, \widetilde{y}_Z\}$ with varying Z in Fig. 2.11 concerning a set of points $y \in \mathbb{R}^2$ in two dimensions. Referring to Fig. 2.11 (a), it can be seen that there are mainly two clusters of points, one more compact and the other more loosely distributed. The median, as marked in red, is identified as the point that has the lowest cumulative distance to all the other points. To study the quality of representation of each point $y = (y_1, y_2)$ in the metric space by its absolute Euclidean distance $\Delta y = \sqrt{(y_1 - \widetilde{y}_{z,1})^2 + (y_2 - \widetilde{y}_{z,2})^2}$ to a given $\widetilde{y}_z = (\widetilde{y}_{z,1}, \widetilde{y}_{z,2})$ as the reference, we calculate the differential entropy $h(\Delta y)$ as a measure of information. It can be seen that with median(y) as the reference, the information contained in the metric

space is comparatively low as the distribution $p(\Delta y)$ of Δy appears unimodal. In other words, the two clusters are less discriminable as they are equidistant from median(y) in accordance to our previous argument that medians makes one of the worse[10] choices of reference for representation in metric space. Referring to Fig. 2.11 (b) and (c), $\widetilde{y}_1$ and $\widetilde{y}_2$ are identified as the pair of points furthest apart. Subsequently, the following $\widetilde{y}_z \; \forall \; z > 2$ are identified iteratively as per eqn. 2.28 (with the spectral difference replaced by Euclidean distance) as illustrated in Fig. 2.11 (d), (i) - (l), and Fig. 2.11 (a) - (d) . Referring to Fig. 2.12 (c), it can be seen that $Conv(y)$ is best approximated with $Z = 10$ beyond which the approximations get worse. As such, from $Z = 11$ onwards we only illustrate the case for $Z = 15, 20, 25$ and 30 which results in bad $Conv(y)$ approximation as in Fig. 2.12 (i) - (l). As for the representation in metric space, the efficiency of each $\widetilde{y}_z$ as reference varies: $\widetilde{y}_{30}$ makes the best choice with the highest $h(\Delta y) = 5.77$. Anyway, the efficiency of $\widetilde{y}_1$ is very close ($h(\Delta y) = 5.75$) to $\widetilde{y}_{30}$ which suggests that there is no need to identify up to $Z = 30$ to obtain good references. In contrast, $\widetilde{y}_7$ makes the worst choice with the lowest $h(\Delta y) = 5.01$ which is actually lower than that of the median ($h(\Delta y) = 5.21$). This is due to its relatively higher degree of equidistance to the majority of points compared to median, which suggests that not all points on the convex hull necesssarily makes the better choices of references. To recall, the identification of spectra forming the convex hull $Conv(\bar{S})$ of the normalized spectral set $\bar{S}$ is only a prior step in our automatic spectral reference selection. Their selection will be refined further through the maximization of joint differential entropy in the spectral shape difference space.

[10] In one-dimensional case, the median is the worst choice as exemplified in Fig. 2.9. However, this is not necessarily the case in higher dimensions because depending on the distribution, some points of the convex hull can be a worse choice than median due to their equidistance to majority of other points.

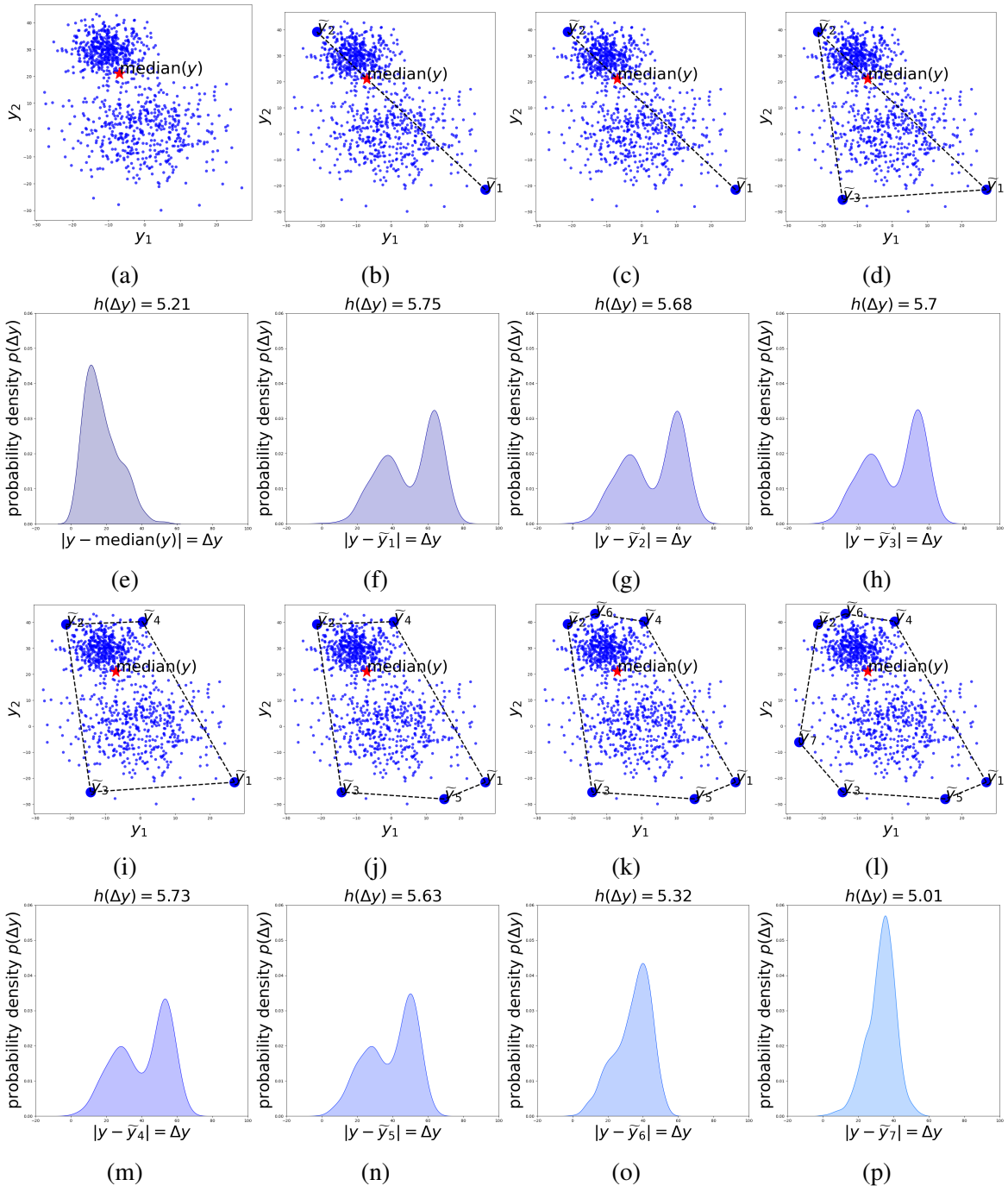


Fig. 2.11 An illustration of convex hull approximation concerning a set of points in two dimensions with varying number of points Z forming the convex hull which must be predefined, with the corresponding probability density of the same set of points in metric (Euclidean distance) space as Δy with the reference being one of the convex hull points. As written at the top of each probability density plot is differential entropy of Δy as a measure of information.

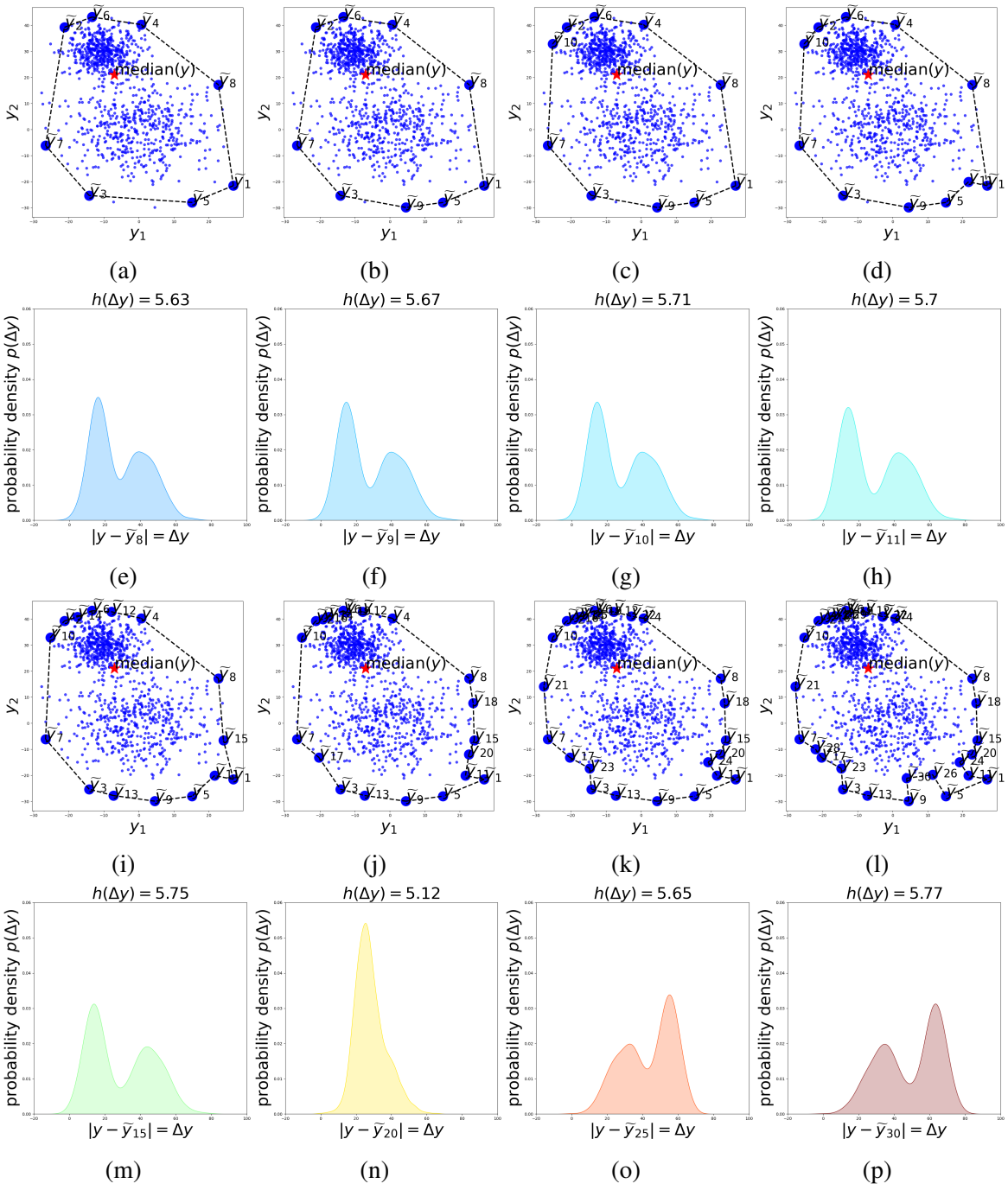


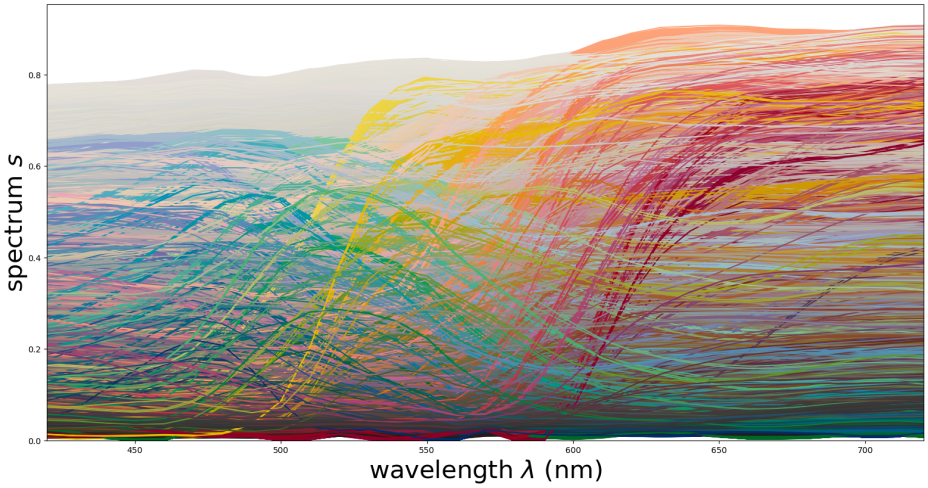
Fig. 2.12 An illustration of convex hull approximation concerning a set of points in two dimensions with varying number of points forming the convex hull which must be predefined, with the corresponding probability density of the same set of points in metric (Euclidean distance) space as Δy with the reference being one of the convex hull points. As written at the top of each probability density plot is differential entropy of Δy as a measure of information.

The novelty of our method lies in the fact that it is distance-based; only pairwise distances of the points are required to approximate the convex hull. As a result, it is relatively low cost. It is also general and gives consistent results in any dimension (number of spectral bands), which is a key requirement in metrology. In fact, we have effectively discovered a general method to approximate the convex hull in any dimensions subjected to adaptation of the distance function for the concerned application. That said, its efficiency is obviously dependent on the choice of distance measure on whether it is a true metric, which is not the case for our spectral difference measure (KL divergence). With a spectral metric or distance function, an improvement in the accuracy of identifying the convex hull can be expected.

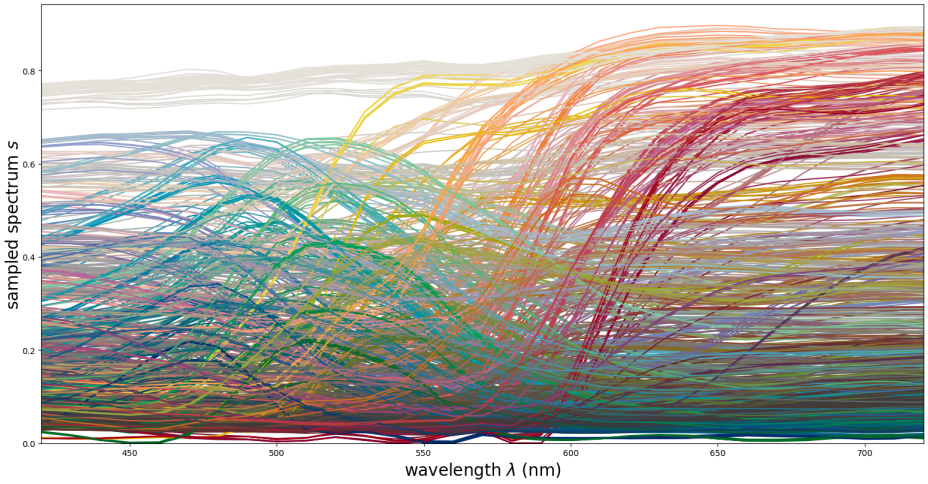
Having detailed our automatic spectral reference selection from entropy maximization to acceleration using convex hull, we demonstrate its entire operation using the hyperspectral image[11] *ChartDC* (860 $\times$ 1650 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval [54] as shown in Fig. 2.13 (a). Consisting of 162 chromatic (colourful) patches and 15 achromatic (grayscale) patches, *ChartDC* can be thought of as having covered all the possible colour spectra that are perceivable by human eyes as shown in Fig. 2.13 (c). To alleviate the huge processing cost of processing the entire spectral set, we randomly select 0.1% of the pixels which amounts to 1341 spectra as displayed in Fig. 2.13 (b) for the following calculations. Visually, it can be seen that the sampled spectra have accounted for most of the spectral shape variation, suggesting its representativeness of the spectral set.

[11] As cropped from the original image [54] by removing the black frame as well as the 60 black and white patches around the colourful patches to avoid dominance by equi-energy i.e. white, black, and gray spectra.

(a)



(b)



(c)

Fig. 2.13 (a) The hyperspectral image *ChartDC* (860 $\times$ 1650 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval, rendered in RGB for illustration; (b) all the (860 $\times$ 1650 = 1419000) spectra s of *ChartDC*; and (c) 0.1 % of the spectra as illustrated in (b), sampled randomly from *ChartDC*.

As per our proposition in this chapter, we start by decomposing the (sampled) spectra s into two parts i.e. spectral shape G and spectral intensity W as shown in Fig. 2.14 (a) and (b) respectively, calculated as per eqn. 2.6 and 2.9 considering wavelengths $\lambda \in [420, 720]$ nm:

$$G = \bar{s} = \frac{s}{W},$$
$$W = \int_{420}^{720} s(\lambda) d\lambda,$$

where $\bar{s} \in \bar{S}$ is the normalized spectrum. Then, we identify the "convex hull spectra" $\widetilde{S} = \{\widetilde{s}_1, \ldots, \widetilde{s}_Z\}$ up to $Z = 30$ as shown in Fig. 2.15 (b) that approximate the convex hull $Conv(\bar{S})$ of the normalized spectral set $\bar{s}$, calculated as per eqn. 2.26 and 2.28:

$$\widetilde{s}_1, \widetilde{s}_2 = \underset{\bar{s}, \bar{s}' \in \bar{S}}{\arg\max}\, \mathrm{KLD}(\bar{s}, \bar{s}'),$$
$$\widetilde{s}_z = \underset{\bar{s} \in \bar{S}}{\arg\max} \sum_{z' < z} \mathrm{KLD}(\bar{s}, \widetilde{s}_{z'}) \quad \forall z > 2.$$

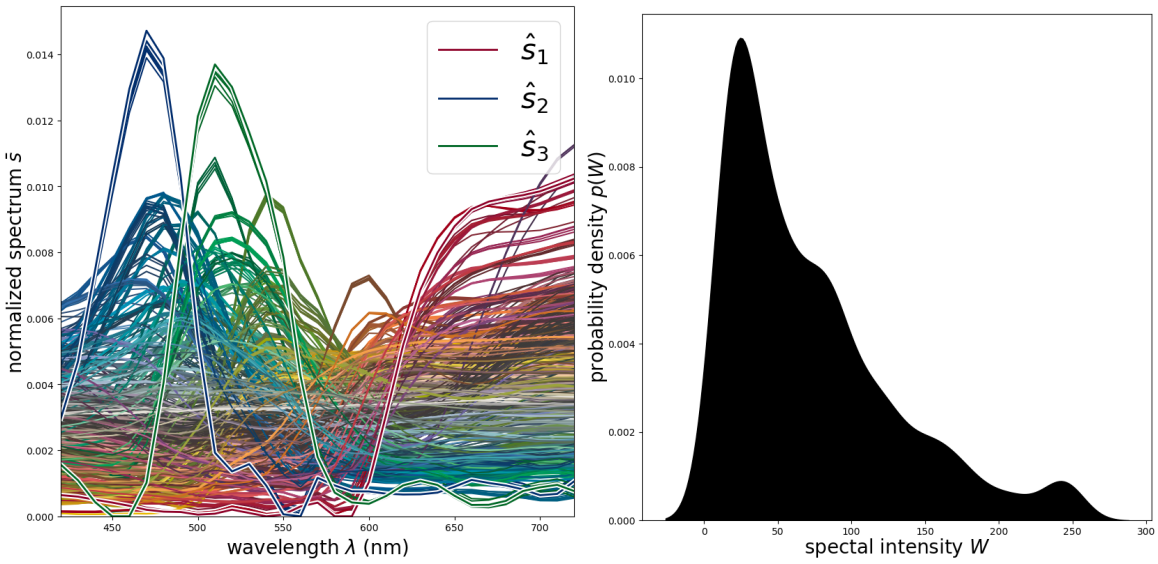
Finally, we rank $\widetilde{s}_z \; \forall \; z \in [1, Z]$ as shown in Fig. 2.15 (c), calculated as per eqn. 2.22 and 2.23:

$$\hat{s}_1 = \underset{\widetilde{s}_z \in \widetilde{S}}{\arg\max} \left\{ h\left(\Delta G\right) \right\},$$
$$\hat{s}_z = \underset{\widetilde{s}_z \in \widetilde{S}}{\arg\max} \left\{ h\left(\Delta G_1, \ldots, \Delta G_z\right) - h\left(\Delta G_1, \ldots, \Delta G_{z-1}\right) \right\} \quad \forall z > 1,$$

where $\Delta G_z = \mathrm{KLD}(\bar{s}, \bar{s}'_z) \; \forall \; z \in [1, Z]$. The optimum number of spectral reference $R_{opt} = 3$, also shown in Fig. 2.15 (c), is identified as per eqn. 2.24 noting that here $R_{max} = Z$:

$$
\begin{aligned}
R_{opt} &= \underset{R\in[1,R_{max}]}{\arg\max} \left\{ \frac{\left|h(\Delta G_1,\ldots,\Delta G_R) - \alpha R - \beta\right|}{\sqrt{\alpha^2+1}} \right\} \\
\text{s.t.} \qquad \alpha &= \frac{h(\Delta G_1,\ldots,\Delta G_{R_{max}}) - h(\Delta G_1)}{R_{max}-1} \\
\beta &= h(\Delta G_1) - \alpha,
\end{aligned}
$$

The resulted choice of the three spectral references $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$ as shown in Fig. 2.15 (a) and (b) correspond to some kind of red, blue, and green spectra respectively. This is reminiscent of the three primary colours in colour addition theory i.e. red, green, and blue that can be mixed in varying proportions to produce a gamut of colours. Thus, the selection of $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$ seems to be logical in case of *ChartDC* for the optimum representation of spectral shape distribution in the metric space (by spectral shape differences). Thanks to the correspondence between $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$ with the three primary colours, our prior assertion that the spectral shape difference of a given (normalized) spectrum can be thought as a spectral projection onto the spectral reference is further strengthened. In other words, the representation in metric space can be seen as expressing a given spectrum in terms of spectral references. Based on the selected spectral references $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$, we illustrate the distribution of spectral shape differences ΔG_1, ΔG_2, and ΔG_3 as well as spectral intensity W in Fig. 2.16.



(a) (b)

Fig. 2.14 The decomposition of the sampled spectra in (a) spectral shapes G (as normalized spectra $\bar{s}$) and spectral intensities W (depicted as probability distribution).

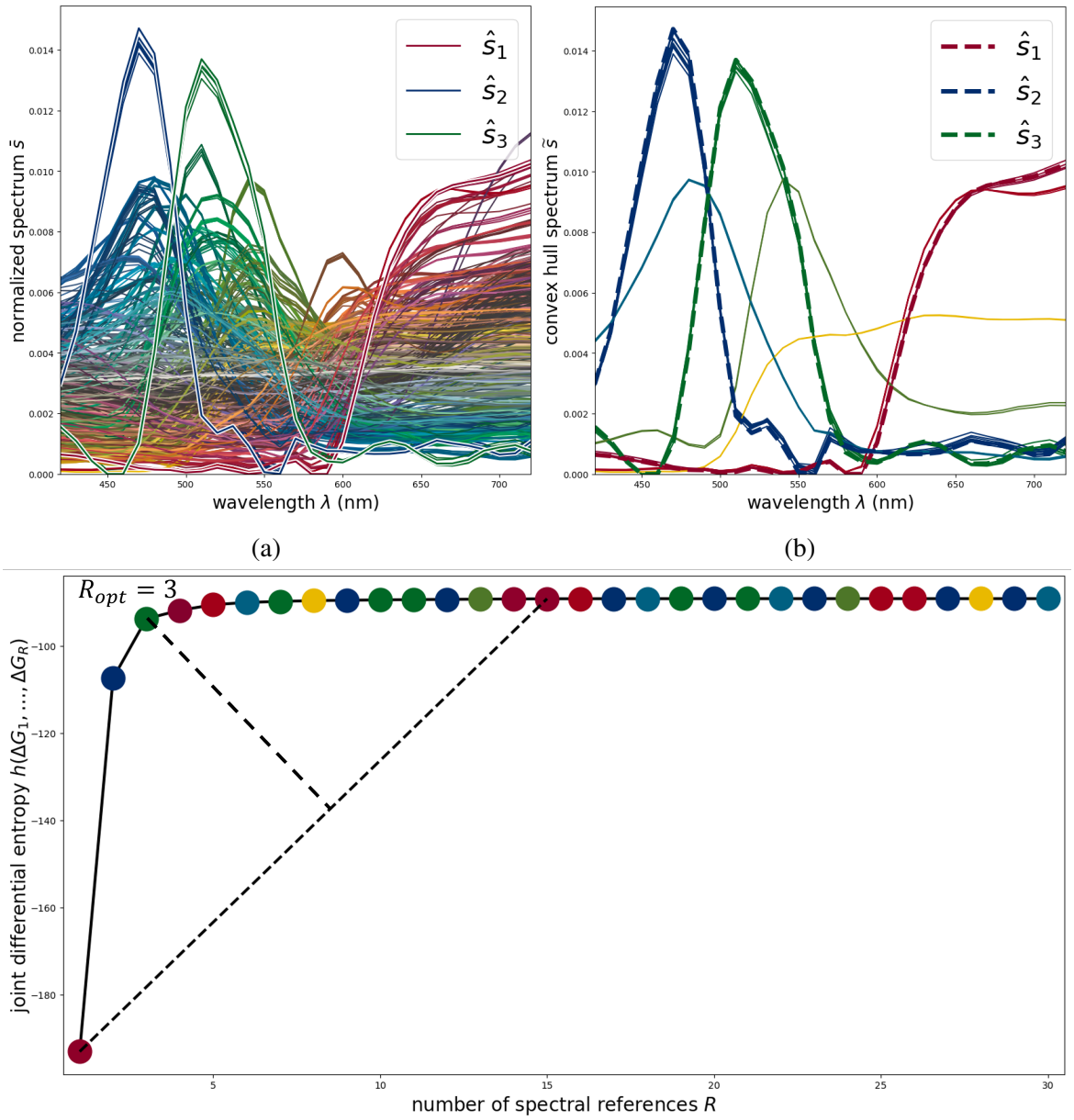
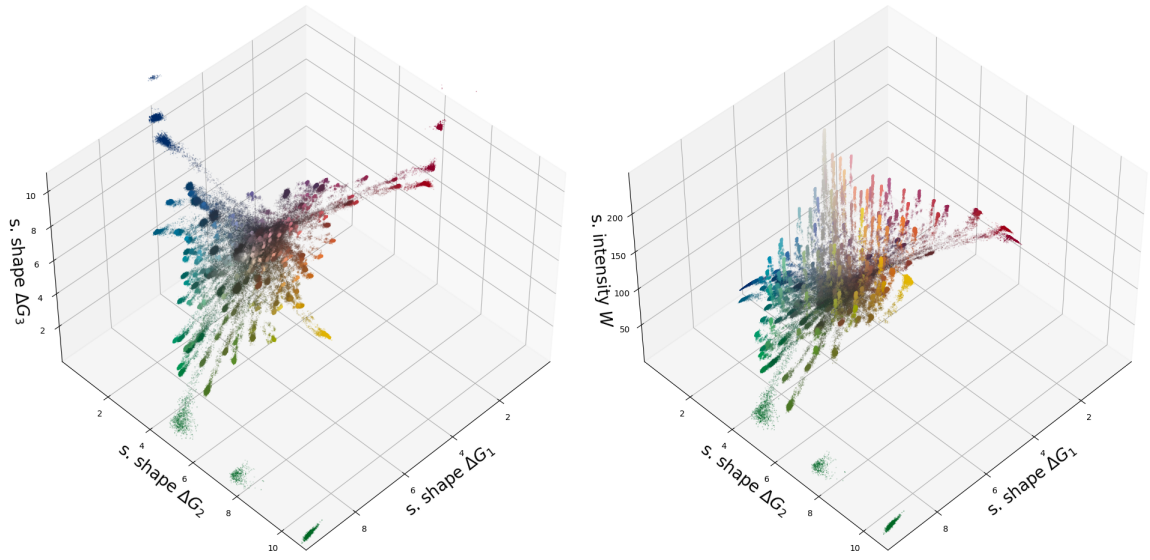


Fig. 2.15 (a) The normalized (sampled) spectra $\bar{s}$; (b) the convex hull spectra $\widetilde{s}$ and the selected spectral references $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$; and (c) the joint differential entropy of spectral shape differences corresponding to the top R spectral references, as a way to rank the convex hull spectra. The optimum number of spectral references is given by $R_{opt} = 3$.



(a) (b)

Fig. 2.16 (a) The three-dimensional scatter plot of spectral shape differences ΔG_1, ΔG_2, and ΔG_3 corresponding to the spectral references $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$; and (b) the three-dimensional scatter plot of ΔG_1, ΔG_2 and spectral intensity W of *ChartDC*.

2.3 Summary

At the end of Chapter 1, we have shown the need for spectral characterization for texture description. In this chapter, we present our approach to achieving that objective. We started by reviewing the physics and definition of spectrum, recalling some measures of spectral differences, and presenting the state of the art for constructing the image histogram facing a large number of spectral channels as in the case of hyperspectral image. Then, we proposed our own approach by first decomposing the spectrum into spectral shape and intensity parts, before resorting to representing the former in metric (spectral shape difference) space. Finally, we proposed the formulation of spectral difference histogram as inspired by [19] and an automatic spectral reference selection thanks to entropy maximization. We also presented a way to accelerate the selection based on the convex hull of the spectral set and subsequently, proposed a general method to approximate the convex hull in any dimensions.

Chapter 3

Spectral-Spatial Characterization of Texture: Differences-Based

Contents

3.1 Background 56

3.2 The proposed formulation 58

3.2.1 Ideas from sum and difference histograms 58

3.2.2 Edge sharpness, scale, and direction 62

3.2.3 Relative spectral difference occurrence matrix (RSDOM) 74

3.2.4 Similarity measurement and statistical modelling 76

3.3 Summary 78

3.1 Background

Julesz conjecture

We begin by considering one of the earliest works in the psychophysics of texture. In 1962, Julesz considered textures as generated by a stochastic process which is specifiable using N^{th}-order statistic i.e. the probability of N selected points (pixels) having certain values. By regarding human texture discrimination as a preattentive process, he asserted that "*whereas textures that differ in their first- and second-order statistics can be discriminated from each other, those that differ in their third- or higher-order statistics usually cannot*" [41]. Termed the Julesz conjecture, it became the basis of many texture analytical tools [30, 26, 33].

In the context of digital images, the first-order statistic refers to the probability distribution of pixels which can be graylevels, colours, spectra, etc. whereas the second-order statistic refers to their joint probability. As such, the first- and second-order statistics are related to the spectral and spatial distributions of pixels respectively. Consistent with our approach, the Julesz conjecture thus implies the need for spatiochromatic characterization of texture.

Graylevel cooccurrence matrix (GLCM)

In 1973, Haralick et. al. developed graylevel cooccurrence matrix (GLCM) [30] which is an operational definition of the second-order statistic in the Julesz conjecture. Formulated in the context of grayscale image $I : \mathbb{R}^2 \to \mathbb{R}$ in two (spatial) dimensions, GLCM records the cooccurrence of pixels $I(x)$ and $I(x')$ separated at distance r and direction θ. Accordingly, GLCM $p\left(I(x), I(x')\right)$ is a two-dimensional matrix (histogram) of $Q \times Q$ elements (where Q is the number of graylevels) from which several texture features can be defined e.g. contrast, entropy, energy, and homogeneity. As summing along the row or column of GLCM gives the graylevel histogram, GLCM also embeds the first-order statistic of image.

We note that feature extraction based on N^{th}-order statistics is limited to stationary textures. If texture stationarity is not respected, the features extracted are less efficient. On

the other hand, further image quantization is necessary prior to calculating GLCM to avoid sparsity for statistical reliability. However, this comes at the expense of possible information loss due to the smoothing of lower contrast textures. Furthermore, it exists no principled approach but empirical means in choosing Q which is typically selected to be 8, 16, 32, etc.

Sum and difference histograms

Considering two random variables a and b, Unser showed in 1986 that the probabilities of their sum $p(a+b)$ and difference $p(a-b)$ define the principal axes of their joint probability $p(a,b)$ [83] as shown in Fig. 3.1. That said, the formulation of GLCM $p\left(I(x),I(x')\right)$ can thus be replaced by that of sum $p\left(I(x)+I(x')\right)$ and neighbouring difference histograms $p\left(I(x)-I(x')\right)$. Considering a grayscale image with Q graylevels, the sizes of sum and neighbouring difference histograms are both given by $2Q$ as:

$$\begin{aligned} range\left(I(x)+I(x')\right) &= \max\left(I(x)+I(x')\right) - \min\left(I(x)+I(x')\right) \\ &= (Q_{\max}+Q_{\max}) - (Q_{\min}+Q_{\min}) \\ &= 2Q, \\ range\left(I(x)-I(x')\right) &= \max\left(I(x)-I(x')\right) - \min\left(I(x)-I(x')\right) \\ &= (Q_{\max}-Q_{\min}) - (Q_{\min}-Q_{\max}) \\ &= 2Q, \end{aligned} \tag{3.1}$$

which is an order less than that of GLCM, thus translating to more efficient computations.

Like GLCM, the sum and difference histograms also embed the first- and second-order statistics. However, they are not directly applicable for hyperspectral or any multivariate image. This is because spectral addition and subtraction are not physically defined. In the following, we shall show how to tackle this issue by slight modifications to their formulation.

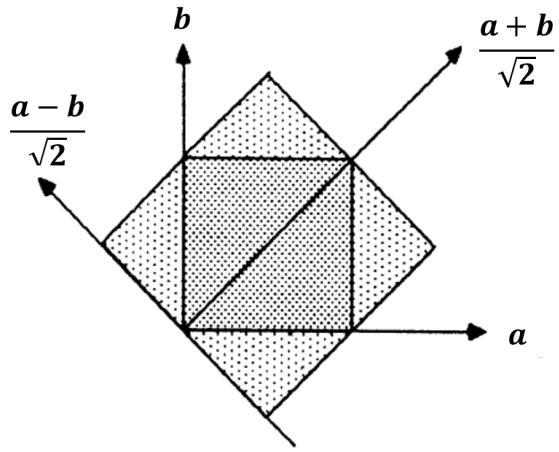


Fig. 3.1 Geometric representation of the sum and difference transformation, concerning two random variables a and b.

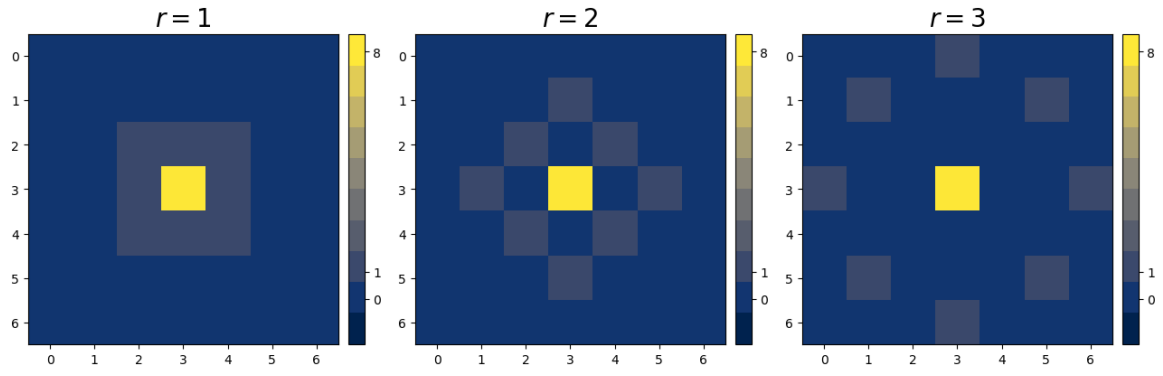
3.2 The proposed formulation

3.2.1 Ideas from sum and difference histograms

Let us begin by considering the sum histogram $p\left(I(x)+I(x')\right)$. We first note that summing neighbouring pixels is a (poor) way to blur the image e.g. the sum $I(x)+I(x')$ averaged over eight directions $\theta = 45°k \;\forall\; k \in [0,7]$ with distance $r = 1$ can be expressed as:

$$I(x)+I(x') = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 8 & 1 \\ 1 & 1 & 1 \end{bmatrix} * I(x) = \mathscr{K}_{+}^{(r=1)} * I(x), \tag{3.2}$$

where $\mathscr{K}_{+}^{(r=1)}$ is clearly a kind of blurring kernel ($*$ denotes convolution). Fig. 3.2 shows the equivalent kernels $\mathscr{K}_{+}$ for $I(x)+I(x')$ averaged over $\theta = 45°k \;\forall\; k \in [0,7]$ with $r = 1,2,3$.



(a) (b) (c)

Fig. 3.2 Equivalent kernels $\mathscr{K}_+K$ for sum $I(x)+I(x')$ averaged over eight directions $\theta = 45°k \;\forall\; k \in [0,7]$ with distance (a) $r = 1$; (b) $r = 2$; and (c) $r = 3$. For fair comparison, all kernels are presented in the same size (7×7 pixels).

That said, we can thus expect that the sum histogram[1] $p\left(\frac{1}{2}\left(I(x)+I(x')\right)\right)$ is similar to the image histogram $p\left(I(x)\right)$ with small r. We demonstrate this in Fig. 3.3 which shows the sums averaged over eight directions $\theta = 45°k \;\forall\; k \in [0,7]$ of the grayscale image *Lena*. With $r = 1$, the sum is virtually indistinguishable from the original image; the corresponding $p\left(\frac{1}{2}\left(I(x)+I(x')\right)\right)$ and $p\left(I(x)\right)$ are also almost identical. With increasing r (blurring), the dissimilarity between $p\left(\frac{1}{2}\left(I(x)+I(x')\right)\right)$ and $p\left(I(x)\right)$ grows as expected.

[1]The sums are halved so that the ranges of the image and sum histograms are identical for fair comparison.

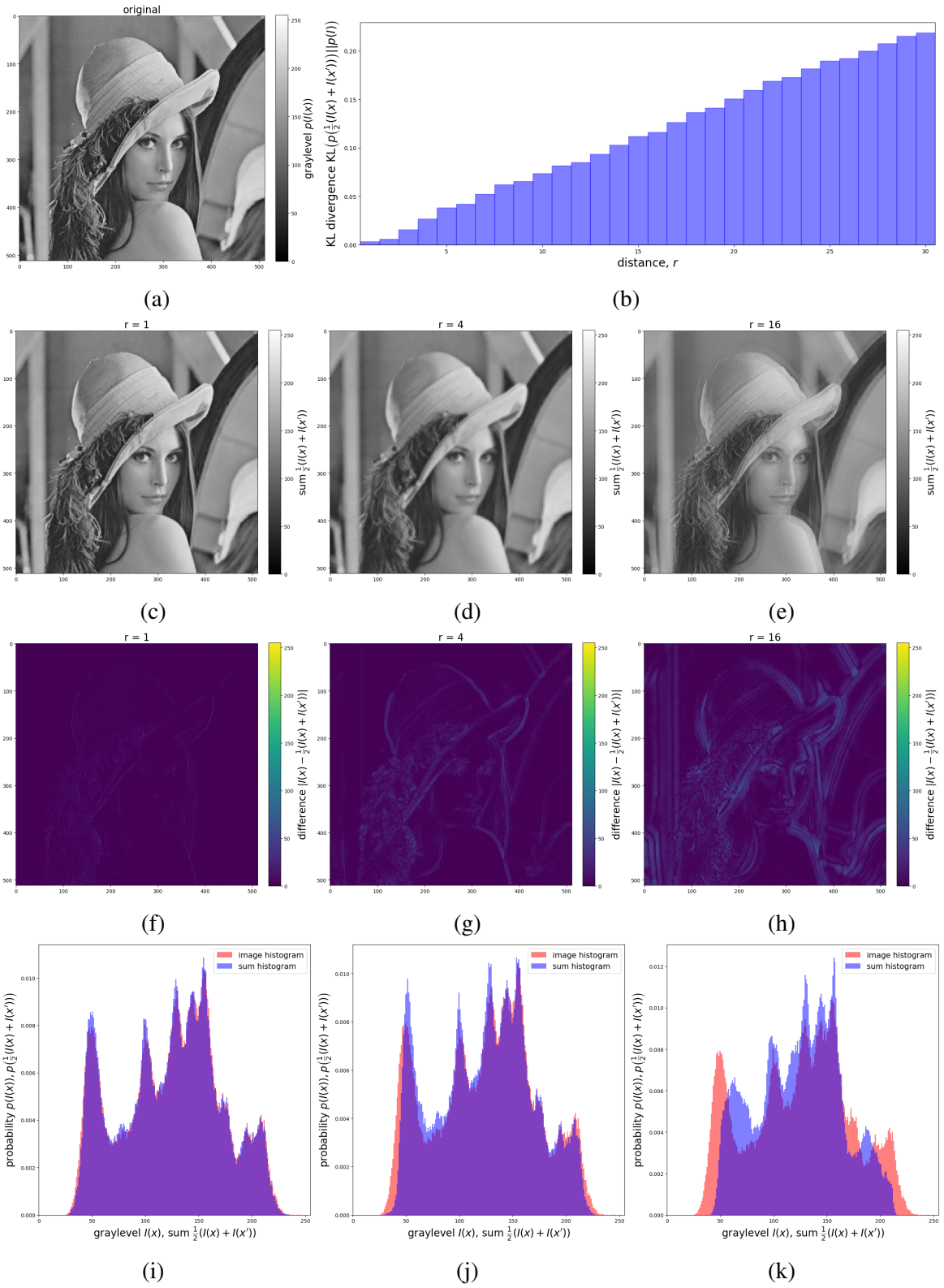


Fig. 3.3 (a) The grayscale image *Lena* (512 × 512 pixels); (b) KL divergence from image histogram $p\left(I(x)\right)$ to sum histograms $p\left(\frac{1}{2}\left(I(x)+I(x')\right)\right)$ for $r \in [1,30]$ averaged over eight directions $\theta = 45°k \;\forall\; k \in [0,7]$; the sum $\frac{1}{2}\left(I(x)+I(x')\right)$ for (c) $r=1$; (d) $r=4$; (e) $r=16$; absolute difference between $I(x)$ and $\frac{1}{2}\left(I(x)+I(x')\right)$ for (f) $r=1$; (g) $r=4$; (h) $r=16$; comparison between $p\left(I(x)\right)$ and $p\left(\frac{1}{2}\left(I(x)+I(x')\right)\right)$ for (i) $r=1$; (j) $r=4$; and (k) $r=16$.

Hence, we can approximate sum histogram by image histogram for small r. In case of hyperspectral image, we can thus express the sum histogram by the spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ as expressed in (2.11) with good approximation provided that r is small:

$$p\Big(\big(I(x)+I(x')\big)\Big) \xrightarrow{\text{small } r} \mathscr{J}_1^{(\hat{S})}(I). \tag{3.3}$$

Next, let us consider the difference histogram $p\big(I(x)-I(x')\big)$ in the context of hyperspectral image. As spectral subtraction is physically undefined, we proposed to consider instead spectral difference $d\big(I(x),I(x')\big)$ where $d(\cdot)$ is a spectral difference measure e.g. RMSE, SAM, and KLPD. Hence, we can express the difference histogram by the neighbouring (spectral) difference histogram $\mathscr{J}_2^{(r,\theta)}(I)$ of spectral differences Δs:

$$p\Big(\big(I(x)-I(x')\big)\Big) \xrightarrow{\text{spectral}} \mathscr{J}_2^{(r,\theta)}(I), \tag{3.4}$$

where $\mathscr{J}_2^{(r,\theta)}(I)$ is defined for a given hyperspectral image $I:\mathbb{R}^2\to\mathbb{R}^L$ with L bands by:

$$\mathscr{J}_2^{(r,\theta)}(I) = p\Big(d\big(I(x),I(x')\big)=\Delta s\Big). \tag{3.5}$$

Employing Kullback-Leibler pseudodivergence (KLPD) as the spectral difference measure, we define $\mathscr{J}_2^{(r,\theta)}(I)$ as the joint probability density function (PDF) of spectral shape difference ΔG and spectral intensity difference ΔW between pixels $I(x)$ and $I(x')$ separated at r and θ as:

$$\begin{aligned} &\mathscr{J}_2^{(r,\theta)}(I) = p\Big(\text{KLPD}\big(I(x),I(x')\big)=(\Delta G,\Delta W)\Big) \\ &\text{s.t.}\quad x' = x + r\cos\theta. \end{aligned} \tag{3.6}$$

For rotational invariance, we can consider the following simplification:

$$\mathscr{J}_2^{(r)}(I) = p\left(\frac{1}{2\pi}\int_0^{2\pi} \text{KLPD}\left(I(x), I(x+r\cos\theta)\right) d\theta = (\overline{\Delta G}, \overline{\Delta W})\right), \tag{3.7}$$

where $\overline{\Delta G}$ and $\overline{\Delta W}$ dictate the average spectral differences over θ.

We note that the formulation of $\mathscr{J}_2^{(r,\theta)}(I)$ is not new; an equivalent formulation for grayscale images is the gray level difference histogram (GLDM) [86] which considers the absolute graylevel difference of pixel pairs separated at a given distance and direction. For texture discrimination, GLDM has been shown to archive similar performance to GLCM [86]. This shows that most of the spatial information in GLCM is captured by the difference histogram, which is understandable as the sum histogram is approximately the image histogram with little spatial information. We can thus also expect $\mathscr{J}_2^{(r,\theta)}(I)$ to work well for hyperspectral image.

3.2.2 Edge sharpness, scale, and direction

Next, we shall demonstrate the relationship between the distribution of neighbouring spectral difference $\text{KLPD}\left(I(x), I(x')\right)$ and the underlying spatial variation. Specifically, we show how $\text{KLPD}\left(I(x), I(x')\right)$ varies with edge sharpness, directionality, and scale (of local spatial variation). For this purpose, we make use of the hyperspectral image *Chart3* (512 $\times$ 512 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval [54] as shown in Fig. 3.4 from which patches 1 - 16 (as bordered and numbered in red) are cropped for further examination. We note that *Chart3* is actually a binary image. However, due to imperfect acquisition (noise, uneven illumination, limited spatial resolution, etc.) it appears in multiple shades of gray.

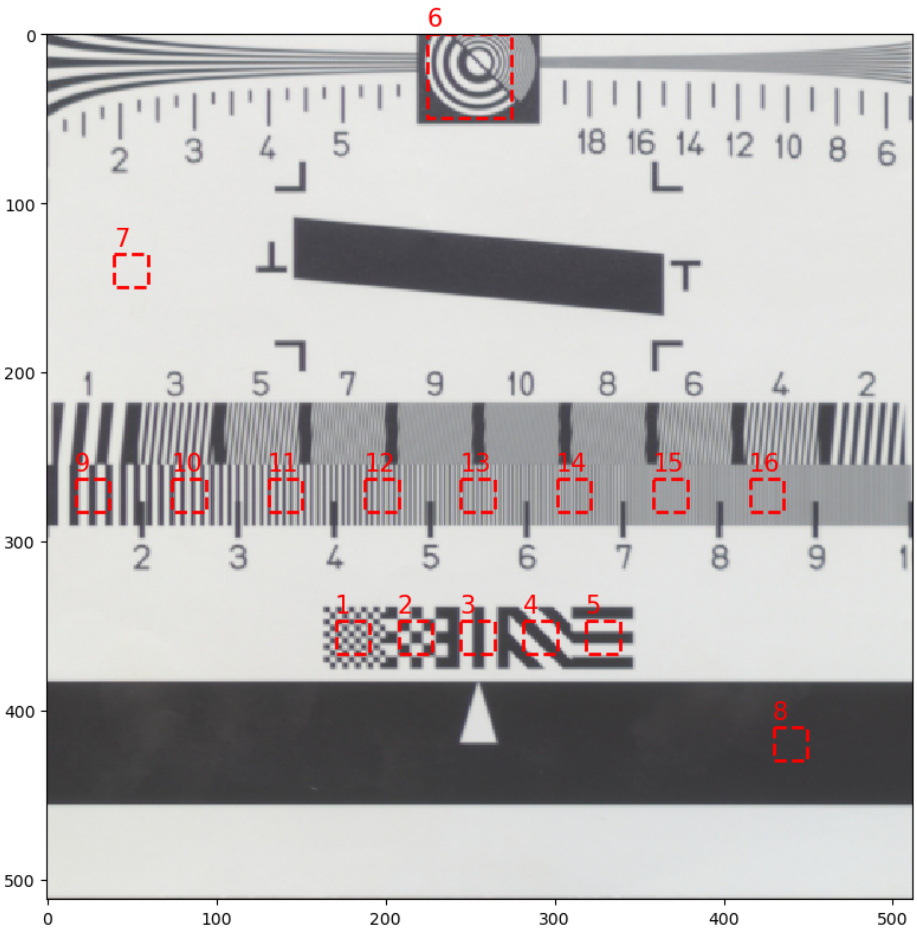
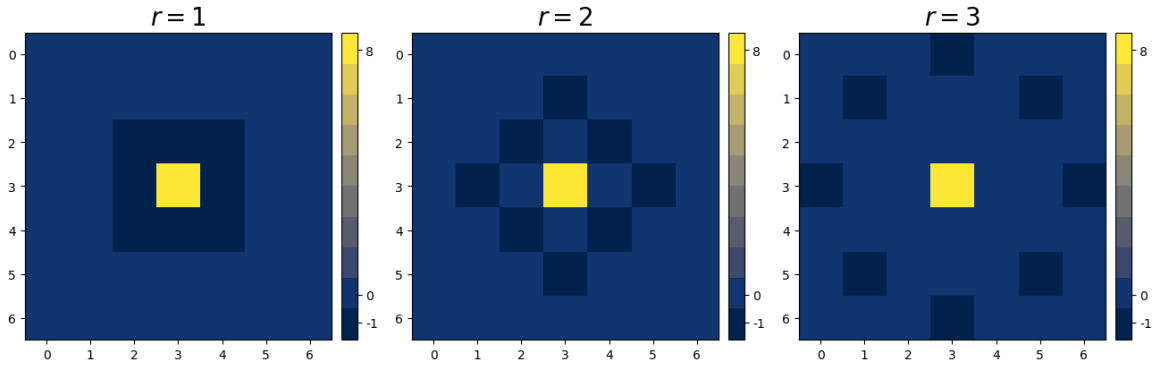


Fig. 3.4 The hyperspectral image *Chart3* (512 × 512 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval, rendered in RGB for illustration. Bordered and numbered in red are the 16 patches considered for neighbouring spectral difference calculation as shown in Fig. 3.7, Fig. 3.8, and Fig. 3.10 - 3.13.

To interpret neigbouring spectral difference KLPD $\big(I(x), I(x')\big)$, let us temporarily revert to grayscale image I for simplicity. Considering the minimum distance $r = 1$, the difference $I(x) - I(x')$ averaged over eight directions $\theta = 45°k \;\forall\; k \in [0,7]$ can be expressed as:

$$I(x) - I(x') = \frac{1}{8}\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} * I(x) = \mathscr{K}_{-}^{(r=1)} * I(x), \tag{3.8}$$

where $\mathscr{K}_{-}^{(r=1)}$ is clearly a kind of Laplacian kernel. Fig. 3.5 shows the equivalent kernels $\mathscr{K}_{-}$ for $I(x)-I(x')$ averaged over $\theta = 45°k \;\forall\; k \in [0,7]$ with $r = 1,2,3$. Therefore, we can expect $I(x)-I(x')$ and thus KLPD $\big(I(x),I(x')\big)$ to act like an edge detector. We demonstrate this in Fig. 3.6 which shows KLPD $\big(I(x),I(x')\big)$ of *Chart3* calculated with $r = 3$ and different θ. It can be seen that with $\theta = 0°$ (or with $\theta = 180°$, not shown), KLPD $\big(I(x),I(x')\big)$ works like a vertical edge detector, whereas with $\theta = 90°$ (or with $\theta = 270°$, not shown) it works like a horizontal edge detector. With KLPD $\big(I(x),I(x')\big)$ averaged over eight directions $\theta = 45°k \;\forall\; k \in [0,7]$, all edges (horizontal, vertical, diagonal, etc.) are visibly detected.



(a) (b) (c)

Fig. 3.5 Equivalent kernels $\mathscr{K}_{-}$ for difference $I(x)-I(x')$ averaged over eight directions $\theta = 45°k \;\forall\; k \in [0,7]$ with distance (a) $r = 1$; (b) $r = 2$; and (c) $r = 3$. For fair comparison, all kernels are presented in the same size (7×7 pixels).

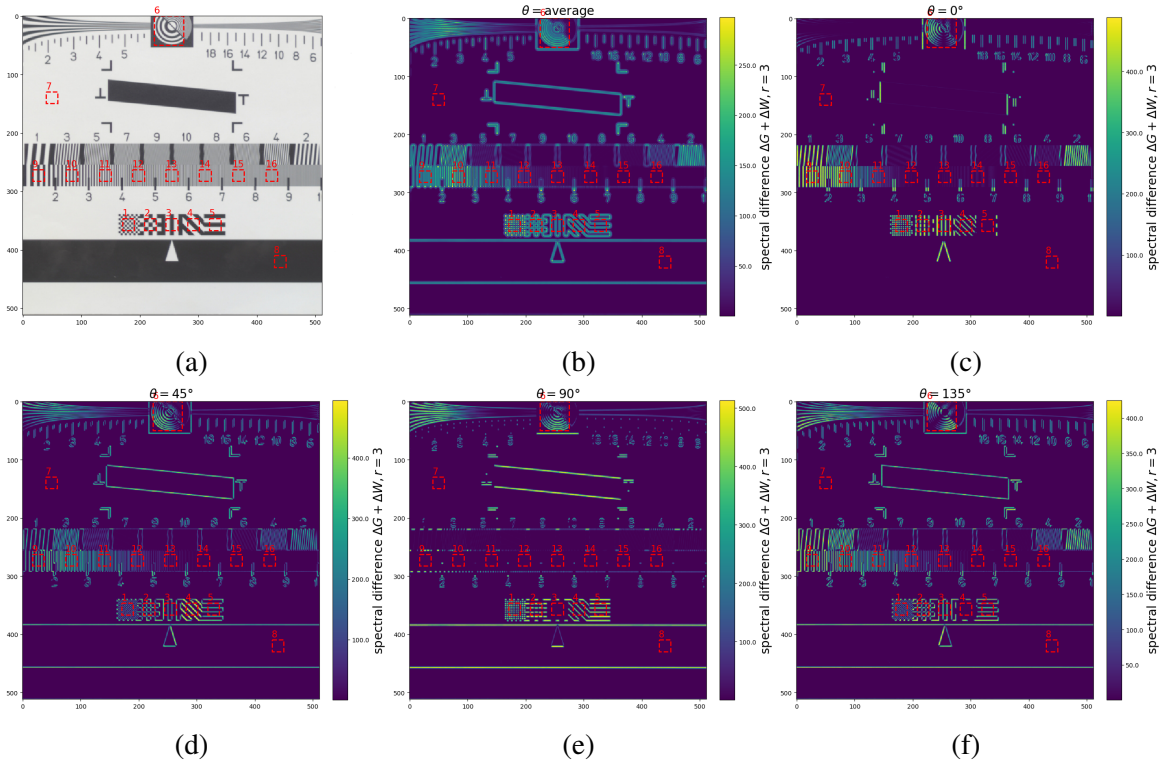


Fig. 3.6 Neighbouring spectral difference of *Chart3* calculated with distance $r = 3$ and (a) averaged over eight directions $\theta = 45°k \; \forall \; k \in [0,7]$; (b) $\theta = 0°$; (c) $\theta = 45°$; (d) $\theta = 90°$; and (e) $\theta = 135°$.

Having established the relationship between neighbouring spectral difference KLPD $\big(I(x), I(x')\big)$ and edge detection, the interpretation of KLPD $\big(I(x), I(x')\big)$ becomes more straightforward. Knowing the fact that *Chart3* is a binary image acquired with limited spatial resolution, we can expect that the regions with similar patterns e.g. patches 9 - 16 have decreasing edge sharpness with increasing rate of spatial variation. To substantiate this, we plot the spectra from each patch and observe that indeed, the spectral variation (overall difference between the white and black spectra) decreases from patch 9 to 16 as shown in Fig. 3.7 and 3.8. We proceed to calculating their KLPD $\big(I(x), I(x')\big)$ with $r = 1$ and $\theta = 0°$ (since there is only horizontal spatial variation). It can be seen that KLPD $\big(I(x), I(x')\big)$ at the edges also decrease from patches 9 to 16, thus confirming the role of neighbouring spectral difference as a measure of edge sharpness.

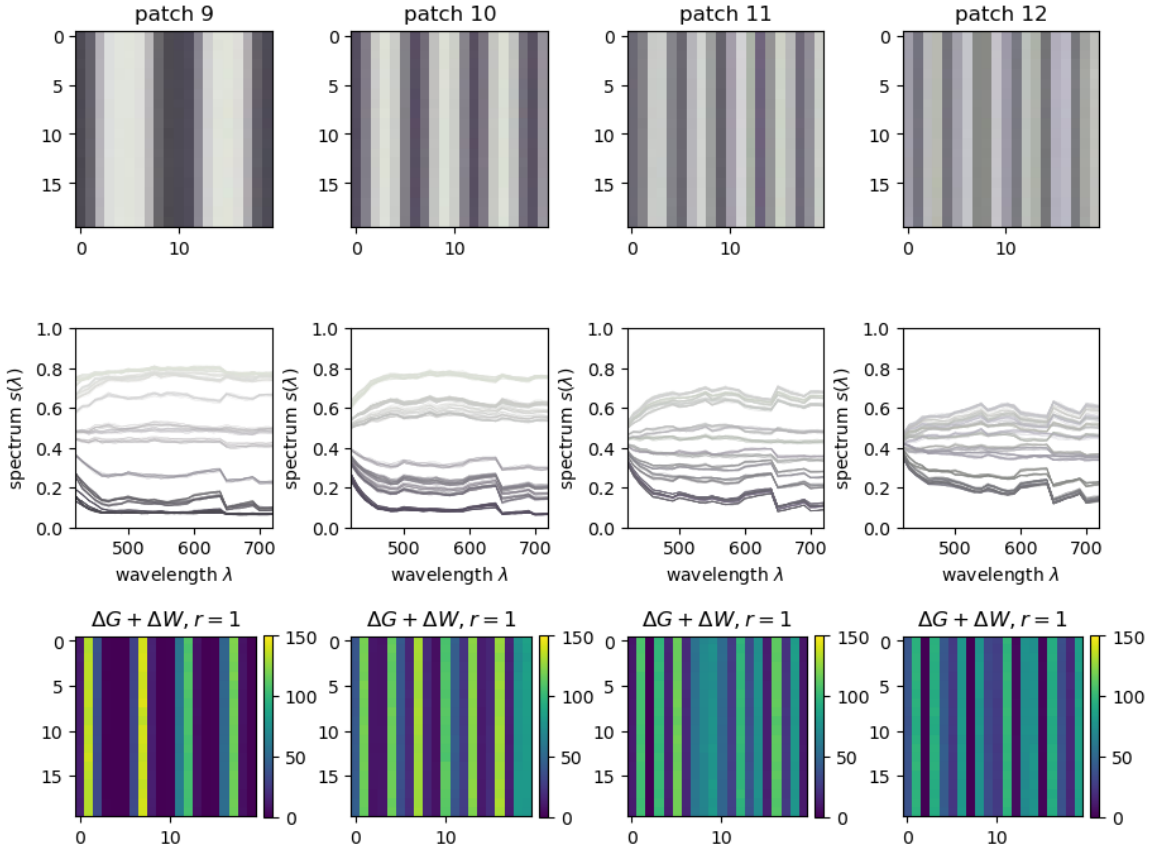


Fig. 3.7 The patches (20×20 pixels) rendered in RGB (top row), followed by the ($20 \times 20 = 400$) spectra of the corresponding patches (middle row) and the neighbouring spectral difference KLPD $\big(I(x), I(x')\big) = \Delta G + \Delta W$ calculated with $r = 1$ and $\theta = 0$ (bottom row) for patches 9 - 12 of *Chart3*.

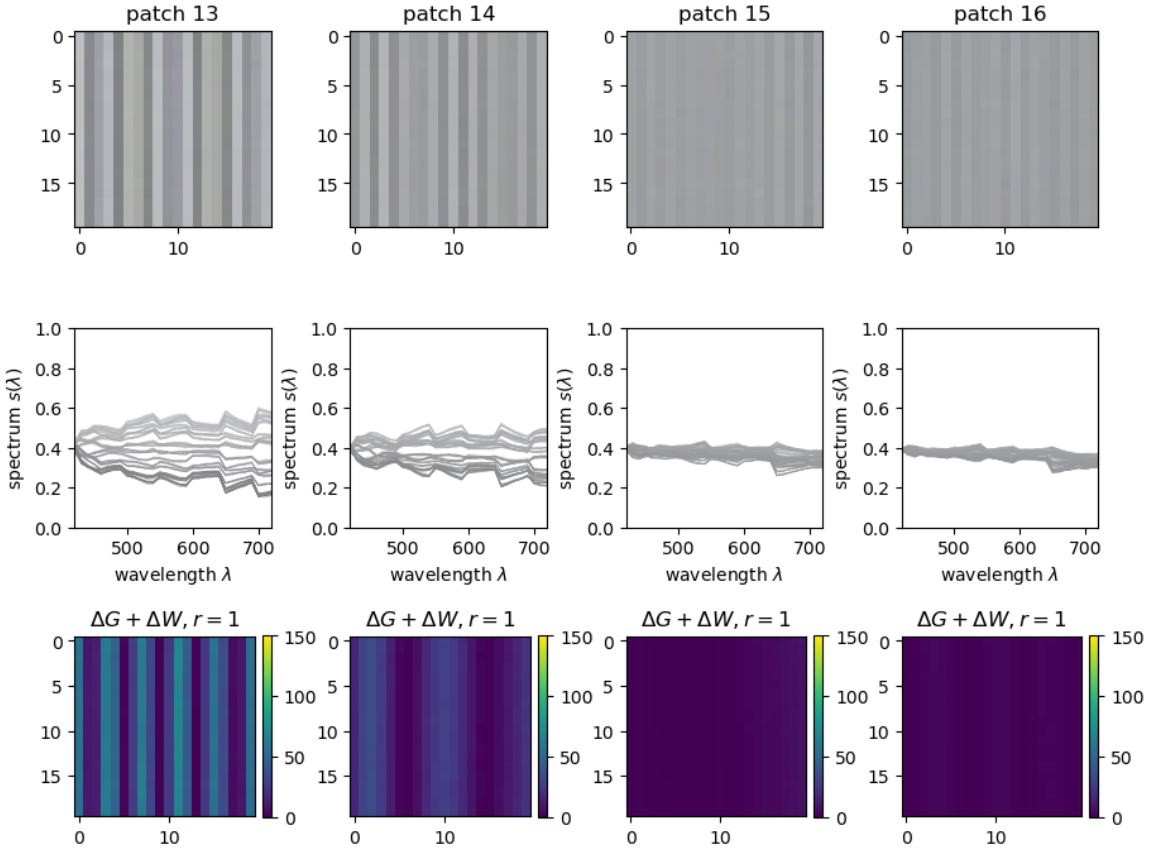


Fig. 3.8 The patches (20×20 pixels) rendered in RGB (top row), followed by the ($20 \times 20 = 400$) spectra of the corresponding patches (middle row) and the neighbouring spectral difference KLPD $\big(I(x), I(x')\big) = \Delta G + \Delta W$ calculated with $r = 1$ and $\theta = 0$ (bottom row) for patches 13 - 16 of *Chart3*.

The distribution of neighbouring spectral difference KLPD $\big(I(x), I(x')\big)$ also reveals information about scale of the spatial variation. Specifically, KLPD $\big(I(x), I(x')\big)$ is mostly high for smaller scale, whereas KLPD $\big(I(x), I(x')\big)$ is mostly low for larger scale. This is because for smaller scale, there is a higher probability of the neighbouring pixels $I(x)$ and $I(x')$ being different. Conversely, for larger scale, the probability of $I(x)$ and $I(x')$ being similar is higher. We note that in this case the absolute value of KLPD $\big(I(x), I(x')\big)$ does not matter (which refers to edge sharpness); it is proportion of the higher values of KLPD $\big(I(x), I(x')\big)$ relative to the lower values that reflects the scale. We demonstrate this in Fig. 3.9 which shows binary images (40×40 pixels) of checkerboard pattern with increasing

block size: 1×1, 2×2, 4×4, and 10×10. Consindering their neighbouring (absolute) pixel difference $|I(x) - I(x')|$ (which can only be either 0 or 1) with $r = 1$ and $\theta = 0°$, it can be seen that $|I(x) - I(x')| = 1 \; \forall x, x' \in \mathbb{R}^2$ in the case for 1×1 blocks. Naturally, this is because all the neigbouring pixels $I(x)$ and $I(x')$ are of different values i.e. $\big(I(x), I(x')\big) = (0,1)$ or $(1,0)$. As size of the blocks increases, the proportion of $|I(x) - I(x')| = 1$ can be seen to decrease. In the case for 10×10 blocks, $|I(x) - I(x')|$ is mostly zero because $I(x)$ and $I(x')$ are mostly of the same values i.e. $\big(I(x), I(x')\big) = (0,0)$ or $(1,1)$. Verifying this with real hyperspectral image, we refer to KLPD $\big(I(x), I(x')\big)$ of patches 9 - 16 as illustrated in Fig. 3.10 and 3.11. Indeed, we observe that the proportion of higher KLPD $\big(I(x), I(x')\big)$ increases from patch 9 to 14 in accordance to their decreasing scale. However, the proportion of higher KLPD $\big(I(x), I(x')\big)$ actually decreases in the case for patch 15 (and less so for patch 16) when it is expected to increase further due to the even smaller scale. We attribute this to the varying degree of blurring throughout patch 15 and 16 due to the limited spatial resolution. In this case, scale interpretation from KLPD $\big(I(x), I(x')\big)$ becomes less straightforward.

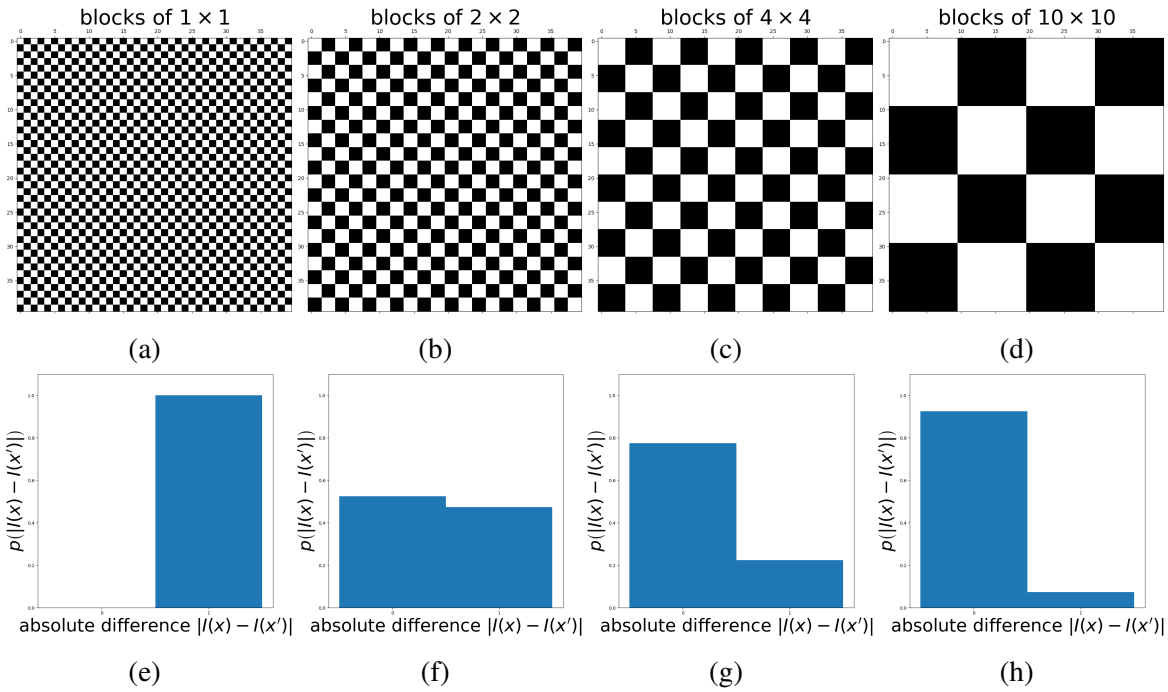


Fig. 3.9 Binary images of checkerboard pattern with varying structure size (top row), followed by their corresponding distribution of neighbouring (absolute) pixel difference as calculated with $r = 1$ and $\theta = 0°$ (bottom row).

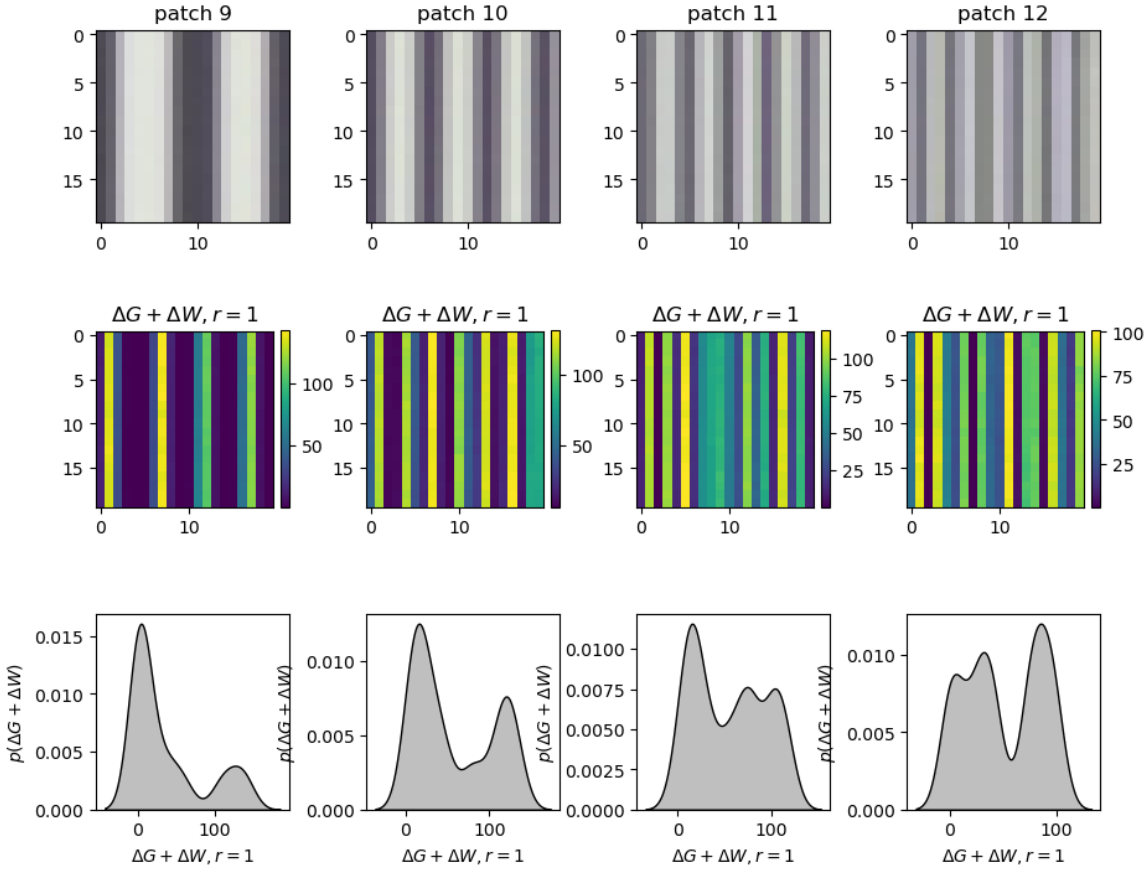


Fig. 3.10 The patches (20 × 20 pixels) rendered in RGB (top row), followed by neighbouring spectral difference KLPD $\left(I(x), I(x')\right) = \Delta G + \Delta W$ calculated with $r = 1$ and $\theta = 0$ (middle row) and distribution of $\Delta G + \Delta W$ (bottom row) for patches 9 - 12 of *Chart3*.

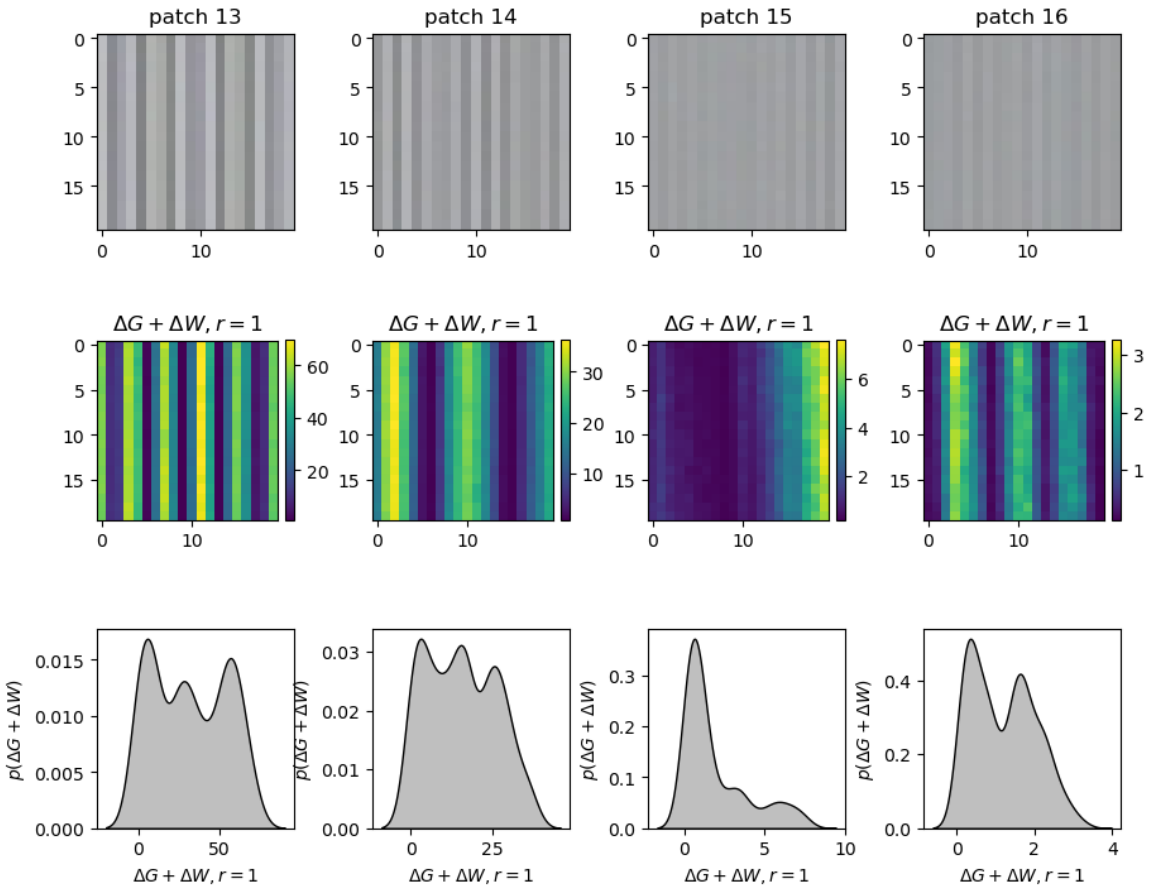


Fig. 3.11 The patches (20×20 pixels) rendered in RGB (top row), followed by neighbouring spectral difference KLPD $\left(I(x), I(x')\right) = \Delta G + \Delta W$ calculated with $r = 1$ and $\theta = 0$ (middle row) and distribution of $\Delta G + \Delta W$ (bottom row) for patches 13 - 16 of *Chart3*.

To obtain information about spatial directionality, we are obliged to calculate neighbouring spectral difference KLPD $\left(I(x), I(x')\right)$ with multiple directions θ. Naturally, KLPD $\left(I(x), I(x')\right)$ is the highest when being calculated in the direction of the spatial variation. Conversely, KLPD $\left(I(x), I(x')\right)$ is the lowest when being calculated in the direction perpendicular to the spatial variation. If the spatial variation is isotropic, then KLPD $\left(I(x), I(x')\right)$ is similar when being calculated with any θ. We demonstrate this in Fig. 3.12 and 3.13 which show KLPD $\left(I(x), I(x')\right)$ of patches 1 - 8 calculated with distance $r = 1, 2, 3$ and $\theta = 0°, 45°, 90°, 135°$. Let us begin by considering the case for $r = 3$. It can be seen that for both patches 1 (smaller blocks) and 2 (larger blocks), KLPD $\left(I(x), I(x')\right)$ is higher with

both $\theta = 0°$ and $\theta = 90°$ in accordance to their dominant spatial variations in horizontal and vertical directions. In contrary, KLPD $\left(I(x), I(x')\right)$ is lower with both $\theta = 45°$ and $\theta = 135°$. For patch 3 (vertical strip), 4 (diagonal strip), and 5 (horizontal strip), KLPD $\left(I(x), I(x')\right)$ are the highest (lowest) with $\theta = 0°$ ($\theta = 90°$), $\theta = 45°$ ($\theta = 135°$), and $\theta = 90°$ ($\theta = 0°$) respectively which corresponds to their direction of spatial variation. As for patches 6 - 8 which are isotropic, KLPD $\left(I(x), I(x')\right)$ is similar with $\theta = 0°, 45°, 90°$, and $135°$ as expected.

In the cases for $r = 1$ and $r = 2$, however, it can be seen that neighbouring spectral difference KLPD $\left(I(x), I(x')\right)$ is not necessarily the highest only in the direction of the spatial variation. In some instances, KLPD $\left(I(x), I(x')\right)$ is similarly high with multiple directions θ it is calculated (implying that the spatial variation is multidirectional) even though the examined patch is monodirectional. For example, in the case for $r = 1$, KLPD $\left(I(x), I(x')\right)$ of patch 5 is identically high with $\theta = 45°$, $\theta = 90°$, and $\theta = 135°$ while it is supposedly the highest only with $\theta = 90°$. A similar scenario can be observed too with patches 3 (in the case for $r = 1$) and 4 (in the case for $r = 2$). Concerning patches 1 and 2, the calculation result of KLPD $\left(I(x), I(x')\right)$ in the case for $r = 1$ is even contradictory as it is higher with both $\theta = 45°$ and $\theta = 135°$ while it is supposedly higher with both $\theta = 0°$ and $\theta = 90°$. We note that in these cases, the calculation of KLPD $\left(I(x), I(x')\right)$ with $\theta = 45°$ and $135°$ is the cause of problems. Due to the discrete nature of digital image, the calculation of KLPD $\left(I(x), I(x')\right)$ other than with $\theta = 0°$ and $\theta = 90°$ can only be approximated. For example, in the case for $r = 1$, the calculation of KLPD $\left(I(x), I(x')\right)$ with $\theta = 45°$ can only be approximated with neigbouring pixel $I(x')$ at one pixel to the left of and one pixel above $I(x)$ by referring to Fig. 3.5. Meanwhile, it should have been done with $I(x')$ at "0.71 pixel" to the left of and "0.71 pixel" above $I(x)$. This explains the problematic calculation of KLPD $\left(I(x), I(x')\right)$ with $\theta = 45°, 135°$, etc. in the case for $r = 1$ and $r = 2$. However, it can be seen that the problem is relieved in the case for $r = 3$ as the approximation error decreases with increasing r. That

said, care should be taken when choosing r for analyzing the spatial directionality with KLPD $\big(I(x), I(x')\big)$. The higher r is, the more accurate the analysis will be in general.

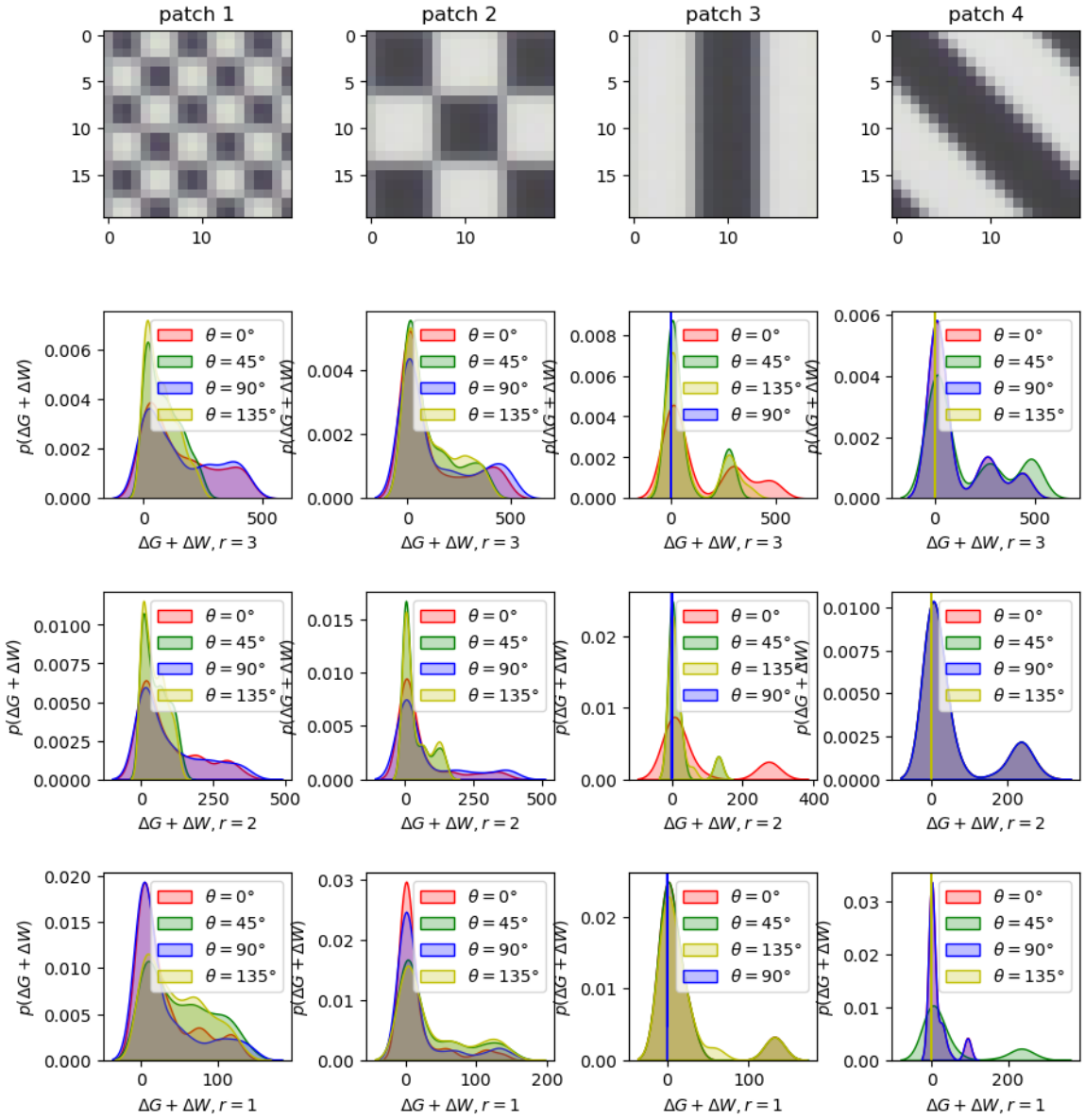


Fig. 3.12 The patches (20×20 pixels) rendered in RGB (top row), followed by distribution of neigbouring spectral difference KLPD $\big(I(x), I(x')\big) = \Delta G + \Delta W$ calculated with $\theta = 0°, 45°, 90°, 135°$ and $r = 3$ (second row), $r = 2$ (third row), or $r = 1$ (bottom row) for patches 1 - 4 of *Chart3*.

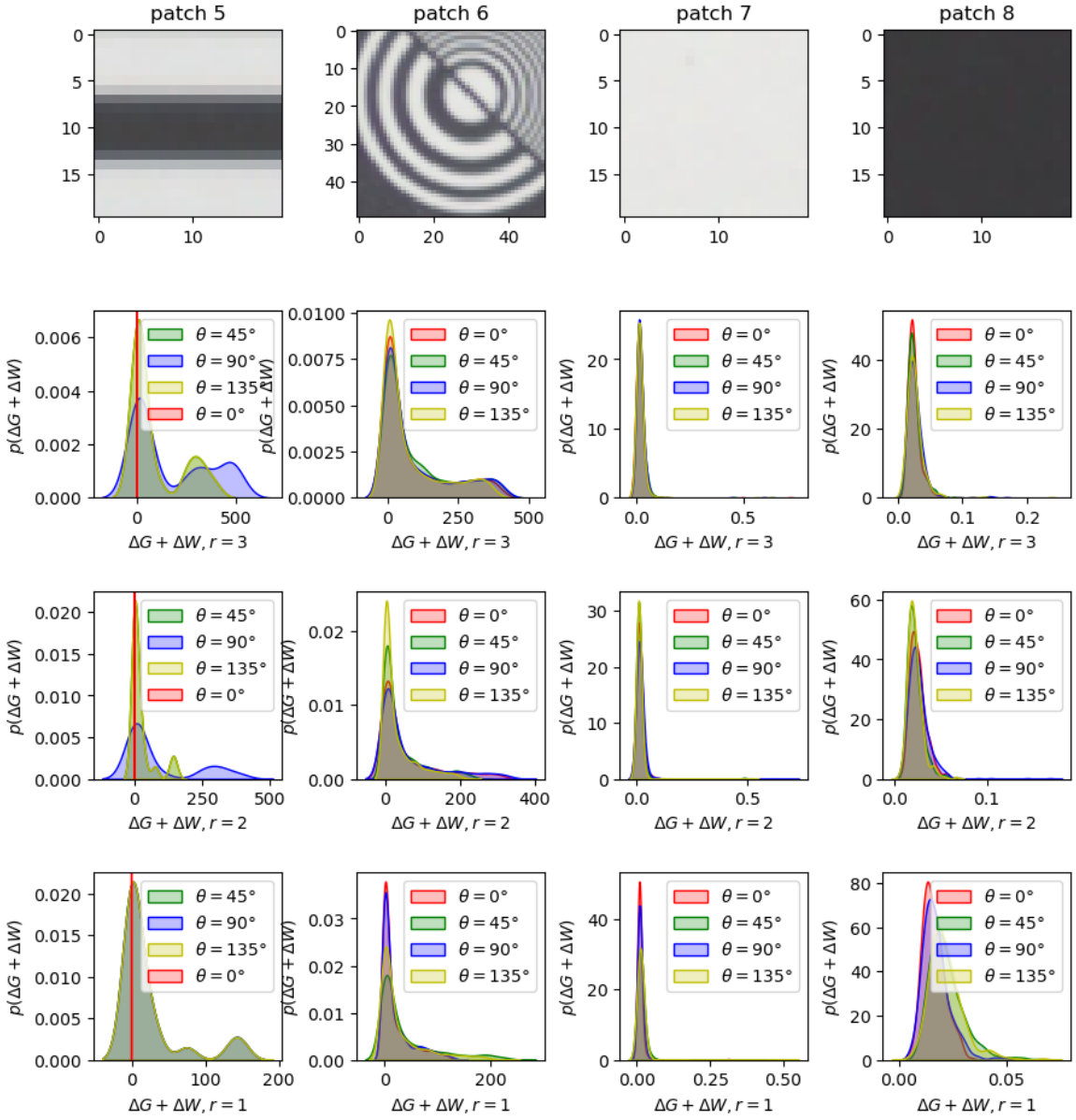


Fig. 3.13 The patches (20×20 pixels, except for patch 6 which measures 50×50 pixels) rendered in RGB (top row), followed by distribution of neigbouring spectral difference $\mathrm{KLPD}\left(I(x), I(x')\right) = \Delta G + \Delta W$ calculated with $\theta = 0°, 45°, 90°, 135°$ and $r = 3$ (second row), $r = 2$ (third row), or $r = 1$ (bottom row) for patches 5 - 8 of *Chart3*.

3.2.3 Relative spectral difference occurrence matrix (RSDOM)

Considering a hyperspectral image I, we have argued that the sum and difference histograms can be expressed by spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ and neighbouring difference

histogram $\mathscr{J}_2^{(r,\theta)}(I)$ respectively in previous section. In the following, we define the texture feature RSDOM (relative spectral difference occurrence matrix) which is expressed as the joint construct between $\mathscr{J}_1^{(\hat{S})}(I)$ and $\mathscr{J}_2^{(r,\theta)}(I)$. Considering R spectral references $\hat{S} = \{\hat{s}_1, \ldots, \hat{s}_R\}$ and a spatial assessment with (r,θ), RSDOM is expressed as a joint PDF of $R+3$ dimensions:

$$\begin{aligned} \mathscr{T}^{(\hat{S},r,\theta)}(I) &= \mathscr{J}_1^{(\hat{S})}(I) \ \cap \ \mathscr{J}_2^{(r,\theta)}(I) \\ &= p(\Delta G_1, \ldots, \Delta G_R, W, \Delta G, \Delta W), \end{aligned} \tag{3.9}$$

where $\mathscr{T}$ is the texture feature as defined by RSDOM in this case. Recall that $\mathscr{J}_1^{(\hat{S})}(I)$ (from Chapter 2) is the joint PDF of $R+1$ random variables i.e. the spectral shape differences $\{\Delta G_1, \ldots, \Delta G_R\}$ and spectral intensity W at a given spatial location x:

$$\mathscr{J}_1^{(\hat{S})}(I) = p\left(\begin{array}{c} \bigcap_{r=1}^{R} \left(\mathrm{KLD}\left(\bar{I}(x), \hat{s}_r\right) = \Delta G_r \right), \\ \left\| I(x) \right\|_1 = W \end{array} \right)$$

$$\text{s.t.} \qquad \left\| \bar{I}(x) \right\|_1 = \left\| \hat{s}_r \right\|_1 = 1 \ \forall \ r \in [1, R],$$

and that $\mathscr{J}_2^{(r,\theta)}(I)$ is the bivariate PDF of spectral shape difference ΔG and spectral intensity difference ΔW between pixels $I(x)$ and $I(x')$ separated at distance r and direction θ as:

$$\mathscr{J}_2^{(r,\theta)}(I) = p\left(\mathrm{KLPD}\left(I(x), I(x')\right) = (\Delta G, \Delta W) \right)$$

$$\text{s.t.} \quad x' = x + r\cos\theta.$$

The formulation of RSDOM as a joint PDF serves three advantages. First, the expression is mathematically sound in contrast to other empirical means such as concatenation or addition of the two statistics. Second, statistical distances can be naturally used for texture similarity measurement thanks to the PDF expression. Third, the feature is highly discriminative thanks to the dual texture representation in terms of spectral and spatial measures.

The expression of RSDOM in difference space has several desirable effects. First of all, this allows the characterization of spectral and spatial properties in a continuous manner. As the spectra are not explicitly modelled (but rather, the spectral differences), quantization is unnecessary as in the case of cooccurrence matrix. Hence, information about the texture is fully preserved. Furthermore, it allows full-band processing of RSDOM without the need for dimensionality reduction e.g. PCA. Therefore, physical fidelity is preserved in line with metrology. Besides, the feature calculation is independent of the sensor resolution. This enables texture comparison across different datasets with varying spectral resolutions.

3.2.4 Similarity measurement and statistical modelling

Having performed feature extraction, texture is discriminated thanks to a feature similarity or distance measure. We note that for the purpose of metrology, the similarity or distance measure must be developed respecting the mathematical nature of the feature. In this regard, the similarity measure must take into account the probabilistic nature of RSDOM.

For maximum discrimination, we decide not to impose any a priori assumptions but to extract information directly from the feature. Given two PDFs, their similarity can be measured using likelihood ratio test [24] (a nonparametric statistical test). It can be shown that maximizing the log-likelihood ratio is equal to minimizing the KL measure of information [20]. Hence, we can use KL divergence [44] for the similarity measurement between RSDOMs:

$$\mathrm{KLD}\left(\mathscr{T}^{(\hat{S},r,\theta)}(I),\mathscr{T}^{(\hat{S},r,\theta)}(I')\right)=\mathrm{KL}\left(\mathscr{T}^{(\hat{S},r,\theta)}(I)\middle\|\mathscr{T}^{(\hat{S},r,\theta)}(I')\right)+\mathrm{KL}\left(\mathscr{T}^{(\hat{S},r,\theta)}(I')\middle\|\mathscr{T}^{(\hat{S},r,\theta)}(I)\right), \tag{3.10}$$

concerning two textures (images) I and I' as assessed using the same set of spectral references $\hat{S}$ as well as with the same spatial relationship between pixel pairs parametrized by distance r and direction θ. Note that $\mathrm{KL}(\cdot,\cdot)$ is the KL measure of information as defined by [44]:

$$\begin{aligned}&\mathrm{KL}\left(\mathscr{T}^{(\hat{S},r,\theta)}(I)\middle\|\mathscr{T}^{(\hat{S},r,\theta)}(I')\right)\\&=\int\dots\int p(\Delta G_1,\dots,\Delta G_R,W,\Delta G,\Delta W)\log\left(\frac{p(\Delta G_1,\dots,\Delta G_R,W,\Delta G,\Delta W)}{p'(\Delta G_1,\dots,\Delta G_R,W,\Delta G,\Delta W)}\right)\mathrm{d}(\Delta G_1)\dots\mathrm{d}(\Delta W),\end{aligned} \tag{3.11}$$

such that $\mathscr{T}^{(\hat{S},r,\theta)}(I)=p(\Delta G_1,\dots,\Delta G_R,W,\Delta G,\Delta W)$ and $\mathscr{T}^{(\hat{S},r,\theta)}(I')=p'(\Delta G_1,\dots,\Delta G_R,W,\Delta G,\Delta W)$. On our choice of KL divergence as the similarity measure, although it exists many other alternatives [2] e.g. α-divergence and Bregman's divergence as well as s-divergence [71], the study of an optimal measure is beyond the focus of our work in this Ph.D. manuscript.

Due to the multidimensionality of RSDOM, it is impractical to process the KL divergence directly on histogram features. Moreover, such approach requires the same PDF support which can vary greatly depending on the texture. As a solution, we propose using the Gaussian mixture model (GMM), a class of density estimation methods that is capable of modelling any complex PDF given sufficient number of components. Using GMM, RSDOM can be parameterized using just a few parameters with a feature size of $M(1+D+D(D+1)/2)$ where M is the number of GMM components and D is the dimensionality:

$$\mathscr{T}^{(\hat{S},r,\theta)}(I)\approx\sum_{m=1}^{M}\pi_m\mathscr{N}(\mu_m,\Sigma_m), \tag{3.12}$$

where π_m is the mixture weight of the m^{th} Gaussian or normal component $\mathcal{N}(\mu_m, \Sigma_m)$ of mean μ_m and covariance matrix Σ_m, such that $\sum_{m=1}^{M} \pi_m = 1$. Note that in case of RSDOM, D is given by $R+3$. As there is no closed form solution for KL measure of information between GMMs, we resort to using the variational approximation [32]:

$$\mathrm{KL}\left(\mathscr{T}^{(\hat{S},r,\theta)}(I) \| \mathscr{T}^{(\hat{S},r,\theta)}(I')\right) \approx \sum_{m=1}^{M} \pi_m \log \frac{\sum_{m'=1}^{M} \pi_{m'} e^{-\mathrm{KLD}\left(\mathcal{N}(\mu_m,\Sigma_m) \| \mathcal{N}(\mu_{m'},\Sigma_{m'})\right)}}{\sum_{m''=1}^{M} \pi'_{m''} e^{-\mathrm{KLD}\left(\mathcal{N}(\mu_m,\Sigma_m) \| \mathcal{N}(\mu'_{m''},\Sigma'_{m''})\right)}}, \tag{3.13}$$

such that $\mathscr{T}^{(\hat{S},r,\theta)}(I) \approx \sum_{m=1}^{M} \pi_m \mathcal{N}(\mu_m, \Sigma_m)$ and that $\mathscr{T}^{(\hat{S},r,\theta)}(I') \approx \sum_{m=1}^{M} \pi'_m \mathcal{N}(\mu'_m, \Sigma'_m)$. Note that the KL measure of information between Gaussians is given analytically by [60]:

$$\mathrm{KL}\left(\mathcal{N}(\mu,\Sigma) \| \mathcal{N}(\mu',\Sigma')\right) = \frac{1}{2}\left(\log \frac{|\Sigma'|}{|\Sigma|} + \mathrm{tr}\left(\left(\Sigma'\right)^{-1}\Sigma\right) - D + (\mu'-\mu)^T \left(\Sigma'\right)^{-1} (\mu'-\mu)\right). \tag{3.14}$$

As RSDOM is a joint PDF of positive measures, it is positively skewed and dense around zero (leptokurtic). To increase compatibility for the GMM fitting, a diffeomorphism is applied by taking the logarithm of each dimension. Such transformation has a desirable effect of stretching the values close to zero while compressing those far from zero, thus reducing the skewness and conforming the distribution towards normality.

3.3 Summary

In this chapter, we focussed on the spatial characterization of texture. We started by recalling the first Julesz conjecture which lies the foundation for our work on describing texture using first- and second-order statistics. We next turned to the formulation of the graylevel cooccurrence matrix before studying its approximation using sum and difference histograms.

Then, we proposed our own approach by first pointing out the correspondence between sum histogram and image histogram, the latter we have accounted for in Chapter 2. Next, we express the difference histogram by neighbouring difference histogram thanks to the consideration of spectral difference between pixel pairs defined for a particular spatial relationship. Subsequently, we analysed the distribution of neighbouring spectral differences facing edge sharpness, scale, and direction of the underlying texture. Finally, we presented our first spectral-spatial texture feature termed relative spectral difference occurrence matrix (RSDOM) that combines the joint distribution of spectral differences (from Chapter 2) and neighbouring differences. As part of the metrological consideration, we also proposed a similarity measurement of RSDOM based on Kullback-Leibler divergence.

Chapter 4

Spectral-Spatial Characterization of Texture: Gradients-Based

Contents

4.1 Background 82

4.2 The proposed formulation 85

4.2.1 Automatic scale selection 85

4.2.2 Spectral gradient magnitude, direction, and coherence 102

4.2.3 Spectral shape and intensity gradients 126

4.2.4 Gradient histogram of spectral texture (GHOST) 136

4.3 Summary 138

4.1 Background

Relationship between derivative and neighbouring difference

In single-variable calculus, the difference quotient is given by.

$$\frac{f(x+h)-f(x)}{h}, \tag{4.1}$$

which when taken to the limit as h approaches 0 gives the derivative of the function f:

$$\frac{df(x)}{dx} = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}. \tag{4.2}$$

In case of digital image analysis which is defined in discrete domain, h are limited to integers. Considering a monovariate image $I : \mathbb{R}^2 \to \mathbb{R}$ with number of spectral bands $L = 1$, one of the definitions of the partial image derivatives is given by:

$$\begin{aligned} \frac{\partial I(x)}{\partial x_1} &= I(x_1+1, x_2) - I(x_1, x_2), \\ \frac{\partial I(x)}{\partial x_2} &= I(x_1, x_2+1) - I(x_1, x_2), \end{aligned} \tag{4.3}$$

obtained as per Robert cross operator with $h = 1$. Note that in this case, $\frac{\partial I(x)}{\partial x_1}$ is equivalent to, in the context of RSDOM, taking the difference between pixel pairs defined with $r = 1$ and $\theta = 0$. Similarly, $\frac{\partial I(x)}{\partial x_2}$ is equivalent to that with $r = 1$ and $\theta = \pi/2$. Indeed, taking the difference between neighboring pixels is a (poor) way to computing the image derivatives.

Having established the relationship between differences and derivatives, it is, therefore, natural to expand the formulation of RSDOM with derivatives. Such approach presents several interests. First, it enables the assessment of the image gradient which provides a natural measure of directionality. Second, it allows the application of scale space theory the

for determination of local structure size. Hence, the need for parameter selection in terms of multiscale and multidirectional assessment, as in the cof for RSDOM, is eliminated elegantly.

Assessment of image gradient

For monovariate image with number of spectral bands $L = 1$, the gradient at $x = (x_1, x_2)$ is a vector of partial derivatives defined with respect to the two spatial dimensions:

$$\nabla I(x) = \left[\frac{\partial I(x)}{\partial x_1} \quad \frac{\partial I(x)}{\partial x_2} \right]^T . \tag{4.4}$$

As derivative is only defined for continuous functions, a low-pass filter is to be applied on discrete images in conjunction with the gradient calculation. According to the scale-space theory, a good choice for the filter would be the Gaussian kernel g [50]. The scale-dependent $\nabla I(x;\sigma)$ as calculated with $g(x,\sigma)$ centered at x with standard deviation of σ is given by:

$$\nabla I(x;\sigma) = \left[\frac{\partial \big(I(x) * g(\sigma)\big)}{\partial x_1} \quad \frac{\partial \big(I(x) * g(\sigma)\big)}{\partial x_2} \right]^T . \tag{4.5}$$

Thanks to the commutativity of convolution, $\nabla I(x;\sigma)$ can be equivalently expressed as:

$$\nabla I(x;\sigma) = \left[\frac{\partial g(x,\sigma)}{\partial x_1} \quad \frac{\partial g(x,\sigma)}{\partial x_2} \right]^T * I(x). \tag{4.6}$$

Assessment of spectral gradient

For multivariate e.g. colour and hyperspectral image with $L > 1$, one is obliged to consider the Jacobian matrix, $\mathrm{J}I(x)$ as the generalization of gradient for vector-valued functions:

$$\mathrm{J}I(x;\sigma) = \begin{bmatrix} \frac{\partial I_1(x;\sigma)}{\partial x_1} & \frac{\partial I_2(x;\sigma)}{\partial x_1} & \cdots & \frac{\partial I_L(x;\sigma)}{\partial x_1} \\ \frac{\partial I_1(x;\sigma)}{\partial x_2} & \frac{\partial I_2(x;\sigma)}{\partial x_2} & \cdots & \frac{\partial I_L(x;\sigma)}{\partial x_2} \end{bmatrix}^T . \tag{4.7}$$

Given the unit vector $v(\theta) = [\cos\theta \;\; \sin\theta]^T$ pointing in the direction θ on $\mathbb{R}^2$ (the spatial dimensions), the Euclidean norm of the directional derivative $\mathrm{J}I(x;\sigma)v(\theta)$ is given by:

$$\|\mathrm{J}I(x;\sigma)v(\theta)\|_2 = \sqrt{\left(\mathrm{J}I(x;\sigma)v(\theta)\right)^T G\left(\mathrm{J}I(x;\sigma)v(\theta)\right)} = \sqrt{v(\theta)^T M v(\theta)}, \tag{4.8}$$

where G is the Gram matrix that takes into account the possible nonorthogonality of spectral channels, defined using the scalar product for integrable functions [10]:

$$G = \begin{bmatrix} \|C_1\|_2^2 & \left\langle C_1, C_2 \right\rangle_2 & \cdots & \left\langle C_1, C_L \right\rangle_2 \\ \left\langle C_2, C_1 \right\rangle_2 & \|C_2\|_2^2 & \cdots & \left\langle C_2, C_L \right\rangle_2 \\ \vdots & \vdots & \ddots & \vdots \\ \left\langle C_L, C_1 \right\rangle_2 & \cdots & \left\langle C_L, C_{L-1} \right\rangle_2 & \|C_L\|_2^2 \end{bmatrix}, \tag{4.9}$$

where C_l is the spectral sensitivity function for the l^{th} band. Note that for orthogonal bands, G reduces to identity matrix. Meanwhile, M, termed spatial covariance matrix, is given by:

$$\begin{aligned} M &= \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \mathrm{J}I(x;\sigma)^T G \mathrm{J}I(x;\sigma), \\ \text{s.t.} \quad M_{ii'} &= \left\langle \frac{\partial I(x;\sigma^*)}{\partial x_i}, \frac{\partial I(x;\sigma^*)}{\partial x_{i'}} \right\rangle_2, \end{aligned} \tag{4.10}$$

where $\frac{\partial I(x;\sigma^*)}{\partial x_i}$ is the vector of marginal image derivatives with respect to x_i:

$$\frac{\partial I(x;\sigma^*)}{\partial x_i} = \begin{bmatrix} \frac{\partial I_1(x;\sigma^*)}{\partial x_i} & \frac{\partial I_2(x;\sigma^*)}{\partial x_i} & \cdots & \frac{\partial I_L(x;\sigma^*)}{\partial x_i} \end{bmatrix} \;\; \forall \;\; i \in [1,2], \tag{4.11}$$

where I_l is the l^{th} spectral channel. Note that $\left\langle \frac{\partial I(x;\sigma^*)}{\partial x_i}, \frac{\partial I(x;\sigma^*)}{\partial x_{i'}} \right\rangle_2$ is the scalar product between $\frac{\partial I(x;\sigma^*)}{\partial x_i}$ and $\frac{\partial I(x;\sigma^*)}{\partial x_{i'}}$ defined as followed with possible nonzero cross terms:

$$\begin{aligned}\left\langle \frac{\partial I(x;\sigma^*)}{\partial x_i}, \frac{\partial I(x;\sigma^*)}{\partial x_{i'}} \right\rangle_2 &= \left(\frac{\partial I(x;\sigma^*)}{\partial x_i} \right)^T G \left(\frac{\partial I(x;\sigma^*)}{\partial x_{i'}} \right) \\ &= \sum_{l=1}^{L} \sum_{l'=1}^{L} \left\langle \frac{\partial I_l(x;\sigma^*)}{\partial x_i}, \frac{\partial I_{l'}(x;\sigma^*)}{\partial x_{i'}} \right\rangle_2 \quad \forall \; i,i' \in [1,2],\end{aligned} \tag{4.12}$$

As $v(\theta)$ is an unit vector, (4.8) is equivalent to the Rayleigh quotient. Therefore, the extrema of (4.8) are defined by (square root of) the eigenvalues $\lambda_\pm$ of M as given by:

$$\lambda_\pm = \frac{1}{2} \left(\text{tr}(M) \pm \sqrt{\text{tr}(M)^2 - 4|M|} \right). \tag{4.13}$$

Accordingly, the (patial) directions $\theta_\pm$ corresponding to the extrema $\lambda_\pm$ are given by [40]:

$$\begin{aligned}\theta_+ &= \text{sgn}(M_{12}) \arcsin \left(\frac{\lambda_+ - M_{11}}{\lambda_+ - \lambda_-} \right)^{\frac{1}{2}}, \\ \theta_- &= \theta_+ + \frac{\pi}{2}.\end{aligned} \tag{4.14}$$

Termed full vector gradient (FVG) [11], the presented spectral gradient assessment will form the basis for our following development of spatial characterization of texture.

4.2 The proposed formulation

4.2.1 Automatic scale selection

The scale with which the gradient is assessed is not trivial. A fine scale provides good (spatial) localization but is susceptible to noise and spurious structures. A coarse scale improves the signal to noise ratio, albeit with poorer localization and risk of distortion

by neighbouring structures. An optimal scale is thus, a tradeoff between detection and localization. To maximize the captured information, the chosen scale has to reflect the local structure size. Whereas smaller structures are optimally assessed using lower scales, larger structures necessitate higher scales for complete detection. However, in practice, the local structure sizes (and thus the optimum scale) cannot be known a priori (unless they are artificially generated with known dimensions). An automatic scale selection mechanism is therefore necessary.

Scale space maximum

A heuristic principle for scale selection was proposed by Lindeberg: "*in the absence of other evidence, assume that a scale level, at which some (possibly non-linear) combination of normalized derivatives assumes a local maximum over scales, can be treated as reflecting a characteristic length of a corresponding structure in the data.*" [52]. The need for scale normalization arises because a derivative tells about the rate of change per pixel which tends to decrease with increasing scale σ. The scale normalization is performed by [51]:

$$\frac{\partial}{\partial x} \xrightarrow{\text{scale}-\text{normalized}} \sigma \frac{\partial}{\partial x} \tag{4.15}$$

In the following analysis, let us temporarily revert to grayscale image $I : \mathbb{R}^2 \rightarrow \mathbb{R}$ for simplicity. Accordingly, we define the (local) optimal scale $\sigma^*(x)$ for I as the one which results in the maximum (scale-normalized[1]) gradient magnitude at x over the scale space:

$$\sigma^*(x) = \arg\max_{\sigma} \sigma \|\nabla I(x;\sigma)\|. \tag{4.16}$$

According to the scale-space theory [50], gradient magnitude is optimally calculated using Gaussian derivatives:

[1]For brevity, in the following we shall refrain from explicitly using the term scale normalized when referring to gradient magnitude. Hereafter, all gradient magnitudes are scale normalized unless stated otherwise.

$$\|\nabla I(x;\sigma)\| = \left\| \begin{bmatrix} \frac{\partial g(x,\sigma)}{\partial x_1} & \frac{\partial g(x,\sigma)}{\partial x_2} \end{bmatrix}^T * I(x) \right\|_2, \tag{4.17}$$

where $*$ denotes convolution. Note that the Gaussian derivatives are given by:

$$\frac{\partial g(x,\sigma)}{\partial x_i} = \frac{x_i' - x_i}{\sigma^2} g(x,\sigma) \quad \forall i \in [1,2], \tag{4.18}$$

where $x_i - x_o$ is the displacement from x_o in the direction of x_i. Note that $g(x_o,\sigma)$ is a two-dimensional Gaussian function centered at $x_o = (x_{o,1}, x_{o,2})$ with standard deviation of σ:

$$g(x,\sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_1' - x_1)^2 + (x_2' - x_2)^2}{2\sigma^2} \right). \tag{4.19}$$

Relationship between scale and structure size

To derive the relationship between σ^* and the structure size, suppose that there is a circular structure centered at x_o with radius r on I. Considering the scale space evolution of the gradient magnitude at x_o, the scale space maximum can be obtained by setting:

$$\frac{\partial}{\partial \sigma} \left[\sigma \left\| \nabla I(x_o;\sigma) \right\| \right] = 0. \tag{4.20}$$

Note that in case of circular structure, the image derivatives at x_o are the same:

$$\frac{\partial g(x_o,\sigma)}{\partial x_1} * I(x_o) = \frac{\partial g(x_o,\sigma)}{\partial x_2} * I(x_o). \tag{4.21}$$

Therefore, (4.20) can be simplified by consideration in either direction x_i for $i = 1$ or $i = 2$:

$$\frac{\partial}{\partial \sigma} \left[\sigma \frac{\partial g(x_o,\sigma)}{\partial x_i} * I(x_o) \right] = 0 \quad \forall i \in [1,2]. \tag{4.22}$$

Substituting (4.18) into (4.22) and then working out the derivative give:

$$\frac{-3\sigma^2(x_i - x_{0,i}) + (x_i - x_{0,i})[(x_1 - x_{0,1})^2 + (x_2 - x_{o,2})^2]}{\sigma^4} g(x_o, \sigma) * I(x_o) = 0. \tag{4.23}$$

Knowing that $g(x_o, \sigma) * I(x_o) \neq 0$ and noting that $(x_1 - x_{o,1})^2 + (x_2 - x_{o,2})^2 = r^2$, we obtain:

$$\sigma^*(x_o) = \frac{r}{\sqrt{3}}, \tag{4.24}$$

which relates the optimal scale for gradient assessment at x_o and the radius r of the circular structure centred at x_o. However, the gradient magnitude at x_o is actually null regardless of scale because the Gaussian derivative is an odd function which is antisymmetric about x_o (compare: Gaussian is an even function which is symmetric about x_o). As such, we modify the relationship between σ^* and r by proposing the following expression:

$$\sigma^*(x_\varepsilon) \approx \frac{r}{\sqrt{3}}, \tag{4.25}$$

where $x_\varepsilon = x_o + \varepsilon$ is slightly displaced from x_o for an extremely small displacement ε.

We demonstrate automatic scale selection in Fig. 4.1 with artificially generated structures. Fig. 4.1 (a) depicts a binary image I (201×201 pixels) composed of a single circular structure with radius $r = 9$ centered at $x_o = (100, 100)$. Convolving I with Gaussian derivaritive filters with 200 scales σ sampled uniformly from 0.1 to 20, we determine the scale space maximum of gradient magnitude $\sigma\|\nabla I(x;\sigma)\|$ (illustrated in Fig. 4.1 (b)) along with the corresponding scale σ^* (illustrated in Fig. 4.1 (c)). To identify the maximum, we check for zero crossings of the scale derivative $\frac{\partial}{\partial\sigma}\sigma\|\nabla I(x;\sigma)\|$. As the scale space evolution of $\sigma\|\nabla I(x;\sigma)\|$ may not be smooth[2] (as exemplified by the jagged scale derivative plotted in gray as illustrated in Fig. 4.1 (e)), we perform polynomial fitting prior to calculating $\frac{\partial}{\partial\sigma}\sigma\|\nabla I(x;\sigma)\|$.

[2]The lack of smoothness is due to the limited precision of numerical calculations as well as the limited spatial filter size of the Gaussian derivative (which theoretically should be infinite) used. Throughout our work, we limit the filter size to $(6\sigma + 1) \times (6\sigma + 1)$. Note that 6σ is chosen in accordance to the three sigma rule.

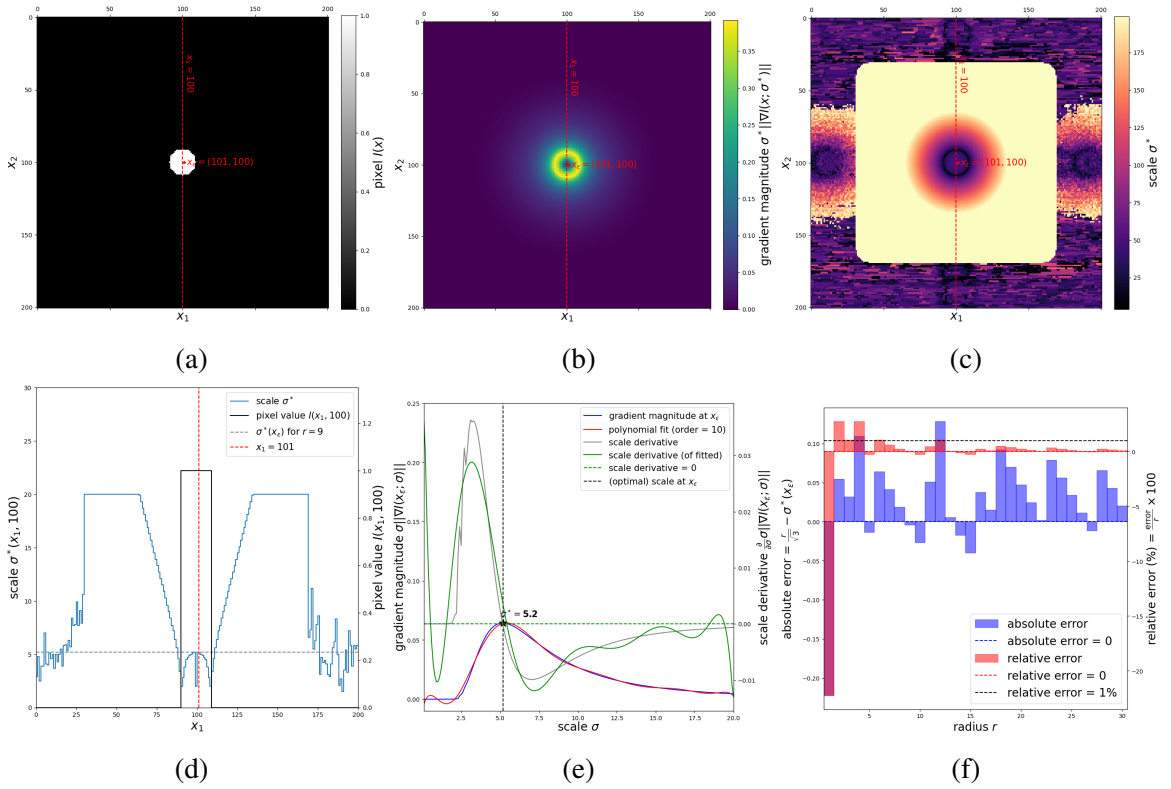


Fig. 4.1 (a) A binary image I (201×201 pixels) depicting a circular structure of radius $r = 9$ pixels centered at $x_o = (100, 100)$, (b) scale space maximum of gradient magnitude $\sigma^* \|\nabla I(x_\varepsilon; \sigma^*)\|$ at $x_\varepsilon = (100, 101)$, (c) scales σ^* of the scale space maximum, (d) scales σ^* at $x_2 = 100$ and $x_1 \in [0, 200]$, and (e) absolute and relative errors of measured scale $\sigma^*(x_\varepsilon)$ at with respect to calculated value based on (4.25)) for circular structures of radius $r \in [1, 30]$. Assessed with 200 scales sampled uniformly from 0.1 to 20.

Referring to Fig. 4.1 (c), it can be seen that σ^* is the lowest near the edge. This is because the area around the edge can be thought of being very small local "structure" (hence small σ^*). Accordingly, σ^* increases with distance from the edge because the local "structure" grows. However, it can be observed that σ^* stops increasing monotonically up to a certain distance from x_o. For example, at $x_2 = 100$, σ^* is fix at 20 for $x_1 \in [30, 64]$ and $x_1 \in [135, 169]$, before taking random values for $x_1 \in [0, 29]$ and $x_1 \in [170, 200]$ as shown in Fig. 4.1 (d). We note that such behaviour is not due to the structure but simply attributed to the fact that we only process scales σ up to 20. If more scales are processed, σ^* would have kept increasing.

The fact that the scale σ^* keeps increasing as it gets further from the edge is actually undesirable. This is because then, edges as detected using image gradient will be smeared (blurred; out of focus) as shown in Fig. 4.1 (b). However, we remind that the image as illustrated in Fig. 4.1 (a) is artificially generated with absolutely zero noise and blurring. In practice, however, acquired images are commonly subjected to noise, illumination variation, the limited spatial resolution of the sensor, etc. As a result, minimal spatial variation can be found in regions which visually appear isotropic. Therefore, the increase of σ^* as a function of distance from the edge will be suppressed as exemplified later in Fig. 4.6 which depicts automatic scale selection in the context of an acquired hyperspectral image.

As explained, the gradient magnitude at the center x_o of circular structure is null. As such, it is more meaningful to examine scale σ^* at x_ε which is minimally displaced from x_o. In the context of discrete image, the smallest possible displacement ε is given by $(1,0)$ or $(0,1)$ i.e. one pixel away in the horizontal or vertical directions in case of two spatial dimensions. As such, we resort to examining $\sigma^*(x_\varepsilon)$ at $x_\varepsilon = (101,100)$. Referring to Fig. 4.1 (e), it can be seen that the $\sigma^*(x_\varepsilon)$ is measured to be 5.2, which is very close to the theoretical value $\frac{r}{\sqrt{3}} = 5.196$ as calculated using (4.25). Considering circular structures of varying $r \in [1,30]$, it can be seen as that the relative error of $\sigma^*(x_\varepsilon)$ with respect to the theoretical value decreases with increasing r as shown in Fig. 4.1 (f). This is because the generated circular structure becomes more accurate (closer to a true circle) with increasing size (radius). Thanks to the low relative errors (mostly $\leq 1\%$), our derivation in (4.25) is thus confirmed.

Scale space stationary (SSS) points

Previously, we have accounted for scale selection concerning single structure based on scale space maximum of gradient magnitude $\sigma\|\nabla I(x;\sigma)\|$. However, the scale space evolution of $\sigma\|\nabla I(x;\sigma)\|$ can be more complex in the presence of other structures. For example, $\sigma\|\nabla I(x;\sigma)\|$ could have multiple maxima. Such scenario is akin to wave interference in

physics as two waves combine to form a greater, lower, or equivalent wave. We demonstrate this in Fig. 4.2 which depict three scenarios. Fig 4.2 (a) and (b) depict a circular structure centered at $(100, 100)$ and $(135, 100)$ with radius $r = 9$ and $r = 15$ respectively, whereas Fig 4.2 (c) shows the two aforementioned structures neighbouring each other. In all cases, we are interested in the scale space evolution of $\sigma\|\nabla I(x_\varepsilon;\sigma)\|$ at $x_\varepsilon = (101, 100)$. Interestingly, it can be seen that $\sigma\|\nabla I(x_\varepsilon;\sigma)\|$ in Fig. 4.2 (f) is like a summation between that in Fig. 4.2 (d) and (e). Whereas the scale selection in Fig. 4.2 (d) and (e) are unmistakenly based on the (only) maximum, it is not so clear for the case of Fig. 4.2 (f) as there are two local maxima. The question then arises: which scale to choose so as to reflect the local structure size at x_ε?

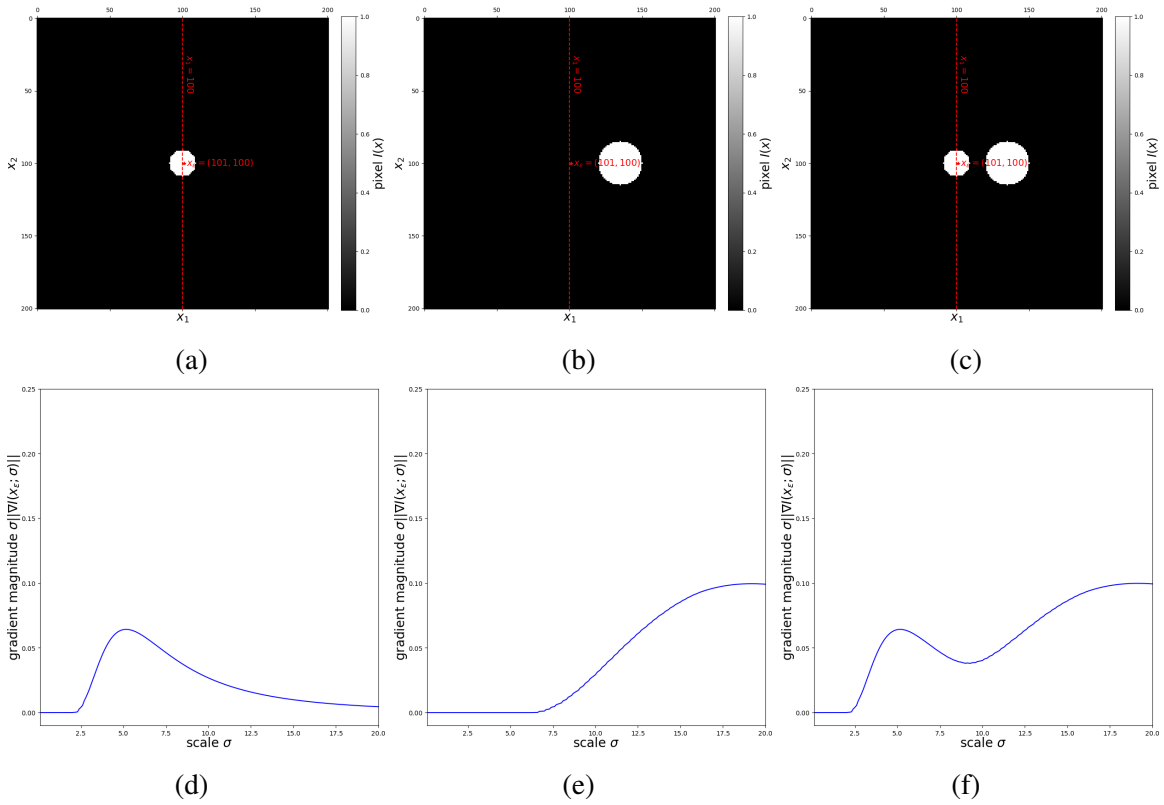


Fig. 4.2 Binary images (a) depicting a circular structure centered at $(100, 100)$ with radius $r = 9$; (b) depicting a circular structure centered at $(135, 100)$ with radius $r = 15$; (c) depicting two circular structures centered at $(100, 100)$ and $(135, 100)$ with radius $r = 9$ and $r = 15$ respectively. The corresponding scale space evolution of (scale-normalized) gradient magnitude at $x_\varepsilon = (101, 100)$ of (a), (b), (c) are depicted in (d), (e), and (f) respectively.

In this case, the heuristic as proposed by Lindeberg is clear: choose the local maximum as the optimal scale, which we interpret as selecting the first maximum. This is based on our assumption that in the absence of noise[3], only the first maximum refers to the local structure size, whereas the other maxima are influences by neighbouring structures. We demonstrate this in Fig. 4.3 which depicts pairs of the structure of different sizes and distances from each other. For brevity, let us denote the circular structure ($r = 9$) centered at $x_o = (100, 100)$ and the neighbouring circular structure as S1 and S2 respectively. In Fig. 4.3 (a), S2 ($r = 15$) is larger than S1, whereas in Fig. 4.3 (b), S2 ($r = 9$) is the same size as S1. By examining the scale space evolution of gradient magnitude $\sigma\|\nabla I(x_\varepsilon;\sigma)\|$ at x_ε, it can be seen that it exists two local maxima. It can be seen that in both cases as illustrated in Fig. 4.3 (d) and (e), the first maximum 5.2 corresponds to the local structure size (of S1) $\frac{r}{\sqrt{3}} = 5.196$. Clearly, 5.2 should thus be taken as the scale $\sigma^*(x_\varepsilon)$. However, having multiple structures does not necessarily result in multiple local maxima. In 4.3 (c), S2 ($r = 3$) is smaller than S1, and $\sigma\|\nabla I(x_\varepsilon;\sigma)\|$ has only one maximum as shown in Fig. 4.3 (f). This is because the gradient of S1 "overwhelms" that of S2 as S2 is smaller (and hence, has a weaker gradient) than S1.

[3]In the presence of noise; the first maximum would have been the "size" of noise which is in general, much smaller than any structures'. A solution would be to ignore scales below some threshold set based on the noise.

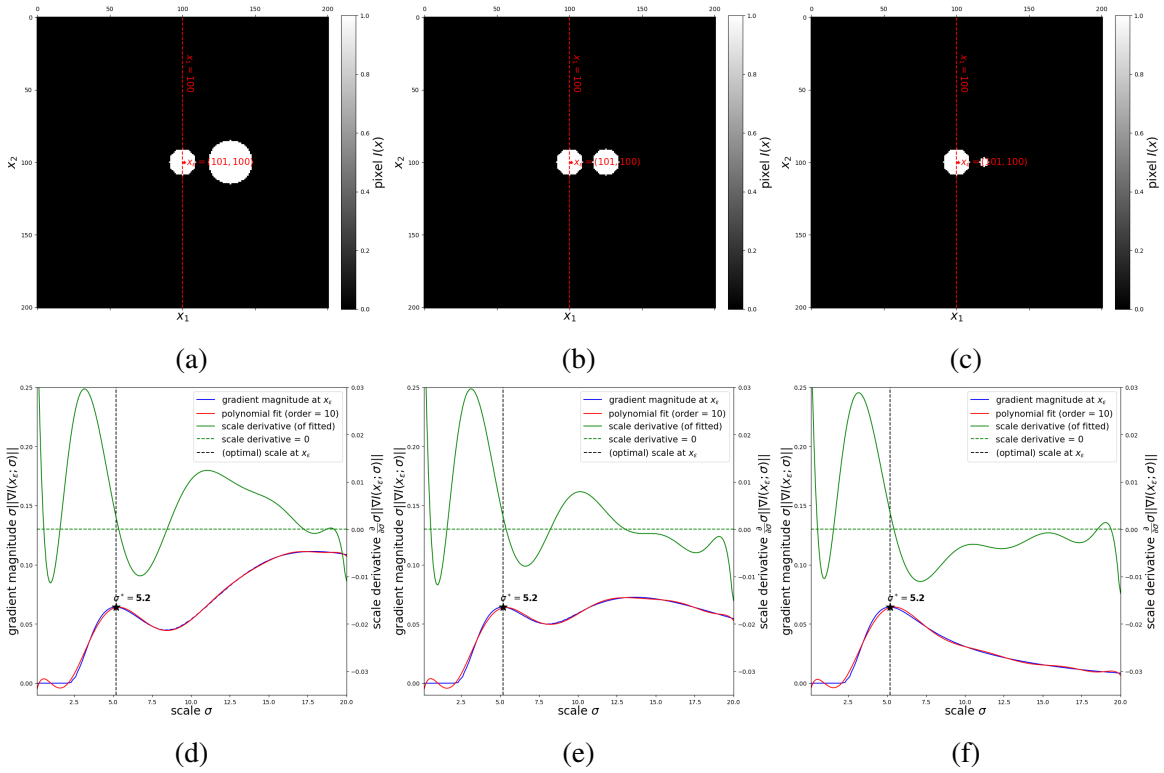


Fig. 4.3 Binary images depicting pairs of structure of different sizes and distances from each other. In all cases, the structure on the left is centered at $x_o = (100, 100)$ with radius $r = 9$, whereas the structure on the right is of radius (a) $r = 15$, (b) $r = 9$, and (c) $r = 3$. The corresponding scale space evolution of (scale-normalized) gradient magnitude at $x_\varepsilon = (101, 100)$ of (a), (b), (c) are depicted in (d), (e), and (f) respectively.

Another complexity arises in the case of neighbouring structures being too close. In this case, the maxima will also get closer and interact with each other. As a result, one of the maxima might disappear, as a rising point of inflection takes shape nearby. We demonstrate this in Fig. 4.4 wdepictsepict two circular structures of different sizes being placed increasingly close. In all cases, the smaller structure ($r = 9$) is centered at $(100, 100)$, whereas the larger structures ($r = 15$) is centered at $(135, 100)$, $(130, 100)$, and $(128, 100)$ in Fig. 4.4 (a), (b), and (c) respectively. Again, we are interested in the scale space evolution of $\sigma\|\nabla I(x_\varepsilon;\sigma)\|$ at $x_\varepsilon = (101, 100)$. Referring to Fig. 4.4 (d) and (e), it can be seen that the two maxima get closer as the distance between the two structures decreases before an inflection point forms as shown in Fig. 4.4 (f). If Lindeberg's heuristic were to be strictly

followed, the scale selection based on the inflection point would have been dismissed as it is not a local maximum. However, it is clear that the scale corresponding to the inflection point is more representative of the local structure size at x_ε. Naturally, a wondering surface: should we too consider the inflection point for scale selection apart from local maxima?

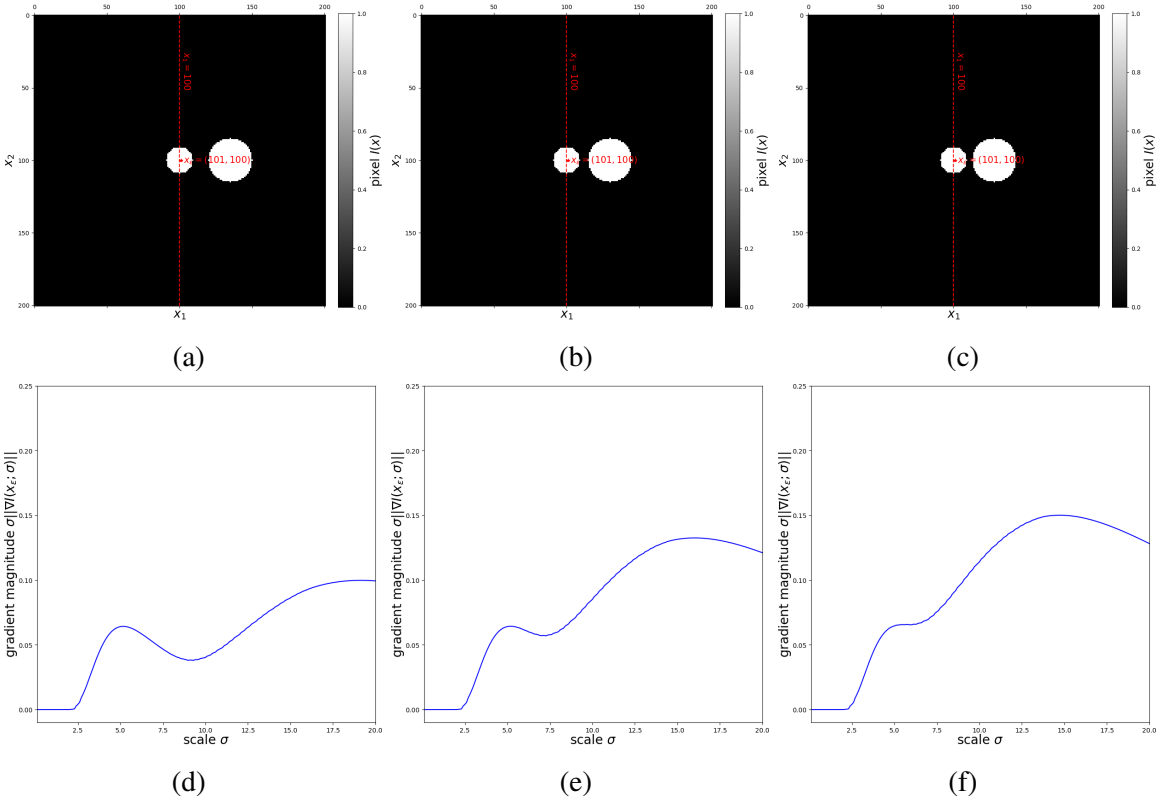


Fig. 4.4 Binary images depicting two circular structures of different sizes being placed increasingly close. In all cases, the structure on the left is centered at $x_o = (100, 100)$ with radius $r = 9$, whereas the structure on the right is centered at (a) $(135, 100)$, (b) $(130, 100)$, and (c) $(128, 100)$ with $r = 15$. The corresponding scale space evolution of (scale-normalized) gradient magnitude at $x_\varepsilon = (101, 100)$ of (a), (b), (c) are depicted in (d), (e), and (f) respectively.

Modifying Lindeberg's heuristic, we hereby propose to **perform automatic scale selection (in the absence of noise) based on the scale space stationary[4] (SSS) point i.e. the**

[4]In fact, stationary points include the extrema (minima and maxima) as well as the rising and falling inflection points. In our context, we are only referring to the maxima and rising inflection points.

first local maximum or (rising) inflection point (whichever comes first) of the scale space evolution of normalized gradient magnitude so as to reflect the local structure size:

$$\sigma^*(x) = \underset{\sigma}{\text{argstat}} \left(\sigma \|\nabla I(x;\sigma)\| \right), \tag{4.26}$$

where σ^* is the locally defined (optimal) scale and "argstat$(\cdot)$" refers to the operation that finds the argument i.e. σ that gives the SSS point. We demostrate this in Fig. 4.5 which shows pairs of structure of different sizes being placed very close (one pixel apart). In all cases, the smaller structure ($r = 9$) is centered at $(100, 100)$, whereas the neigbouring structures are centered at $(128, 100)$, $(120, 100)$, and $(114, 100)$ with $r = 15$, $r = 9$, and $r = 3$ in Fig. 4.5 (a), (b), and (c) respectively. Again, we are interested in the scale space evolution of $\sigma\|\nabla I(x_\varepsilon;\sigma)\|$ at $x_\varepsilon = (101, 100)$. To identify the SSS points, we check for zero crossings (for local maxima) or valleys which are above zero (for rising inflection point) of the scale derivative $\frac{\partial}{\partial\sigma}\sigma\|\nabla I(x_\varepsilon;\sigma)\|$. It can be seen that the scales selected based on the inflection point for Fig. 4.5 (d) ($\sigma^*(x_\varepsilon) = 5.8$) and (e) ($\sigma^*(x_\varepsilon) = 6.8$) are indeed closer to the true scale $\frac{r}{\sqrt{3}} = 5.2$ based on the local structure size ($r = 9$). If the maxima were to be chosen (14.8 and 10.9 in Fig. 4.5 (d) and (e) respectively), they would have highly overestimated the local structure size. In this case, the scales corresponding to the inflection points are clearly of better choice. Referring to Fig. 4.5 (f), no inflection point but a local maximum can be found. This is due to the stronger gradient of the larger structure which "overwhelms" that of the smaller structure. However, the effect of the smaller structure can still be observed: the local maximum has been slightly shifted to 5.4 in comparison to the true scale 5.2.

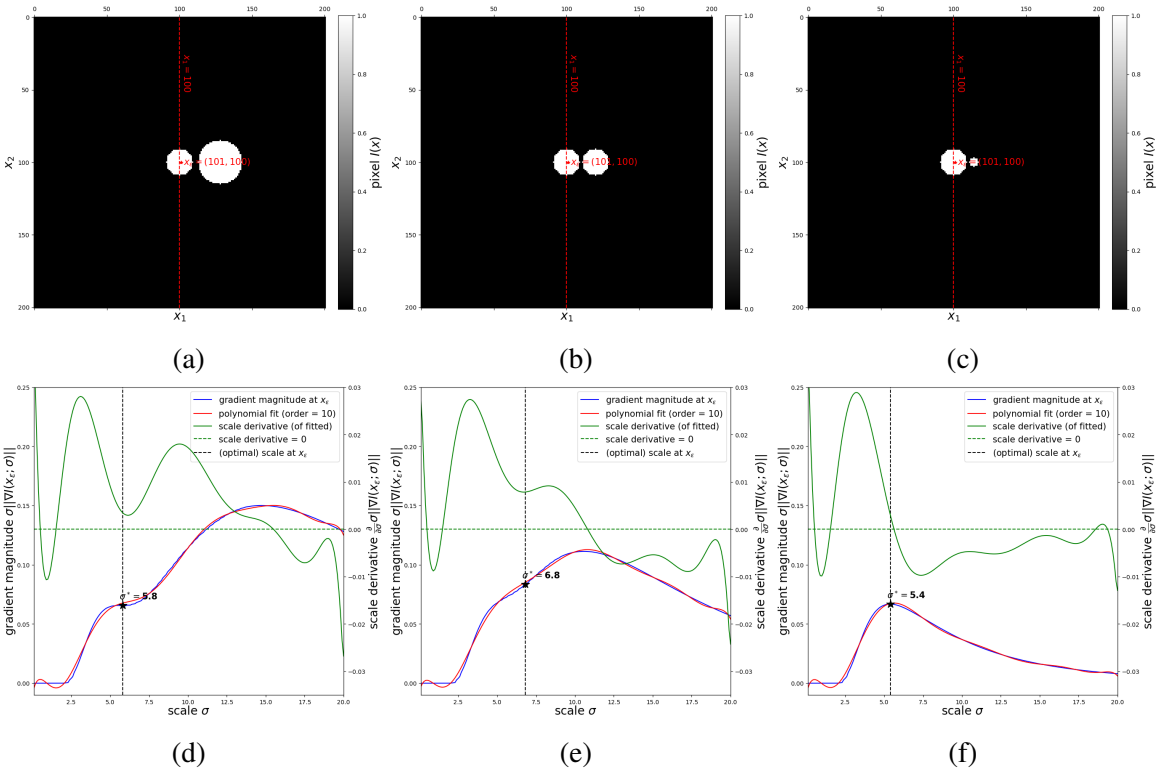


Fig. 4.5 Binary images depicting pairs of structure of different sizes being placed at one pixel apart. In all cases, the structure on the left is centered at $x_o = (100, 100)$ with radius $r = 9$, whereas the structure on the right is centered at (a) $(128, 100)$ with $r = 15$, (b) $(120, 100)$ with $r = 9$, and $(114, 100)$ with $r = 3$. The corresponding scale space evolution of (scale-normalized) gradient magnitude at $x_\varepsilon = (101, 100)$ of (a), (b), (c) are depicted in (d), (e), and (f) respectively.

Demonstration with hyperspectral image

Previously, we have considered scale selection in the context of monovariate, binary images for ease of explainability. In the following, we demonstrate the validity of the developed method with hyperspectral image in which the same rationale applies. For reasons that will be apparent later[5], we determine the (optimal) scale σ^* based on the SSS point of $\sqrt{\lambda_+(\sigma)}$:

$$\sigma^* = \underset{\sigma}{\text{argstat}} \left(\sigma \sqrt{\lambda_+(\sigma)} \right), \tag{4.27}$$

[5]Later, we shall show that $\sqrt{\lambda_+}$ is the spectral gradient magnitude.

where $\sqrt{\lambda_+}$ is the larger eigenvalue of the spatial covariance matrix M as defined in (4.13).

We first consider a small patch as shown (rendered in RGB) in Fig. 4.6 (a) from a hyperspectral image with wavelengths ranging from 420 nm to 720 nm at 10 nm intervals. Specifically, we refer to the seven pixels (c) - (i) as bordered and labelled in black. They are chosen such that those labelled by (c) - (e), (f) - (h), and (i) are representatives of edge, flat (neither edge nor corner), and corner pixels respectively. Referring to Fig. 4.6 (c) - (h), it can be seen that scale of the edge pixels ($\sigma^* \in [3.8, 4.8]$) are considerably higher than that of the flat pixels ($\sigma^* \in [0.6, 0.8]$). This seems to be opposite to the case of Fig. 4.1 which shows that the edge pixels have considerably lower σ^* than the flat pixels. However, we note that Fig. 4.1 (a) on which the gradients are calculated is artificially generated with absolutely zero noise and blurring. Real images, however, as in the case oimagessge from which pixels (c) - (i) are taken, are subjected to noise, sensor limitations, etc. As a result, minimal spatial variation can be found in regions which visually appear isotropic. This explains the case for the flat pixels (f) - (h) as their σ^* actually refers to spatial variation due to noise. This, however, is a "blessing" as it avoids the smearing of gradient magnitude (as the case in Fig. 4.1 (b)) at the edge. On the other hand, we note that the edge of a real image is unlikely to be as sharp as artificially generated ones. Due to limited spatial resolution, the edge will be manifested more like a smoothed step function. This is clearly the case for the edge pixels (c) - (e) as their σ^* is influenced by the level of edge diffusion. Indeed, a higher scale is required for diffedgesedge compared to sharp ones to improve the signal (gradient magnitude) to noise ratio. On the other hand, it can be seen that $\sigma^* = 8.7$ of the corner pixel is considerably higher than that of edge and flat pixels as depicted in Fig. 4.6 (i). We note that this is due to the low spectral gradient coherence which will be explained further in the coming subsections.

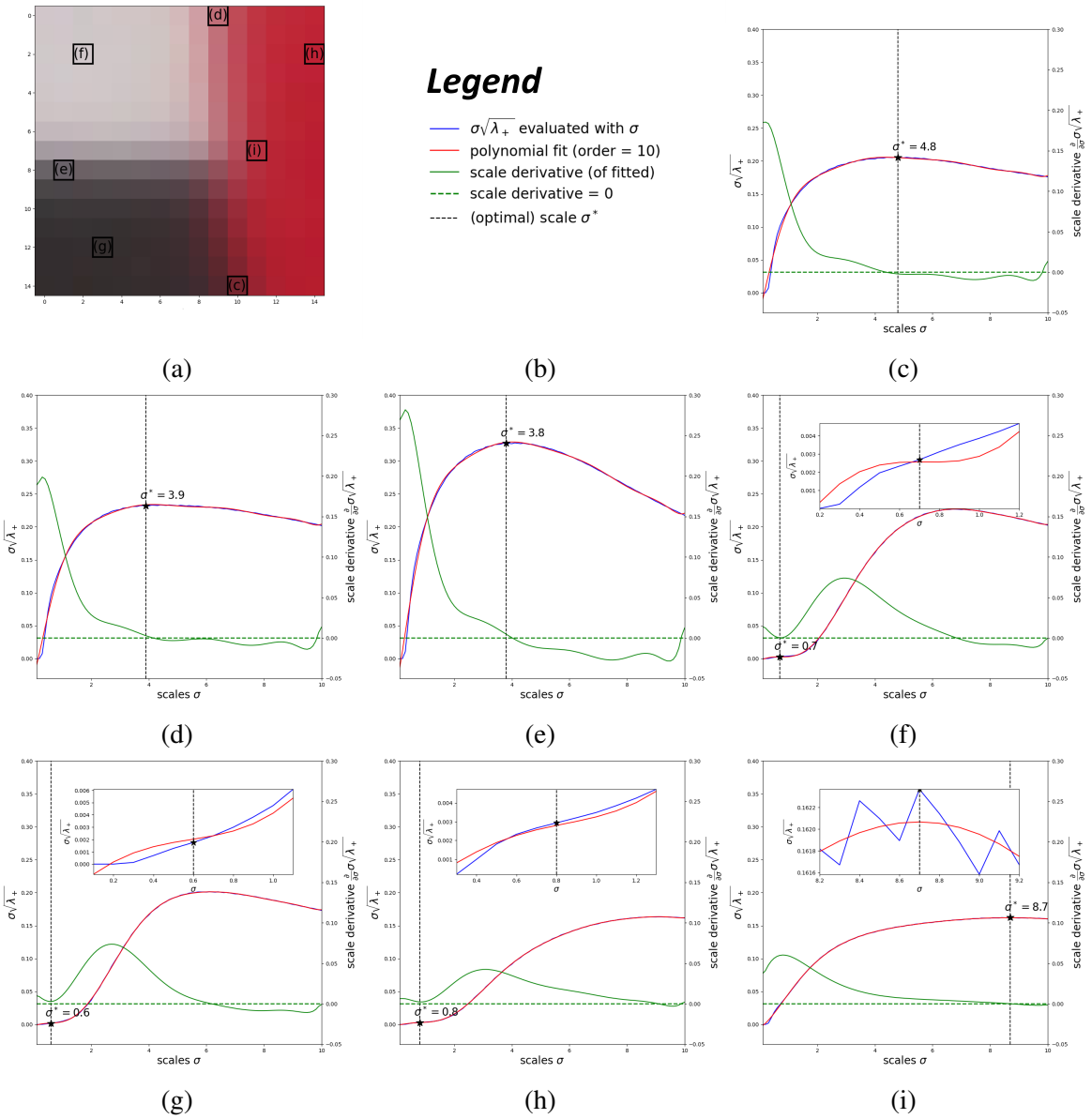


Fig. 4.6 Scale evolution of FVG magnitude $\sigma\sqrt{\lambda_+}$ for the selected seven pixels in patch 3 from *Character*. Following the convention in Fig. 4.1, the solid plots in blue, red, green are the FVG magnitude, FVG magnitude fitted using polynomials, and scale derivative (of the FVG magnitude fitted using polynomials) $\frac{\partial}{\partial\sigma}\sigma\sqrt{\lambda_+}$ respectively. The horizontal green dotted line is $\frac{\partial}{\partial\sigma}\sigma\sqrt{\lambda_+} = 0$, whereas the vertical black dotted line is the FVG scale $\sigma = \sigma^*$.

Next, we examine the hyperspectral image *Chart3* (512 $\times$ 512 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval [54] as shown in Fig. 4.7. It can be seen that the scales σ^* do reflect the local scale of variation. For example, σ^* of patch 1 is lower

than that of patch 2 because the former has smaller (block-like) structures. Clearly, σ^* is not affected by spatial directionality as σ^* is the same for patches 3 - 5 which have the same scale of variation. The same can be observed for patch 6 as σ^* is independent of the local edge directions. In the case of "flat" regions e.g. patches 7 and 8, σ^* refers to the scale of spatial variation due to noise which is in general small (hence σ^* is also low).

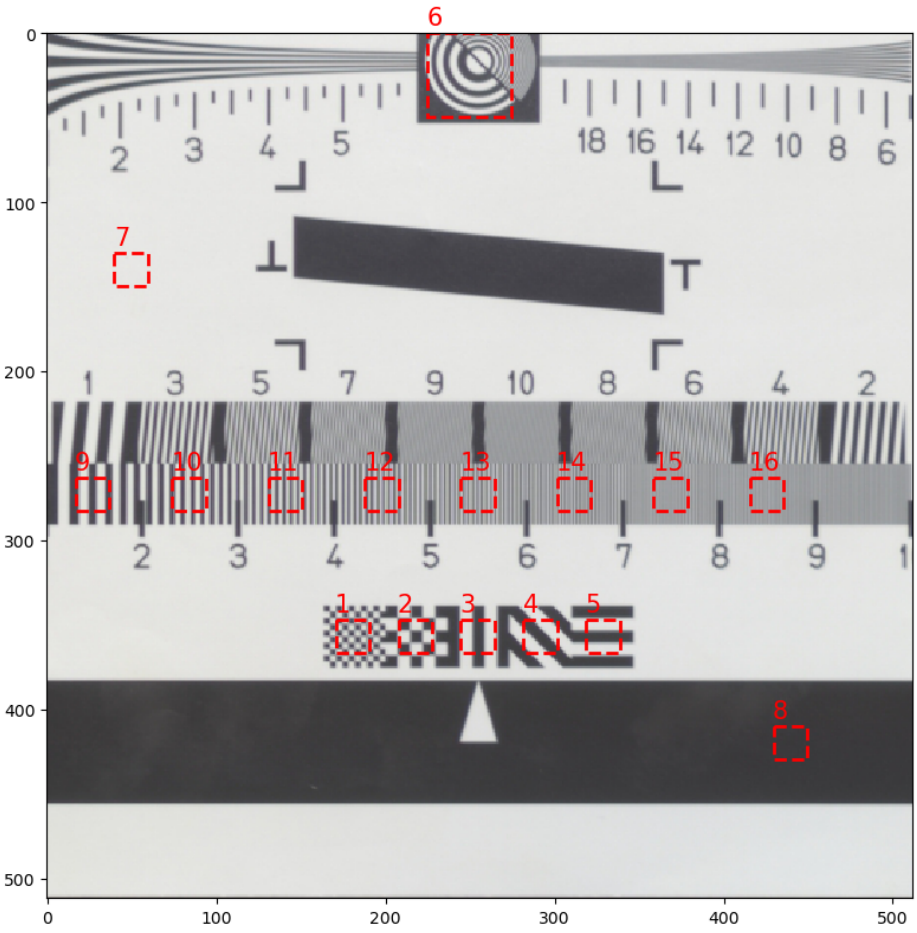


Fig. 4.7 The hyperspectral image *Chart3* (512 $\times$ 512 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval, rendered in RGB for illustration. Bordered and numbered in red are the 16 patches partly considered for FVG calculation in Fig. 4.9.

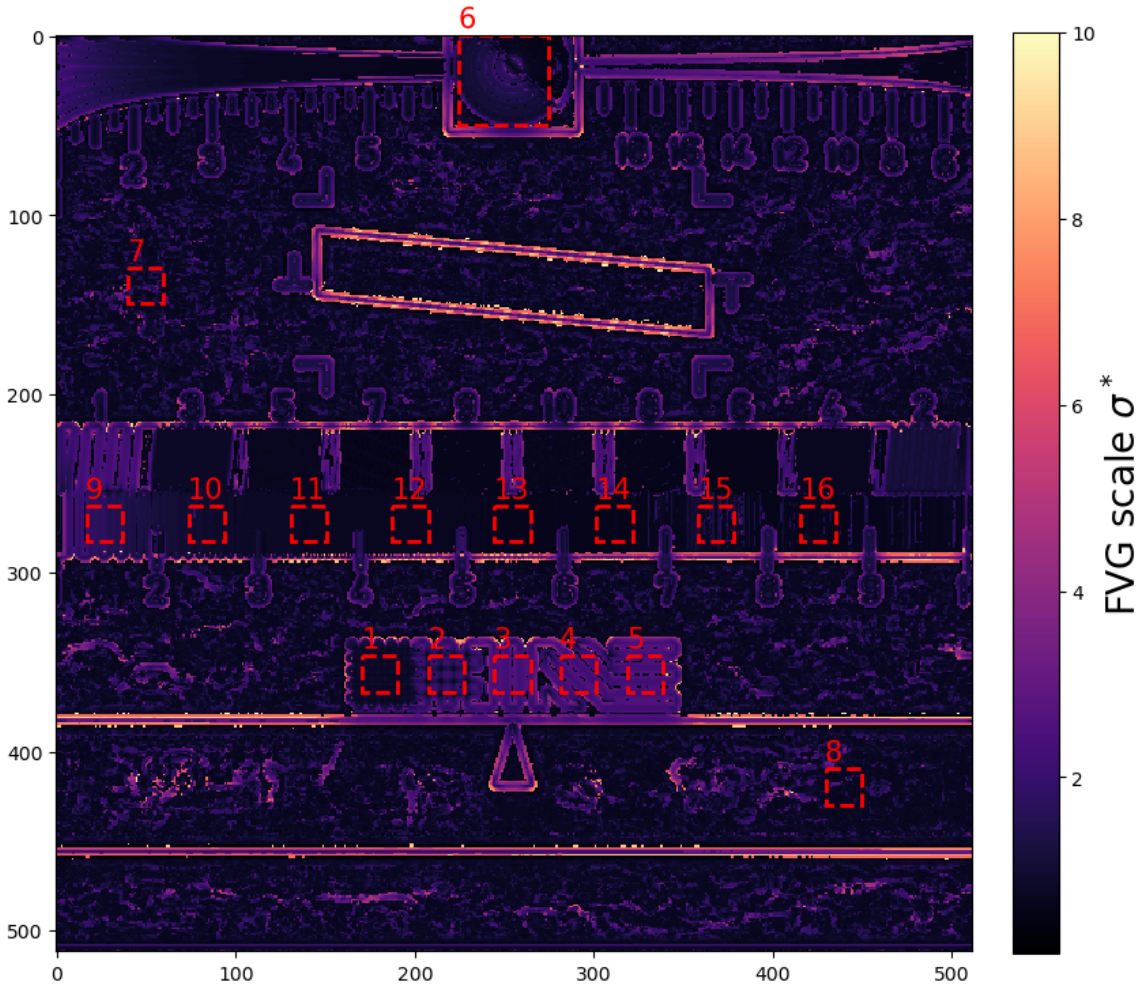
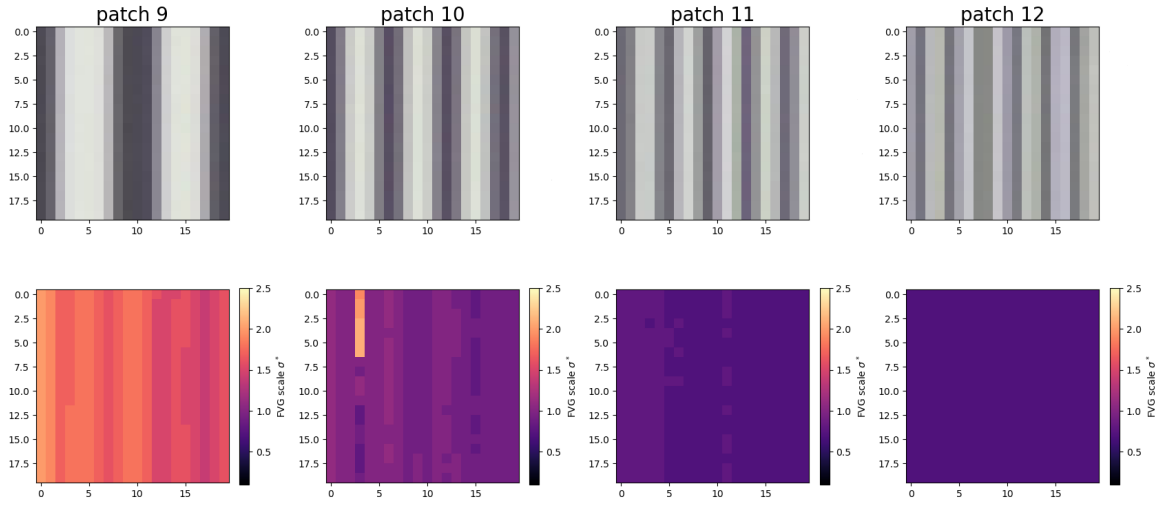


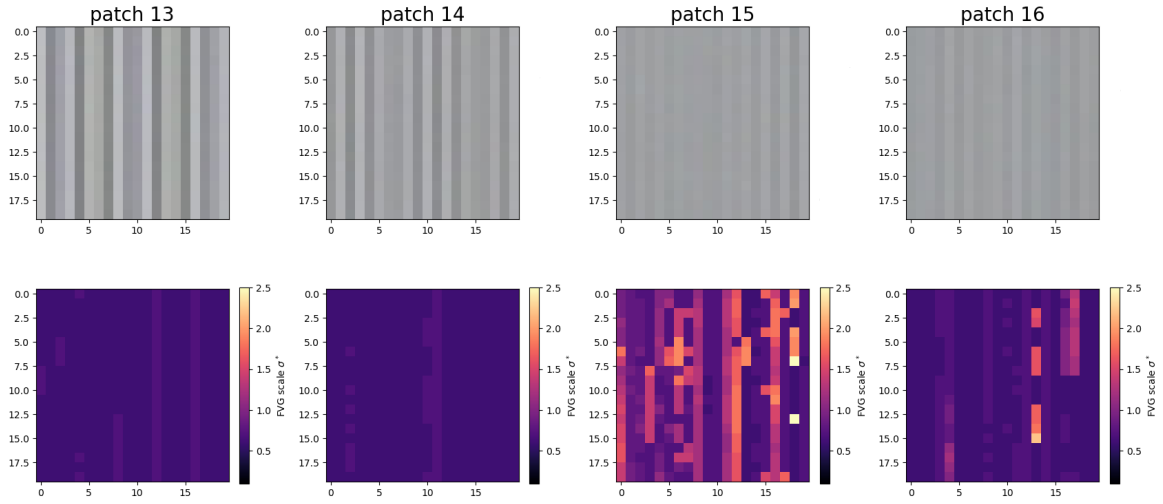
Fig. 4.8 FVG scale σ^* of *Chart3*.

Referring to Fig. 4.9, it can be observed that in general, the scale σ^* decreases from patches 9 - 14 in accordance with their decreasing scale of spatial variation. Note that to improve visibility, we have capped the maximum σ^* at 2.5 which corresponds to the maximum value of σ^* among patches 9 - 16. However, it can be seen that σ^* actually increases in patch 15 (and less so in patch 16) when it is expected to decrease further as their scale of spatial variation is even smaller. We attribute this to much-reduced contrast in patches 15 and 16 due to the limited spatial resolution which results in σ^* overestimation. Such argument is supported by fact that σ^* is observed to be irregularly distributed in patches 15 and 16 which have a regulapatternsrn. Meanwhile, referring to Fig. 4.10 which is taken

from the top left corner of *Chart3*, it can be seen that the scale σ^* varies continuously from right to left as the strips deviate from each other. Note that to improve visibility, we have capped the maximum σ^* at 5. Indeed, the local spatial variation can be said to decrease as the distance between the strips grows. This further validates our method for automatic scale selection.



(a)



(b)

Fig. 4.9 The patches (20 × 20 pixels) rendered in RGB (top row), followed by the FVG scale σ^* (bottom row) for (a) patches 9 - 12; and (b) patches 13 - 16 of *Chart3*.

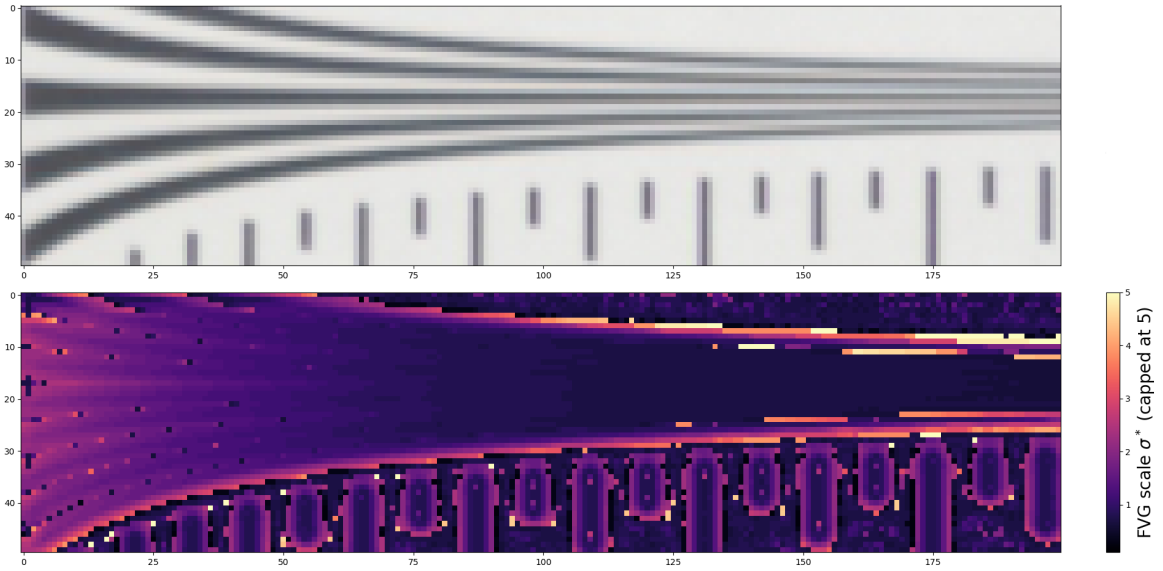


Fig. 4.10 The patch as taken from the top left corner of *Chart3* rendered in RGB (top row), followed by the FVG scale σ^* (bottom row). To improve visibility, the maximum σ^* is capped at 5.

4.2.2 Spectral gradient magnitude, direction, and coherence

Having developed automatic scale selection, we are now ready to calculate the gradient for hyperspectral images. As detailed in t previous section, the expression of image gradient $\nabla I(x)$ is generalized by that of Jacobian matrix $\mathrm{J}I(x)the$ for the multivariate image. In the following, we shall determine the measures for spectral gradient magnitude and direction before exploiting them for texture characterization. On the other hand, we also shall explore the notion of spectral gradient coherence which has no correspondence in monovariate images.

Spectral gradient magnitude and direction

In the literature, it exists several definitions of spectral gradient magnitude based on the eigenvalues $\lambda_{\pm}$ of the spatial covariance matrix $M = \mathrm{J}I(x)^T G \mathrm{J}I(x)$ as expressed in (4.13):

$$\lambda_+ = \frac{1}{2}\left(\mathrm{tr}(M) \pm \sqrt{\mathrm{tr}(M)^2 - 4|M|}\right), \tag{4.28}$$

recalling that G is the Gram matrix. In particular, Di Zenzo designated $\sqrt{\lambda_+}$ as the "maximum rate of change" or "edge strength" [89], an approach also echoed by Cumani [17]. On the other hand, Sapiro argued for measures of $\lambda_+ - \lambda_-$ as edge "strength" for image contour extraction [65], whereas Chatoux et. al. favored Frobenius norm $\lambda_+ + \lambda_-$ of the Jacobian as the "gradient magnitude" [10]. In the following, we shall clarify the meanings of $\lambda_\pm$ before arriving at a suitable choice of spectral gradient magnitude and thus direction.

To recall, $\sqrt{\lambda_\pm}$ are the extrema of the directional derivative $\mathrm{J}I(x)v(\theta)$'s Euclidean norm concerning a unit vector $v(\theta)$ pointing in the direction θ on $\mathbb{R}^2$ (the spatial dimensions):

$$\begin{aligned}
\sqrt{\lambda_+} &= \max_\theta \|\mathrm{J}I(x;\sigma)v(\theta)\|_2 = \|\mathrm{J}I(x;\sigma)v(\theta_+)\|_2, \\
\sqrt{\lambda_-} &= \min_\theta \|\mathrm{J}I(x;\sigma)v(\theta)\|_2 = \|\mathrm{J}I(x;\sigma)v(\theta_-)\|_2,
\end{aligned} \tag{4.29}$$

where σ is the standard deviation of the Gaussian derivative by which the image derivatives are calculated; θ_+ and θ_- are the (spatial) directions corresponding to the extrema of $\|\mathrm{J}I(x;\sigma)v(\theta)\|_2$ as defined in (4.14). Geometrically, $\|\mathrm{J}I(x;\sigma)v(\theta)\|_2$ is thus an ellipse with $\sqrt{\lambda_+}$, $\sqrt{\lambda_-}$, and θ_+ as its semi-major axis, semi-minor axis, and rotation angle respectively.

Hereafter, we shall denote measures evaluated with scale σ^* with an asterisk *. For example, $\sqrt{\lambda_+^*} = \sqrt{\lambda_+(\sigma^*)}$ is the larger eigenvalue as evaluated with σ^*, whereas θ_+^* is the corresponding direction. For reasons that will be apparent later[6], σ^* is chosen by:

$$\sigma^* = \operatorname*{argstat}_\sigma \left(\sigma\sqrt{\lambda_+(\sigma)}\right), \tag{4.30}$$

[6]In previous subsection, we propose to perform scale selection based on (scale-normalized) gradient magnitude for monovariate image. Later, we shall show that $\sqrt{\lambda_+}$ is the spectral gradient magnitude.

where argstat$(\cdot)$ is the operation which finds the scale space stationary (SSS) point of $\overset{\sigma}{\sigma\sqrt{\lambda_+(\sigma)}}$ as detailed in previous subsection.

In the following, we shall consider a small patch as shown (rendered in RGB) in Fig. 4.11 (a) from a hyperspectral image with wavelengths ranging from 420 nm to 720 nm at 10 nm interval. Specifically, we refer to the seven pixels (c) - (i) as bordered and labelled in black. They are chosen such that those labelled by (c) - (e), (f) - (h), and (i) are representatives of edge, flat (neither edge nor corner), and corner pixels respectively. As gradient analysis at a given pixel location is impossible without referring to its neighbouring pixels, we consider each of the seven pixels (c) - (i) within a 25×25 neighbourhood as shown in Fig. 4.11 (c) - (i). For clarity, the seven pixels (as located at the center of the patches) are buldged like a big dot. For visibility, we overlay the directional derivative's Euclidean norm $\|\mathrm{J}I(x;\sigma^*)v(\theta)\|_2$ (with axes $\|\mathrm{J}I(x;\sigma^*)\cos\theta\|_2$ and $\|\mathrm{J}I(x;\sigma^*)\sin\theta\|_2$) on top of the patches (with axes x_1 and x_2). Referring to Fig. 4.11 (c) - (e), it is clear that $\|\mathrm{J}I(x;\sigma^*)v(\theta)\|_2$ are rotated in the direction of the highest rate of pixel changes. However, it is arguably less clear in cases of Fig. 4.11 (d) - (i) as $\|\mathrm{J}I(x;\sigma^*)v(\theta)\|_2$ have lower eccentricity. Referring to Fig. 4.11 (i), $\|\mathrm{J}I(x;\sigma^*)v(\theta)\|_2$ even appears like a circle which suggests similar rate of pixel change in all directions.

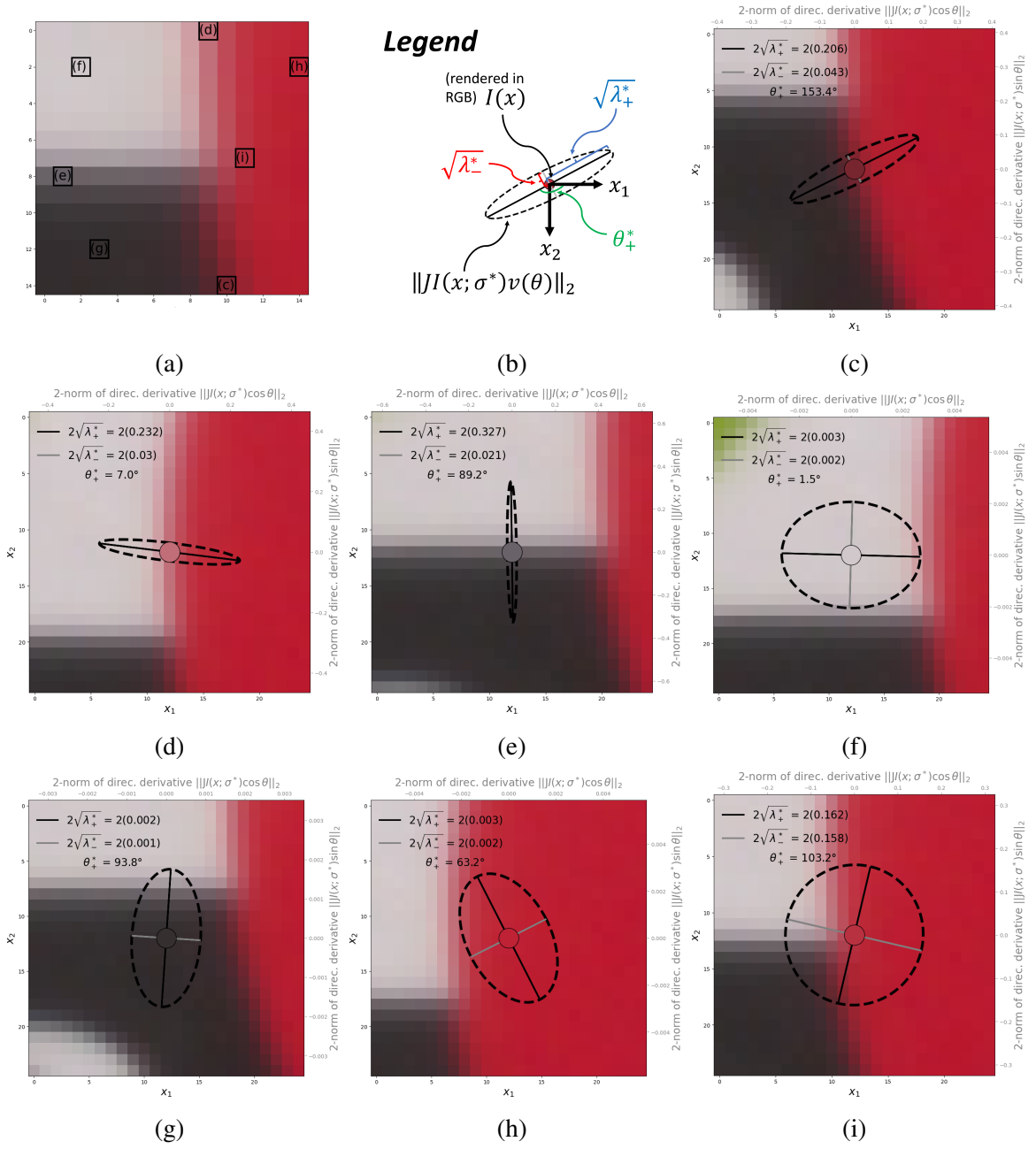


Fig. 4.11 (a) Bordered and labelled in black are the seven pixels that will be considered for the assessment of directional derivative's Euclidean norm $\|JI(x;\sigma^*)v(\theta)\|_2$ as illustrated in (c) - (i). The corresponding legend can be found in (b).

To understand the behaviour of the directional derivative's Euclidean norm $\|JI(x;\sigma^*)v(\theta)\|_2$, let us examine the marginal (band by band) gradients $\nabla I_l(x;\sigma^*)$ as defined by:

$$\nabla I_l(x;\sigma^*) = \left[\frac{\partial I_l(x;\sigma^*)}{\partial x_1} \quad \frac{\partial I_l(x;\sigma^*)}{\partial x_2}\right]^T \quad \forall \ l \in [1,L], \tag{4.31}$$

where I_l is the l^{th} spectral channel and L is the number of spectral bands. The marginal gradient magnitudes $\|\nabla I_l(x;\sigma^*)\|_2$ and directions θ_l^* are given by:

$$\begin{aligned} \|\nabla I_l(x;\sigma^*)\|_2 &= \sqrt{\left(\frac{\partial I_l(x;\sigma^*)}{\partial x_1}\right)^2 + \left(\frac{\partial I_l(x;\sigma^*)}{\partial x_2}\right)^2}, \\ \theta_l^* &= \arctan\left(\frac{\partial I_l(x;\sigma^*)}{\partial x_2} \Big/ \frac{\partial I_l(x;\sigma^*)}{\partial x_1}\right), \end{aligned} \tag{4.32}$$

We illustrate the marginal gradients $\nabla I_l(x;\sigma^*)$ (for 420 nm - 720 nm; 31 spectral channels in total) in Fig. 4.12 - 4.18 which correspond to pixels (c) - (i) respectively. For visibility, we overlay $\nabla I_l(x;\sigma^*)$ (with axes $\frac{\partial I_l(x;\sigma^*)}{\partial x_1}$ and $\frac{\partial I_l(x;\sigma^*)}{\partial x_2}$) on top of the patches (with axes x_1 and x_2). Referring to Fig. 4.12 - 4.18, it can be seen that the spatial organization of pixels varies with different wavelengths. Specifically, the red pixels are visible only from 590 nm to 720 nm, whereas the white pixels are visible with all wavelengths i.e. from 420 nm to 720 nm. In contrast, the black pixels are invisible with all wavelengths. As a result, $\nabla I_l(x;\sigma^*)$ also vary as they point to the direction of highest variation which changes with the wavelengths.

We begin by examining the edge pixels (c) - (e) as dictated in Fig. 4.11 (a). Referring to Fig. 4.12, it can be seen that the marginal gradient directions vary with $\theta_l^* \in [49.1°, 76.9°]$ from 420 nm to 580 nm, and $\theta_l^* \in [122.2°, 154.5°]$ from 590 nm to 720 nm. However, the marginal gradient magnitudes $\|\nabla I_l(x;\sigma^*)\|_2$ from 420 nm to 580 nm are much lower than that from 590 nm to 720 nm, and as a result the latter dominates. Consequently, the (overall) directionality is strong with $\theta_+^* = 153.4°$ as there is little opposing gradients. The same can also be observed in Fig. 4.13 as $\theta_l^* \in [6.1°, 7.1°]$ from 420 nm to 580 nm, $\theta_l^* \in [9.1°, 52.2°]$ from 590 nm to 620 nm, and $\theta_l^* \in [88.3°, 113.4°]$ from 630 nm to 720 nm. As the marginal

gradients $\nabla I_l(x;\sigma^*)$ from 420 nm to 580 nm have much higher magnitudes, they thus dominate. As a result, the directionality is also strong with $\theta_+^* = 7.0°$. Next, Fig. 4.14 shows that θ_l^* are coherent with all wavelengths. This is due to the similar spatial organization of pixels with all wavelengths: the black pixels are consistently absent whereas the white pixels are always present. Thanks to the high coherence (little variation) of $\theta_+^* \in [85.1°, 93.8°]$, the directionality is even stronger with $\theta_+^* = 89.2°$. Directionality is however, weak for the flat and corner pixels (f) - (i) as shown in Fig. 4.15 - 4.18. Pointing in opposing directions, $\nabla I_l(x;\sigma^*)$ have similar magnitudes. This results in overall, similar rate of pixel change in all directions especially in case of pixel (i). As such, θ_+^* of pixels (f) - (i) should be taken lightly.

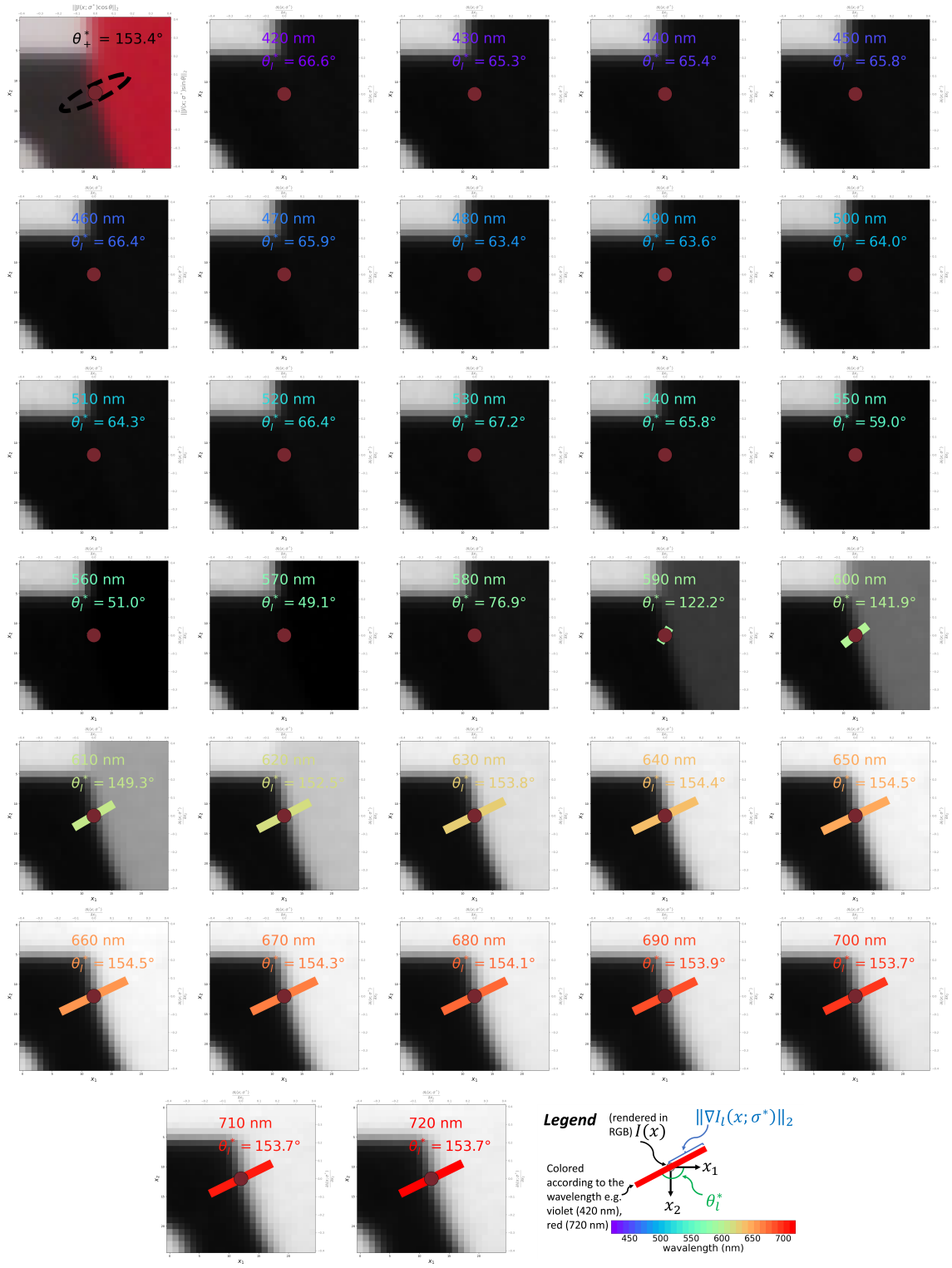


Fig. 4.12 The marginal gradients for the pixel labelled as (c) in Fig. 4.11 (a).

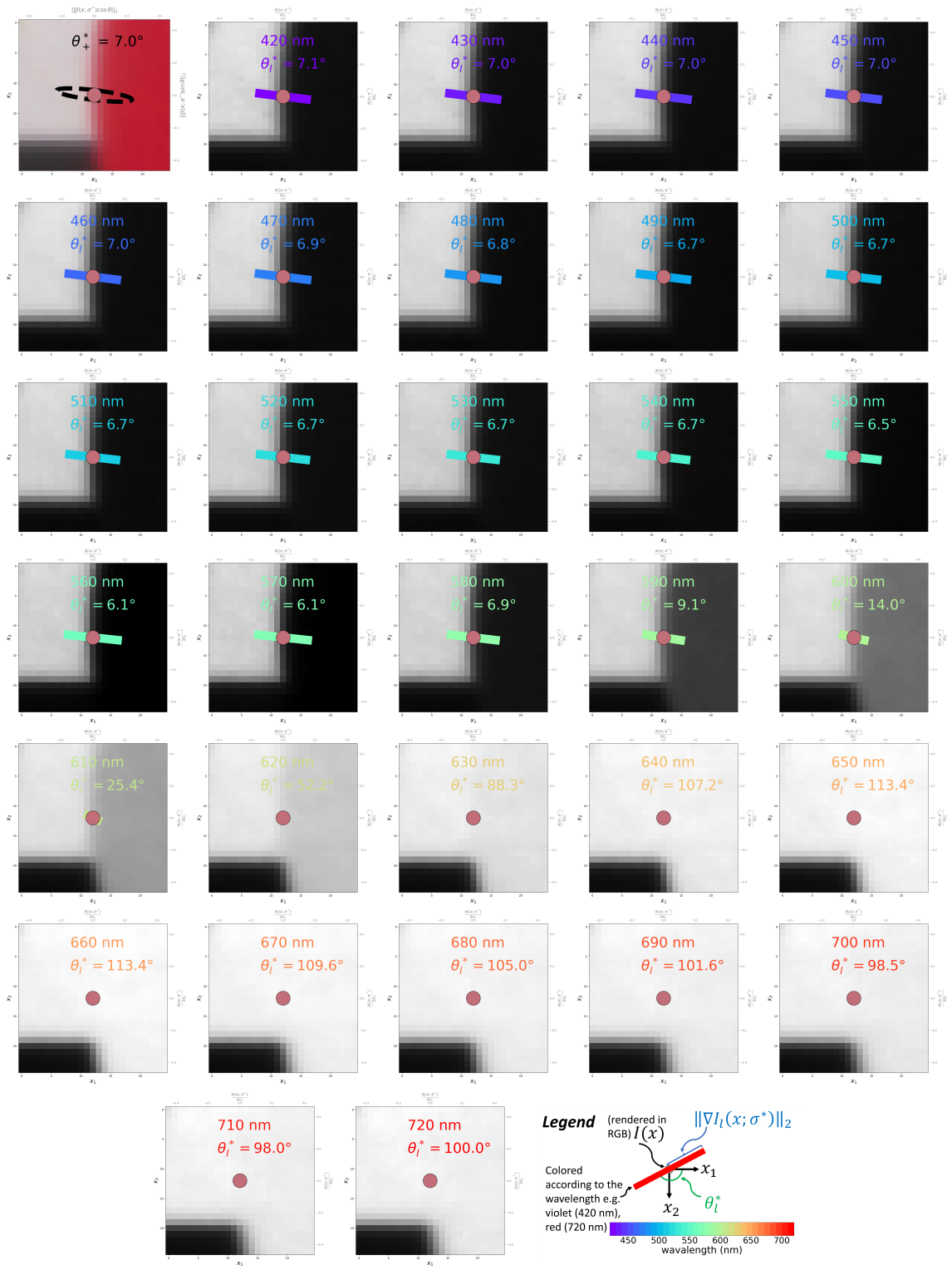


Fig. 4.13 The marginal gradients for the pixel labelled as (d) in Fig. 4.11 (a).

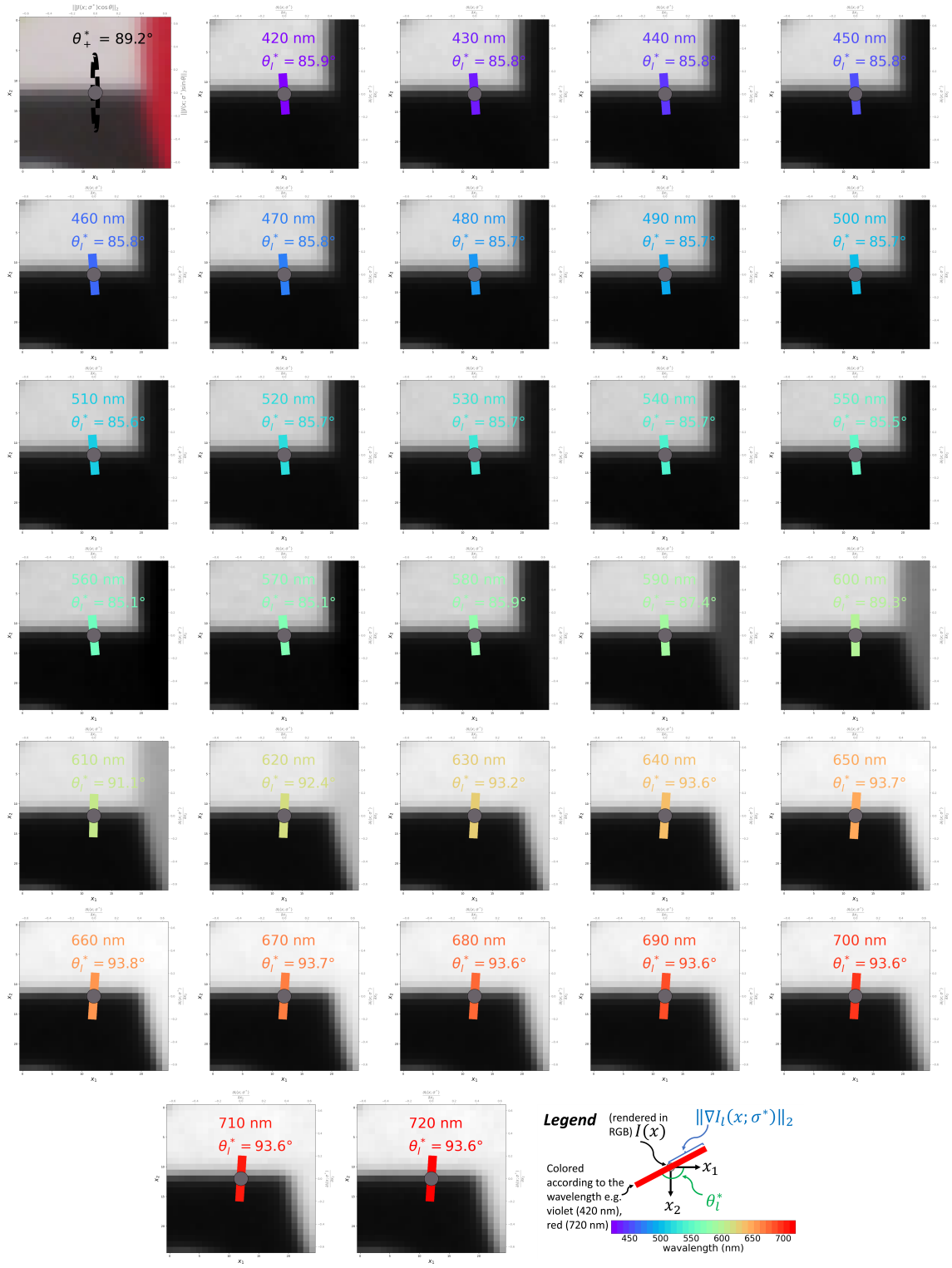


Fig. 4.14 The marginal gradients for the pixel labelled as (e) in Fig. 4.11 (a).

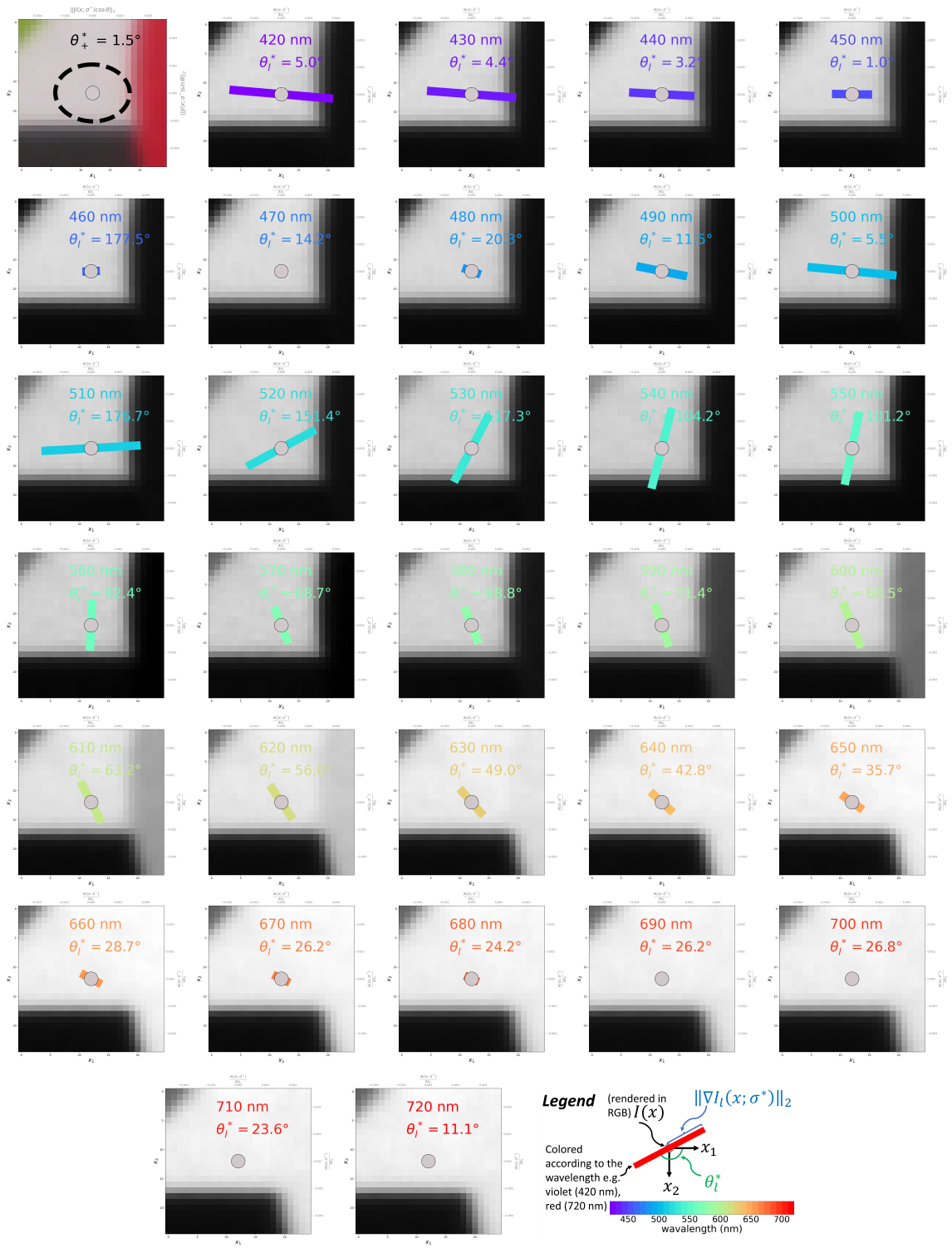


Fig. 4.15 The marginal gradients for the pixel labelled as (f) in Fig. 4.11 (a).

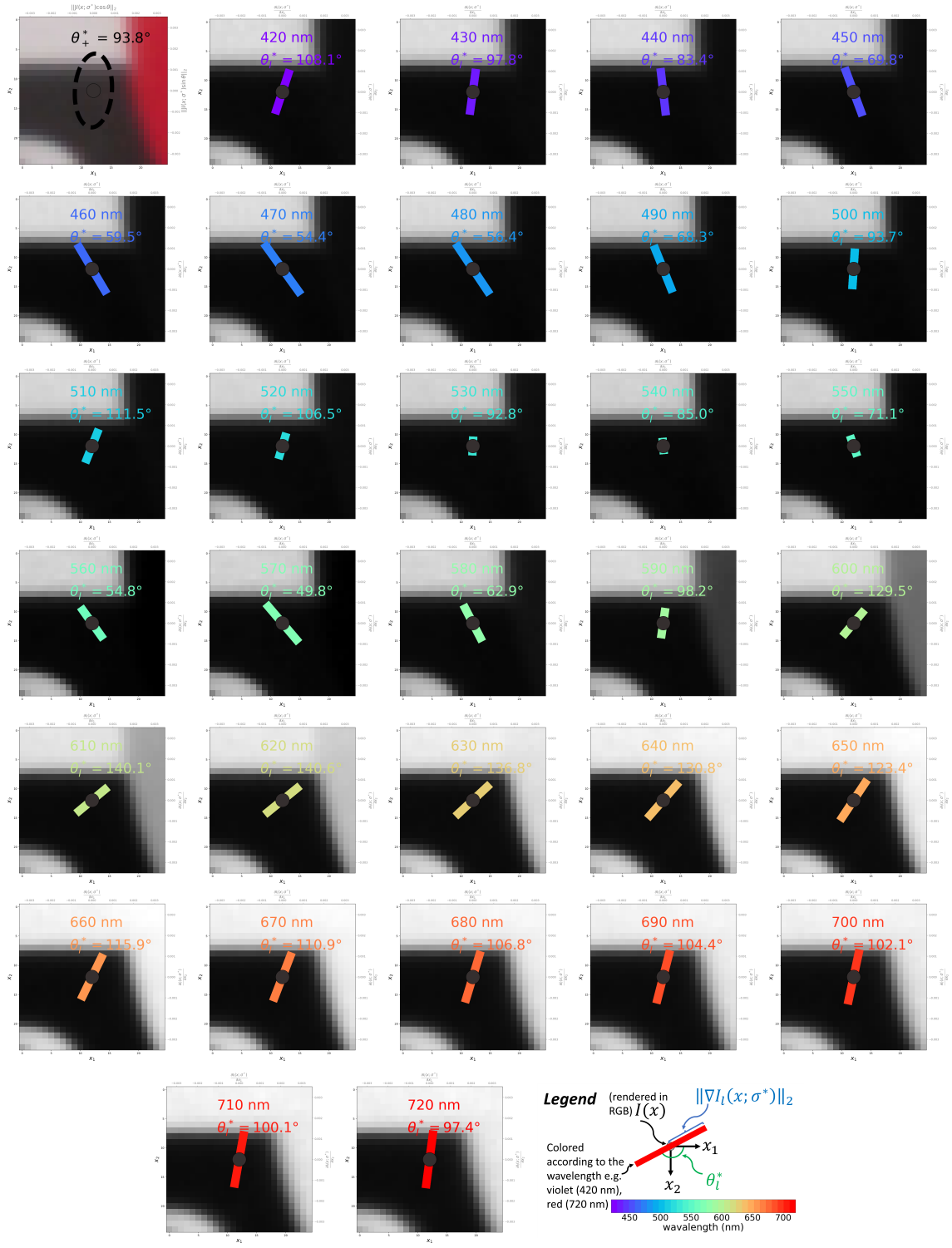


Fig. 4.16 The marginal gradients for the pixel labelled as (g) in Fig. 4.11 (a).

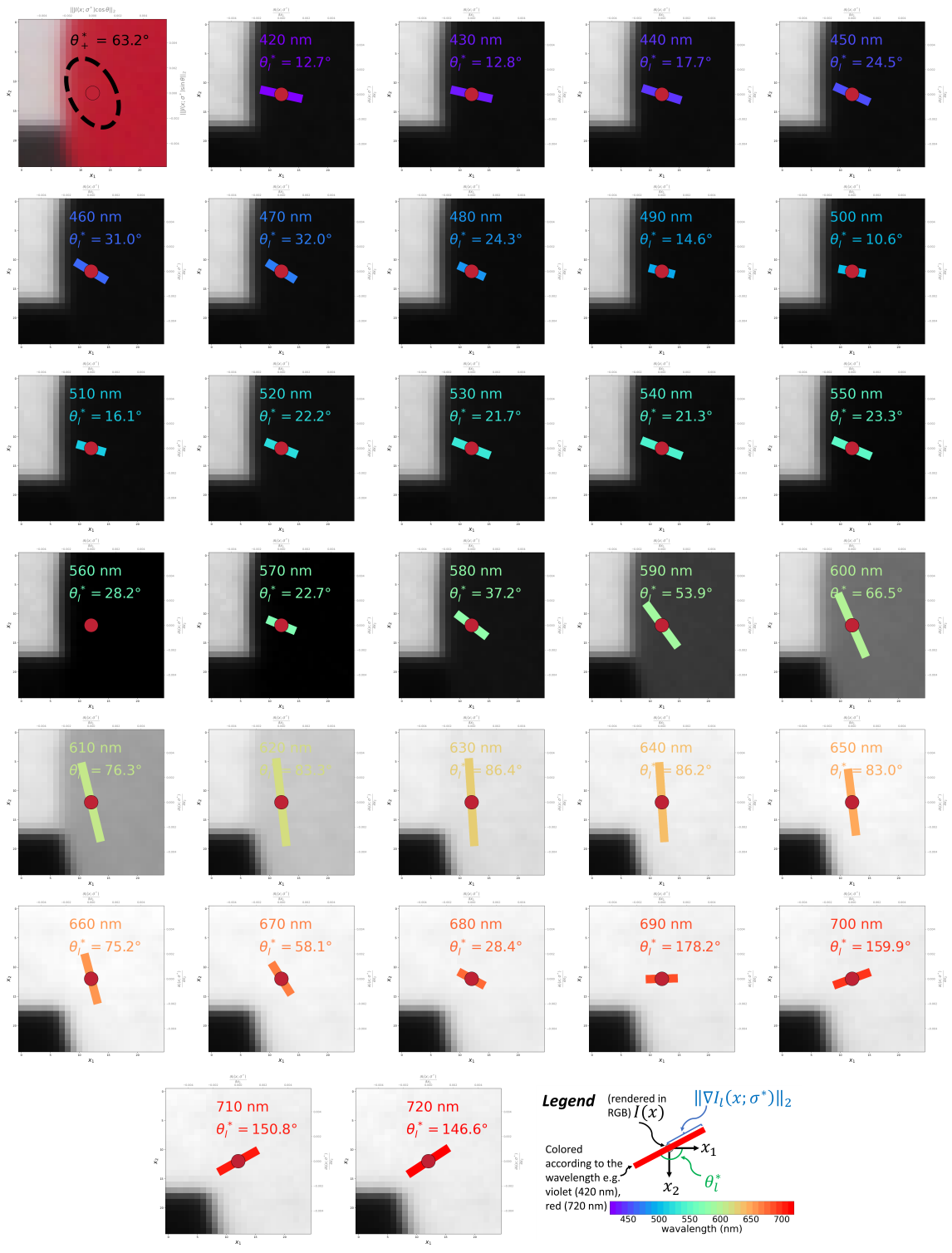


Fig. 4.17 The marginal gradients for the pixel labelled as (h) in Fig. 4.11 (a).

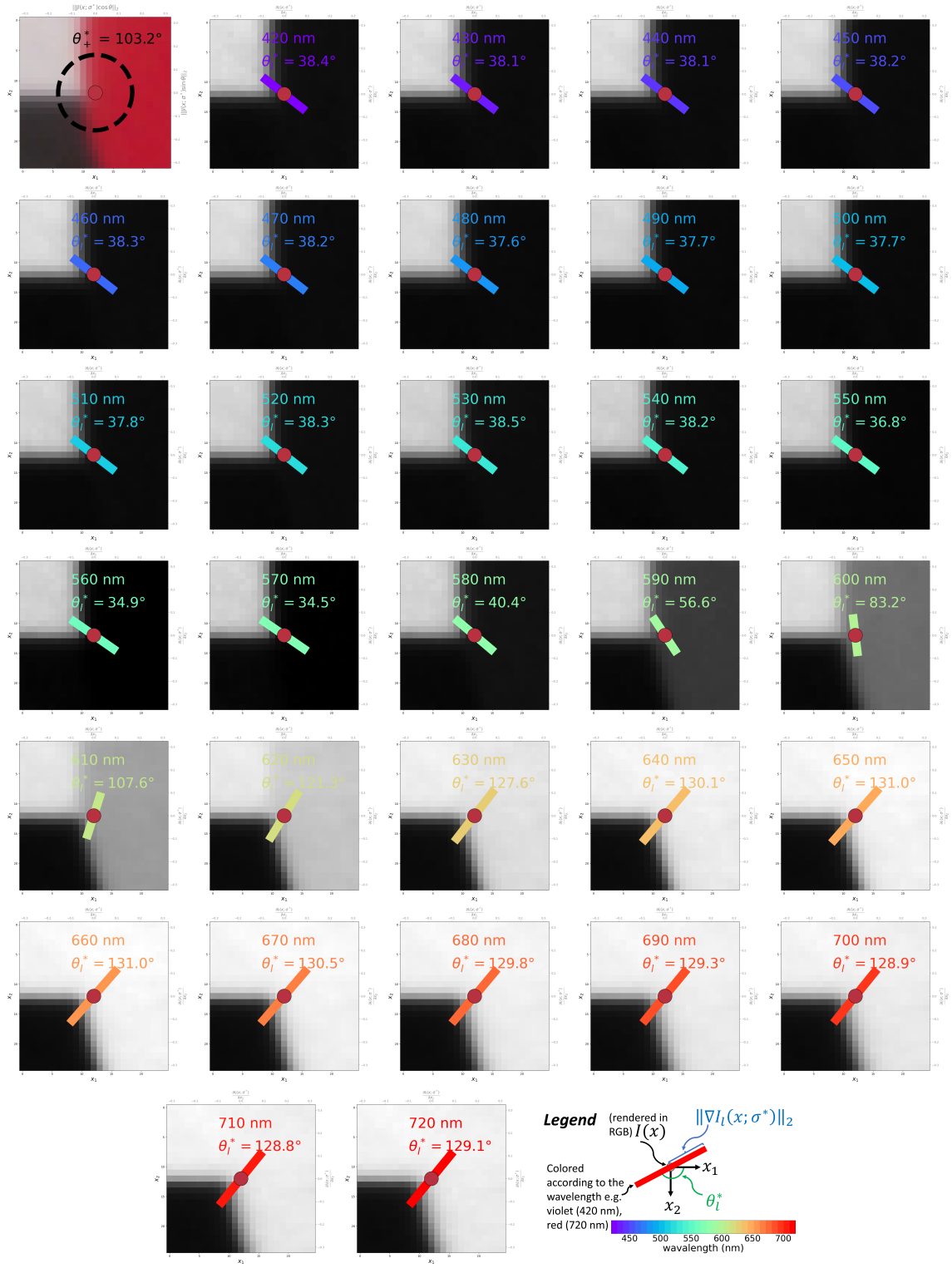


Fig. 4.18 The marginal gradients for the pixel labelled as (i) in Fig. 4.11 (a).

To analyze further, we display the marginal gradients $\nabla I_l(x;\sigma^*) = \left[\frac{\partial I_l(x;\sigma^*)}{\partial x_1} \;\; \frac{\partial I_l(x;\sigma^*)}{\partial x_2}\right]^T$ together in the same plot. Referring to Fig. 4.19 (c) - (i), we note that in all cases the major axis $2\sqrt{\lambda_+^*}$ is like the line of best fit for $\nabla I_l(x;\sigma^*)$ i.e. it is directed at the highest variation. This is akin to principal component analysis (PCA) from which the first principle component is chosen such that it maximizes data variance in certain direction. Such similarity is indeed expected as λ_+ is the largest eigenvalue of the spatial covariance matrix $M = \mathrm{J}I(x;\sigma^*)^T G \mathrm{J}I(x;\sigma^*)$, where $\mathrm{J}I(x)$ is the collection of $\nabla I_l(x;\sigma^*)$. At this stage, we can thus envision a correspondence between $\sqrt{\lambda_+^*}$ and spectral gradient magnitude. To verify this, let us consider M in case of single spectral channel ($L = 1$) in which $\mathrm{J}I(x;\sigma^*) = \nabla I(x;\sigma^*)^T$:

$$
\begin{aligned}
M_{(L=1)} &= \nabla I(x;\sigma^*)\nabla I(x;\sigma^*)^T \\
&= \begin{bmatrix} \left(\frac{\partial I(x;\sigma^*)}{\partial x_1}\right)^2 & \left(\frac{\partial I(x;\sigma^*)}{\partial x_1}\right)\left(\frac{\partial I(x;\sigma^*)}{\partial x_2}\right) \\ \left(\frac{\partial I(x;\sigma^*)}{\partial x_1}\right)\left(\frac{\partial I(x;\sigma^*)}{\partial x_2}\right) & \left(\frac{\partial I(x;\sigma^*)}{\partial x_2}\right)^2 \end{bmatrix}.
\end{aligned} \tag{4.33}
$$

Evidently, $\lambda_+^* = \frac{1}{2}\left(\mathrm{tr}(M) + \sqrt{\mathrm{tr}(M)^2 - 4|M|}\right) = \left(\frac{\partial I(x;\sigma^*)}{\partial x_1}\right)^2 + \left(\frac{\partial I(x;\sigma^*)}{\partial x_2}\right)^2 = \|\nabla I(x;\sigma^*)\|_2^2$. It is clear, then, $\sqrt{\lambda_+^*}$ **is the spectral gradient magnitude** which reduces to image gradient magnitude $\|\nabla I(x;\sigma^*)\|_2$ when $L = 1$. Accordingly, θ_+ **is the spectral gradient direction**.

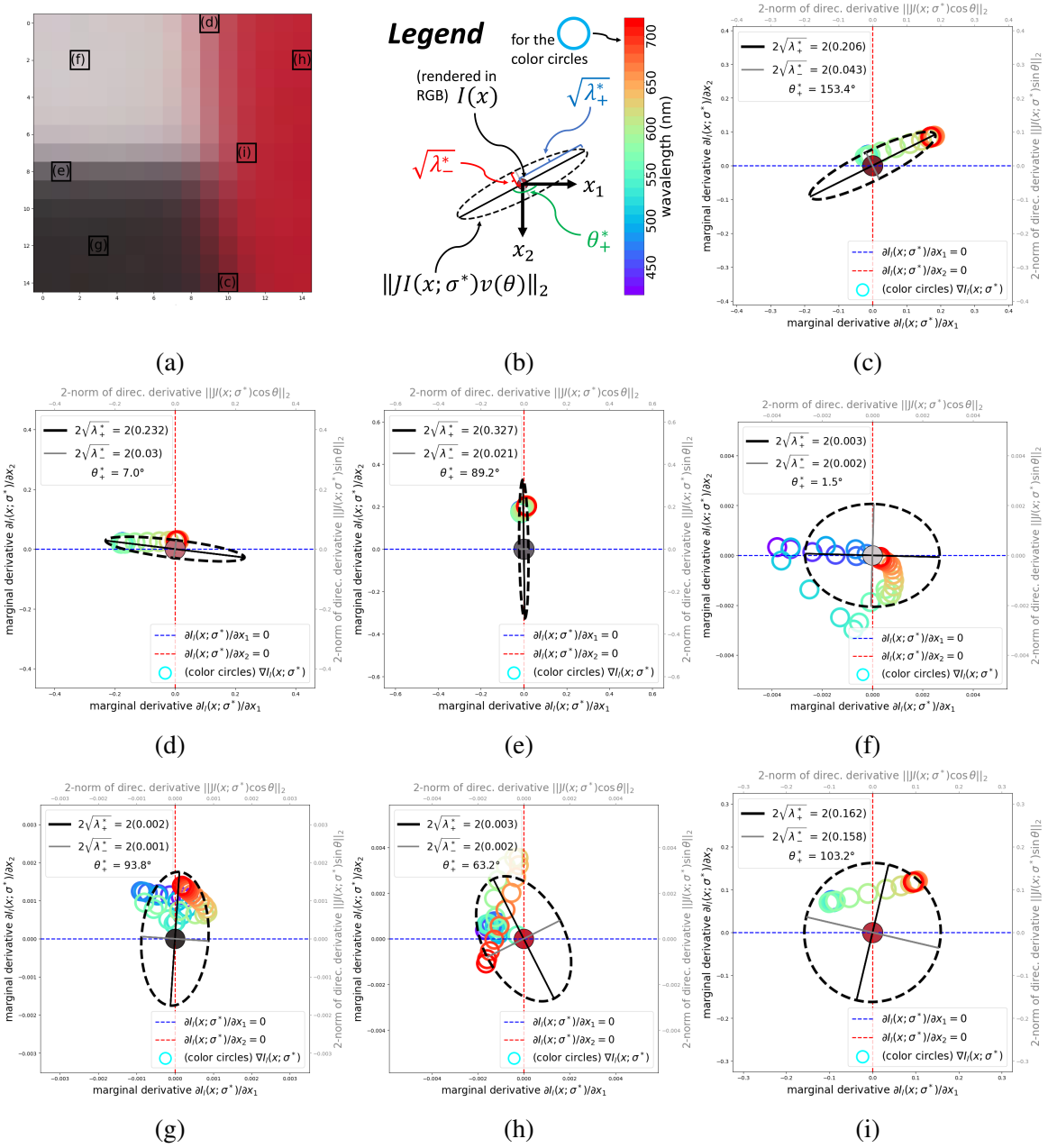


Fig. 4.19 (a) Bordered and labelled in black are the seven pixels that will be considered for the assessment of marginal gradients as illustrated in (c) - (i). The corresponding legend can be found in (b).

Demonstration with real hyperspectral image

Having defined the spectral gradient magnitude $\sqrt{\lambda_+^*}$ and direction θ_+^*, in the following we demonstrate their calculation with a real hyperspectral image. For this purpose, we shall again examine the hyperspectral image *Chart3* (512 $\times$ 512 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval [54] as shown in Fig. 4.20. In particular, we shall refer to the 16 patches as bordered and numbered in red (or black in Fig. 4.23) for further discussion.

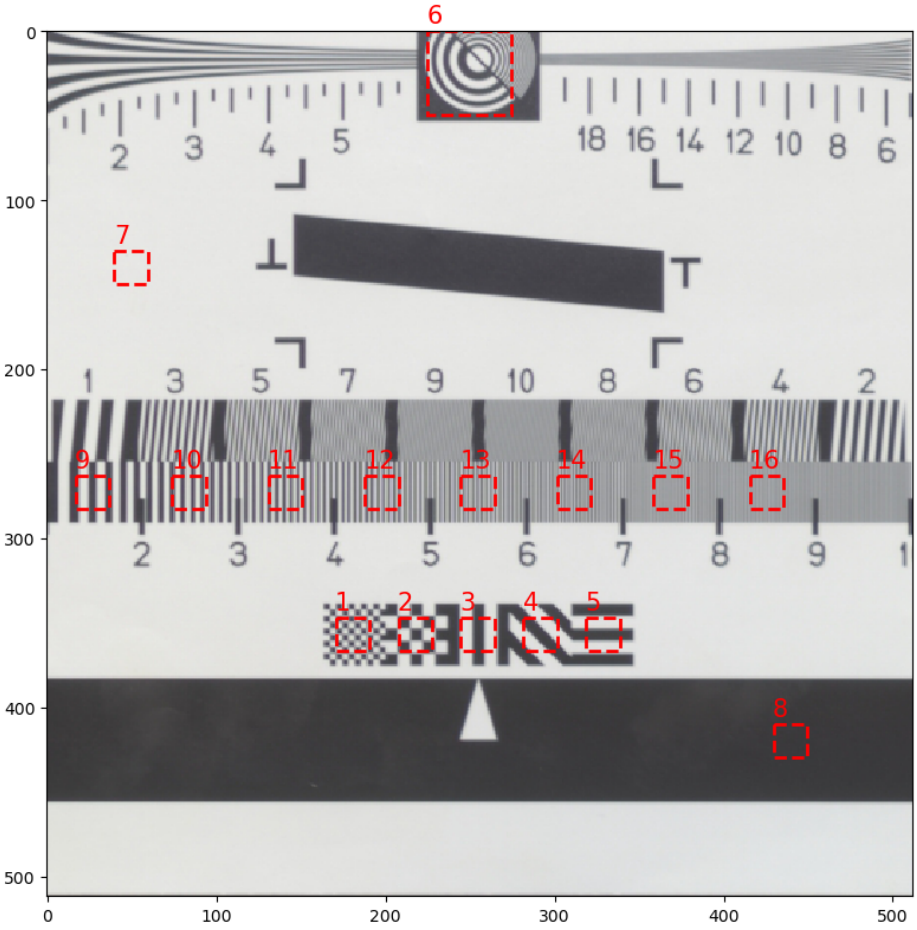


Fig. 4.20 The hyperspectral image *Chart3* (512 $\times$ 512 pixels) with wavelengths ranging from 420 nm to 720 nm at 10 nm interval, rendered in RGB for illustration. Bordered and numbered in red are the 16 patches considered for FVG calculation in Fig. 4.22 and 4.24.

Referring to Fig. 4.21, it can be seen that the edge information in *Chart3* is readily revealed thanks to the spectral gradient magnitude $\sqrt{\lambda_+^*}$. As expected, $\sqrt{\lambda_+^*}$ is strong at the interface between black and white pixels, while being weak elsewhere. Overall, the edges extracted are clean and well-connected. However, we do not observe the same $\sqrt{\lambda_+^*}$ throughout the image; $\sqrt{\lambda_+^*}$ varies in the image depending on the rate of spatial variation. For example, $\sqrt{\lambda_+^*}$ decreases from patch 9 to 16 in which the spatial variation increases. At first, this seems suspicious because *Chart3* is a binary image with supposedly identical edge strength anywhere in the image. Nevertheless, due to imperfect acquisition (noise, uneven illumination, limited spatial resolution, etc.) it actually appears in multiple shades of gray (spectrum). We illustrate this in Fig. 4.22 which shows the spectra for patches 9 - 16. Clearly, the variation of spectra decreases from patch 9 to 16 and as a result, $\sqrt{\lambda_+^*}$ also decreases. Such behaviour confirms the validity of $\sqrt{\lambda_+^*}$ as the spectral gradient magnitude.

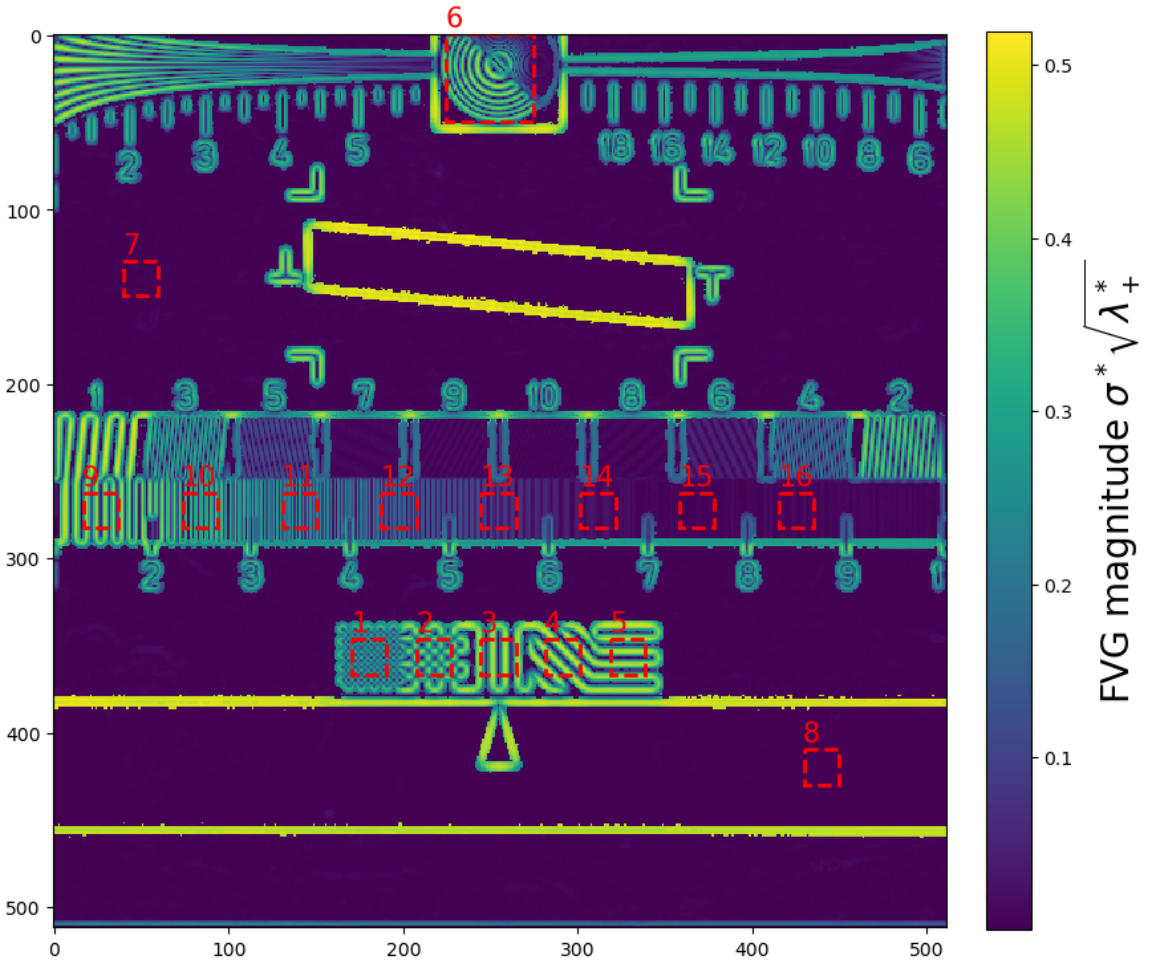
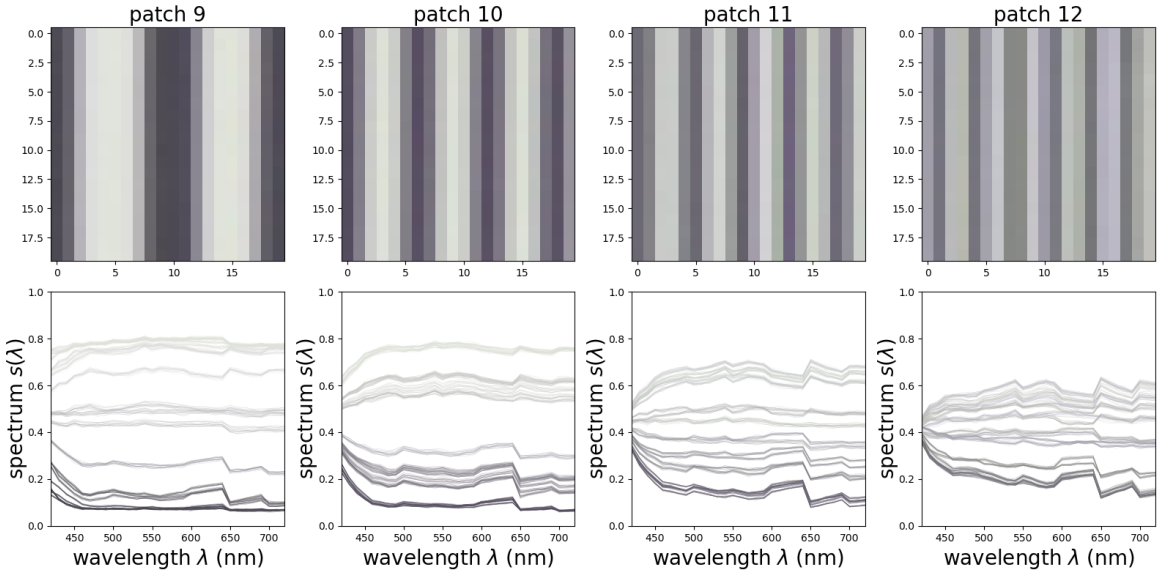
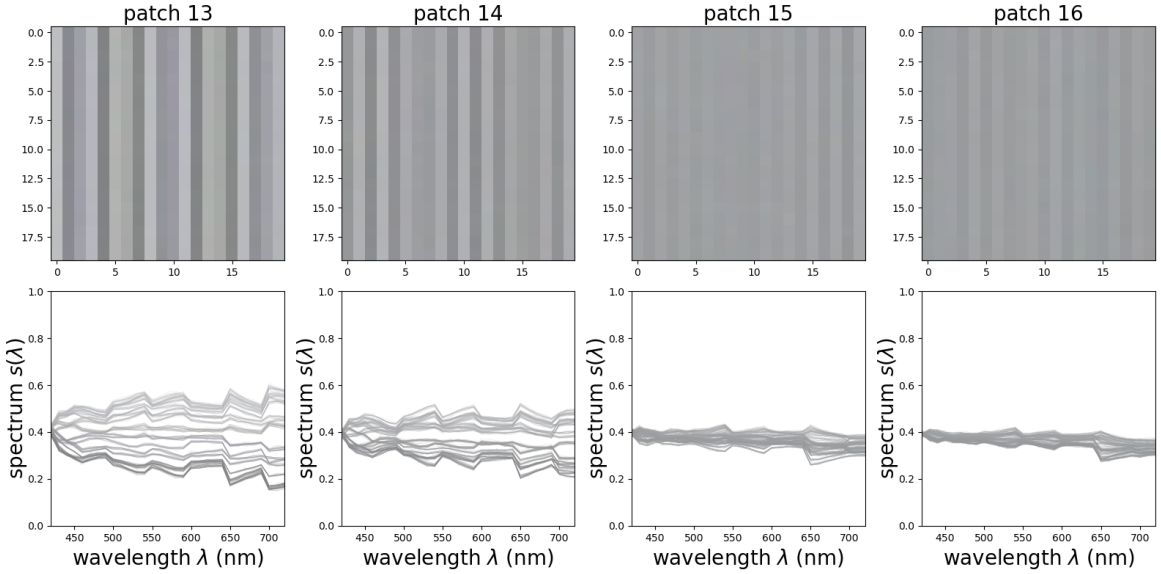


Fig. 4.21 FVG magnitude $\sigma^* \sqrt{\lambda_+^*}$ of *Chart3*.



(a)



(b)

Fig. 4.22 The patches (20 $\times$ 20 pixels) rendered in RGB (top row), followed by the (20 $\times$ 20 = 400) spectra of the corresponding patches (bottom row) for patches 9 - 12 of *Chart3*.

Referring to Fig. 4.23, it can be be seen that the spatial directionality of *Chart3* is also readily revealed thanks to the spectral gradient direction θ_+^*. Indeed, θ_+^* appears coherent for the unidirectional regions e.g. patches 3 - 5 and 9 - 16. For spatially isotropic regions e.g. patches 7 and 8, θ_+^* is random. As for the block-like regions e.g. patches 1 and 2, θ_+^* takes

predominantly two values (horizontal and vertical) whereas for the circular pattern in patch 6, θ_+^* varies continuously according to local edge directions. Next, we illustrate the distribution of θ_+^* (along with the spectral gradient magnitude $\sqrt{\lambda_+^*}$ in the form of scatter plot) for patches 1 - 8 in Fig. 4.24. Evidently, the distribution provides another means for interpreting the spatial directionality of a given region. In particular, the number of peaks specifies the number of dominant directions; the location of the peaks indicates the dominant direction; the spread for a given peak (if any) refers to the dominance of the direction. For example, a unimodal distribution of θ_+^* indicates unidirectionality as exemplified in the cof for patches 3 - 5. A bimodal distribution as in the case of patches 1 and 2 points to bidirectionality. By looking closer, it can be seen that the bimodal distribution for patch 1 is more dispersed compared to that for patch 2. This is because the blocks in patch 1 are smaller in size than that in patch 2, and as a result, there are more corner pixels (which "point" in directions further from the two dominant ones i.e. vertical and horizontal). In contrast, the blocks in patch 2 are larger and as a result, there are more edge pixels (which "point" closer in the two dominant directions). Hence, θ_+^* is less spread out. Meanwhile, a uniform distribution of θ_+^* as in the case for patches 6 - 8 indicates that the spatial variation is isotropic i.e. has no directionality.

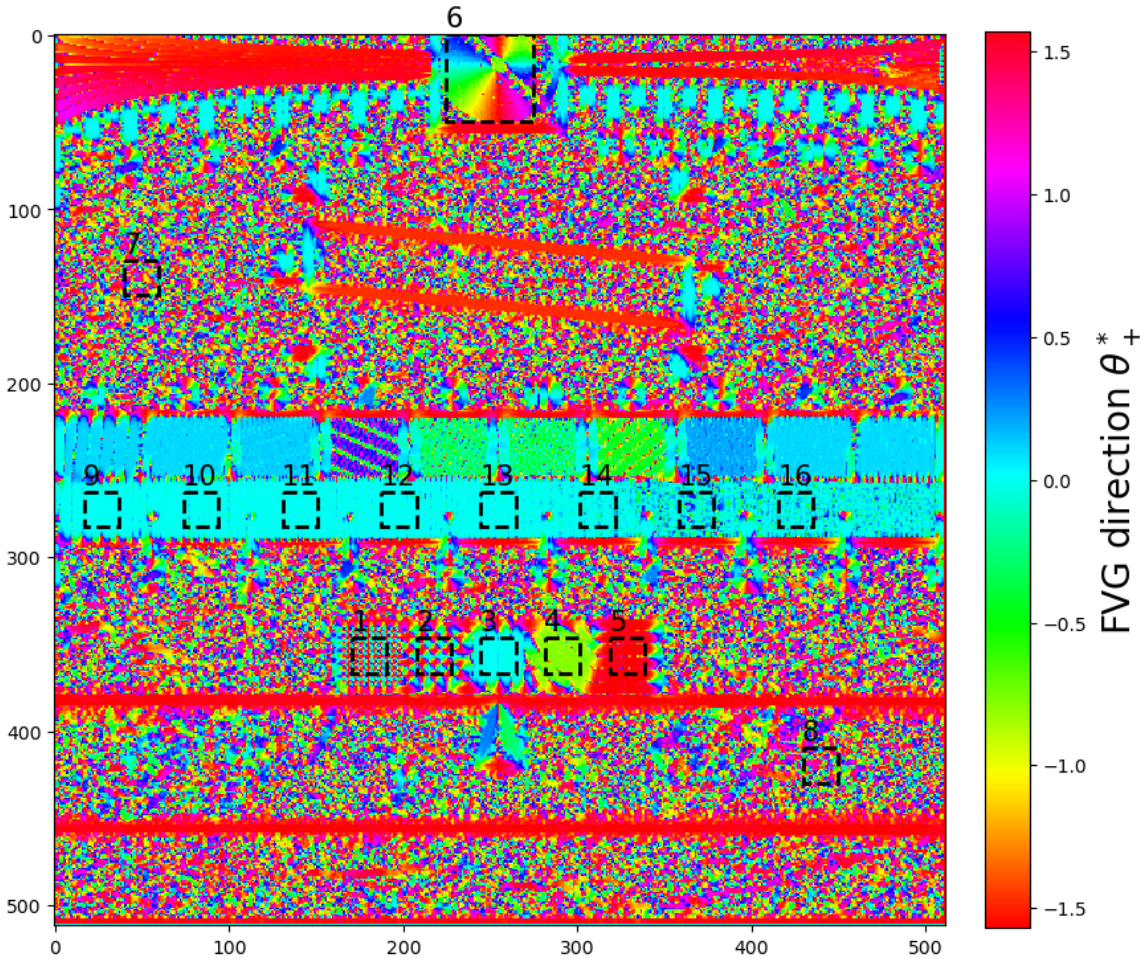
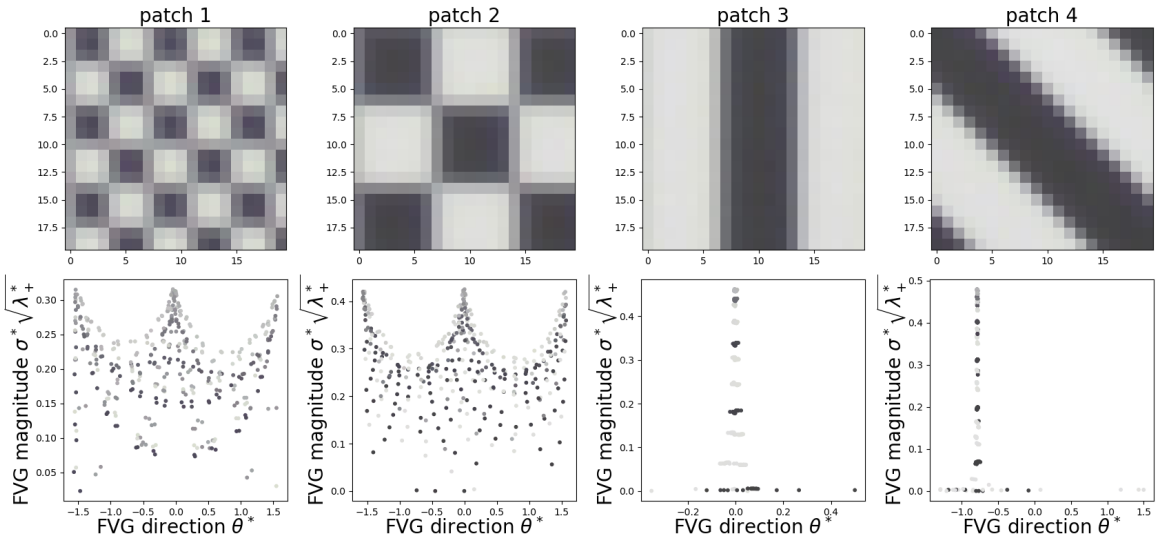
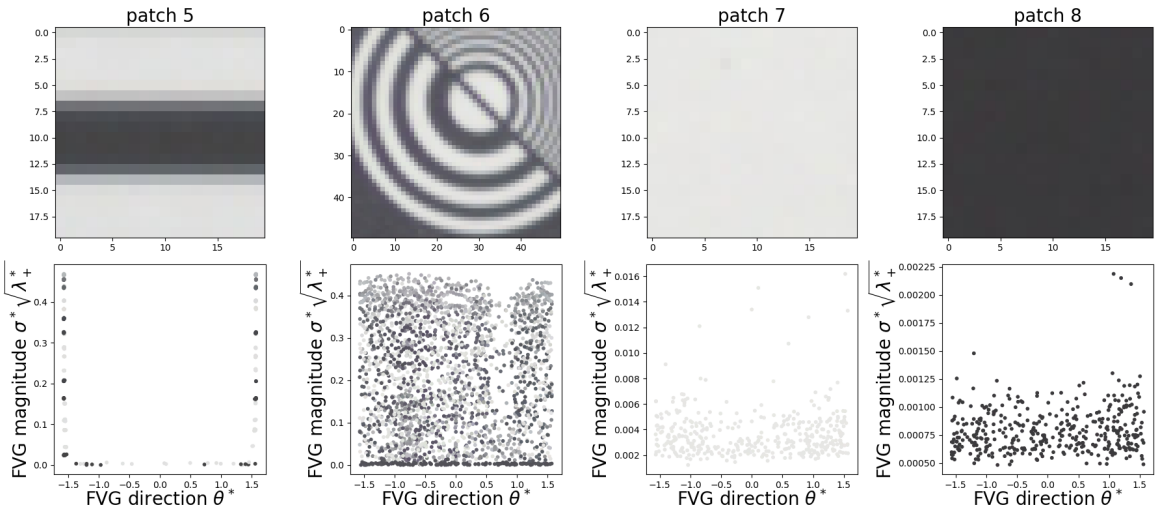


Fig. 4.23 FVG direction θ^* of *Chart3*.



(a)



(b)

Fig. 4.24 The patches (20 × 20 pixels, except for patch 6 which measures 50 × 50 pixels) rendered in RGB (top row), followed by the two-dimensional scatter plot of FVG direction θ_+^* and FVG magnitude $\sqrt{\lambda_+^*}$ for patches 1 - 8 of *Chart3*.

Spectral gradient coherence

Referring to Fig. 4.11 - 4.19, it can be observed that $\sqrt{\lambda_-^*}$ grows as the marginal gradient directions θ_l^* becomes increasingly incoherent. Furthermore, θ_-^* which is associated with λ_-^* is defined to be orthogonal to the spectral gradient direction θ_+^* as dictated in (4.14). Thus,

we can envision a relationship between $\sqrt{\lambda_-^*}$ and coherence of θ_l^*. To verify this, let us consider[7] the following "system of equations": $\lambda_+^*\lambda_-^* = |M|$ and the definition of $|M|$:

$$\begin{aligned} |M| &= \left| JI(x;\sigma^*)^T G JI(x;\sigma^*) \right| \\ &= \left| \begin{bmatrix} \left\langle \frac{\partial I(x;\sigma^*)}{\partial x_1}, \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\rangle_2 & \left\langle \frac{\partial I(x;\sigma^*)}{\partial x_1}, \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\rangle_2 \\ \left\langle \frac{\partial I(x;\sigma^*)}{\partial x_1}, \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\rangle_2 & \left\langle \frac{\partial I(x;\sigma^*)}{\partial x_2}, \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\rangle_2 \end{bmatrix} \right| \\ &= \left\| \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\|_2^2 \left\| \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\|_2^2 - \left\langle \frac{\partial I(x;\sigma^*)}{\partial x_1}, \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\rangle_2^2, \end{aligned} \tag{4.34}$$

where $\frac{\partial I(x;\sigma^*)}{\partial x_i}$ is the vector of marginal image derivatives with respect to x_i. Considering the geometric definition of scalar product of two Euclidean vectors:

$$\left\langle \frac{\partial I(x;\sigma^*)}{\partial x_1}, \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\rangle_2^2 = \left\| \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\|_2^2 \left\| \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\|_2^2 \cos^2\phi, \tag{4.35}$$

where ϕ is the angle[8] between $\frac{\partial I(x;\sigma^*)}{\partial x_1}$ and $\frac{\partial I(x;\sigma^*)}{\partial x_2}$, (4.34) can be expressed further as:

$$\begin{aligned} |M| &= \left\| \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\|_2^2 \left\| \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\|_2^2 \left(1 - \cos^2\phi\right) \\ &= \left\| \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\|_2^2 \left\| \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\|_2^2 \sin^2\phi, \end{aligned} \tag{4.36}$$

Substituting (4.36) into $\lambda_+^*\lambda_-^* = |M|$ and rearranging gives:

$$\lambda_-^* = \frac{\left\| \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\|_2^2 \left\| \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\|_2^2}{\lambda_+^*} \sin^2\phi. \tag{4.37}$$

[7] The product of the eigenvalues of matrix A is the same as the determinant of A.

[8] ϕ is not to be confused with θ in (4.29), as the latter refers to the angle on $\mathbb{R}^2$ (the spatial dimensions.)

Thus, for $\lambda_-^* = 0$ it requires that $\left\| \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\|_2 = 0$ or $\left\| \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\|_2 = 0$ or $\phi = 0$. Note that $\left\| \frac{\partial I(x;\sigma^*)}{\partial x_1} \right\|_2 = 0$ and $\left\| \frac{\partial I(x;\sigma^*)}{\partial x_2} \right\|_2 = 0$ imply that all the nonzero marginal gradients are pointing at the direction of x_2 and x_1 respectively, and in both cases the marginal gradient directions θ_l^* are fully coherent. Meanwhile, ϕ (angle between $\frac{\partial I(x;\sigma^*)}{\partial x_1}$ and $\frac{\partial I(x;\sigma^*)}{\partial x_2}$) being zero implies:

$$\begin{aligned} \left(\frac{\partial I_l(x;\sigma^*)}{\partial x_2} \Big/ \frac{\partial I_l(x;\sigma^*)}{\partial x_1} \right) &= \text{constant } \forall \ l \in [1,L] \\ \therefore \arctan \left(\frac{\partial I_l(x;\sigma^*)}{\partial x_2} \Big/ \frac{\partial I_l(x;\sigma^*)}{\partial x_1} \right) &= \theta_l^* = \text{constant } \forall \ l \in [1,L]. \end{aligned} \tag{4.38}$$

Thus, $\phi = 0$ implies that θ_l^* are identical i.e. fully coherent. It is clear, then, $\sqrt{\lambda_-^*}$ **is the spectral gradient coherence**. Accordingly, a low $\sqrt{\lambda_-^*}$ indicates high θ_l^* coherence which results in stronger spatial directionality as in the case of Fig. 4.11 (c) - (e). Conversely, a high $\sqrt{\lambda_-^*}$ indicates low θ_l^* coherence which results in weaker directionality as in the case of Fig. 4.11 (f) - (i). When $\sqrt{\lambda_-^*} = \sqrt{\lambda_+^*}$, there is no directionality as the spatial variation is isotropic. Considering the expression for the spectral gradient direction θ_+^* in (4.39) (recalled from (4.14)), in this case θ_+^* would be undefined[9] as the denominator $\lambda_+^* - \lambda_-^*$ becomes zero:

$$\begin{aligned} \theta_+^* &= \text{sgn}(M_{12}) \arcsin \left(\frac{\lambda_+^* - M_{11}}{\lambda_+^* - \lambda_-^*} \right)^{\frac{1}{2}} \\ \text{s.t. } M_{ii'} &= \left\langle \frac{\partial I(x;\sigma^*)}{\partial x_i}, \frac{\partial I(x;\sigma^*)}{\partial x_{i'}} \right\rangle_2. \end{aligned} \tag{4.39}$$

Referring to the expression of M in (4.33) in case of single spectral channel ($L = 1$), it is evident that $\lambda_- = \frac{1}{2} \left(\text{tr}(M_{(L=1)}) - \sqrt{\text{tr}\left(M_{(L=1)}\right)^2 - 4|M_{(L=1)}|} \right) = 0$ as $|M_{(L=1)}| = 0$. In

[9] Numerically, this is rarely the case for any acquired images due to noise and quantization. Nevertheless, a large value of λ_-^* comparable to that of λ_+^*, does suggest some sort of θ_+^* indeterminacy.

other words, λ_- is always null for monovariate image. Indeed, this is because there is no incoherence "among" the gradient direction (as there is only one spectral channel!).

4.2.3 Spectral shape and intensity gradients

Like spectral difference, spectral gradient too can be observed at either spectral shape or intensity levels or both. To demonstrate the interest of such distinction in texture analysis, consider a piece of plain paper without any pattern, line, or mark, the one which can be found in any photocopy room. Under standard viewing conditions, we expect to perceive a texture with negligible contrast (though under microscopic view, we expect to perceive a highly fibrous texture, hence the interest in multiscale analysis). Now, if we were to crumple the paper, we then expect to perceive a creased and wrinkled texture due to its physical surface's nonuniformity. Obviously, the contrast, in this case, would be registered only at the spectral intensity level, whereas at the spectral shape level it is null. This illustrates the interest in such distinction as it allows richer and more discriminative texture characterization.

That said, we propose to characterize contrast at two levels: spectral shape and intensity. Before to do so, a selection of image normalization is required for the definition of image norm $\|I(x)\|$ and normalized image $\bar{I}(x) = I(x)/\|I(x)\|$. As the formulation of gradient is defined in the context of vector space, we resort to using 2-norm. Thus, $\|I(x)\|$ is given by:

$$\|I(x)\|_2 = \left(I(x)^T G I(x)\right)^{\frac{1}{2}}, \tag{4.40}$$

recalling that G is the Gram matrix which takes into account the possible non-orthogonality of spectral channels. By taking the derivative of $\|I(x)\|_2 \in \mathbb{R}$, we obtain the spectral intensity gradient, hereafter abbreviated as SIG. An application of the chain rule (from calculus) gives:

$$\begin{aligned}\nabla\|I(x;\sigma)\|_2 &= \nabla\left(I(x)^T GI(x)\right)^{\frac{1}{2}} \\ &= \frac{1}{2}\left(I(x)^T GI(x)\right)^{-\frac{1}{2}}\left(2I(x)^T GJI(x;\sigma)\right) \\ &= \bar{I}(x)^T GJI(x;\sigma),\end{aligned} \tag{4.41}$$

which is a vector in $\mathbb{R}^2$ (like image gradient $\nabla I(x;\sigma)$) consisting of:

$$\nabla\|I(x;\sigma)\|_2 = \begin{bmatrix}\frac{\partial\|I(x;\sigma)\|}{\partial x_1} & \frac{\partial\|I(x;\sigma)\|}{\partial x_2}\end{bmatrix}^T. \tag{4.42}$$

Similarly, the spectral shape gradient (which is in fact, a Jacobian matrix), hereafter abbreviated as SSG can be obtained by taking the Jacobian of $\bar{I}(x) \in \mathbb{R}^L$. An application of the quotient rule, followed by the substitution of the SIG defintion from (4.41) give:

$$\begin{aligned}\mathrm{J}\bar{I}(x;\sigma) &= \mathrm{J}\left(\frac{I(x)}{\|I(x)\|}\right) \\ &= \frac{\|I(x)\|JI(x;\sigma) - I(x)\nabla\|I(x;\sigma)\|}{\|I(x)\|^2} \\ &= \frac{1_L - \bar{I}(x)G\bar{I}(x)^T}{\|I(x)\|}GJI(x;\sigma),\end{aligned} \tag{4.43}$$

which is a matrix in $\mathbb{R}^{L\times 2}$ consisting of:

$$\mathrm{J}\bar{I}(x;\sigma) = \begin{bmatrix}\frac{\partial\bar{I}_1(x;\sigma)}{\partial x_1} & \frac{\partial\bar{I}_2(x;\sigma)}{\partial x_1} & \cdots & \frac{\partial\bar{I}_L(x;\sigma)}{\partial x_1} \\ \\ \frac{\partial\bar{I}_1(x;\sigma)}{\partial x_2} & \frac{\partial\bar{I}_2(x;\sigma)}{\partial x_2} & \cdots & \frac{\partial\bar{I}_L(x;\sigma)}{\partial x_2}\end{bmatrix}^T. \tag{4.44}$$

Accordingly, the SSG magnitude $\sqrt{\lambda_{G+}}$, coherence $\sqrt{\lambda_{G-}}$, and direction θ_G can be defined in similar manner as the case for FVG:

$$\lambda_{G\pm} = \frac{1}{2}\left(\mathrm{tr}(M_G) \pm \sqrt{\mathrm{tr}(M_G)^2 - 4|M_G|}\right), \tag{4.45}$$

$$\theta_G = \mathrm{sgn}(M_{G,12}) \arcsin\left(\frac{\lambda_{G+} - M_{G,11}}{\lambda_{G+} - \lambda_{G-}}\right)^{\frac{1}{2}}, \tag{4.46}$$

where $M_G = \mathrm{J}\bar{I}(x;\sigma)^T G \mathrm{J}\bar{I}(x;\sigma)$. On the other hand, the SIG magnitude $\sqrt{\lambda_W}$ and direction θ_W can be defined in similar manner as the case for (monovariate) image gradient:

$$\lambda_W = \left(\frac{\partial \|I(x;\sigma)\|}{\partial x_1}\right)^2 + \left(\frac{\partial \|I(x;\sigma)\|}{\partial x_2}\right)^2, \tag{4.47}$$

$$\theta_W = \arctan\left(\frac{\partial \|I(x;\sigma)\|}{\partial x_2} \Big/ \frac{\partial \|I(x;\sigma)\|}{\partial x_1}\right). \tag{4.48}$$

Having decomposed spectral gradient into spectral shape and intensity parts, it is straightforward to also perform scale selection separately based on SSG and SIG magnitudes:

$$\begin{aligned} \sigma_G^*(x) &= \underset{\sigma}{\mathrm{argstat}}\left(\sigma\sqrt{\lambda_{G+}(\sigma)}\right), \\ \sigma_W^*(x) &= \underset{\sigma}{\mathrm{argstat}}\left(\sigma\sqrt{\lambda_W(\sigma)}\right), \end{aligned} \tag{4.49}$$

where $\underset{\sigma}{\mathrm{argstat}}(\cdot)$ is the operation which finds the scale space stationary (SSS) point of $\sigma\sqrt{\lambda_{G+}(\sigma)}$ in case of σ_G^* or that of $\sigma\sqrt{\lambda_W(\sigma)}$ in case of σ_W^*.

We demonstrate the decompostion of FVG into SSG and SIG in Fig. 4.26 - 4.28 concerning nine hyperspectal images with wavelengths ranging from 420 nm to 720 nm at 10 nm interval. Each measures 500×500 pixels with the exception of *Butterfly* (512×512 pixels). They are selected to illustrate advantages of the decomposition not limited to: separation of chromatic and achromatic components (*Colorchart*, *Character*, *Butterfly*), sharper edge detection by virtue of SSG (*Cloth6*, *Fan*, *Tshirts2*), and surpression of illumination effects

(*Cloth5*, *Party*, *Fan2*). For joint representation, we shall employ the two dimensional colormap as illustrated in Fig. 4.25 with the hue and transparency referring to the SSG / SIG / FVG direction (ranges from $-90°$ to $90°$) and magnitude (scaled from 0 to 1) respectively.

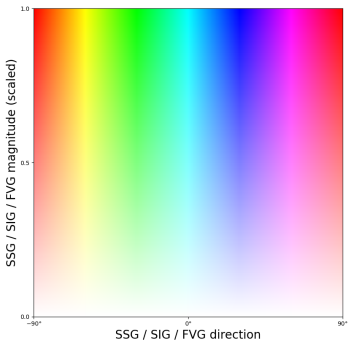
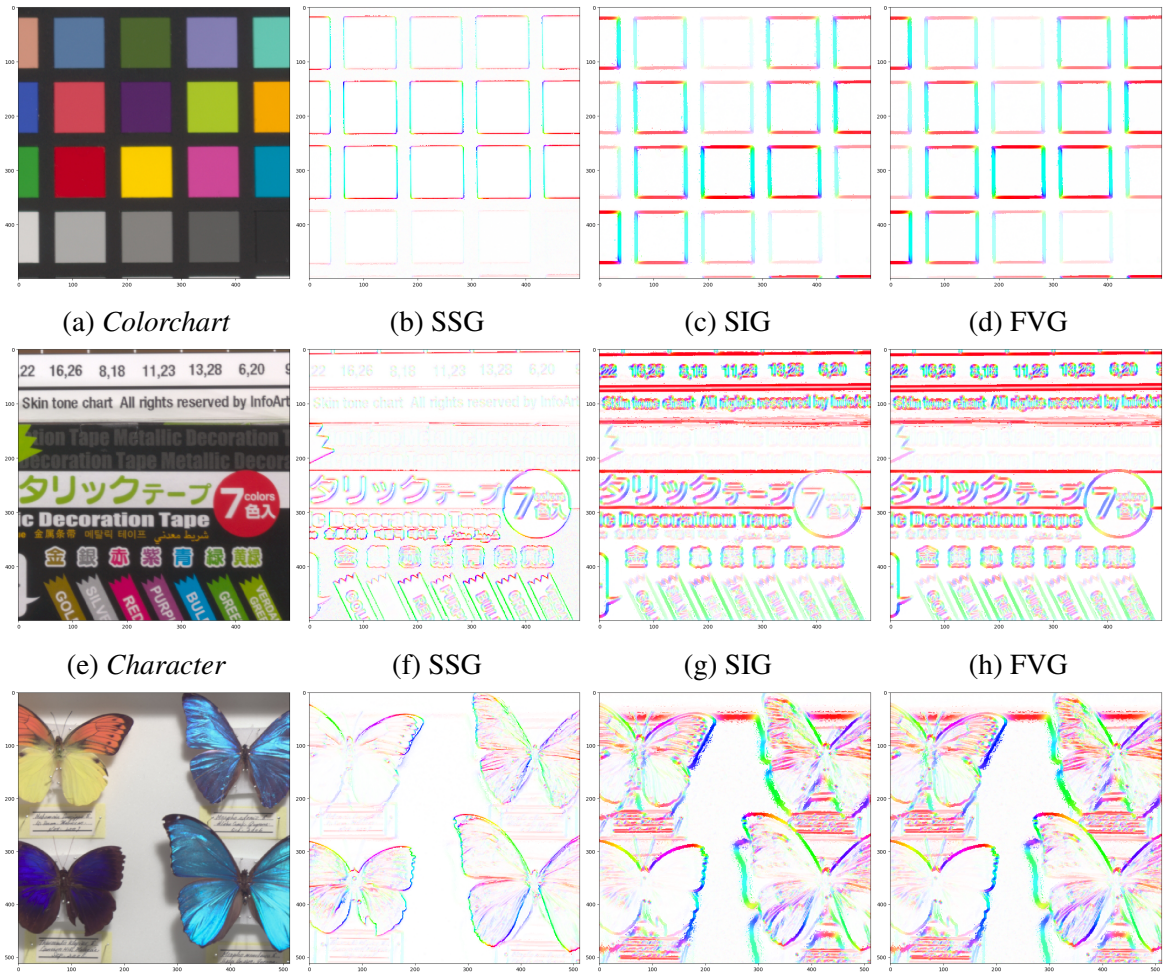


Fig. 4.25 The two dimensional colormap as used for Fig. 4.26 - 4.28, with the hue and transparency referring to the SSG / SIG / FVG direction (ranges from $-90°$ to $90°$) and magnitude (scaled from 0 to 1) respectively.

Referring to the first image *Colorchart* in Fig. 4.26, it can be seen that the edges around the grayscale patches (against the black backdrop) in are less visible with SSG. Indeed, SSG is invariant to intensity by definition. On the contrary, they are detected by SIG accordingly with the SIG magnitude around the white patch is the strongest compared to that around gray and black patches. Edges around the colourful patches e.g. red, yellow, and blue are, however, detected by both SSG and SIG due to simultaneous changes of spectral shape (from black to coloured) and intensity (from dark to bright). As with FVG, it can be seen that resulted edge detection is very similar to SIG's. Referring to the second image *Character*, it can be seen that the black (against the white backdrop) or white (against the black backdrop) structures (including characters and numbers) are better detected (with higher visibility) by SIG, whereas the colourful structures are better detected (with cleaner, sharper edges) by SSG. This further demonstrates the interest in spectral gradient (FVG) decomposition into SSG and SIG. On the other hand, the results of FVG (which detects both spectral shape and intensity gradients, though the latter sometimes dominates) differ slightly from SIG. Specifically, the word " 7 colours" (and the Chinese characters) at the centre-right of the

image as well as the word “RED” are more visible in FVG compared to SIG. They are in turn, less visible with SIG as their intensity difference with the background (red/dark pink) is reduced compared to others. Referring to the third image *Butterfly*, it can be seen that the edges around the butterfly specimens’ wings are neatly detected with SSG. This is in stark contrast to SIG with also detects the edges resulting from the specimens’ shadow on top of the wings’. On the other hand, the wordings (written in black against the white backdrop) underneath each specimen are more visibly detected by SIG. The results of FVG are again very similar to that of SIG, except around the yellow labels (on which the wordings are written) The fact that they are detected by FVG (and SSG) but less so by SIG shows that they are mainly due to spectral shape difference. Evidently, the **distinction between chromatic and achromatic edges is made possible thanks to SSG and SIG**.



(a) *Colorchart* (b) SSG (c) SIG (d) FVG

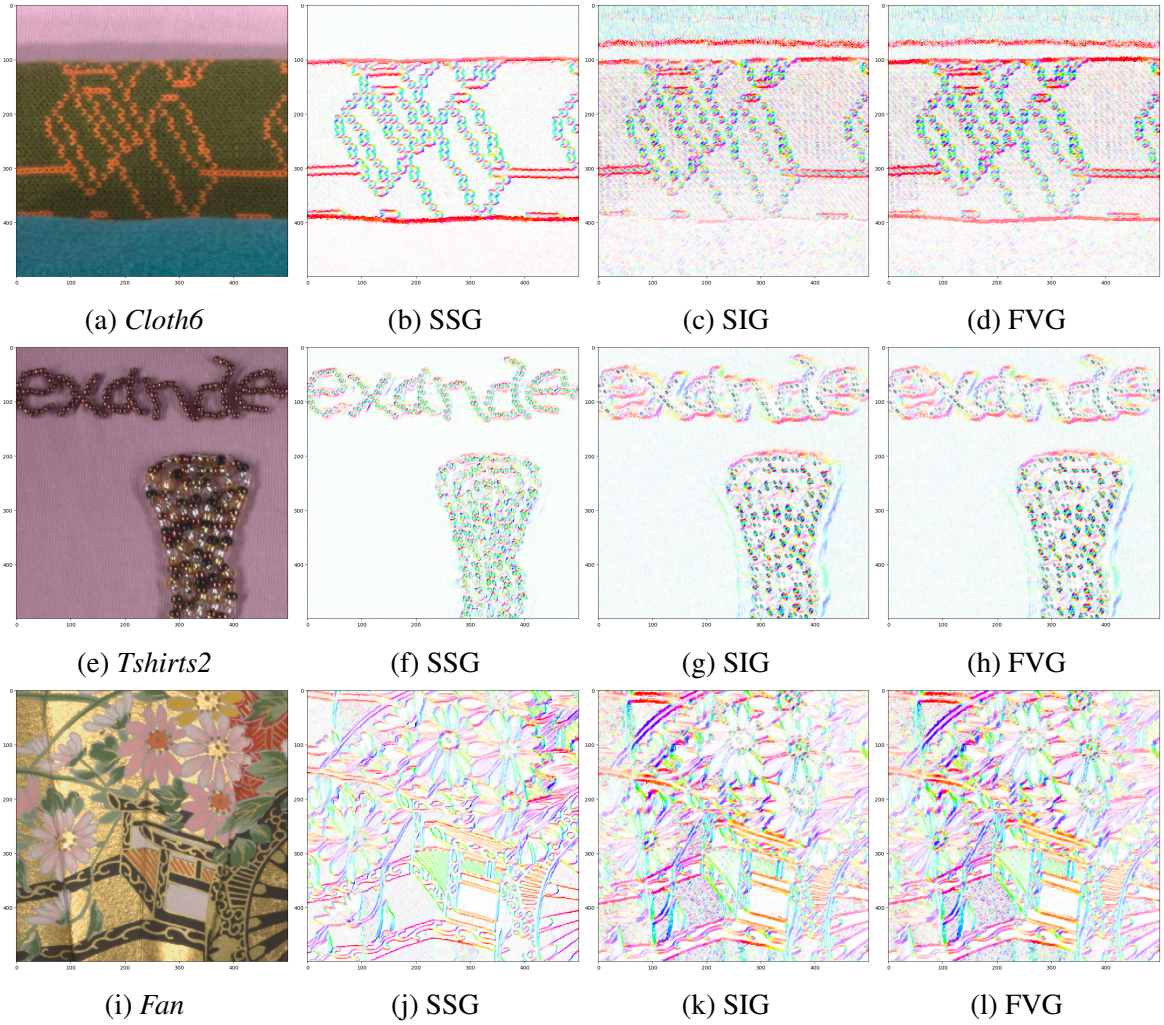
(e) *Character* (f) SSG (g) SIG (h) FVG

(i) *Butterfly* (j) SSG (k) SIG (l) FVG

Fig. 4.26 The FVG, SSG, and SIG magnitude and directions of *Colorchart*, *Character*, and *Butterfly*. The used colourmap is illustrated in Fig. 4.25.

Next, referring to the first image *Cloth6* in Fig. 4.27, it can be seen that the edges as extracted with SSG are "cleaner" than that with SIG. This is because thanks to SSG's invariance to intensity, spatial variation due to intensity variation is suppressed. Furthermore, the shadow on the pink fabric is also virtually invisible with SSG, which is otherwise detected by SIG. Moreover, the edge between green and blue fabric is also well detected by SSG. It is, however, not visibly detected by SIG due to similar intensity between the two fabrics which is otherwise detected by FVG (though less so compared to SSG). Referring to the second image *Tshirts2*, again the object edges as detected with SSG appear sharper and neater. In

contrast, the object edges as detected with SIG are blurred by the shadows (around the beads) and specular reflection (on the beads). The same can also be observed in the case of FVG. Referring to the third image *Fan*, the same scenario can also be observed in which the object edges are best detected with SSG. In comparison, the object edges as detected with SIG and FVG are less clean due to local intensity variations. In all cases, the **object edges detected with SSG is noticeably sharper, cleaner, and neater**. It is even more remarkable to note that the detection is performed without undergoing any preprocessing or postprocessing steps.



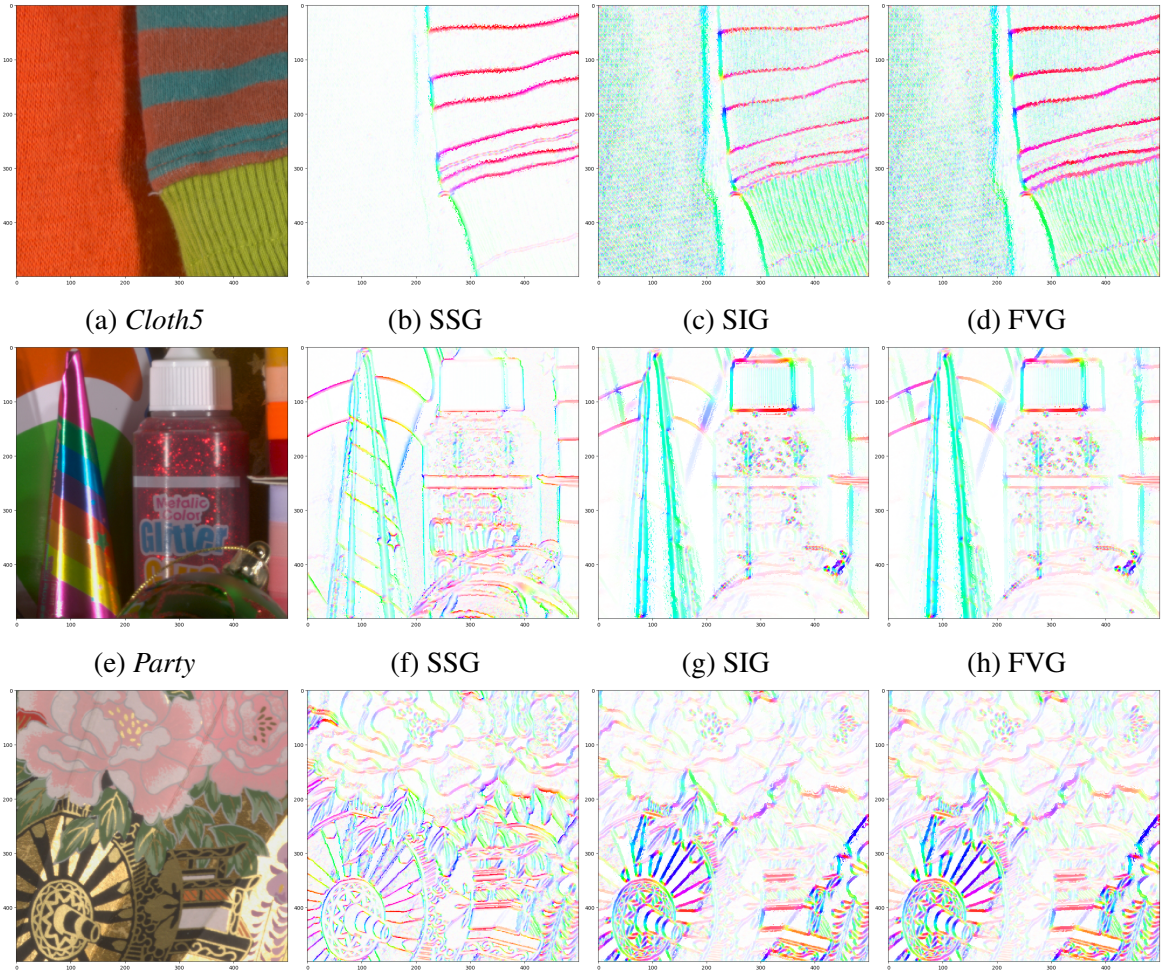
(a) *Cloth6* (b) SSG (c) SIG (d) FVG

(e) *Tshirts2* (f) SSG (g) SIG (h) FVG

(i) *Fan* (j) SSG (k) SIG (l) FVG

Fig. 4.27 The FVG, SSG, and SIG magnitude and directions of *Cloth6*, *Tshirts2*, and *Fan*. The used colourmap is illustrated in Fig. 4.25.

Finally, we demonstrate the robustness of edge detection with SSG facing various illumination effects in Fig. 4.28. Due to the blocking of light by the fabric on the right (presumably part of a sock, scarf or shirt sleeve) which is laid on top of the red fabric, a long strip of shadow can be seen at the centre of the first image *Cloth5*. Its edges are less visible with SSG but are manifested clearly with SIG and FVG. On the other hand, two "lines" of specular reflection can be seen on the party cone (with pink, green, blue, purple, red, orange, and yellow strips). In this case, it can be observed that the object edge detection with SSG is less affected by the specularities. As a result, the diagonal edges (between the colourful strips) on the cone and the wordings "Glitter Glue" on the bottle is visible with SSG, which is not the case with SIG and FVG. Lastly, the surface of *Fan2* is clearly not flat due to the folding (it's a fan) which results in it being unevenly illuminated. The effect of the uneven illumination is, however, not visible with SSG as the object edges are detected as if the surface is flat; the edges share similar SSG magnitude throughout the image. In contrast, SIG and FVG are clearly affected as the edges on the brighter surface are stronger compared to that on the darker surface. Clearly, **SSG is more robust facing the effects of shadow, specularities, and uneven illumination**.



(a) *Cloth5* (b) SSG (c) SIG (d) FVG

(e) *Party* (f) SSG (g) SIG (h) FVG

(i) *Fan2* (j) SSG (k) SIG (l) FVG

Fig. 4.28 The FVG, SSG, and SIG magnitude and directions of *Cloth5*, *Party*, and *Fan2*. The used colourmap is illustrated in Fig. 4.25.

Relationship between FVG, SSG, and SIG

The relationship between FVG, SSG, and SIG can be obtained by incorporating the image normalization into $\mathrm{J}I(x;\sigma)$ and applying the product rule:

$$\begin{aligned} \mathrm{J}I(x;\sigma) &= \mathrm{J}\left(\|I(x)\|\bar{I}(x)\right) \\ &= \|I(x)\|\mathrm{J}\bar{I}(x;\sigma) + \bar{I}(x)\nabla\|I(x;\sigma)\|. \end{aligned} \tag{4.50}$$

In words, we can thus express that FVG is the sum of SSG and SIG which are "weighted" by image norm $\|I(x)\|$ and normalized image $\bar{I}(x)$ respectively. Considering the Frobenius norm of FVG, $\lambda_{\pm}$, $\lambda_{G\pm}$, and λ_W too can be related. We begin by substituting in the FVG expression in terms of SSG and SIG from (4.50) into $\|\mathrm{J}I(x;\sigma)\|_F$:

$$\begin{aligned}\|\mathrm{J}I(x;\sigma)\|_F^2 &= \Big\|\|I(x)\|\mathrm{J}\bar{I}(x;\sigma)+\bar{I}(x)\nabla\|I(x;\sigma)\|\Big\|_F^2 \\ &= \|I(x)\|^2\|\mathrm{J}\bar{I}(x;\sigma)\|_F^2+\Big\|\bar{I}(x)\nabla\|I(x;\sigma)\|\Big\|_F^2+2\Big\langle\|I(x)\|\mathrm{J}\bar{I}(x;\sigma),\bar{I}(x)\nabla\|I(x;\sigma)\|\Big\rangle_F.\end{aligned} \tag{4.51}$$

To process the Frobenius inner product, we invoke the fact that for two given matrices A and B, $\langle A,B\rangle_F = \mathrm{tr}(A^TB)$. Substituting in the definition of SSG or of $\mathrm{J}\bar{I}(x;\sigma)$ from (4.43) gives:

$$\begin{aligned}\Big\langle\|I(x)\|\mathrm{J}\bar{I}(x;\sigma),\bar{I}(x)\nabla\|I(x;\sigma)\|\Big\rangle_F &= \mathrm{tr}\Big(\big[\|I(x)\|\mathrm{J}\bar{I}(x;\sigma)\,\big]^T G\bar{I}(x)\nabla\|I(x;\sigma)\|\Big) \\ &= \mathrm{tr}\left(\left[\left(1_L-\bar{I}(x)G\bar{I}(x)^T\right)GJI(x;\sigma)\right]^T G\bar{I}(x)\nabla\|I(x;\sigma)\|\right) \\ &= \mathrm{tr}\left(\mathrm{J}I(x;\sigma)^T G\left(1_L-\bar{I}(x)^TG\bar{I}(x)\right)G\bar{I}(x)\nabla\|I(x;\sigma)\|\right) \\ &= 0,\end{aligned} \tag{4.52}$$

noting that $\bar{I}(x)^TG\bar{I}(x)=1_L$. Invoking the fact that $\|A\|_F=\sqrt{\mathrm{tr}(A^TA)}$, we can next process:

$$\begin{aligned}\Big\|\bar{I}(x)\nabla\|I(x;\sigma)\|\Big\|_F^2 &= \mathrm{tr}\Big(\big(\bar{I}(x)\nabla\|I(x;\sigma)\|\big)^T G\bar{I}(x)\nabla\|I(x;\sigma)\|\Big) \\ &= \mathrm{tr}\Big(\nabla\|I(x;\sigma)\|^T\bar{I}(x)^TG\bar{I}(x)\nabla\|I(x;\sigma)\|\Big) \\ &= \lambda_W\end{aligned} \tag{4.53}$$

Finally, by substituting in (4.52) and (4.53) into (4.51) and again using the fact that $\|A\|_F = \sqrt{\mathrm{tr}(A^T A)}$ so that $\|\mathrm{J}I(x)\|_F^2 = \lambda_+ + \lambda_-$ and that $\|\mathrm{J}\bar{I}(x;\sigma)\|_F^2 = \lambda_{G+} + \lambda_{G-}$, it can be shown that:

$$\lambda_+ + \lambda_- = \|I(x)\|^2 (\lambda_{G+} + \lambda_{G-}) + \lambda_W \tag{4.54}$$

4.2.4 Gradient histogram of spectral texture (GHOST)

Considering a hyperspectral image I, in Chapter 3 we have defined the neighbouring difference histogram $\mathscr{J}_2^{(r,\theta)}(I)$ as the joint probability density function (PDF) of finding a particular spectral shape difference and spectral intensity difference between pixel pairs separated at distance r and direction θ. Now, having made the connection between gradient and neighbouring difference as presented at the beginning of this chapter, we propose the spectral gradient histogram $\mathscr{D}(I)$ as the joint PDF of spectral gradient magnitude, direction, and coherence. As we also make the distinction between spectral shape gradient (SSG) and spectral intensity gradient (SIG), we express $\mathscr{D}(I)$ as the joint PDF of seven dimensions as:

$$\begin{aligned} \mathscr{D}(I) = p \left(\begin{array}{c} \mathrm{J}\bar{I}(x) = \left(\sqrt{\lambda^*_{G\pm}}, \theta^*_G, \sigma^*_G \right), \\ \nabla \|I(x)\|_2 = \left(\sqrt{\lambda^*_W}\ , \theta^*_W, \sigma^*_W \right) \end{array} \right) \\ \text{s.t.} \qquad \left\| \bar{I}(x) \right\|_2 = 1, \end{aligned} \tag{4.55}$$

where $\sqrt{\lambda^*_{G+}}$, $\sqrt{\lambda^*_{G-}}$, and θ^*_G are the SSG magnitude, coherence, and direction as evaluated with its optimal scale σ^*_G, whereas $\sqrt{\lambda^*_W}$ and θ^*_W are the SIG magnitude and direction as evaluated with its optimal scale σ^*_W. In a similar fashion as RSDOM (in Chapter 3), we thus

define the second texture feature, gradient histogram of spectral texture (GHOST) as the joint construct between the spectral difference histogram $\mathscr{J}_1^{(\hat{S})}(I)$ (from Chapter 2) and $\mathscr{D}(I)$:

$$\begin{aligned}\mathscr{T}^{(\hat{S})}(I) &= \mathscr{J}_1^{(\hat{s})}(I) \ \cap \ \mathscr{D}(I) \\ &= p\left(\Delta G_1,\ldots,\Delta G_R, W, \sqrt{\lambda^*_{G\pm}}, \sqrt{\lambda^*_W}, \theta^*_G, \theta^*_W, \sigma^*_G, \sigma^*_W\right),\end{aligned} \tag{4.56}$$

where $\mathscr{T}$ is the texture feature as defined by GHOST which is a joint PDF of $R+8$ dimensions. Dictating the spectral distribution, recall that $\mathscr{J}_1^{(\hat{S})}(I)$ is the joint PDF of $R+1$ random variables i.e. the spectral shape differences $\{\Delta G_1,\ldots,\Delta G_R\}$ (corresponding to the R spectral references $\hat{S}=\{\hat{s}_1,\ldots,\hat{s}_R\}$) and spectral intensity W at a given spatial location x:

$$\mathscr{J}_1^{(\hat{S})}(I) = p\left(\begin{array}{c}\bigcap_{r=1}^{R}\left(\mathrm{KLD}\left(\bar{I}(x),\hat{s}_r\right)=\Delta G_r\right), \\ \\ \left\|I(x)\right\|_1 = W\end{array}\right)$$

$$\text{s.t.} \qquad \left\|\bar{I}(x)\right\|_1 = \|\hat{s}_r\|_1 = 1 \ \forall\ r\in[1,R].$$

Also in a similar fashion as the case of RSDOM, the similarity measurement between GHOSTs of two textures (images) I and I' is assessed using Kullback-Leibler divergence:

$$\mathrm{KLD}\left(\mathscr{T}^{(\hat{S})}(I),\mathscr{T}^{(\hat{S})}(I')\right) = \mathrm{KL}\left(\mathscr{T}^{(\hat{S})}(I)\middle\|\mathscr{T}^{(\hat{S})}(I')\right) + \mathrm{KL}\left(\mathscr{T}^{(\hat{S})}(I')\middle\|\mathscr{T}^{(\hat{S})}(I)\right), \tag{4.57}$$

where $\mathrm{KL}(\cdot,\cdot)$ is the KL measure of information. Similarly, the Gaussian mixture model (GMM) is used to model GHOST. As GHOST is a joint PDF of positive measures, it is also positively skewed and dense around zero as with RSDOM. To increase the compatibility to GMM fitting, a diffeomorphism is applied by taking the logarithm of each dimension. This

has a desirable effect of stretching the values close to zero while compressing those far from zero, thus reducing the skewness and conforming the distribution towards normality.

4.3 Summary

In this chapter, we pointed out the relationship between gradient and neighbouring difference and thus showed that the spatial characterization of texture could be carried out using image gradients (apart from neighbouring differences as described in Chapter 3). We reviewed the assessment of the image gradient using Gaussian derivative kernels before extending the discussion within the context of the multivariate image with full vector gradient (FVG). Then, we proposed our own approach by addressing the first issue i.e. the automatic scale selection for image gradient assessment based on scale space theory. Particularly, we showed the correspondence between scale space maximum and local structure size. Taking care of the interaction of neighbouring structures, we improved Lindeberg's heuristic by including scale space stationary (SSS) points for the local optimal scale selection. Next, we clarified the measures of spectral gradient magnitude, direction, and coherence. Another major part of our contribution is the mathematical decomposition of spectral gradient into spectral shape and intensity parts, which thus allows numerous applications which we strived to illustrate. Finally, we proposed our second texture feature termed gradient histogram of spectral texture (GHOST) that combines the joint distribution of spectral differences (from Chapter 2) and gradient measures. Its similarity measurement is also given by Kullback-Leibler divergence.

Chapter 5

Performance Assessment of The Proposed Texture Features

Contents

5.1 Texture classification on *HyTexiLa* 140

5.1.1 Dataset and classification scheme 140

5.1.2 Feature computation 144

5.1.3 Result analysis 146

5.2 Texture retrieval on *HSI-ICONES* 148

5.2.1 Dataset and retrieval scheme 148

5.2.2 Feature computation 149

5.2.3 Result analysis 149

5.1 Texture classification on *HyTexiLa*

5.1.1 Dataset and classification scheme

Composed of 112 reflectance images, *HyTexiLa* is a dataset [43] of spectrally and spatially high resolution texture from five categories i.e. food (10 images), stone (4 images), textile (65 images), vegetation (15 images), and wood (18 images) as illustrated in Fig. 5.1 - 5.6. Each image measures 1024×1024 pixels with $L = 186$ spectral bands. Sampled at an interval of 3.19 nm, the wavelengths range from 405.37 nm to 995.83 nm while covering both visible and near infrared regions.

Fig. 5.1 The images from *HyTexiLa*. Image adapted from [43].

Fig. 5.2 (continued) The images from *HyTexiLa*. Image adapted from [43].

Fig. 5.3 (continued) The images from *HyTexiLa*. Image adapted from [43]

Fig. 5.4 (continued) The images from *HyTexiLa*. Image adapted from [43]

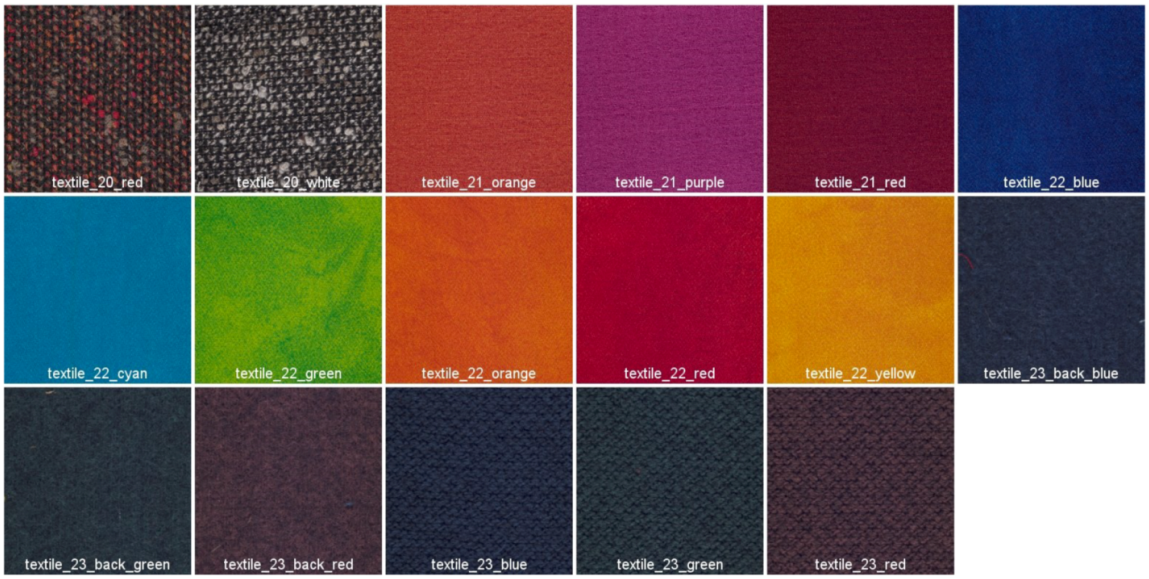


Fig. 5.5 (continued) The images from *HyTexiLa*. Image adapted from [43]

Fig. 5.6 (continued) The images from *HyTexiLa*. Image adapted from [43]

For the classification, we employ the nearest neighbour (1-NN) search whereby the classification is performed based on texture similarity alone. Such nonparametric approach allows us to attribute the classification performance directly to the feature's discriminability instead of the classifier's efficiency. Furthermore, this also enables us to deal with the issue of nonstationarity as there is a versatility of texture representation from which each class can be identified. Each image is separated into 25 patches without overlapping of which 12 and 13 of them are randomly selected for training and testing respectively. To reduce bias, we repeat the classification 100 times and consider the average performance. We report the accuracy (percentage correct classification) for both intercategorical (on the entire dataset) and intracategorical classification.

5.1.2 Feature computation

As per Chapter 2, we perform an analysis of the entire dataset and identify the optimal number of spectral references R to be three. For RSDOM, we compute the neighbouring spectral differences with three distances (scales) i.e. $r = 1$, 2, and 3 and eight directions i.e. $\theta = 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°$. However, we average the computations over the eight directions so as to produce a rotation invariant assessment. As a result, the final RSDOM feature constitutes a joint probability density function (PDF) of three spectral shape differences corresponding to the three spectral references + spectral intensity + (number of scales)(two neighbouring spectral differences i.e. of spectral shape and spectral intensity) = $3 + 1 + 3(2) = 10$ dimensions. To relieve the curse of dimensionality, which from our experience is found to be affecting the performance of RSDOM, we approximate the similarity measurement between RSDOMs by:

$$\begin{aligned}
&\mathrm{KLD}\left(\mathscr{T}^{(\hat{S},R)}(I), \mathscr{T}^{(\hat{S},R)}(I')\right) \\
&\approx \mathrm{KLD}\left(\mathscr{J}_1^{(\hat{S})}(I), \mathscr{J}_1^{(\hat{S})}(I')\right) + \mathrm{KLD}\left(\mathscr{J}_{2,G}^{(R)}(I), \mathscr{J}_{2,G}^{(R)}(I')\right) + \mathrm{KLD}\left(\mathscr{J}_{2,W}^{(R)}(I), \mathscr{J}_{2,W}^{(R)}(I')\right),
\end{aligned} \tag{5.1}$$

where $\mathscr{T}^{(\hat{S},R)}(I)$ here is texture feature concerning texture (patch) I (I' refers to another texture) as defined by RSDOM. Note that $\mathscr{J}_1^{(\hat{S})}(I)$ is a joint PDF of $R+1 = 4$ dimensions referring to the spectral difference histogram (from Chapter 2) based on the set of (three) spectral references $\hat{S}$. On the other hand, $\mathscr{J}_{2,G}^{(R)}(I)$ and $\mathscr{J}_{2,W}^{(R)}(I)$ are both joint PDFs of three dimensions referring to the neigbouring difference histogram (from Chapter 3) of I with three scales $R = \{1,2,3\}$ concerning only the spectral shape differences (hence the subscript G) and only the spectral intensity differences (hence the subscript W) respectively. Note that $\mathrm{KLD}(\cdot,\cdot)$ refers to Kullback-Leibler (KL) divergence.

As for GHOST, we compute the spectral shape gradients (SSGs) and spectral intensity gradients (SIGs) with 30 scales ranging from 0.333... to 10. As a result, the final GHOST feature constitutes a joint PDF of three spectral shape differences corresponding to the three spectral references + spectral intensity + SSG magnitude + SSG direction + SSG coherence + SSG scale + SIG magnitude + SIG direction + SIG scale = $3+1+7=11$ dimensions. Similarly, we approximate the similarity measurement between GHOSTs by:

$$\begin{aligned}&\mathrm{KLD}\left(\mathscr{T}^{(\hat{S})}(I),\mathscr{T}^{(\hat{S})}(I')\right)\\&\approx \mathrm{KLD}\left(\mathscr{J}_1^{(\hat{S})}(I),\mathscr{J}_1^{(\hat{S})}(I')\right)+\mathrm{KLD}\left(\mathscr{D}_G(I),\mathscr{D}_G(I')\right)+\mathrm{KLD}\left(\mathscr{D}_W(I),\mathscr{D}_W(I')\right),\end{aligned} \tag{5.2}$$

where $\mathscr{T}^{(\hat{S})}(I)$ here is texture feature concerning texture I as defined by GHOST. Note that $\mathscr{D}_G(I)$ is a joint PDF of four dimensions referring to the spectral gradient histogram (from Chapter 4) concerning only the SSG measures (hence the subscript G), whereas $\mathscr{D}_W(I)$ is a joint PDF of three dimensions referring to the spectral gradient histogram concerning only the SIG measures (hence the subscript W). For both RSDOM and GHOST, the number of Gaussian mixture model (GMM) components is selected such that the classification accuracy is maximized.

In comparison, we also consider the following state-of-the-art:

- GLCM: Three spatial distances of $r=1,2,3$ is used with 32 quantization levels. As in [31, 80, 45], five Haralick features i.e. energy, entropy, contrast, correlation, and homogeneity are considered. For feature similarity measurement, normalized Euclidean distance [57] is used.

- Gabor filter: The frequency is set at $\frac{1}{\sqrt{2}}$, $\frac{2}{\sqrt{2}}$, and $\frac{4}{\sqrt{2}}$ with a bandwidth of one octave. For each spectral band, the Gabor energy is defined using 2-norm of the filter responses [70]. Normalized Euclidean distance [57] is used.

- LBP: A neighborhood of eight pixels ($K = 8$) with radius $r = 1,2,3$ is considered. Histogram intersection [75] is used.

As with RSDOM, the GLCM features and Gabor energies are averaged over the eight directions $\theta = 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°$ for each scale. We consider only the implementation of GLCM, Gabor filter, and LBP in marginal (m-) and cross-channel (cc-) processing [57], [70] and [43] as their vectorial formulation are metrologically impractical due to computational complexity. For dimensionality reduction, we employ band selection with a uniform spectral spacing of 10 bands so that the interchannel correlation is at least 0.9. The *HyTexiLa* images are thus spectrally downsampled to $L' = 18$ bands for the computation of GLCM, Gabor filter, and LBP. In contrast, both RSDOM and GHOST are processed in full-band ($L = 186$).

5.1.3 Result analysis

Referring to Table 5.1 and 5.2, it can be seen that both RSDOM (99.2%) and GHOST (99.1%) score comparatively good in the texture classification. They perform better than m-GLCM (88.6%), m-Gabor (92.2%), m-LBP (89.6%), cc-GLCM (94.7%), cc-Gabor (95.3%), and cc-LBP (98.3%). Compared to cc-LBP (98.3%), RSDOM (99.2%) performs slightly better (+0.9%) but with a feature size of 1990 times smaller (125 scalars for RSDOM vs. 248832 for cc-LBP). GHOST (99.1%) also performs slightly better (+0.8%) than cc-LBP (98.3%) but with a feature size 1345 times smaller (185 scalars). Dimensionality reduction is also unnecessary for RSDOM and GHOST which is otherwise required for efficient implementation of cc-LBP. Full-band processed, light, and performant, RSDOM and GHOST are clearly adapted for the designated texture classification task based on *HyTexiLa*.

Instead of performing full directional analysis as in GHOST, RSDOM only assesses spatial information in eight directions $\theta = 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°$. Besides, RSDOM only assesses spatial information with three randomly chosen scales $r = 1,2,3$,

in contrast to GHOST which performs automatic scale selection based on local structure size. At this stage, we would expect GHOST to be more performant than RSDOM. However, from Table 5.1 and 5.2 it can be observed that RSDOM actually performs slightly better than GHOST. This may be attributed to the fact that the *HyTexiLa* is a relatively simple dataset with reduced spatiochromatic complexity. As a result, the relative comprehensiveness of GHOST is not fully taken advantage of.

Table 5.1 COMPARISONS OF ACCURACY FOR TEXTURE CLASSIFICATION ON *HyTexiLa*. BOLD NUMBERS INDICATE THE BEST PERFORMANCE. RESULTS AVERAGED OVER 100 CLASSIFICATIONS. INDICATED IN BRACKETS ARE THE CORRESPONDING NUMBER OF GMM COMPONENTS.

		Full-band	***Marginal (m-)***		
Category	**No. class**	**RSDOM**	**GLCM**	**Gabor**	**LBP**
Food	10	**100.0±0.2** (2,2,7)	98.6±1.0	95.6±1.6	94.0±2.5
Stone	4	100.0±0.2 (1,1,2)	86.2±3.4	85.9±4.7	92.5±3.1
Textile	65	**100.0±0.0** (5,2,1)	96.6±0.6	98.5±0.6	93.0±0.8
Vegetation	15	**98.2 ±1.2** (9,4,2)	80.4±2.4	82.7±2.4	80.6±3.0
Wood	18	97.1 ±1.4 (5,2,9)	75.9±3.1	81.0±3.5	90.9±2.92
HyTexiLa	112	**99.2±0.3** (5,3,2)	88.6±1.0	92.2±1.1	89.6±1.2

Table 5.2 COMPARISONS OF ACCURACY FOR TEXTURE CLASSIFICATION ON *HyTexiLa*. BOLD NUMBERS INDICATE THE BEST PERFORMANCE. RESULTS AVERAGED OVER 100 CLASSIFICATIONS. INDICATED IN BRACKETS ARE THE CORRESPONDING NUMBER OF GMM COMPONENTS.

		Full-band	***Cross-channel (cc-)***		
Category	**No. class**	**GHOST**	**GLCM**	**Gabor**	**LBP**
Food	10	**100.0±0.2** (6,10, 2)	98.4±1.1	98.2±1.1	99.8±0.4
Stone	4	**100.0±0.0** (2, 1, 5)	94.6±3.0	97.3±1.9	99.7±0.7
Textile	65	**100.0±0.0** (5, 4, 1)	99.2±0.3	99.4±0.2	99.9±0.1
Vegetation	15	98.0 ±1.3 (8, 6, 3)	89.5±2.0	89.9±2.4	97.2±1.4
Wood	18	**97.2 ±1.4** (4,10, 4)	84.1±2.8	85.2±3.0	93.8±2.4
HyTexiLa	112	99.1±0.4 (5,4,5)	94.7±0.7	95.3±0.7	98.3±0.5

5.2 Texture retrieval on *HSI-ICONES*

5.2.1 Dataset and retrieval scheme

Composed of 486 radiance images, *ICONES-HSI* is a dataset [4] of remote sensing images collected using AVIRIS sensor. The images are organized into nine classes i.e. agriculture (50 images), cloud (29 images), desert (54 images), dense urban (73 images), forest (69 images), mountain (53 images), ocean (68 images), snow (55 images), and wetland (35 images) as illustrated in Fig. 5.7. Each image measures 300×300 pixels with $L = 224$ spectral bands ranging from 365 nm to 2497 nm, covering both visible and near-infrared regions.

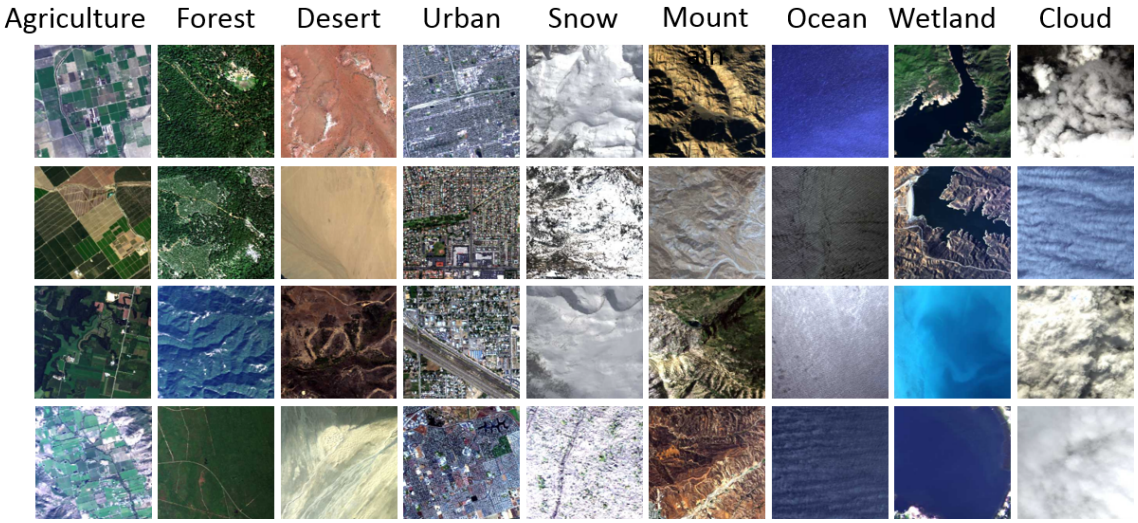


Fig. 5.7 Some of the images from *ICONES-HSI*. Image adapted from [4].

We assess the performance of RSDOM and GHOST in a content-based image retrieval (CBIR) framework in which the candidate images are ranked according to their relevance to the query image. The relevance is defined by the similarity measurement between features as extracted from the images. In the case of RSDOM and GHOST, KL divergence is used for the similarity measurement. In the end, we report the mean average precision (MAP).

5.2.2 Feature computation

After performing analysis on the entire dataset, the optimal number of spectral references R is also identified to be three. However, we disregard the spectral intensity information W as it offers no help for scene recognition in this CBIR context. This is because the same scene could exhibit different brightness depending on the time of acquisition e.g. brighter if taken at the noon, or darker if acquired in the early morning or late evening. For the neighbouring differences (in case of RSDOM) and gradients (in case of GHOST) computation, the same parameters as in texture classification on *HyTexiLa* are employed. As for GLCM, Gabor filter and LBP approaches, we consider their marginal processing (m-) as applied on the top three principal components (PCs) in line with the common practice in the community [73, 39, 34]. For *ICONES-HSI*, this corresponds to an average (over all images) explained variance of 98.0%. Other than that, the same implementation and similarity measurement as in texture classification on *HyTexiLa* are employed. Note that we do not calculate GLCM, Gabor, and LBP features using cross-channel processing because PCs are décor-related by definition. Hence, they are no interest to process cc-GLCM, cc-Gabor, and cc-LBP as they exist no added information between the PCs.

5.2.3 Result analysis

Referring to Table 5.3, it can be seen that GHOST performs the best (63.2%) compared to all the other approaches. On the other hand, RSDOM scores 62.1%, followed by Gabor (43.6%), GLCM (37.4%), and LBP (37.3%). As expected, GHOST performs better (+1.1%) than RSDOM due to its relative comprehensiveness of texture characterization. Indeed, the directional assessment of GHOST covers all the possible directions thanks to the gradient assessment. In comparison, RSDOM only assess spatial information in eight directions $\theta = 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°$. Besides, the gradient calculation in GHOST is based on optimal scales as chosen based on local structure size. In contrast, RSDOM only

assess spatial information with three randomly chosen scales $r = 1, 2, 3$. Although RSDOM actually slightly outperforms GHOST in texture classification on *HyTexiLa*, the performance of GHOST surpasses that of RSDOM in texture retrieval on *ICONES-HSI*. This may be attributed to the fact that the spatiochromatic complexity of *ICONES-HSI* is higher than that of *HyTexiLa*. As a result, the relative comprehensiveness of GHOST is fully taken advantage of.

To compare with deep learning approaches, we also include results from [4] concerning convolutional neural network (CNN) that is pretrained from ResNet. It can be seen that both RSDOM (62.1%) and GHOST (63.2%) outperform CNN (59.1%). In terms of feature size, RSDOM (75 scalars) and GHOST (140 scalars) are also 27 times and 15 times smaller respectively than the CNN features (2048 scalars). Furthermore, RSDOM and GHOST are highly interpretable which is obviously not the case for CNN features. We have thus shown that traditionally handcrafted features like RSDOM and GHOST can potentially outperform deep learning approaches thanks to metrological considerations.

Table 5.3 COMPARISONS OF MEAN AVERAGE PRECISION (MAP) FOR CONTENT-BASED IMAGE RETRIEVAL ON *ICONES-HSI*. BOLD NUMBERS INDICATE THE BEST PERFORMANCE. INDICATED IN BRACKETS ARE THE CORRESPONDING NUMBER OF GMM COMPONENTS.

Class	**No. img.**	**RSDOM**	**GHOST**	**CNN**	**GLCM**	**Gabor**	**LBP**
Agriculture	50	58.2 (1, 1, 1)	**59.3** (1, 4, 5)	48.9	45.9	35.0	26.8
Cloud	29	31.4 (10, 6, 5)	32.9 (10,10,10)	**66.5**	29.7	25.6	50.4
Desert	54	**89.3** (1, 2, 2)	53.4 (1, 9, 6)	50.7	25.6	27.0	26.8
Dense urban	73	53.2 (1, 1, 9)	**92.4** (1, 5, 5)	86.5	71.1	64.2	51.0
Forest	69	71.2 (1,10,10)	**71.8** (1, 5, 5)	57.9	41.3	34.2	38.8
Mountain	53	44.5 (2,10, 5)	47.2 (2, 4, 4)	**64.5**	34.7	46.2	31.5
Ocean	68	78.3 (10, 1,10)	78.7 (10,10, 6)	**83.1**	20.6	79.6	43.4
Snow	55	**65.5** (1, 1, 1)	65.4 (1,10, 7)	57.8	30.0	33.6	44.3
Wetland	35	**35.5** (1, 1, 1)	35.4 (1, 1, 5)	23.9	20.0	14.3	11.6
ICONES-HSI	486	62.1 (1,1,5)	**63.2 (1, 5, 5)**	59.1	37.4	43.6	37.3

Chapter 6

Conclusion

This Ph.D. manuscript is critical of the state of the art in visual texture analysis which leads to its obligation to propose metrological and fully interpretable texture features which are adapted for hyperspectral image analysis, hence the title *Hyperspectral Texture Metrology*. Throughout this work, we are not interested in developing yet another texture feature. Instead, we approached the problem from the metrological point of view. In line with the science of measurement, we first defined the measurand i.e. what texture is. Then, we developed features in strict accordance with the definition. Accordingly, the similarity measure was designated by respecting the mathematics of the features. In the following, we shall briefly review the six chapters of this manuscript with the emphasis being put on our contributions.

Naturally, our inspiration was drawn by reviewing the state of the art. In Chapter 1, we considered the three major aspects of texture analysis: feature extraction, similarity measurement, and adaptation in the hyperspectral domain. Whereas the first two components are self-explanatory, the last component deserves some elaboration. In the literature, hyperspectral texture features are mainly formulated by adapting features originally developed for monovariate applications in a multivariate context. As a result, they are at times constrained by the initial feature construction which leads to numerous ad hoc adaptations. In our work, we decided to free ourselves from such limitations by developing general features which are applicable to any number of spectral channels. The trick is to work in metric space e.g. with spectral differences and gradients which then allows fullband processing. Thanks to such general construction, our features are applicable for both monovariate and multivariate (e.g. hyperspectral) applications. Without the need for dimensionality reduction, we can thus take full advantage of the wealth of information provided by hyperspectral images.

Our next endeavour is impossible without first setting a valid texture definition. Inspired by psychophysical findings in human visual perception, we defined texture as the joint distribution of spectral and spatial properties. Akin to the particle-wave duality, we argued that spectral and spatial properties are intertwined; they are complementary aspects that make

up a given texture. In other words, texture cannot be described by the spatial arrangement of the texture primitives alone. The spectral characteristics of the texture primitives also play a role in texture discrimination. Describing the spectral-spatial properties from a statistical point of view, such texture definition must result in a probability distribution. This thus allows the exploitation of various information-theoretic theories and measures in the feature manipulation which we regard as being a less visited research territory as of now.

In Chapter 2, we proceeded to develop metrological approaches to characterizing the spectral properties of hyperspectral images. Considering spectrum as a continuous function, we first decomposed it into two parts: spectral shape and spectral intensity. Whereas the latter is a scalar, the former is a function i.e. a normalized spectrum which renders its manipulation trickier. Inspired by Hilda's work [19], we approached the problem by representing the normalized spectrum in metric (spectral difference) space thanks to spectral references. Next, we constructed the spectral difference histogram which is really a joint probability density function (PDF) of spectral (shape) differences and spectral intensity. Recognizing its probabilistic nature, we then proposed a method to measure spectral diversity using entropy. Taking benefit of the diversity measurement, we detailed the selection of optimal spectral references based on entropy maximization. Drawing correspondences with principal component analysis (PCA), we showed that our approach shares a similar spirit with PCA but with an important caveat: that the physics of spectrum as function is respected.

In Chapter 3, we set out to develop metrological approaches to characterizing the spatial properties of hyperspectral images. Inspired by Julesz's [41], Haralick's [30], and Unser's work [83], we proposed to capture spatial information by assessing the neighbouring differences. Under the designated spatial relationship as parametrized by distance and direction, we assessed the probability distribution of spectral differences of pixel pairs, termed neighbouring difference histogram. Then, we provided interpretations of the distribution of neighbouring differences facing edge sharpness, scale, and directionality of the underlying

texture. In its simplicity lies its efficiency as demonstrated later in Chapter 6. Next, we proposed our first texture feature, termed relative spectral difference occurrence matrix (RSDOM) which is a joint construct between spectral difference histogram (from Chapter 2) and neighbouring difference histogram. Finally, we presented Kullback-Leibler (KL) divergence for its similarity measurement. To aid its storage and manipulation, we modelled RSDOM using the Gaussian mixture model (GMM) which proved to be adequate and efficient.

Our motivation for Chapter 4 stemmed from our dissatisfaction with RSDOM which for now necessitates parameter selection for optimal performance. Envisioning a parameterless formulation, we showed the link between image gradient and neighbouring differences. The usage of gradient is an interesting one because it allows a natural description of spatial directionality as well as the application of scale space theory. We began our proposition by first dictating automatic scale selection for gradient assessment originally based on Lindeberg's heuristic [50]. Then, we related the selected scales with the local structure size which lends information to granulometry. Realizing the interaction of neighbouring structures, we improve Lindeberg's heuristic for scale selection by taking into account not just the local maximum but also the inflection point of the scale space evolution of gradient magnitude. Next, we clarified the measures of spectral gradient magnitude, direction, and coherence based on Chatoux et. al.'s work [10]. Inspired by Richard's work [62] with spectral differences, we demonstrated a mathematical decomposition of spectral gradient into spectral shape gradient (SSG) and spectral intensity gradient (SIG). Some of the applications include separation of chromatic and achromatic components, sharper edge detection by virtue of SSG, and suppression of illumination effects. Finally, we expressed the spectral gradient histogram as the joint PDF of gradient magnitude, direction, and coherence at both SSG and SIG levels. Accordingly, we proposed our second texture feature, termed gradient histogram of spectral texture (GHOST) which is a joint construct between spectral difference histogram (from

Chapter 2) and spectral gradient histogram. In a similar fashion as RSDOM, we presented KL divergence for GHOST's similarity measurement and GMM for its statistical modelling.

In Chapter 5, we very briefly presented two tasks to assess the performance of RSDOM and GHOST in comparison to graylevel cooccurrence matrix (GLCM), Gabor filter, and local binary pattern (LBP). In the first task, we made use of the *HyTexiLa* dataset on which we performed texture classification of food, stone, textile, vegetation, and wood images. In order to assess the efficiency of our features, we did not employ sophisticated classification algorithms e.g. machine learning methods. Rather, we stuck with 1-nearest neighbour (1-NN) which is highly dependent on the efficiency of feature and similarity measure. In the second task, we made use of the *HSI-ICONES* dataset on which we performed texture retrieval on remote sensing e.g. agriculture, wetland, and urban scenes. In both tasks, we demonstrated the efficiency of RSDOM and GHOST compared to GLCM, Gabor filter, and LBP. For the second task, their performance even exceeded that of the CNN-based approach. Thanks to metrological considerations, we have thus shown that the performance of traditionally handcrafted features can potentially be at the same level as deep learning approaches.

To conclude, we have developed two metrological texture features, RSDOM and GHOST for hyperspectral image analysis. Their formulation is metrological, general, and independent of spectral resolution thanks to their operation in spectral difference and gradient spaces. Their application extends beyond hyperspectral images; with adapted difference or gradient measures, their formulation can be readily extended to both monovariate (grayscale or intensity) and multivariate (e.g. colour and magnetic resonance) images. As both RSDOM and GHOST are processed in a fullband manner, they can take full benefit of the wealth of spectral information. Furthermore, dimensionality reduction e.g. PCA is not required prior to their calculation which could otherwise alter the semantic and physical meaning of data. Thanks to their separate assessment of spectral and spatial properties of texture, RSDOM and GHOST are also adaptive. Depending on the current application, their formulation can

be altered e.g. by disregarding the spectral assessment in case of texture with known spectral properties or dropping the directionality measures in case of isotropic texture. Finally, thanks to the statistical modelling using GMM, RSDOM and GHOST are relatively light which translates to lower manipulation and storage costs.

Evidently, a research never ends. Naturally, a lot of future works can be undertaken based on our current construction. First of all, we have to admit that our work on similarity measurement could have been given more attention than we had. In particular, we are in search of true metrics as KL divergence does not respect triangle inequality. Furthermore, the similarity measure that is less affected by the curse of dimensionality is also expected. In contrast, KL divergence is subjected to density estimation which is less efficient in high dimensions as with RSDOM and GHOST. On the other hand, we are also in search of more advanced statistical modelling e.g. Generalized Gaussian mixture model which can be executed efficiently. On the other hand, we envisioned an approach similar to GHOST's in performing automatic scale selection for RSDOM. For example, the distance between pixel pairs could be selected such that their difference is maximized so as to reflect the local structure size. As for GHOST, we settled for automatic scale selection within a predefined set of scales. In future, we envisioned an automatic filterbank design i.e. with the minimum and maximum scales as well as scale intervals adapted automatically to the texture content of a given dataset. Furthermore, we also expect to develop further the automatic scale selection in order to relate to granulometric analysis which has widespread applications.

References

[1] J. Astola, P. Haavisto, and Y. Neuvo. Vector median filters. Proceedings of the IEEE, 78(4):678–689, 1990. doi: 10.1109/5.54807.

[2] M. Basseville. Divergence measures for statistical data processing—an annotated bibliography. Signal Processing, 93(4):621–633, 2013.

[3] T. C. Bau, S. Sarkar, and G. Healey. Hyperspectral region classification using a three-dimensional gabor filterbank. IEEE Trans. Geosci. and Remote Sens., 48(9): 3457–3464, 2010.

[4] O. Ben-Ahmed, T. Urruty, N. Richard, and C. Fernandez-Maloigne. Toward content-based hyperspectral remote sensing image retrieval (cb-hrsir): A preliminary study based on spectral sensitivity functions. Remote Sensing, 11(5), 2019.

[5] T. Berrett. Modern k-Nearest Neighbour Methods in Entropy Estimation, Independence Testing and Classification. PhD thesis, 10 2017.

[6] D. Bhattacharjee and R. Hiranmoy. Pattern of local gravitational force (<italic>plgf</italic>): A novel local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2):595–607, 2021. doi: 10.1109/TPAMI.2019.2930192.

[7] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(2):354–379, 2012.

[8] C. I. Chang. Spectral information divergence for hyperspectral image analysis. In IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99, volume 1, pages 509–511, 1999.

[9] C. I. Chang. An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. IEEE Transactions on Information Theory, 46(5):1927–1932, 2000.

[10] H. Chatoux, N. Richard, F. Lecellier, and C. Fernandez-Maloigne. Gradient in spectral and color images: from the di zenzo initial construction to a generic proposition. journal of the Optical Society of America A, 36:C154, 11 2019.

[11] H. Chatoux, N. Richard, F. Lecellier, and C. Fernandez-Maloigne. Full-vector gradient for multi-spectral or multivariate images. IEEE Transactions on Image Processing, 28 (5):2228–2241, 2019.

[12] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, and W. Gao. Wld: A robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1705–1720, 2010. doi: 10.1109/TPAMI.2009.155.

[13] Z. Chunsen, Z. Yiwei, and F. Chenyi. Spectral–spatial classification of hyperspectral images using probabilistic weighted strategy for multifeature fusion. IEEE Geoscience and Remote Sensing Letters, 13(10):1562–1566, 2016. doi: 10.1109/LGRS.2016.2596039.

[14] J. M. Coggins and A. K. Jain. A spatial filtering approach to texture analysis. Pattern Recognit. Letters, 3(3):195 – 203, 1985.

[15] Pierre Comon. Independent component analysis, a new concept? Signal Processing, 36(3):287–314, 1994. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(94)90029-9. URL https://www.sciencedirect.com/science/article/pii/0165168494900299. Higher Order Statistics.

[16] G. R. Cross and A. K. Jain. Markov random field texture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(1):25–39, 1983. doi: 10.1109/TPAMI.1983.4767341.

[17] A. Cumani. Edge detection in multispectral images. CVGIP: Graphical Models and Image Processing, 53(1):40–51, 1991. ISSN 1049-9652. doi: https://doi.org/10.1016/1049-9652(91)90018-F. URL https://www.sciencedirect.com/science/article/pii/104996529190018F.

[18] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. Computational Geometry: Algorithms and Applications. Springer-Verlag, second edition, 2000.

[19] H. Deborah. Towards Spectral Mathematical Morphology. PhD thesis, Norwegian Univ. of Science and Technology and Univ. of Poitiers, 2016.

[20] S. Eguchi and J. Copas. Interpreting kullback–leibler divergence with the neyman–pearson lemma. journal of Multivariate Analysis, 97(9):2034 – 2040, 2006.

[21] M. D. Fairchild. Color Appearance Models, 3rd Edition. Wiley, 2013.

[22] J. Feng, J. Zhang, and Y. Zhang. A multiview spectral–spatial feature extraction and fusion framework for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 19:1–5, 2022. doi: 10.1109/LGRS.2021.3066613.

[23] C. Fernandez-Maloigne. Segmentation et caractérisation d'images de texture à l'aide d'informations statistiques : application à l'industrie agro-alimentaire. PhD thesis, 1989. Thèse de doctorat dirigée par Smolarz, André Contrôle des Systèmes Compiègne 1989.

[24] R. V. Foutz and R. C. Srivastava. The performance of the likelihood ratio test when the model is incorrect. The Annals of Statistics, 5(6):1183–1194, 1977.

[25] D. Freedman and P. Diaconis. On the histogram as a density estimator:l2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57:453–476, 1981.

[26] M. M. Galloway. Texture analysis using gray level run lengths. Computer Graphics and Image Processing, 4(2):172–179, 1975. ISSN 0146-664X. doi: https://doi.org/10.1016/S0146-664X(75)80008-6. URL https://www.sciencedirect.com/science/article/pii/S0146664X75800086.

[27] L. Gan, J. Xia, P. Du, and J. Chanussot. Multiple feature kernel sparse representation classifier for hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 56(9):5343–5356, 2018. doi: 10.1109/TGRS.2018.2814781.

[28] L. Gan, J. Xia, P. Du, and J. Chanussot. Class-oriented weighted kernel sparse representation with region-level kernel for hyperspectral imagery classification. IEEE journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(4):1118–1130, 2018. doi: 10.1109/JSTARS.2017.2757475.

[29] X. Guo, X. Huang, and L. Zhang. Three-dimensional wavelet texture feature extraction and classification for multi/hyperspectral imagery. IEEE Geosci. Remote Sens. Lett., 11(12):2183–2187, 2014.

[30] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6):610–621, 1973. doi: 10.1109/TSMC.1973.4309314.

[31] M. Hauta-Kasari, J. Parkkinen, T. Jaaskelainen, and R. Lenz. Generalized co-occurrence matrix for multispectral texture analysis. In Proceedings of 13th International Conference on Pattern Recognition, volume 2, pages 785–789, 1996.

[32] J. R. Hershey and P. A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, volume 4, pages IV–317–IV–320, 2007.

[33] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih. Image indexing using color correlograms. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 762–768, 1997. doi: 10.1109/CVPR.1997.609412.

[34] W. Huang, Y. Huang, H. Wang, Y. Liu, and H. J. Shim. Local binary patterns and superpixel-based multiple kernels for hyperspectral image classification. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 13:4550–4563, 2020.

[35] W. Huang, Y. Huang, Z. Wu, J. Yin, and Q. Chen. A multi-kernel mode using a local binary pattern and random patch convolution for hyperspectral image classification. IEEE journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14:4607–4620, 2021. doi: 10.1109/JSTARS.2021.3076198.

[36] A. Humeau-Heurtier. Texture feature extraction methods: A survey. IEEE Access, 7:8975–9000, 2019. doi: 10.1109/ACCESS.2018.2890743.

[37] E. T. Jaynes. Information theory and statistical mechanics. Phys. Rev., 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620.

[38] S. Jia, J. Hu, J. Zhu, X. Jia, and Q. Li. Three-dimensional local binary patterns for hyperspectral imagery classification. IEEE Trans. Geosci. and Remote Sens., 55(4): 2399–2413, 2017.

[39] S. Jia, K. Wu, J. Zhu, and X. Jia. Spectral–spatial gabor surface feature fusion approach for hyperspectral imagery classification. IEEE Trans. Geosci. and Remote Sens., 57(2): 1142–1154, 2019.

[40] L. Jin, H. Liu, X. Xu, and E. Song. Improved direction estimation for di zenzo's multichannel image gradient operator. Pattern Recognition, 45(12):4300–4311, 2012. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2012.06.003. URL https://www.sciencedirect.com/science/article/pii/S0031320312002737.

[41] B. Julesz. Visual pattern discrimination. IRE Transactions on Information Theory, 8 (2):84–92, 1962. doi: 10.1109/TIT.1962.1057698.

[42] B. Julesz. Textons, the elements of texture perception, and their interactions. Nature, 290(5802):91–97, 1981. doi: 10.1038/290091a0. URL https://app.dimensions.ai/details/publication/pub.1040887630.

[43] H. A. Khan, S. Mihoubi, B. Mathon, J. B. Thomas, and J. Y. Hardeberg. Hytexila: High resolution visible and near infrared hyperspectral texture images. Sensors, 18(7), 2018.

[44] S. Kullback and R. A. Leibler. On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

[45] B. Kumar and O. Dikshit. Spectral–spatial classification of hyperspectral imagery based on moment invariants. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 8(6): 2457–2463, 2015.

[46] N. N. Leonenko L. F. Kozachenko. Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii, 23:9–16, 1987.

[47] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'00, page 535–541, Cambridge, MA, USA, 2000. MIT Press.

[48] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. International journal of Computer Vision, 43:29–44, 06 2001. doi: 10.1023/A:1011126920638.

[49] W. Li, F. Feng, H. Li, and Q. Du. Discriminant analysis-based dimension reduction for hyperspectral image classification: A survey of the most recent advances and an experimental comparison of different techniques. IEEE Geoscience and Remote Sensing Magazine, 6(1):15–34, 2018.

[50] T. Lindeberg. Scale-space theory: a basic tool for analyzing structures at different scales. journal of Applied Statistics, 21(1-2):225–270, 1994.

[51] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. In Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 465–470, 1996. doi: 10.1109/CVPR.1996.517113.

[52] T. Lindeberg. Feature detection with automatic scale selection. International journal of Computer Vision, 30:77–116, 09 1998. doi: 10.1023/A:1008045108935.

[53] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell., 11(7):674–693, 1989.

[54] Y. Monno, S. Kikuchi, M. Tanaka, and M. Okutomi. A practical one-shot multispectral imaging system using a single image sensor. IEEE Transactions on Image Processing, 24(10):3048–3059, 2015. doi: 10.1109/TIP.2015.2436342.

[55] J. M. P. Nascimento and J. M. B. Dias. Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 43 (4):898–910, 2005. doi: 10.1109/TGRS.2005.844293.

[56] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In Proceedings of 12th International Conference on Pattern Recognition, volume 1, pages 582–585 vol.1, 1994. doi: 10.1109/ICPR.1994.576366.

[57] C. Palm. Color texture classification by integrative co-occurrence matrices. Pattern Recognit., 37(5):965–976, 2004.

[58] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and journal of Science, 2(11): 559–572, 1901. doi: 10.1080/14786440109462720.

[59] A. P. Pentland. Fractal-based description of natural scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6(6):661–674, 1984. doi: 10.1109/TPAMI.1984.4767591.

[60] C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005. ISBN 026218253X.

[61] G. Rellier, X. Descombes, F. Falzon, and J. Zerubia. Texture feature analysis using a gauss-markov model in hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 42(7):1543–1551, 2004. doi: 10.1109/TGRS.2004.830170.

[62] N. Richard, D. Helbert, C. Olivier, and M. Tamisier. Pseudo-divergence and bidimensional histogram of spectral differences for hyperspectral image processing. Journal of Imaging Science and Technology, 60(5):1–13, 2016.

[63] A. S. Robert. Chapter 5 - spectral transforms. In A. S. Robert, editor, Remote Sensing (Third Edition), pages 183–228. Academic Press, Burlington, third edition edition, 2007. ISBN 978-0-12-369407-2. doi: https://doi.org/10.1016/B978-012369407-2/50008-5. URL https://www.sciencedirect.com/science/article/pii/B9780123694072500085.

[64] A. Rosenfeld. Visual texture analysis: An overview. 1975.

[65] G. Sapiro. Color snakes. Computer Vision and Image Understanding, 68(2):247–253, 1997. ISSN 1077-3142. doi: https://doi.org/10.1006/cviu.1997.0562. URL https://www.sciencedirect.com/science/article/pii/S1077314297905622.

[66] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In 2011 31st International Conference on Distributed Computing Systems Workshops, pages 166–171, 2011. doi: 10.1109/ICDCSW.2011.20.

[67] C. Schmid. Constructing models for content-based image retrieval. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 2, pages II–II, 2001. doi: 10.1109/CVPR.2001.990922.

[68] C. E. Shannon. A mathematical theory of communication. The Bell System Technical journal, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[69] Steven K. Shevell. 4 - color appearance. In S. K. Shevell, editor, The Science of Color (Second Edition), pages 149–190. Elsevier Science Ltd, Amsterdam, second edition edition, 2003. ISBN 978-0-444-51251-2. doi: https://doi.org/10.1016/B978-044451251-2/50005-2. URL https://www.sciencedirect.com/science/article/pii/B9780444512512500052.

[70] M. Shi and G. Healey. Hyperspectral texture recognition using a multiscale opponent representation. IEEE Transactions on Geoscience and Remote Sensing, 41(5):1090–1095, 2003. doi: 10.1109/TGRS.2003.811076.

[71] S. Simić, S. S. Alzaid, and H. Aydi. On the symmetrized s-divergence. Open Mathematics, 18(1):378–385, 2020.

[72] A. H. Sturges. The choice of a class interval. journal of the American Statistical Association, 21(153):65–66, 1926.

[73] H. Su, B. Zhao, Q. Du, P. Du, and Z. Xue. Multifeature dictionary learning for collaborative representation classification of hyperspectral imagery. IEEE Trans. Geosci. and Remote Sens., 56(4):2467–2484, 2018.

[74] W. Sun and Q. Du. Hyperspectral band selection: A review. IEEE Geoscience and Remote Sensing Magazine, 7(2):118–139, 2019.

[75] M. J. Swain and D. H. Ballard. Color indexing. International journal of Computer Vision, 7(1):11–32, 1991.

[76] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics, 8(6):460–473, 1978. doi: 10.1109/TSMC.1978.4309999.

[77] S. Tanimoto. An optimal algorithm for computing fourier texture descriptors. IEEE Transactions on Computers, 27(01):81–84, jan 1978. ISSN 1557-9956. doi: 10.1109/TC.1978.1674956.

[78] X. Tolsa. Principal values for the cauchy integral and rectifiability. Proceedings of the American Mathematical Society, 128, 01 2000. doi: 10.2307/119706.

[79] F. Tomita and S. Tsuji. Statistical Texture Analysis, pages 13–36. Springer US, Boston, MA, 1990.

[80] F. Tsai and J. Lai. Feature extraction of hyperspectral image cubes using three-dimensional gray-level cooccurrence. IEEE Trans. Geosci. and Remote Sens., 51 (6):3504–3513, 2013.

[81] M. Tuceryan and A. K. Jain. Texture Analysis, chapter 2, pages 235–276. 1993.

[82] Y. C. Tzeng, K. T. Fan, and K. S. Chen. A parallel differential box-counting algorithm applied to hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 9(2):272–276, 2012. doi: 10.1109/LGRS.2011.2166243.

[83] M. Unser. Sum and difference histograms for texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(1):118–125, 1986. doi: 10.1109/TPAMI.1986.4767760.

[84] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In Computer Vision — ECCV 2002, pages 255–271, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[85] F. M. Vilnrotter, R. Nevatia, and K. E. Price. Structural analysis of natural textures. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(1):76–89, 1986. doi: 10.1109/TPAMI.1986.4767754.

[86] J. S. Weszka, C. R. Dyer, and A. Rosenfeld. A comparative study of texture measures for terrain classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6 (4):269–285, 1976. doi: 10.1109/TSMC.1976.5408777.

[87] Z. Xue, J. Li, L. Cheng, and P. Du. Spectral–spatial classification of hyperspectral data via morphological component analysis-based image separation. IEEE Transactions on Geoscience and Remote Sensing, 53(1):70–84, 2015. doi: 10.1109/TGRS.2014.2318332.

[88] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop, pages 147–149, 1992.

[89] S. Di Zenzo. A note on the gradient of a multi-image. Comput. Vis. Graph. Image Process., 33:116–125, 1986.

[90] E. Zhang, X. Zhang, H. Liu, and L. Jiao. Fast multifeature joint sparse representation for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 12(7):1397–1401, 2015. doi: 10.1109/LGRS.2015.2402971.

[91] L. Zhang, L. Zhang, D. Tao, and X. Huang. On combining multiple features for hyperspectral remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 50(3):879–893, 2012. doi: 10.1109/TGRS.2011.2162339.

[92] C. Zhou, B. Tu, N. Li, W. He, and A. Plaza. Structure-aware multikernel learning for hyperspectral image classification. IEEE journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14:9837–9854, 2021. doi: 10.1109/JSTARS.2021.3111740.