Illumination Invariant Deep Learning for Hyperspectral Data

Lloyd Windrim

A thesis submitted in fulfillment
of the requirements of the degree of
Doctor of Philosophy

THE UNIVERSITY OF
SYDNEY

Australian Centre for Field Robotics
School of Aerospace, Mechanical and Mechatronic Engineering
The University of Sydney

Submitted January 2018; revised August 2018

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the University or other institute of higher learning, except where due acknowledgement has been made in the text.

Lloyd Windrim

August 2018

Abstract

Motivated by the variability in hyperspectral images due to illumination and the difficulty in acquiring labelled data, this thesis proposes different approaches for learning illumination invariant feature representations and classification models for hyperspectral data captured outdoors, under natural sunlight. The approaches integrate domain knowledge into learning algorithms and hence does not rely on *a priori* knowledge of atmospheric parameters, additional sensors or large amounts of labelled training data.

Hyperspectral sensors record rich semantic information from a scene, making them useful for robotics or remote sensing applications where perception systems are used to gain an understanding of the scene. Images recorded by hyperspectral sensors can, however, be affected to varying degrees by intrinsic factors relating to the sensor itself (keystone, smile, noise, particularly at the limits of the sensed spectral range) but also by extrinsic factors such as the way the scene is illuminated. The appearance of the scene in the image is tied to the incident illumination which is dependent on variables such as the position of the sun, geometry of the surface and the prevailing atmospheric conditions. Effects like shadows can make the appearance and spectral characteristics of identical materials to be significantly different. This degrades the performance of high-level algorithms that use hyperspectral data, such as those that do classification and clustering.

If sufficient training data is available, learning algorithms such as neural networks can capture variability in the scene appearance and be trained to compensate for it. Learning algorithms are advantageous for this task because they do not require *a priori* knowledge of the prevailing atmospheric conditions or data from additional sensors. Labelling of hyperspectral data is, however, difficult and time-consuming, so acquiring enough labelled samples for the learning algorithm to adequately capture the scene appearance is challenging. Hence, there is a need for the development of techniques that are invariant to the effects of illumination that do not require large amounts of labelled data.

In this thesis, an approach to learning a representation of hyperspectral data that is invariant to the effects of illumination is proposed. This approach combines a

physics-based model of the illumination process with an unsupervised deep learning algorithm, and thus requires no labelled data. Datasets that vary both temporally and spatially are used to compare the proposed approach to other similar state-of-the-art techniques. The results show that the learnt representation is more invariant to shadows in the image and to variations in brightness due to changes in the scene topography or position of the sun in the sky. The results also show that a supervised classifier can predict class labels more accurately and more consistently across time when images are represented using the proposed method.

Additionally, this thesis proposes methods to train supervised classification models to be more robust to variations in illumination where only limited amounts of labelled data are available. The transfer of knowledge from well-labelled datasets to poorly labelled datasets for classification is investigated. A method is also proposed for enabling small amounts of labelled samples to capture the variability in spectra across the scene. These samples are then used to train a classifier to be robust to the variability in the data caused by variations in illumination. The results show that these approaches make convolutional neural network classifiers more robust and achieve better performance when there is limited labelled training data.

A case study is presented where a pipeline is proposed that incorporates the methods proposed in this thesis for learning robust feature representations and classification models. A scene is clustered using no labelled data. The results show that the pipeline groups the data into clusters that are consistent with the spatial distribution of the classes in the scene as determined from ground truth.

Acknowledgements

I would like to thank my supervisors Arman Melkumyan, Richard Murphy and Anna Chlingaryan for their invaluable guidance over the last few years. Between the three of you there was such a wide breadth of knowledge and I have always felt wiser after my discussions with you. Thank you for all of the advice you gave me - whether it was related to thesis work, research philosophy, careers or anything in general.

A special thank you to my friend Rishi Ramakrishnan who has been my unofficial mentor since I was an undergraduate. You were the one who got me interested in the research topics I chose to pursue. I am infinitely grateful for the great many things I have learnt from you over the years, and for keeping me sane throughout the process (you were my go to for procrastination).

Thank you to my parents and sister for always pushing me, supporting me and checking up on my progress. It is thanks to you that I am where I am today. Thank you to Abbey for all of her love and support, and for helping me to maintain balance.

Thank you to everyone at the Australian Centre for Field Robotics and Sydney University. In particular, thank you to Andrew, Zach, Victor, Rishi, Tatsumi, Phil, Nader, Troy, Eric, Alex, Suchet, Jono, Steve and John for all of the lunches, coffee and discussions. I am also very grateful for all of the financial support and resources provided by the Rio Tinto Centre for Mining Automation, without which this work would not have been possible. Thank you to Steve Scheding and Salah Sukkarieh for your great leadership.

Thank you to the rest of my friends and family for all of the positive distractions.

A special thank you to my original supervisor Juan Nieto for opening the doors to the world of academia to me. You took me on as your student and in a short amount of time instilled in me a great research philosophy.

Contents

List of Figures

List of Tables

List of Algorithms

Nomenclature

List of Symbols

$a_i^{(l)}$	activation of neuron i in layer l
$b_j^{(l)}$	bias of neuron j in layer *l+1*
C_{ml}	number of rows in each kernel in the l-th convolutional layer
C_{nl}	number of columns in each kernel in the l-th convolutional layer
C_{dl}	number of dimensions in each kernel in the l-th convolutional layer
E	cost function for neural network
$\mathbf{E}_{ind}$	indirect illumination irradiance spectrum (function of wavelength)
$\mathbf{E}_{sky}$	diffuse skylight irradiance spectrum (function of wavelength)
$\mathbf{E}_{sun}$	extraterrestrial sunlight irradiance spectrum (function of wavelength)
$\mathbf{E}_{sun}\boldsymbol{\tau}$	terrestrial sunlight irradiance spectrum (function of wavelength)
$f(\cdot)$	activation function
K	number of spectral channels
$\mathbf{L}$	radiance or digital number of spectra (function of wavelength)
l	layer number of a neural network
V	line-of-sight visibility between region and sun
$W_{ji}^{(l)}$	the weight connecting neuron i in layer l with neuron j in layer *l+1*
X	log-chromaticity
x_i	feature value i of the input to the neural network
y_i	feature value i of the target of the neural network
$z_i^{(l)}$	weighted sum of neurons in layer *(l-1)* and the bias term going into the i-th neuron in layer l
$\boldsymbol{\alpha}$	learning rate
Γ	sky (or view) factor
$\boldsymbol{\omega}$	direction of the illumination invariant axis in log-chromaticity space
$\boldsymbol{\rho}$	albedo (function of wavelength)
$\boldsymbol{\tau}$	solar path transmittance (function of wavelength)

θ	angle between the surface normal and vector towards the sun

List of Acronyms

ARI	adjusted rand index
AVIRIS	airborne visible/infrared imaging spectrometer
CNN	convolutional neural network
CRF	conditional random field
CSA-SAE	cosine spectral angle-stacked autoencoder
DAE	denoising autoencoder
DN	digital number
FA	factor analysis
GP	Gaussian process
GPS	global positioning system
GPU	graphics processing unit
IARR	internal average relative reflectance
ICA	independent component analysis
ILSVRC	ImageNet large scale visual recognition challenge
KNN	k-nearest-neighbour
KSC	Kennedy space station
LDA	linear discriminant analysis
LiDAR	light detection and ranging *(also commonly known as a 'laser range scanner')*
LLE	local linear embedding
MLP	multi-layer perceptron
MSE	mean squared error
PCA	principal component analysis
PSNR	peak signal-to-noise ratio
ReLU	rectified linear unit
ROSIS-3	reflective optics system imaging spectrometer
RSA-SAE	relit spectral angle-stacked autoencoder
SA-SAE	spectral angle-stacked autoencoder
SAE	stacked autoencoder
SAM	spectral angle mapper
SID	spectral information divergence
SID-SAE	spectral information divergence-stacked autoencoder
SSE-SAE	sum of squared errors-stacked autoencoder
SWIR	short-wave infrared
SVM	support vector machine
USGS	United States geological survey
VIS	visible
VNIR	visible and near infrared

Glossary

Autoencoder: A type of neural network which learns to reconstruct an input in the output layer.

Convolutional Neural Network: A type of neural network which learns localised kernels which convolve over the data.

Diffuse: A material that reflects light uniformly in all directions. Also called Lambertian.

Diffuse Skylight: Extraterrestrial sunlight that has been scattered by particles in the atmosphere. Predominantly blue in colour.

Digital Number: The units of a pixels intensity as measured by a sensor.

Incident Illumination: The light that illuminates a region in the scene.

Neural Network: A computational network of mathematical units capable of learning a mapping from an input to an output. Related to the field of deep learning.

Radiometric Normalisation: The process of converting the digital number of each pixel to reflectance.

Sky Factor: Proportion of the sky dome hemisphere that is visible from a region in the scene.

Terrestrial Sunlight: Extraterrestrial sunlight that has not been scattered by particles in the atmosphere and reaches the earths surface.

Chapter 1

Introduction

The use of hyperspectral data in supervised applications is constrained by the lack of labelled training data. Thus, the aim of this thesis is to develop illumination invariant representations and classification models for outdoor hyperspectral data that use limited or no labelled training samples. When collected in the outdoor environment, where the light source is the sun, much of the variability in hyperspectral data spanning the visible to short-wave infrared (SWIR) wavelengths is related to how the scene is illuminated and collecting and annotating enough data to capture this variability is difficult. Through the use of learning algorithms that leverage models of the physical processes involved with scene illumination, this thesis presents an approach to robustly representing and classifying hyperspectral data acquired under natural light with little to no annotated training data.

1.1 Motivation

Sensors which perceive the environment provide a means in which machines can quantify and mimic human understanding of the physical world. The perceptual data collected from sensors on-board robots and remote-sensing platforms provide a wealth of information which can be harvested and interpreted in order to carry out high-level processes such as classification (e.g. Camps-Valls et al. (2014)), odometry (e.g. Nistér

et al. (2004)), planning (e.g. Baltzakis et al. (2003)), mapping (e.g. Se et al. (2002)), detection (e.g. Ren et al. (2015)) and recognition (e.g. Turk and Pentland (1991)). Imaging sensors in particular have received considerable attention in the development of these platforms due to their ability to non-destructively capture information at high spatial resolution and their ability to work from a variety of viewing distances and under a variety of different environmental conditions (e.g. rain). Hyperspectral sensors are a particularly powerful variety of imaging sensor that are capable of measuring the spectral reflectance of objects in numerous, contiguous band-passes in the visible to SWIR range. Unlike multispectral sensors which measure reflectance in a small number of broad, discontiguous bands of variable width, hyperspectral sensors capture the entire spectral curve, seamlessly across the observed range. This makes it possible to resolve the shape, intensity and wavelength location of absorptions within the spectrum that are relevant to a broad range of quantitative tasks including geological mapping, agriculture and defence (Schowengerdt, 2007).

A major problem with using an imaging system operating outdoors, in particular a hyperspectral imaging system, is that the appearance of the scene is highly variable (Ramakrishnan, 2016). When capturing the appearance of a scene, variations (e.g. in brightness) are translated to the image resulting in degradations in the performance of high-level tasks using the image data (Corke et al., 2013). This is because objects and materials that are intrinsically similar can appear to be differently in the image and objects and materials that are intrinsically different can appear to be similarly in the image (Figure 1.1). The appearance of a scene is determined largely by the ways in which light interacts with the environment. Consequently, along with non-Lambertian factors, one of the largest sources of variability in scene appearance is related to the variability of the incident illumination.

The appearance of the scene as measured by a camera is based on how the scene reflects incoming light, that is, the incident illumination. When light is emitted from the sun it transmits through the atmosphere and is absorbed and scattered across certain wavelengths by atmospheric gasses, water and other particles (Schowengerdt, 2007). This modifies the colour temperature and intensity of the spectrum of incident

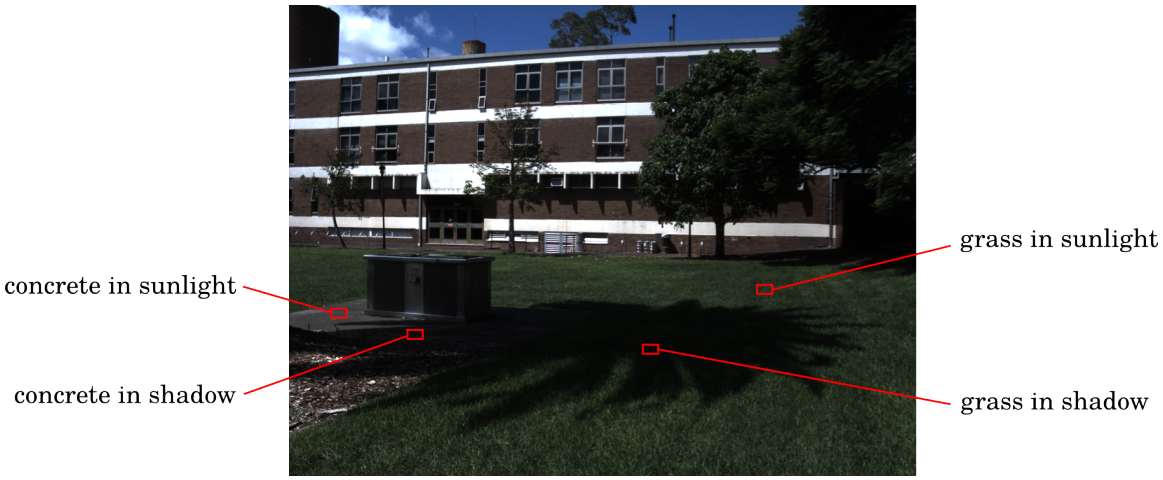


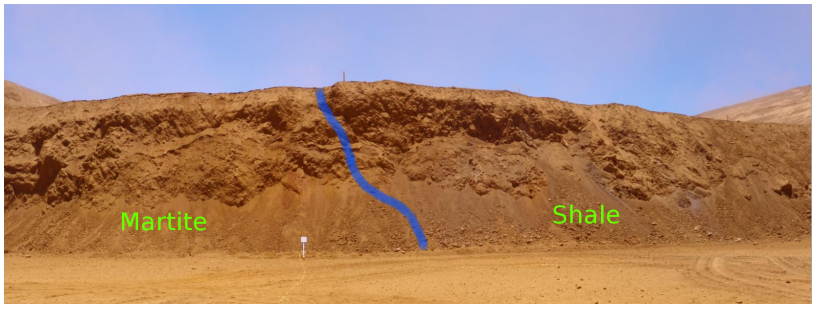
Figure 1.1 – Example of illumination variability in an RGB image. Scene constituents that are semantically similar appear differently under shadow and sunlight (e.g. the grass), and constituents that are semantically different appear similarly in shadow (e.g. the grass and concrete).

light based on the ambient atmospheric composition. Hence, obtaining measurements under consistent conditions over long periods of time is difficult due to the variability in the primary light source illuminating the scene (i.e. sunlight).

Additionally, there are factors which alter how the incident light illuminates the scene in a single image. The spectrum of the incident light source is modified by the geometry of the scene (Schowengerdt, 2007). The angle at which the illuminant strikes the surface effects the intensity at which it is reflected back (Hapke, 1981). Parts of the scene may also occlude other regions of the scene from the primary light source. However, these occluded regions may still be illuminated by indirect sources of light, such as rays from the primary source that are reflected off of the sky and other materials in the scene. This phenomenon results in shadows. The indirect incident illuminant bears the wavelength-intensity distribution of light of the materials from which it is reflected and hence differs significantly from the primary light source. This effect is also possible when there are clouds occluding the primary light source. The amount of sky visible to the surface influences how much indirect skylight it is illuminated by.

Thus, the appearance of the scene is dependent on the atmospheric composition and

(a) Almonds in a tree (Bargoti and Underwood, 2017).



(b) A mine face containing Martite and Shale (Murphy et al., 2012).

Figure 1.2 – It is difficult to label the constituents of both of these images.

the azimuth and elevation of the sun, which can vary over time, the geometry of the scene, which varies across a single image, and the material properties of elements of the scene. Estimating material properties is valuable for use in high-level algorithms, however, isolating them from the other factors of variation remains an ill-posed problem due to the number of unknown variables and complex ways in which they interact. If high-level algorithms using hyperspectral data are to work reliably in the outdoors, it is important to find representations and classification models that are robust to these complex interactions.

As the development of learning algorithms moves forward at a rapid pace, a trend is emerging in their application to image data. There is a strong reliance on using large datasets to teach learning algorithms to solve problems. For example, very

Figure 1.3 – Use of a spectrometer in the field to obtain ground truth information.

good classification results can be achieved by training on a very large dataset which captures all of the variability in the data (e.g. Krizhevsky and Hinton (2012), Szegedy et al. (2015)). In doing so, it is less tempting to incorporate domain knowledge into state-of-the-art techniques, for example, physics-based models that have been developed over many years which mathematically model the interactions of light with the environment (e.g. Hapke (1981)). The strong reliance on large datasets is only feasible when large labelled datasets are available for learning and unfortunately for applications which use hyperspectral data, labelled data is limited due to the difficulty in acquiring it.

Once captured, there are multiple ways of assigning labels to hyperspectral data (i.e. annotating pixel data). One approach is by visually inspecting RGB colour imagery, individual bands or false-colour imagery constructed from other wavelengths beyond the visible range and then manually labelling regions of pixels. In many scenarios, particularly when there is a high degree of variation in illumination, this can be challenging because different materials can appear similar (Figure 1.1). For example, in an agriculture context, labelling green apples, avocados or almonds in a tree is problematic due to their similar appearance to leaves, both in colour and texture (Figure 1.2a). Another example is in an open pit mining operation where different materials can appear very similar (Murphy et al., 2014a). In Figure 1.2b, it is very difficult to distinguish between the shale region and the martite region of the mine face. An alternative to labelling regions of pixels by inspection of the image is to look at the spectral information of individual pixels, and have an expert annotate them. The spectral information is much more informative of the material's class, but it is

challenging and time-consuming to label individual pixels. In reality, this is because the pixel spectra from the image are often affected by illumination, noise and spectral mixing among other things. The spectra that are the most unrecognisable, such as the ones in shadow, can be difficult to label. However, these examples are necessary to train the classifier to be robust to the sorts of effects expected in the image. For these reasons it is usually necessary to take a field spectrometer into the environment (Figure 1.3) and measure the spectra of individual targets at high spectral resolution, and then tag the location or target so that it can be matched to the image. The targets can be artificially illuminated and spectra collected are generally less noisy than there image counterparts making it easier to label, however, collecting the readings is extremely laborious because of the difficulties in aligning the spatial scale of the samples with the spatial scale of the image, co-locating the measurements with pixels in the image and collecting a large number of measurements at the sufficient spatial resolution. In environments such mine pits, interacting with the scene can also be hazardous with the possibility of rock falls, explosives and heavy machinery in the environment. In summary, annotating hyperspectral data is difficult, thus classifiers and feature learners are often limited to using small amounts of labelled training data that do not accurately capture the variability of the scene. If the labelled data does not capture the large and complex variability in the outdoor scene, then it cannot be modelled by learning algorithms for applications which use hyperspectral data. Because of the difficulty in labelling hyperspectral data for training, there is a need to develop methods which are robust to large variations in hyperspectral data due to illumination that do not require a large amount of labelled data.

1.2 Contributions of Thesis

In this thesis, an approach is proposed for learning illumination invariant representations and classification models for outdoor hyperspectral data. The approach leverage's domain knowledge in the form of physics-based models, and learns the elements to the problem that are unknown. By integrating domain knowledge into learning al-

gorithms (specifically, deep learning algorithms), fewer labelled samples are required for learning, but the representations and classification models can still account for large amounts of variation in the data. Specifically, the illumination effects targeted are the influence of the geometry of the surface with respect to the sun and sky on the incident illumination, including variations in brightness and shadows caused by occlusions from the primary light source.

The specific contributions are:

- A set of unsupervised approaches to learning feature representations specifically designed for hyperspectral data. The approaches use remote sensing methods to improve a deep learning technique, and the representations are designed to be invariant to variations in the intensity (i.e. brightness) of the incident illuminant and improve performance when scenes have a variable surface geometry. These techniques are useful for dimensionality reduction or feature extraction.

- An extended unsupervised approach to learning a low-dimensional, illumination invariant feature representation of hyperspectral data. This representation retains the discriminative information of the original spectra, but is also invariant to shadows as well as illumination intensity effects. Hence, materials in an image that are the same will appear the same, regardless of the illumination effects. The technique requires no labelled data, *a priori* knowledge or additional sensors and the representation is useful for classification, clustering or any high-level task where robustness to the illumination is a desirable property.

- A transfer learning scheme for utilising training samples from well-labelled hyperspectral images in order to pre-train a classifier for better performance on poorly labelled images. This is useful for learning better spectral features when there are limited labelled training samples available. The pre-trained features are also faster to train. Using this scheme, it is possible to transfer knowledge from airborne datasets, of which there exist well-labelled, public datasets, to field-based datasets, which typically must be labelled from scratch.

- An image-based approach for estimating the terrestrial sunlight-diffuse skylight ratio. This is important for spectral relighting (e.g. Marschner and Greenberg (1997)), which involves altering the appearance of an image such that it is illuminated differently. Relighting is used heavily in the context of this thesis.
- A relighting-based approach to learning a robust classification model using limited labelled training data. This classification model can predict labels for each pixel in a hyperspectral image, given a training set where the labels have poor spatial coverage of the scene. Samples acquired from localised, sunlit regions can train the classifier to accurately predict labels for regions with large variations in geometric orientation and regions in shadow.

1.3 Publications

The work in this thesis has led to the following publications:

- Lloyd Windrim, Arman Melkumyan, Richard Murphy, Anna Chlingaryan and Juan Nieto. Unsupervised Feature Learning for Illumination Robustness. In *Proceedings of the International Conference on Image Processing*, pages 4453-4457, IEEE, 2016. - This paper presents one of the unsupervised approaches proposed in Chapter 4 for learning feature representations specifically for hyperspectral data that are invariant to the illumination intensity.
- Lloyd Windrim, Rishi Ramakrishnan, Arman Melkumyan and Richard Murphy. A Physics-based Deep Learning Approach to Shadow Invariant Representations of Hyperspectral Images. In *Transactions on Image Processing* 27.2(2018):665-677, IEEE. - This paper extends the unsupervised approach proposed in Chapter 4 for learning hyperspectral feature representations that are invariant to both shadows and the illumination intensity.
- Lloyd Windrim, Rishi Ramakrishnan, Arman Melkumyan and Richard Murphy. Hyperspectral CNN classification with Limited Training Samples. In *Proceedings of the British Machine Vision Conference*, pages 2.1-2.12, BMVA Press,

2017. - This paper presents the image-based approach to extracting the terrestrial sunlight-diffuse skylight ratio as well as the relighting-based approach to learning a robust hyperspectral classification model when there is limited labelled training samples, proposed in Chapter 5.

- Lloyd Windrim, Arman Melkumyan, Richard Murphy, Anna Chlingaryan and Rishi Ramakrishnan. Pretraining for Hyperspectral Convolutional Neural Network Classification. In *Transactions on Geoscience and Remote Sensing*, IEEE. Accepted for inclusion in a future issue. Pre-print available as of 03 January 2018. - This paper presents the transfer learning scheme proposed in Chapter 5.

1.4 Structure of Thesis

This thesis is structured as follows:

Chapter 2 introduces hyperspectral sensing, the outdoor illumination model and associated relighting equations and deep learning, followed by an overview of the relevant literature. The literature covered includes how illumination invariance has been approached using methods from the fields of computer vision, remote-sensing, multi-modal sensing and statistical learning, with a focus on approaches for hyperspectral imagery. The literature tracking the progress of the convolutional neural network (CNN) is presented, along with a review of how deep learning has been utilised for hyperspectral applications. Finally, the use of data augmentation in machine learning is briefly reviewed.

Chapter 3 presents the datasets and evaluation metrics used throughout the thesis.

Chapter 4 proposes unsupervised approaches to learning low-dimensional, illumination invariant feature representations for hyperspectral images. The approaches are extensively evaluated and the results and their implications are presented and analysed.

Chapter 5 proposes a transfer learning scheme, image-based approach to estimating the terrestrial sunlight-diffuse skylight ratio and the relighting-based approach to

learning robust hyperspectral classification models. This is for the scenario where there is limited labelled training data. The results of rigorous evaluation of these approaches are presented and analysed.

Chapter 6 develops a pipeline for a case study which utilises the proposed unsupervised feature representation and robust classification model to cluster a hyperspectral image without using any labelled data. The performance of the pipeline is discussed.

Chapter 7 discusses the conclusions of this thesis and future work.

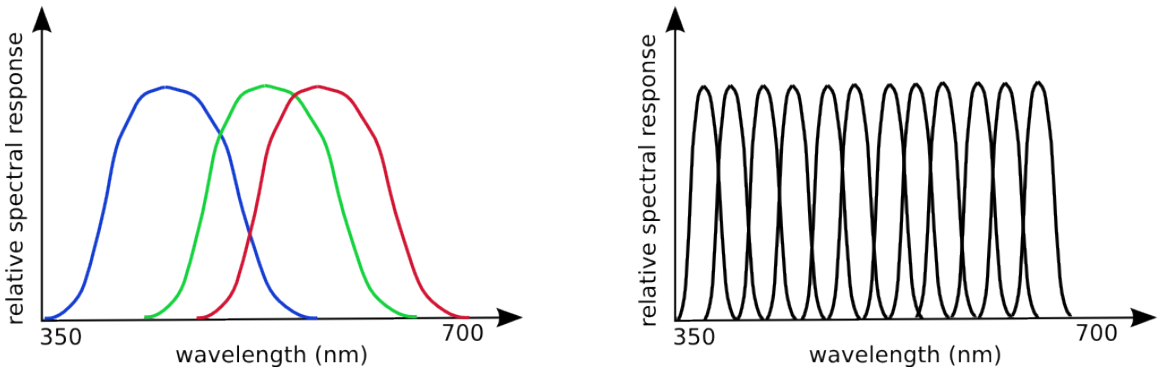
Chapter 2

Background

This chapter presents an overview of the theory most relevant to this thesis, followed by a summary of the literature. The relevant theory includes the basic principles of hyperspectral imagery, the illumination model and relighting that is used throughout this thesis, and deep learning, including the workings of an autoencoder and CNN. The literature addressing the problem of variability in illumination is reviewed, followed by the progression of deep learning and how it has been used for hyperspectral applications.

2.1 Hyperspectral Sensing

A digital RGB camera captures a high spatial resolution image over only three channels, where each channel consists of an integration of the incident light over many wavelengths using a wide-band filter. Conversely, hyperspectral sensors capture high spatial resolution images with numerous channels (tens to hundreds) each covering a narrow range of wavelengths (Figure 2.1). Each pixel in a hyperspectral image contains the spectrum of the material it captures across the visible, near infrared and SWIR portions of the electromagnetic spectrum (e.g. in the range 400 nm to 2500 nm). Hence, each pixel spectrum provides a more detailed measurement of the reflectance behaviour of a material. Because reflectance is measured across narrow



(a) RGB sensor response. **(b)** Hyperspectral sensor response.

Figure 2.1 – An example of the differences between the sensor response of an RGB camera and a hyperspectral camera in terms of the width and number of filters.

intervals of wavelength, subtle changes in reflectance can be detected that would be impossible to pick up with an RGB camera.

A hyperspectral image can be conceptualised as a cube, with two dimensions attributing to the x and y dimensions of the scene and the third dimension formed by many different wavelengths sequentially stacked one upon the other (Figure 2.2).

Hyperspectral images are typically acquired from sensors on-board airborne platforms such as aircraft and balloons and more recently data has been acquired from field-based platforms such as ground-based systems and robots. The higher the altitude of the sensor, the larger the area of ground described by each pixel is and thus the easier it is acquire scans with greater swath. However, as the pixel size increases, the spatial resolution decreases. If multiple spectrally distinct materials present in the scene are mixed together in a pattern that is too fine to be resolved by the spatial resolution of the sensor, then the pixels of the cube will be a composite (sum) of the reflectance signals from all of these materials. This process is called spectral mixing (Bioucas-dias et al., 2012; Mustard and Pieters, 1987). It is particularly prominent in hyperspectral data collected from airborne or spaceborne platforms because of the larger pixel sizes. There are also numerous ways in which environmental factors can affect the signal detected by hyperspectral sensors including the atmospheric conditions, the geometry of the scene and the season.

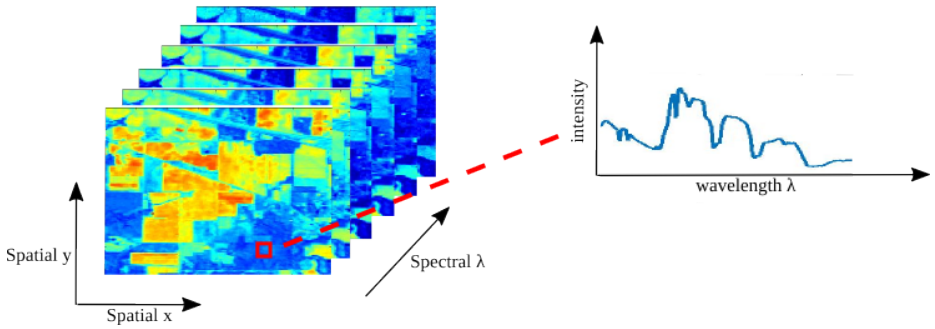


Figure 2.2 – A hyperspectral image can be conceptualised as a hypercube with two spatial dimensions and a spectral dimension.

Due to the high spatial and spectral resolution of hyperspectral images, it is possible to identify surface materials from a scene through an analysis of their diagnostic spectral features. These diagnostic spectral absorption features appear due to processes relating to the chemical and structural composition of each material. This allows hyperspectral cameras to be used in many diverse applications such as food safety (Elmasry et al., 2012; Feng and Sun, 2012), military surveillance (Eismann et al., 1996; Stein et al., 2001; Yuen and Richardson, 2010), medical imaging (Lu and Fei, 2014), environmental monitoring (Adam et al., 2010; Govender et al., 2007), geological mapping (Murphy et al., 2012; Van der Meer et al., 2017) and precision agriculture (Haboudane et al., 2002).

2.1.1 Classification

Using classification algorithms, it is possible to automate the process of surface material identification of a scene from a hyperspectral image (Camps-Valls et al., 2014). The advantage of hyperspectral imagery is that each pixel contains sufficient information so that it is possible to classify a scene at the pixel level using these algorithms. There are a number of ways in which this is typically done.

One way is to build up a library of reference spectra, ideally collected under laboratory conditions using a non-imaging spectrometer. If the library contains several entries per class, then the pixel spectra in an image can be matched to the most similar entry in this library (Murphy et al., 2012). This determines its class. There are

many ways to match the spectra to the reference library, but the most common way is the spectral angle mapper (SAM) (Kruse et al., 1993; Yuhas et al., 1992). SAM computes the inverse cosine of the normalised dot product between the target spectra and all reference spectra in the library. A pixel is assigned the class of a reference spectra if it lies within a certain angular threshold (normally expressed in radians). Alternatively, a spectrum is assigned the class of the reference spectrum with which it has the smallest angle. The advantage of SAM is that it is robust to scaler multiples of the spectra that are constant across the wavelength, which can occur due to differences in brightness (Hecker et al., 2008).

Another approach is to collect training spectra from the image being examined, either as pixels in the image or samples collected from the scene that are scanned separately, annotate them and train a supervised classification model which can be used for inference. There are many methods for training a classification model, including a support vector machine (SVM) (Melgani and Bruzzone, 2004), k-nearest-neighbour (KNN) (Yang et al., 2010), Gaussian process (GP) (Schneider et al., 2010), decision tree (Ham et al., 2005) and neural networks (Ratle et al., 2009). Once the model has been trained to sufficiently minimise a loss function using enough samples to capture the variability of each class in the training set, the model can be used to predict the class label of new data drawn from the same distribution as the training data.

Typically, to use the second approach, a small fraction of the pixels from each class in the scene to be classified are annotated and used to train the classifier. Then, predictions for the rest of the pixels in the image are computed using the classifier.

The classifier requires each input pixel be described by a set of features. Features are the way in which the hyperspectral data are represented. The goal of the features is to make high-level tasks such as classification more efficient by maximising the separation between classes. It is possible to simply use the intensity or reflectance at each wavelength as the feature space. However, this is very simplistic and can often lead to problems with overfitting due to the dimensionality of the data (Bishop, 2006). Also, there is a lot of redundancy in this representation of the data due to the high degree of correlation between the channels (Demarchi et al., 2014). Traditionally,

it has been popular to use hand-crafted features which are designed using domain knowledge of the data. For example, layers of clay are mapped on a mine face using the width and depth of spectral absorption features at 2200 nm (Murphy et al., 2014a). It is often difficult, time-consuming and requires sufficient expertise to hand-craft these features, which is why an alternative approach is to use feature learning. Feature learning is an autonomous means of extracting features from the data. Methods for learning features can be supervised such as linear discriminant analysis (LDA) (Du, 2007) and kernel methods (Kuo et al., 2009). There are also unsupervised methods such as principal component analysis (PCA) (Cheriyadat and Bruce, 2003; Rodarmel and Shan, 2002) or independent component analysis (ICA) (Chiang et al., 2000). As an alternative to feature extraction, there are also feature selection techniques which find a subset of bands to use as features which optimise some criteria, such as class separation (Backer et al., 2005). Dimensionality reduction techniques (e.g. PCA) can be interpreted as feature extraction/selection methods as they are finding an abstraction of the raw data that is useful for classification.

Many of the hyperspectral classifiers mentioned above only use spectral information. However, many modern classification techniques also use the spatial dimension of the hyperspectral image. The spatial dimension provides contextual information about a pixel, which is usually correlated with its material class. A variety of approaches to spectral-spatial classification have been proposed. Some approaches combine morphological features with spectral vectors (Fauvel et al., 2012, 2008). Other approaches use segmentation to spatially regularise a pixel-wise classification map. In Tarabalka et al. (2009), segmentation using partial clustering is combined with a spectral SVM classifier using majority voting. Similarly, watershed segmentation was used in Tarabalka et al. (2010). The classification maps obtained using spectral-spatial classification methods are usually more homogeneous than those obtained from purely spectral-based methods.

2.1.2 Dimensionality Reduction

Dimensionality reduction is the statistical process of reducing the number of dimensions (or variables) required to describe a dataset. Hyperspectral datasets are of high dimensionality as the reflectance at each wavelength for a single pixel can be interpreted as a separate dimension. The high dimensionality of hyperspectral data is linked to problems such as the curse of dimensionality (Donoho et al., 2000; Hughes, 1968; Lee and Landgrebe, 1993). With more dimensions, exponentially more data points are required to accurately represent the data's distribution. Hence, a low ratio between the number of data points and the number of dimensions may limit many algorithms from working well where the data has a large number of dimensions. Most of the mass of a high dimensional multivariate Gaussian distribution is near its edges. Thus, many algorithms designed around an intuitive idea of 'distance' in two or three dimensional space, cease to work at higher dimensions, where those intuitions no longer hold. Other problems with having a high number of dimensions are the high storage requirements for the data and slow processing speeds (Bioucas-dias et al., 2012). For these reasons, many algorithms perform poorly when the data has too many dimensions due to the increased complexity of the task (Pal and Foody, 2010). Hyperdimensionality also makes it difficult to visualise the data without advanced software tools. As previously mentioned, because proximal wavelengths in hyperspectral data are often highly correlated, there can be redundant information in the datacube. Therefore, it makes sense to reduce the number of dimensions in the data without loss of information.

The goal of dimensionality reduction is to compress the data whilst preserving all of its relevant information. As these criteria are also important for finding features, there is significant overlap in the techniques used for feature extraction/selection and dimensionality reduction. For this reason, many of the papers proposing dimensionality techniques for hyperspectral data use a classification application as a means of evaluating them. Besides the basic techniques (e.g. PCA, factor analysis (FA) and ICA) there are many dimensionality reduction techniques that have been appropriated for use with hyperspectral data, such as local linear embedding (LLE) (Chen and

Qian, 2007; Han and Goodenough, 2005; Kim and Finkel, 2003), ISOMAP (Guangjun, Dong Yongsheng and Song, 2007; Sun et al., 2014) and Laplacian eigenmaps (Qian and Chen, 2007). Approaches also exist for dimensionality reduction of spectral-spatial features (Zhang et al., 2013). The need for reducing the dimensionality of hyperspectral data motivates the development of low-dimensional representations.

2.2 Outdoor Illumination Model and Relighting

Many of the techniques proposed in this thesis incorporate a model describing processes of illumination into deep neural networks. The model is described in this section. Firstly, the illumination sources assumed to be present in an outdoor scene are described. Then, the outdoor illumination model from which the relighting equations can be derived is explained.

2.2.1 Sources of Illumination

An outdoor scene has several sources of illumination, with the two most dominant sources being direct terrestrial sunlight and diffuse skylight (Gijsenij et al., 2012; Sato and Ikeuchi, 1995). The terrestrial sunlight source consists of the extraterrestrial light emitted from the sun that passes through earth's atmosphere. The light takes on a spherical pattern when it is first emitted, but by the time it reaches earth it has travelled such a long distance that is can be considered a parallel light source (Schowengerdt, 2007). When the sunlight passes through the atmosphere, it is absorbed and scattered across specific ranges of wavelengths. The proportion of light that penetrates through the atmosphere and reaches the earth's surface (called the solar path transmittance (Schowengerdt, 2007)) is wavelength dependent and extremely variable. This is the terrestrial sunlight. The diffuse skylight occurs due to the Rayleigh scattering of light by particles that are smaller than the wavelength of the visible light. This results in blue light being preferably scattered (during daylight hours). By assuming consistent cloud coverage, the sky is considered to be a dome,

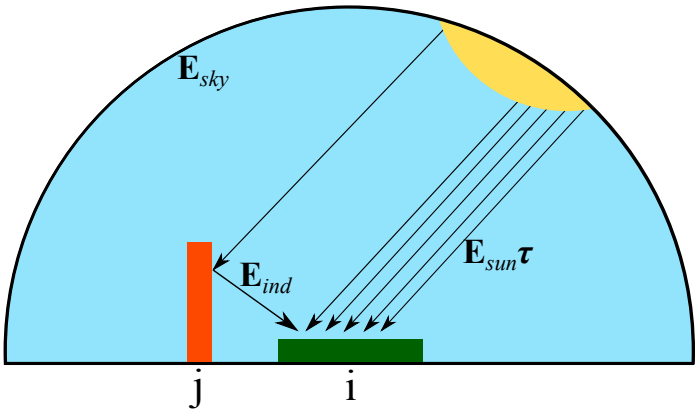
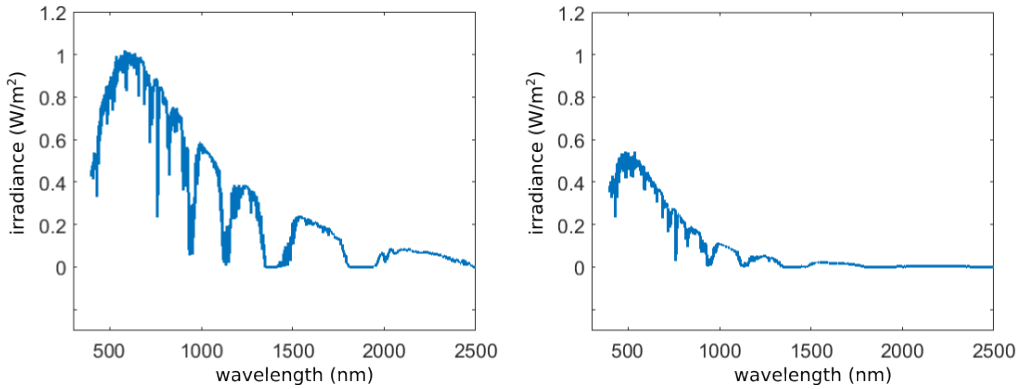


Figure 2.3 – The three sources of illumination assumed to be present in an outdoor scene, illuminating a region i. They are the terrestral sunlight $\mathbf{E}_{sun}\boldsymbol{\tau}$, the diffuse skylight $\mathbf{E}_{sky}$ and the indirect illumination $\mathbf{E}_{ind}$ from the region j.



(a) Terrestrial sunlight $\mathbf{E}_{sun}\boldsymbol{\tau}$. **(b)** Diffuse skylight $\mathbf{E}_{sky}$.

Figure 2.4 – Example of differences in intensity and spectral power distribution of terrestrial sunlight and diffuse skylight as simulated by an atmospheric modeller (Gueymard, 2001).

and the diffuse skylight is modelled as a hemispherical light source (Sato and Ikeuchi, 1995). The intensity and spectral power distribution of the terrestrial sunlight differs significantly from the diffuse skylight, as seen in the example in Figure 2.4. Additional to these sources of illumination, there is indirect illumination. Light that is reflected off of surfaces in the scene can illuminate other surfaces, but with a much weaker intensity. The sources of illumination in outdoor scenes are shown in Figure 2.3.

2.2.2 Physics-Based Illumination Model

The following outdoor illumination model (Ramakrishnan, 2016) for the radiance of spectra reflected from a scene consists of the parallel, terrestrial sunlight source $\mathbf{E}_{sun}\boldsymbol{\tau}$, hemispherical diffuse skylight source $\mathbf{E}_{sky}$ and indirect illumination source $\mathbf{E}_{ind}$ each described above. Assuming all materials in the scene diffusely reflect light, the radiance L of a region i as captured by a camera can be approximated as:

$$L_i(\lambda) = L_{e,i}(\lambda) + \frac{\rho_i(\lambda)}{\pi}\left[V_i E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda) + E_{ind}\right], \tag{2.1}$$

where ρ_i is the albedo of the material, V_i is a binary variable indicating whether there is line-of-sight visibility between the region and the sun position, τ is the solar path transmittance, $L_{e,i}$ is the emitted radiance, θ_i is the angle between the surface normal and the line-of-sight vector towards the sun, and Γ_i is the sky (or view) factor ranging from 0 to 1 indicating the portion of the sky dome that is visible. The emitted radiance occurs when objects heat up and begin to glow. In this work, it is assumed for simplicity that there is no emission in the scene and that indirect illumination is negligible. Thus, the model simplifies to:

$$L_i(\lambda) = \frac{\rho_i(\lambda)}{\pi}\left[V_i E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)\right]. \tag{2.2}$$

Through inspection of this model, it can be seen that the spectral variability in the appearance of a material is related to the sources of illumination ($\mathbf{E}_{sun}$ and $\mathbf{E}_{sky}$), the geometric factors (V_i, θ_i and Γ_i) that control the intensity of the sources of illumination, as well as the atmospheric conditions ($\boldsymbol{\tau}$). The advantage of this model is that it allows each illumination source to be treated independently.

2.2.3 Relighting

Relighting is a technique for scaling the appearance of a material by a wavelength dependent function such that it appears to be under different illumination conditions.

Relighting has predominantly been used in the computer vision and remote sensing literature to relight color and spectral images (Beauchesne and Sbastien, 2003; Marschner and Greenberg, 1997; Ramakrishnan et al., 2015; Troccoli and Allen, 2005).

In Ramakrishnan (2016), relighting equations are derived from the model (2.2) for different scenarios. To relight the radiance $\mathbf{L}$ of a region i in sunlight with respect to diffuse skylight, the scaling factor is calculated as:

$$L_j(\lambda) = L_i(\lambda)\frac{1}{\frac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_i + \Gamma_i}. \tag{2.3}$$

Relighting a region to have only a diffuse skylight component is equivalent to relighting a region to be in shadow. Similarly, to relight the radiance $\mathbf{L}$ of a region i in sunlight with respect to full terrestrial sunlight and diffuse skylight exposure, the scaling factor is calculated as:

$$L_j(\lambda) = L_i(\lambda)\frac{\frac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)} + 1}{\frac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_i + \Gamma_i}. \tag{2.4}$$

This is equivalent to relighting a scene to be in sunlight, where the angle between the surface normal and the line-of-sight to the sun is zero, and the full sky dome is visible, producing maximum exposure. The derivation for these relighting equations is in Appendix D.

2.3 Deep Learning

Deep neural networks comprise a subset of machine learning algorithms, which learn parametrised models from data. Neural networks often have many parameters compared to most other machine learning algorithms, and hence require a significant amount of data to train them in order to learn generalisable models. Because of their

large number of parameters, multiple layers and non-linear components, neural networks can learn very powerful models, and thus have had success in a range of tasks including speech recognition (Dahl et al., 2012; Graves et al., 2013; Hinton et al., 2012), text recognition (Wang et al., 2012), digit recognition (LeCun et al., 1990) and object recognition (He et al., 2016; Simonyan and Zisserman, 2014). In many of these tasks, deep neural networks achieve state-of-the-art results on benchmark datasets, outperforming hand-crafted feature-based techniques. This section provides a brief theoretical background to deep neural networks.

2.3.1 Multi-layer Perceptron

The most simple type of multi-layered neural network is the multi-layer perceptron (MLP) (Ng, 2011), which learns a non-linear mapping from data at the input layer to values in the output layer. The output layer can be a layer of classification labels, but this is not always the case as will be explained in Section 2.3.2. The fundamental unit of an MLP is a neuron (Figure 2.5a). The neuron takes a set of inputs $\mathbf{x}$, computes a weighted addition of them with weights $\mathbf{W}$ and an additional bias term $\mathbf{b}$, and then passes the result through an activation function f to compute the output a as follows:

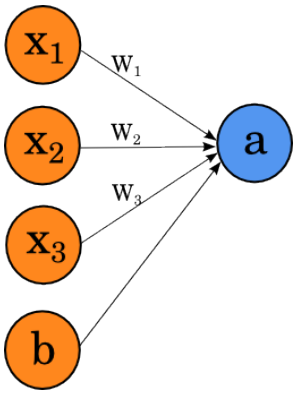
$$a = f(\sum_{i=1}^{N} W_i x_i + b), \tag{2.5}$$

where N is the number of inputs, and the activation function could be a sigmoid:

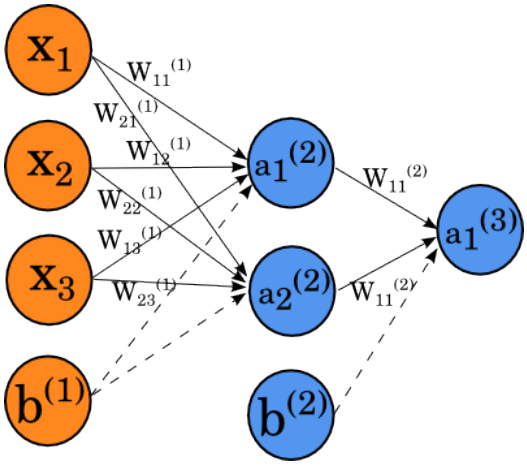
$$f(\omega) = \frac{1}{1 + \exp^{-\omega}}, \tag{2.6}$$

Although the activation function is not limited to a sigmoid. Several of these neurons form a layer, and several layers combine to form a network (Figure 2.5b). The value of each neuron is then calculated as:

$$a_1^{(2)} = f(\sum_{i}^{N} W_{1i}^{(1)} x_i + b_1^{(1)}), \tag{2.7}$$



(a) The fundamental unit of a neural network, a neuron. This particular neuron takes three inputs.



(b) Multiple layers form a network. Note that the dashed line indicates that the bias term is different for the calculation of each output unit.

Figure 2.5

$$a_2^{(2)} = f(\sum_i^N W_{2i}^{(1)} x_i + b_2^{(1)}), \tag{2.8}$$

$$a_1^{(3)} = f(\sum_i^N W_{1i}^{(2)} x_i + b_1^{(2)}), \tag{2.9}$$

where the first subscript indice of the weight **W** refers to the output unit it is related to and the second subscript indice refers to the input unit it is related to (the subscript

for $\mathbf{b}$ refers to the output unit that the bias is related to). The superscript indice refers to the layer number, with the input data $\mathbf{x}$ being the first layer. Note that each neuron is connected to every other neuron in its adjacent layers, but not to the neurons in its own layer. These equations can be simplified to a matrix form and generalised to any number of layers. By letting $z_i^{(l)}$ be the weighted sum of the input units and bias term going into the i-th neuron in layer l, then for the first layer:

$$\mathbf{z}^{(2)} = \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}, \tag{2.10}$$

$$\mathbf{a}^{(2)} = f(\mathbf{z}^{(2)}), \tag{2.11}$$

and in every subsequent layer up to L layers:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l-1)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l-1)}, \tag{2.12}$$

$$\mathbf{a}^{(l)} = f(\mathbf{z}^{(l)}), \tag{2.13}$$

for $l = L, L-1, L-2, ..., 3$. By letting $\mathbf{a}^{(1)} = \mathbf{x}$, the equations 2.12 and 2.13 can be further generalised for $l = L, L-1, L-2, ..., 3, 2$. The layers in between the input and output layer are often referred to as the hidden layers (Deng et al., 2010; Schmidhuber, 2014). The MLP can have many different architectures where the number of layers and width of each layer changes.

The previous set of equations determines the feed-forward value of each neuron once the parameters for $\mathbf{W}$ and $\mathbf{b}$ have been learnt. The backpropagation algorithm is used to train the network from the data by learning the parameters for $\mathbf{W}$ and $\mathbf{b}$. In order to do backpropagation, a cost function must be defined on the output layer that is a function of all of the parameters in the network. Hence the parameters can be learnt by minimising this cost function. If a training label $\mathbf{y}^{(m)}$ exists for each input $\mathbf{x}^{(m)}$, then the cost function of the MLP for all observations, including a regularization term, can be defined as:

$$E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} (\frac{1}{2} \|\mathbf{a}^{(L)(m)} - \mathbf{y}^{(m)}\|)^2 + \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2, \tag{2.14}$$

where $\mathbf{a}^{(L)(m)}$ is the values of the neurons in the output layer L, which are dependent on $\mathbf{x}^{(m)}$. M is the number of observations, λ is the regularization parameter, and I and J are the number of units in layers l and $l+1$ respectively. The regularization term prevents overfitting of the parameters, which is when the network does not generalise to new data (Bishop, 2006). A squared error cost function such as this one is useful for both regression and classification problems. For regression problems, $\mathbf{y}$ takes on real, continuous values. For classification tasks, each $\mathbf{y}$ becomes a one-hot vector (e.g. [0 0 1 0] for a four class problem). Of course, MLPs are not limited to a squared error cost function. Other measures of error can also be used, such as cross-entropy (Golik et al., 2013), which is popular for classification.

In order to use optimisation to find the parameters that minimise the cost function in equation 2.14, the partial derivatives of equation 2.14 must be calculated:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial W_{ji}^{(l)}} (\frac{1}{2} \|\mathbf{a}^{(L)(m)} - \mathbf{y}^{(m)}\|)^2 + \lambda W_{ji}^{(l)}, \tag{2.15}$$

$$\frac{\partial}{\partial b_{j}^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial b_{j}^{(l)}} (\frac{1}{2} \|\mathbf{a}^{(L)(m)} - \mathbf{y}^{(m)}\|)^2, \tag{2.16}$$

where for a single observation m:

$$\frac{\partial}{\partial W_{ji}^{(l)}} (\frac{1}{2} \|\mathbf{a}^{(L)} - \mathbf{y}\|)^2 = \delta_j^{(l+1)} a_i^{(l)}, \tag{2.17}$$

$$\frac{\partial}{\partial b_{j}^{(l)}} (\frac{1}{2} \|\mathbf{a}^{(L)} - \mathbf{y}\|)^2 = \delta_j^{(l+1)}, \tag{2.18}$$

for $l = 1, 2, 3, ..., L-1$, with the value of δ dependent on the layer number l. For the output layer $l = L$, in which there are K units, the δ for each output unit k is calculated as:

$$\delta_k^{(L)} = -(y_k - a_k^{(L)}) \cdot f'(z_k^{(L)}), \tag{2.19}$$

and for all other layers $l = L-1, L-2, L-3, ..., 2$,

$$\delta_i^{(l)} = \sum_J (\delta_j^{(l+1)} W_{ji}^{(l)}) f'(z_i^{(l)}).,\tag{2.20}$$

where i is the index of the unit in layer l and j represents the index of the unit in layer $l+1$. In this way, the gradient of the error is propagated back through the network via the weights. The parameter update equations for gradient descent optimisation are:

$$W_{ji}^{(l)} := W_{ji}^{(l)} - \alpha \frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}),\tag{2.21}$$

$$b_j^{(l)} := b_j^{(l)} - \alpha \frac{\partial}{\partial b_j^{(l)}} E(\mathbf{W}, \mathbf{b}),\tag{2.22}$$

where α is the learning rate.

The MLP is trained by first initialising all of the parameters, either randomly or by some other means, then doing a feed-forward pass of the data through the network, measuring the error between the target data and the neuron activation values in the output layer, and then backpropagating the error through the network. This process is repeated until the error - or cost - converges. If the parameters are updated iteratively to minimise the cost function, then the MLP learns to map the $\mathbf{x}$ values to their corresponding $\mathbf{y}$ values via a highly non-linear function comprising the weights and biases at each layer of the network.

2.3.2 Autoencoder

A deterministic autoencoder (Bourlard and Kamp, 1988; Hinton and Salakhutdinov, 2006; Kramer, 1991) is a special case of the regression MLP, where the target vector is set as the input data:

$$\mathbf{y} := \mathbf{x},\tag{2.23}$$

and hence the MLP learns to reconstruct its input layer in the output layer. This is an unsupervised learning process as no labelled training data are required. The

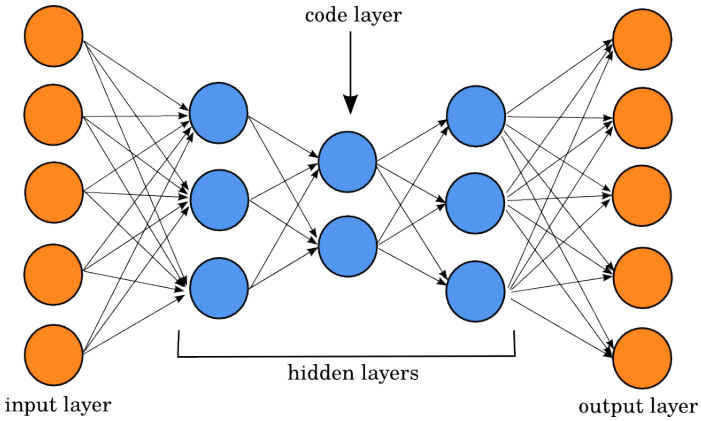


Figure 2.6 – An example of a stacked autoencoder. The MLP reconstructs the input in the output layer and has a symmetric architecture, with a code layer in the middle of the hidden layers.

autoencoders are often structured such that the width of the hidden layers gets progressively smaller until a bottleneck point, after which they get larger again, such that the width of the layers is symmetric about the smallest hidden layer (Figure 2.6). The benefit of this is that the layer with the smallest width, often called the code layer or bottleneck layer, becomes a condensed representation of the input data. This is because the code layer is forced to encode any important structure in the input data so that it can accurately reconstruct it. The weights and biases that map the input data to the code layer are called the encoder stage of the network, and the weights and biases that reconstruct the input from the code layer are called the decoder stage. When an autoencoder contains multiple hidden layers it is called a stacked autoencoder (SAE) (Larochelle et al., 2007). To train the parameters in each layer of an SAE, a greedy pre-training step, in which layers are trained in turn whilst keeping other layers frozen, usually precedes an end-to-end fine-tuning step, whereby all layers are trained at the same time. Because SAEs can learn a condensed form of the data, they find use as an unsupervised method for finding low dimensional encodings or feature representations of the data.

There are variants of the basic deterministic autoencoder, including the sparse autoencoder (Ng, 2011) and the contractive autoencoder (Rifai et al., 2011). These autoencoders have additional constraints imposed on them to promote sparsity and

robustness, respectively. Imposing a constraint is a form of regularisation, similar to what the penalty term in the cost function of equation 2.14 is doing in order to prevent overfitting. These regularised autoencoders can also be stacked to form deeper architectures, just like the basic SAEs.

Another type of autoencoder exists called the denoising autoencoder (DAE). For the DAE, a stochastic corruption process is applied to the input layer only (Figure 2.7), forcing the network to learn an encoder-decoder mapping to reconstruct the uncorrupted, clean input, so that it preserves the input information and reverses the effect of the corruption process (Vincent et al., 2008).

In the DAE cost function, the error is computed between the clean input $\mathbf{x}$ and the output neurons $\mathbf{a}^{(L)(m)}$ which are a function of the corrupted input $\tilde{\mathbf{x}}$. The training process is exactly the same as with SAEs, but the mapping learnt is more complex because it must denoise the corrupted input which often results in being able to learn more robust features.

There are several common methods of corrupting the input. These include the addition of Gaussian noise, forcing a randomly masked fraction of the input elements to be zero (masking noise), and forcing a randomly masked fraction of the input elements to be zero or one (salt-and-pepper noise)(Vincent et al., 2010). Masking noise is equivalent to having missing elements in a given input sample, and the DAE is trained to fill in these missing values which is possible by capturing the high-dimensional dependencies in the data.

2.3.3 Convolutional Neural Network

Another popular deep learning algorithm is a CNN (LeCun et al., 1990). CNNs are structured slightly differently to MLPs. They often consist of a number of convolutional and pooling/subsampling layers followed by fully connected layers (Figure 2.8). Whilst the neurons of the MLP are connected to all neurons in the adjacent layers, the neurons in the convolutional layers of the CNN are only locally connected, and they share weights with other neurons. This can be interpreted as the weights acting

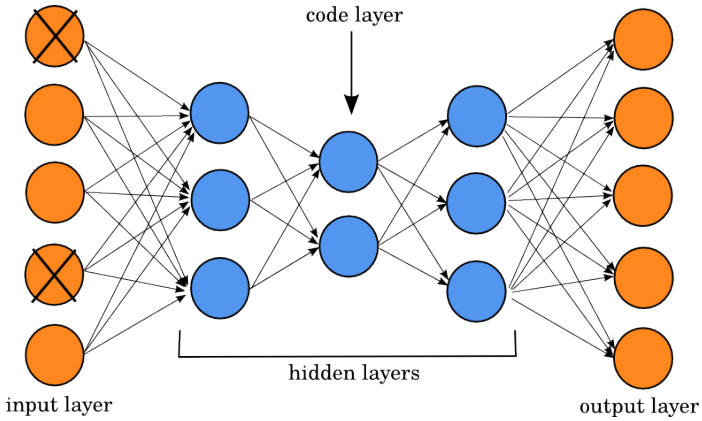


Figure 2.7 – An example of a stacked DAE. The input layer is corrupted in some way (e.g. by masking out some of the units), and the network must reconstruct the clean input in the output layer.

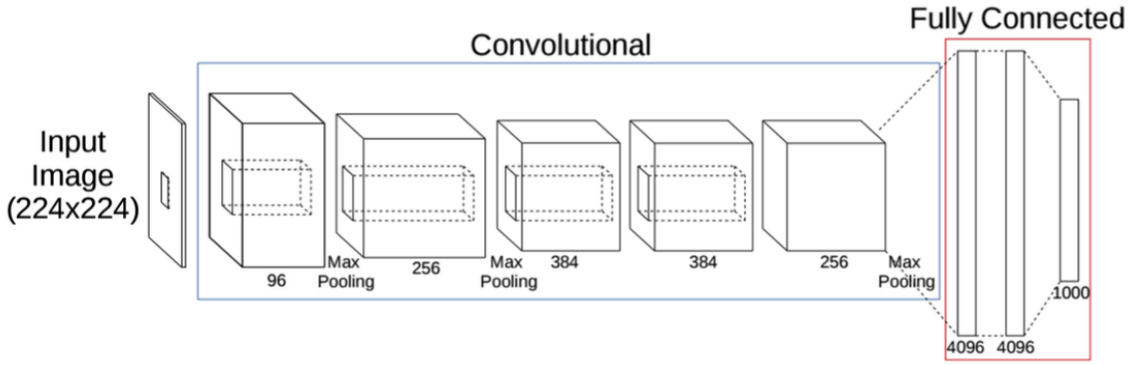


Figure 2.8 – An example of a CNN architecture - AlexNet (Krizhevsky and Hinton, 2012). There are convolutional layers, activation functions, pooling layers and fully connected layers. AlexNet was designed to classify 224×244 pixel images as one of 1000 possible classes.

as filters which convolve over the input data, thus outputting the neuron values in the next layer (Figure 2.9). When the CNN is trained, the filters or kernels, as they are sometimes called, are refined to detect localised features in the data. The benefit of the shared weights in the convolutional layers is that there are fewer parameters to learn. Multiple filters can be trained to filter the data, and the result is one feature map output for each filter/kernel. Non-linear activation functions are applied to each element of the feature map independently just as with the neurons in the hidden layers of the MLP. The rectified linear unit (ReLU) is a common choice of activation function for CNNs (Nair and Hinton, 2010). Other activation functions include the

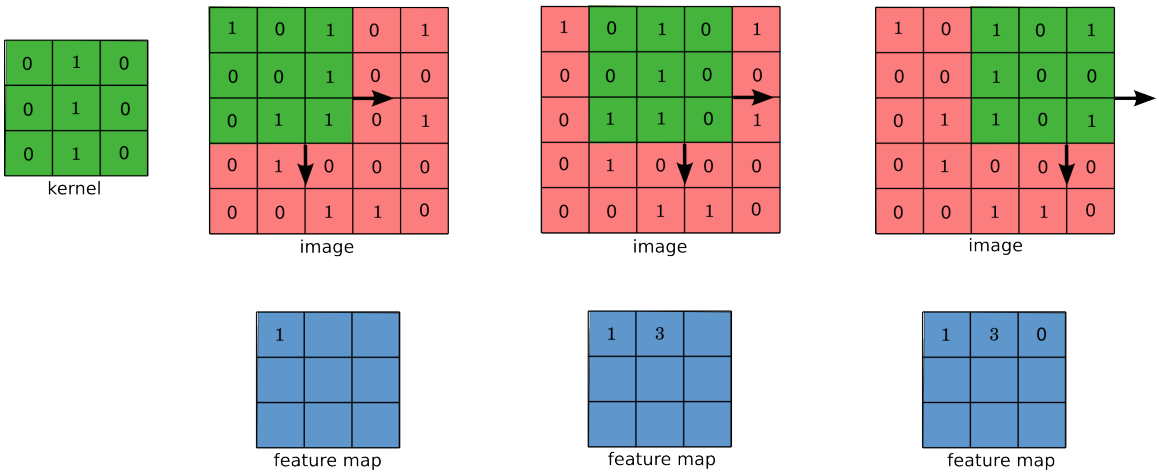


Figure 2.9 – The process of a 3×3 kernel filtering a 2D image to produce a feature map.

leaky ReLU (Maas et al., 2013), parametric ReLU (He et al., 2015), sigmoid (i.e. logistic) and TanH.

The pooling layers are designed to aggregate the statistics of the data output from the convolutions at various locations, effectively summarising the feature maps. This process might involve finding the average or maximum over a specified area of the feature map. By taking advantage of the stationary nature of the data, the number of elements in the feature maps are reduced, reducing the likelihood of overfitting, and some translational invariance is built into the features.

The fully connected layers are simply an MLP where all neurons in adjacent layers are connected to each other. The feature maps output by the kernels in the final convolutional layer must be vectorised before they are passed into the fully connected layers. After the fully connected layers, there is usually a classification layer such as softmax which outputs a vector of scores over the label space. These scores can be used to predict the class label.

If a CNN classifier is applied to image data, then the dimensions of the data input into the first convolutional layer are $m \times n \times d \times 1$ where m is the number of rows, n is the number of columns and d is the number of channels in the image (e.g. $d = 3$ for an RGB image). The k_1 kernels in the first convolutional layer will have size

$C_{m1} \times C_{n1} \times C_{d1} \times 1$ where C_{m1} is the number of rows in each kernel, C_{n1} is the number of columns in each kernel and C_{d1} is the number of dimensions of each kernel. Also, $C_{m1} < m$, $C_{n1} < n$ and $C_{d1} \leq d$. The size of each of the k_1 feature maps produced is $(m - C_{m1} + 1) \times (n - C_{n1} + 1) \times (d - C_{d1} + 1)$. These dimensions are usually further reduced by pooling. The k_2 kernels in the second convolutional layer will have size $C_{m2} \times C_{n2} \times C_{d2} \times k_1$, and the k_l kernels in the l-th convolutional layer will have size $C_{ml} \times C_{nl} \times C_{dl} \times k_{l-1}$.

CNNs are typically supervised learners, so they require some form of label associated with each of the training points (e.g. a class label associated with each image). The parameters of the CNN are learnt by establishing an error measure with a cost function, using the backpropagation algorithm to compute the error derivatives at each layer and then iteratively minimising the error with an optimisation technique, such as stochastic gradient descent.

2.4 Literature Review

A key problem being addressed in this thesis is the variability in hyperspectral data caused by illumination. There are a range of approaches in the literature to solving a similar problem under the name of illumination invariance. An illumination invariant image is one which is independent of the effects of illumination. In this review, many of these approaches are compared and discussed in the context of outdoor imaging with a hyperspectral sensor.

In this thesis, deep learning algorithms are employed for learning feature representations and classification models for hyperspectral data that are invariant to the illumination. Thus, the advancements of CNNs and the corresponding utilisation of deep learning in hyperspectral applications is reviewed. Finally, data augmentation in the literature is briefly reviewed.

2.4.1 Illumination Invariance

The problem of illumination invariance in images has received attention from several different research communities. Solutions have been formulated using image-based, learning-based, multi-modal and remote sensing techniques. Each of the techniques have strengths and weaknesses. Many of these techniques have either been extended or have inspired the development of algorithms for finding illumination invariant hyperspectral images. Many new techniques devised specifically for hyperspectral images have also been proposed.

In the field of computer vision, illumination invariance has been tackled by considering images as a combination of the intrinsic material properties of their constituents and the illumination in the scene. By removing the illumination, an invariant image can be found. The benefit of image-based approaches for finding illumination invariant representations is that they often do not rely on information from additional external sensors.

Colour constancy techniques aim to remove the effects of varying illumination such that materials under different illumination conditions appear to be constant (Agarwal et al., 2006). One such technique is the Grey-World approach (Buchsbaum, 1980), which assumes that the average colour over the entire image is grey, and any deviation from grey is due to the scene illuminant. Hence, the image is corrected to reduce alterations in the color due to the illuminant. In a similar way, the Scale by Max (Agarwal et al., 2006) approach assumes that there is something white in the scene which reflects all of the illumination at each channel. An estimate for the illuminant can be found as the maximum intensity value of each channel across the image, and this estimate can be used to correct the colours in the image.

Whilst colour constancy can correct the effects of illumination for simple colour images of the scene, these approaches are limited because they often do not account for geometric variability and multiple light sources in an image. Such is the case when there are shadows in the image. This makes them unsuitable for most outdoor scenarios where shadows and geometric variability almost always occur. Methods

developed by Lynch et al. (2013) and Gijsenij et al. (2012), which work for multiple light sources, out-perform the more basic colour constancy techniques. They do not, however, account for variability in the surface geometry of the scene. These techniques also have not been extended for use with hyperspectral data and rely on limiting assumptions in order to work, such as the average colour in an image being grey.

Some very successful approaches to illumination invariance in computer vision and image processing are derived from a Lambertian model for photodetector response:

$$\rho_j = \sigma \int E(\lambda)S(\lambda)Q_j(\lambda)d\lambda \tag{2.24}$$

where ρ_j is the camera sensor response of the jth sensor, σ is a constant based on the scene geometry given by the dot product of the surface normal and the incident illumination direction, $E(\lambda)$ is the spectral power distribution of the illuminant, $S(\lambda)$ is the surface reflectance and Q_j is the spectral sensitivity of the camera sensor. Illumination invariant features can be derived from this model by assuming Planckian illumination, narrow-band camera sensor functions and Lambertian surfaces. One of these approachs (Marchant and Onyango, 2000) derives an invariant feature from RGB images, and extends this to multiple invariant features for spectra covering more than three wavelengths (Marchant and Onyango, 2002). Other methods take the logarithm of photodetector responses to separate the surface reflectance components from the illuminant dependent components of the response. After taking the logarithm, Ratnasingam and McGinnity (2012) find an optimal linear combination of sensor responses to remove the effects of the illuminant and scene geometry. In Finlayson et al. (2006, 2002, 2004, 2009), prior to taking the logarithm, the chromaticity of the sensor responses are computed to normalize out the illuminant intensity and scene geometry effects. It can be shown that the log-chromaticity vectors for a given surface will move along a camera-dependent direction as the illuminant changes, independent of the surface reflectance. Hence, it is possible to remove the effects of the illuminant by projecting the log-chromaticity vectors into the subspace orthogonal to that direction. When dealing with RGB images, it is difficult to determine the

exact direction of the orthogonal subspace as the narrow-band assumption does not hold true. However, other methods for determining the camera-dependent direction to project onto have been investigated. A Macbeth ColourChecker (Finlayson et al., 2002) was imaged using an RGB camera under different daylight illumination conditions so that the direction of change can be found for that camera. Another approach (Finlayson et al., 2004, 2009) looks at many different projections and selects the one with the smallest entropy. There have been numerous applications which have used these illumination invariant images, including road segmentation (Álvarez and Antonio, 2011), image-based outdoor localisation (Corke et al., 2013; Mcmanus et al., 2014) and urban street classification (Upcroft et al., 2014). Shadows are removed from RGB images by Yang et al. (2012) by computing an intrinsic image found with a similar approach to Finlayson et al. (2006) and then doing bilateral filtering.

The approach of Finlayson et al. (2006) has been extended from RGB images to multispectral and hyperspectral images (Drew and Salekdeh, 2011; Salekdeh, 2011). This involves log-chromaticity vectors with much higher dimensionality, covering wavelengths from the visible and near infrared (VNIR) range ($\backsim 400 - 1000nm$) to the SWIR range ($\backsim 1000 - 2500nm$). However, given that every channel in the hyperspectral sensor is narrow-band, it is easier to determine the orthogonal projection than it is with RGB sensors. The orthogonal projection matrix can be computed directly as a function of the sensor wavelengths and various constants, removing the need to use Macbeth colour checkers or entropy minimisation. The result is that an invariant spectrum can be found, having the same number of channels as the original spectrum.

The biggest problem with the approaches based on the Lambertian model of equation 2.24, is related to one of the assumptions that are made in order to derive the illumination invariant features - the assumption of Planckian illumination. In this assumption, the spectral power distribution of the incident light is modelled by Wein's approximation to Planck's law. As the illuminant transmits through the atmosphere, deep absorption features are introduced into its spectrum as it encounters molecules such as oxygen, carbon dioxide and water. These absorptions features are

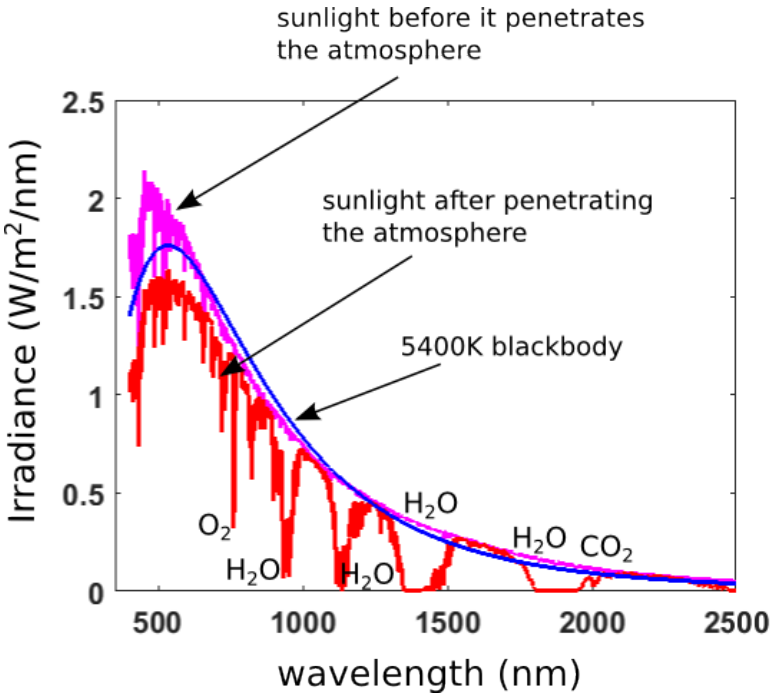


Figure 2.10 – An illustration of how the solar radiation spectrum, which has a spectral power distribution similar to that of a 5400 K black body, is absorbed at specific wavelengths by gas molecules as it is transmitted through the atmosphere. After this occurs, the spectral power distribution no longer resembles the black body distribution due to the additional atmospheric absorption features. These curves were generated using the SMARTS atmospheric modeller (Gueymard, 2001).

not accounted for by the approximation even though they have a large impact on the incident light spectra (Schowengerdt, 2007). Whilst these absorption features do not have a significant effect on three channel RGB images, they have a much more prominent influence on the shape of a hyperspectral curve (Figure 2.10). Another potential problem is that these illumination invariant features are not optimised for class separability. Hence there might be a trade-off between the illumination invariance and the discriminability of different materials.

Many remote sensing techniques turn the problem of finding illumination invariant representations of imagery into a radiometric normalisation task. An estimate for reflectance spectra, which hypothetically are independent of the incident illumination, is found by dividing the observed signal by the incident illumination. This is conceptually similar to colour correction from the computer vision literature (discussed

above).

Empirical normalisation techniques, including residual images, internal average relative reflectance (IARR), continuum removal and empirical line are often used to correct overhead imagery acquired by aircraft or satellites. This is due to the difficulty of measuring or estimating the illumination sources for these cases. Continuum removal requires the spectra to be in reflectance and adjusts individual pixels by fitting a convex hull over the peaks of each spectrum before the spectra are divided by their respective hull (Clark and Roush, 1984). Whilst this approach can normalise out the effects of solar irradiance, it cannot remove the unwanted atmospheric absorption features. In the IARR approach, each digital number (DN) pixel spectrum is divided by the mean DN spectrum calculated over the entire image (Kruse, 1988) in order to compute an approximate for reflectance. Whilst this can remove multiplicative atmospheric absorption features, it cannot compensate for the effects of topographic variation or additive effects such as path radiance. The empirical line method uses multiple reflectance measurements obtained from bright and dark targets from within the scene with a spectrometer. Linear relationships between the observed reflectance spectra and the raw image data are determined. The slope and intercept of this relationship are used to convert image data in DN or radiance to reflectance (Smith and Milton, 1999). Unlike the convex hull and IARR methods, the empirical line method can remove the effects associated with path radiance. However, it cannot compensate for topographic variations and also requires extensive field measurements. For the residual image approach, a wavelength is chosen and then all spectra are scaled such that the pixels of the chosen wavelength all have an intensity equal to the maximum. Then, the mean intensity of each channel over the image is subtracted from all of the pixels in each channel, producing an estimate for reflectance (Marsh and McKeon, 1983). This method can remove the effects of atmospheric absorption features, view path radiance and topographic variations.

The empirical normalisation methods have the advantage of being simple to implement and, apart from the spectrometer for the empirical line method, no additional sensors are required. However, they are limited in that they cannot account for the

changes in spectra induced by shadows and many of the methods do not remove the effects related to variability in the scene geometry.

An alternative to empirical normalisation that also does not require direct measurement of the illumination sources is through the use of a computational atmospheric radiative transfer model to simulate the illumination sources. These models can produce an estimate of the terrestrial sunlight and diffuse skylight via simulation of light propagation through the atmosphere given a particular aerosol composition, turbidity and moisture among other parameters. There are several radiative transfer models available, from the relatively simple ones that are popular in the field of computer graphics (Hošek and Wilkie, 2013; Perez et al., 1993; Preetham et al., 1999) to the more complex ones, SMARTS (Gueymard, 2001) and MODTRAN (Berk et al., 1987), used in remote sensing applications.

The problem with radiative transfer models is that they have many parameters which must be known *a priori*. If the illumination sources are to be accurately simulated, the various atmospheric parameters must either be measured or estimated from the location and season (e.g. based on closest meteorological station data). The data needed to set these parameters is not always available and if the geographical location of the sensor is unknown then the models cannot be used to correct a captured image. Also, whilst the radiative transfer model can simulate how the illumination sources interact with the atmosphere, how they interact with the scenes with complex surface geometry remains largely unknown.

Of course, rather than simulating the illumination sources or using empirical methods, a much more reliable method for radiometric normalisation is to simply measure the atmospheric sources using hardware of some kind. One approach is to normalise the raw digital values observed by the sensor against the material spectrum of a material of known reflectance such as a diffuse calibration board placed in the scene (i.e. flat-field correction) (Murphy et al., 2008; Rast et al., 1991). Another approach is to measure the incoming incident illumination using a downwelling irradiance sensor and use this to normalise the digital values of all pixels in the scene (Borengasser et al., 2007). The limitation of these methods is that they only provide a correct

measurement of the illumination for the region that the calibration board or sensor is placed. They cannot capture the variation in the incident illumination due to the scene geometry and occlusions (i.e. shadows). The ideal scenario would be to measure the incident illumination at every point in the scene, however this is not practical.

Instead of trying to measure the illumination and correct for it at every pixel in the image, physics-based models for remotely-sensed data can be formulated to derive wavelength-dependent scaling terms which de-shadow the spectra, correcting them as if they were illuminated by sunlight (Adler-golden et al., 2001, 2002; Richter and Müller, 2005). The limitation of these methods is similar to the radiometric normalisation approaches, as they require accurate measurements or estimations with radiative transfer models of the direct terrestrial sunlight and diffuse skylight which are difficult to acquire. Also, since much of the remotely-sensed data is captured from satellites or airborne platforms, many of the models disregard the influence of scene geometry because the spatial resolution is so low. For close-range sensing, the geometric parameters must be factored into the models.

The remote sensing approaches to illumination invariance largely fail to account for variations due to the geometry and occlusions in the scene. These variations affect both the amplitude and spectral shape of the incident illumination, and hence have a significant impact on the spectra observed by the sensor. This limitation can be overcome using multi-modal approaches that incorporate geometric information (Broadwater and Banerjee, 2013; Friman et al., 2011; Ientilucci, 2012; Ramakrishnan, 2016). Mutli-modal approaches utilise additional sensors such as light detection and ranging (LiDAR) and GPS to form geometry-based illumination models of the scene making it possible to compensate for shadows in spectra captured over the VNIR and SWIR spectrum. They often do not require any labelled data and can take into account the impact of the atmosphere when used in conjunction with radiative transfer models.

The approach in Ramakrishnan (2016) uses equation 2.3 to relight the spectrum of each pixel to a common illuminant (skylight) to remove the variability due to illumination. This required a hyperspectral image to be fused with LiDAR data,

so that all of the geometric parameters required for relighting were known, as well as whether or not each pixel had a direct line-of-sight with the sun. Although the multi-modal approaches can compensate for the illumination variability such that the scene-geometry is accounted for, the limitation is that the additional sensors are not always available. Also, when they are available they rely on accurate spatial registration with the hyperspectral image data.

Learning approaches (Guo et al., 2011; Lalonde and Efros, 2010; Zhang et al., 2015; Zhu et al., 2010) for detecting and removing shadows from RGB images have been proposed as an alternative means of finding illumination invariant representations of scenes. These methods rely on fewer assumptions than the purely image-based approaches to illumination invariance, but instead require labelled training data or user interaction to learn a model. Likewise, labelled data of the classes of interest are used to train Support Vector Machines and CNNs to accurately classify hyperspectral pixels, regardless of the illumination conditions (Chen et al., 2016; Izquierdo-Verdiguier et al., 2013; Makantasis et al., 2015). In Healey and Slater (1999), a library of known reflectance for each class is used with a physical model incorporating the illumination conditions to learn a set of orthonormal basis functions for that class. Using these class-specific basis functions, an illumination insensitive likelihood for each material can then be calculated for a given spectral vector. Despite not having to make any assumptions about the image data, the reliance of the learning-based methods on labelled data limits their application as labelled hyperspectral data is time-consuming and expensive to acquire (as explained in Chapter 1). Regions under very low lighting can also be challenging to label correctly and sometimes samples must be physically collected from the scene for analysis in order to label them. In many applications (such as mining), this can be both difficult and hazardous.

Unsupervised learning techniques have the advantage of not requiring labelled data. Robust Principle Component Analysis (Wright et al., 2009) treats shadows as outliers or sparse errors in a set of images and recovers an uncorrupted low-rank matrix where the shadows are removed. This method, which requires no labelled data, worked well on a set of greyscale images of human faces that were illuminated differently. It

does however require that multiple images are captured of the same scene under different conditions which is an undesirable and impractical constraint for outdoor imaging. Another unsupervised method (Zheng et al., 2015) separates illumination and reflectance spectra in hyperspectral images in a low dimensional subspace, using a low-rank matrix factorization method. This approach was tested with a range of illuminants such as daylight, fluorescence and LED, but its ability to separate the illuminant when parts of the scene are illuminated by direct irradiance and diffuse skylight (sunlit regions) and others by diffuse skylight only (shadowed regions) must still be evaluated.

In summary, illumination invariance for hyperspectral imagery has been approached using techniques from several different research fields. Each field has associated advantages and disadvantages. Image-based approaches originally designed for finding illumination invariant colour images and extended to work for hyperspectral images do not properly account for the effects of the atmosphere on the incident illumination. Remote sensing methods overcome this through the use of either radiative transfer models which simulate the propagation of light through the atmosphere or direct measurement of the incident illumination using hardware. The limitation of radiative transfer models is that they require the estimation of many atmospheric parameters, which is often difficult to acquire data for. Remote sensing techniques also rarely take into account small-scale changes in the geometry of the scene due to the low spatial resolution of the data. Multi-modal techniques overcome this by utilising geometric data fused with remotely-sensed imagery, but at the cost of additional sensors which are not always available. Finally, learning-based techniques take a more data-driven approach, and hence do not rely on complex atmospheric models or additional sensors, but can still account for geometry and the atmosphere. The disadvantage of supervised approaches is that they require labelled data which is laborious to acquire, and unsupervised techniques either work under impractical constraints or have not been shown to work for multiple illuminants, which is the case when there is shadows in the image.

Hence, there is a need for the development of methods for determining illumination

invariant algorithms for outdoor hyperspectral data that do not rely on *a priori* knowledge of atmospheric parameters, additional sensors, vast amounts of labelled data and do not make limiting or impractical assumptions. These algorithms must compensate for the effects of geometry, including its influence on the intensity of the incident illuminant and shadows.

2.4.2 Advancements in Convolutional Neural Networks

In recent years, CNNs (a supervised learning technique) have revolutionised the field of computer vision and, in particular, image classification. This technique is not new, for example, it was applied to the problem of hand written digit recognition in the 90s (LeCun et al., 1990). In 2012, however, it made a comeback when AlexNet (Krizhevsky and Hinton, 2012) was used to win the ImageNet large scale visual recognition challenge (ILSVRC) (Deng et al., 2009) competition by an impressive margin from the previous year and the second place (scoring 16.4% in top-5 classification error, approximately 10% better than the Fisher vector based approaches). This started a new trend where image classification techniques moved away from more traditional, hand-engineered features and moved towards features that were learnt from the data. To achieve robust classification results, AlexNet was trained by augmenting the input data with translations, horizontal reflections and colour adjustments.

In ILSVRC 2014, most of the entries were based on CNN architectures (Russakovsky et al., 2015). In ILSVRC 2014, GoogLeNet (Szegedy et al., 2015), another CNN approach, won the competition with a classification error of 6.7%. Their success came from using 1×1 convolutional layers to increase the width and depth of the network. Another notable performer that year was VGGNet (Simonyan and Zisserman, 2014) which also used smaller filter sizes for the convolutional layers than AlexNet. The success of both GoogLeNet and VGGNet is attributed to these small window sizes. Several stacked layers of small filters have the same receptive field as one large filter, but with more non-linear activation layers in between them, which makes the output much more discriminative. This is expected to be the reason for its superior

performance.

The ILSVRC 2016 winner of the image classification competition was the residual network (ResNet) (He et al., 2016). By introducing residual layers, the network had fewer filters to learn and hence could be made much deeper (34 layers compared to the 19 layer VGGNet) without running into the problem of vanishing gradients. ResNet reduced the top-5 classification error to 4.5%, and an ensemble of ResNets further reduced this error to 3.6%.

Although previously considered too expensive to train, CNNs have grown in popularity with the increase in large annotated image databases and high performance graphics processing unit (GPU)s to reduce the computation time. From the ILSVRC results of the last few years, CNNs have been shown to outperform traditional hand crafted features in classification tasks. CNNs are now achieving state-of-the-art results in many different applications such as stereo-matching (Pang et al., 2017), object detection (Ren et al., 2015), classifying remotely-sensed images (Castelluccio et al., 2015; Zhang et al., 2016) and learning hand-eye coordination for robotic grasping (Levine et al., 2017).

2.4.3 Deep Learning Models for Hyperspectral Data

With their great success in other domains, deep learning techniques such as deep MLPs, autoencoders and CNNs are now being adopted for hyperspectral research. The hyperspectral application that deep learning algorithms have been utilised most strongly in is pixel-wise classification of hyperspectral images. Many of the early deep learning classifiers for hyperspectral data utilised SAEs. An early work to employ deep learning in hyperspectral images was Licciardi et al. (2009) who extracted unsupervised features from pixel spectra using a SAE before classifying them with an MLP. Following this an SAE was assessed as an unsupervised feature extractor for an MLP spectral classifier in Demarchi et al. (2014). Other notable works were in Chen et al. (2014) and Lin et al. (2013), where pixels were classified using a multi-layer spectral-spatial SAE combined with a logistic regression layer. In this network, the

input vector was a concatenation of the pixel spectral vector and a vectorized window of surrounding pixels from a principle component. Later, deep belief nets (Chen et al., 2015; Li et al., 2014) and DAEs (Xing et al., 2015) with a logistic regression layer and sparse SAEs with an SVM (Tao et al., 2015) were used for the spectral-spatial feature extraction and classification of hyperspectral pixels. Liu et al. (2015) used SLIC superpixels to impose spatial constraints on a DAE spectral feature extractor for spectral-spatial classification. The significance of these early works was that they proposed a way to train a deep neural network on hyperspectral data. Some of these works also paved the way for deep neural network frameworks to learn features from both the spectral and spatial dimensions.

With the success of these initial works, and with the trends in other domains such as computer vision, CNNs quickly became popular to use for classifying hyperspectral data. The difference between the CNNs and the SAEs was that CNNs convolved filters (shared weights) rather than having fully connected layers. However, many of the earlier proposed CNN classifiers only convolved in the spatial domain and not the spectral domain. The CNN in Makantasis et al. (2015) convolved a small window over spatial patches extracted from a reduced dimensionality hyperspectral cube. A similar approach was taken by Zhao and Du (2016), but the spatial features were combined with spectral features learnt using a method based on Local Discriminant Embedding. In Lee and Kwon (2016) spatial patches were extracted from a non-reduced hyperspectral cube, but there were still no convolutions occurring over the spectral channel (i.e. high dimensional spatial kernels were learnt).

Progressively, CNNs were used to learn features in the spectral domain. A very simple CNN was proposed by Hu et al. (2015a) which learnt features by convolving over the spectral channel. This approach was shown to perform favourably against other types of neural networks as well as SVMs. Following this, other CNN approaches where the filters were convolved in the spectral domain were proposed, many of which incorporated spatial information as well. The CNN in Mei et al. (2016) used the architecture of Hu et al. (2015a) with modifications to exploit both the spectral and the spatial information. The filters convolved over an input layer which concatenated the re-

flectance at all channels and the spatial context. Similarly, Mei et al. (2017), inspired by the spectral feature learning of Hu et al. (2015a), added spatial context to the learning pipeline. Yang et al. (2016) and Yang et al. (2017) took a different approach of learning the spectral and spatial filters separately and then combining the high level information. Using a two-arm architecture where one arm had filters convolving in the spectral domain and the other arm had filters convolving in the spatial domain, the two arms were combined in a fully-connected layer. Chen et al. (2016) pioneered another completely different approach to learning spectral and spatial features with the CNN. A 3-D CNN was proposed where the kernel convolved in both the spectral and spatial dimensions in spatially smaller hypercubes extracted from the hyperspectral image. Cao et al. (2016) trained a CNN to learn spectral features which were fused with a SLIC segmentation under a Bayesian framework to impose spatial constraints on the classification. In a unique approach, Li et al. (2016) convolved filters spectrally to train a CNN on pixel pairs, and used a voting strategy with pixel neighbourhoods to predict the classification label. Recently, deep residual networks, popular in the computer vision literature (He et al., 2016), have been used to classify hyperspectral pixels using spectral information (Zhong et al., 2017).

Deep learning approaches have also been used for hyperspectral applications other than classification. Methods for unsupervised feature extraction from hyperspectral data have been developed using CNNs (Romero et al., 2014, 2016). A conditional random field (CRF) was used with a CNN to segment a hyperspectral image in Alam et al. (2016). An SAE has been used for unsupervised non-linear spectral unmixing (Licciardi et al., 2012a; Licciardi and Del Frate, 2011), to enhance the quality of hyperspectral images (Licciardi and Chanussot, 2015; Licciardi et al., 2014) and to extract features for finding extended morphological profiles (Licciardi et al., 2012b).

Many of the mentioned deep learning approaches to processing hyperspectral data in the literature have shown how deep, learnt features can improve results over traditional hand-crafted features for various high-level tasks. They have also tackled the problem of combining spatial and spectral information in a unified framework. However, few have considered the problem of limited labelled training data. Because of the

variability in hyperspectral data arising from the complex interaction of illumination and scene geometry, limited training data is a significant problem as this variability cannot be captured. This prevents algorithms from being robust to variations. In most of the current literature on applications of deep learning to hyperspectral data, a sufficient amount of labelled data is used to capture the variability. There are also other problems associated with limited training data, including non-discriminative features and overfitting.

Different aspects of the limited training samples problem have received some attention. The work of Yu et al. (2017) and Ghamisi et al. (2016) identified the potential for overfitting when training a CNN on hyperspectral data due to limited training samples. The solution of Yu et al. (2017) to this problem was to reduce the number of parameters in the network (e.g. removing the fully connected layers, using 1×1 spatial kernels) and by doing basic spatial data augmentation such as rotations and flips. The solution of Ghamisi et al. (2016) was to use feature selection methods to reduce the number of bands in the dataset. Whilst these help to prevent overfitting, they do not solve the problem of not having enough variability captured in the training dataset. In Kemker and Kanan (2017), a self-taught learning framework is used to approach the limited training data problem. Networks were pre-trained on unlabelled data before they were trained on a small amount of labelled data for a target classification task. The benefit of this approach is that if there is a sufficient quantity of unlabelled data, then generalised features can be learnt with unsupervised learning that can be fine-tuned with limited amounts of labelled data to effectively discriminate between different classes. The problem however is that the features cannot be made robust to variations unless there is enough labelled training samples to capture the variability. To address this problem, Chen et al. (2016) took a virtual sample approach to adding variability to the training data in order to make the features more robust without needing additional labelled data. The virtual samples are scalar multiples of the original spectra in the scene, designed to emulate variable illumination. The problem with this approach is that the process for creating the virtual sample is based on a very simplified model, with no modelling of the illuminant or

scene geometry, and hence cannot generate new training samples that are an accurate representation of those in the scene.

A related problem is that a limited number of datasets were used for the evaluation of many of these deep learning algorithms, captured from satellite or airborne platforms. In many of the above works, particularly those with a classification application, the proposed methods are tested on well-researched benchmark hyperspectral datasets (e.g. Indian Pines, Pavia University, Salinas, etc.). Illumination variability might not be as prevalent in data acquired from satellite or airborne platforms due to the spatial scale. Also, relative to many other scenarios, there is more labelled data available for these images and hence there is less motivation to develop learning algorithms that require limited amounts of labelled data. This is evident in the trend of many of these works to explore and develop novel deep learning architectures which rely on lots of labelled training data, rather than to incorporate domain knowledge into deep learning algorithms. Thus, it appears as though the research community is largely stagnating on a solution to the problem of making learning algorithms robust to illumination variability in the presence of limited labelled data.

2.4.4 Data augmentation

Augmentation of training data is a common strategy used in conjunction with machine learning algorithms, including neural networks, to build invariance into the learnt models when there are limited amounts of labelled training data available. Various types of label-preserving transformations are popular for training CNNs to classify RGB images, such as pixel translation, rotation, scaling/zooming, shearing, noise injection, elastic deformations and rendering a scene from novel viewpoints (Ciresan et al., 2010; Gupta et al., 2014; Jaderberg et al., 2014; Zhang et al., 2014). By training with augmented datasets, the networks learn to be invariant to these types of transformations in the test data, even if they are poorly represented in the training data.

Most of the augmentations utilised by these approaches are spatial in nature, as they

have been used for RGB applications where CNNs were learning spatial features. There are some methods which use augmentation in the colour/spectral domain, which is useful for training algorithms to be robust to the effects of illumination. To train AlexNet (Krizhevsky and Hinton, 2012), the intensities of the RGB channels are altered with a very basic approach of applying PCA on the pixel colour values and adding multiples of the principal components that are found to each pixel. This made the final classifier slightly more robust to changes in intensity and colour. In the hyperspectral literature, Chen et al. (2016) took a similar approach in devising the scheme of virtual samples. However, as discussed previously, these virtual samples are not generated by modelling the interaction of the illuminant with the scene geometry and, therefore, cannot accurately simulate the variability in the scene (especially the effects of shadows). In Izquierdo-Verdiguier et al. (2013), hyperspectral signals from data in shadow are exponentially modulated to improve SVM classification performance in the shadow. This approach cannot be used to simulate what a normal sunlit pixel would appear as in shadow, so its usefulness is limited as labelled data is still required to capture most of the variability in the scene.

By accurately simulating the influence of the incident illumination on spectra, data augmentation could be used as a strategy to make learning algorithms, trained on hyperspectral data, robust to illumination variability despite having a limited amount of labelled data.

2.4.5 Summary

Different research communities have attempted to develop representations and models for hyperspectral data that are robust to the effects of illumination. However, there are drawbacks of the approaches from each field, including the reliance on limiting assumptions, *a priori* knowledge of the atmospheric conditions which is difficult to obtain, additional sensors which are not always available or large amounts of labelled data to train on which are usually difficult to acquire. Thus, the problem still requires new solutions which overcome these drawbacks.

Learning algorithms typically do not make limiting assumptions and do not rely on *a priori* knowledge about the atmospheric conditions or additional sensors. However, at present, they largely rely on sufficient amounts of labelled training data to capture the variability in the image. A variety of learning algorithm, the deep neural network, has been gaining momentum in many fields related to image and signal processing over the last few years, including hyperspectral data processing. Researchers have proposed novel strategies to training neural networks to learn spectral and spatial features for a range of applications including classification, non-linear unmixing and image quality enhancement. With all of the progress being made in this field, it remains clear from the literature that the problem of compensating for variability in illumination with limited labelled training samples still needs addressing.

Through the use of data augmentation, it has been shown to be possible to incorporate domain knowledge into learning algorithms to make them invariant to unwanted variations such as spatial and spectral transformations, despite having limited labelled data to capture these variations.

Chapter 3

Datasets and Metrics

This chapter describes the datasets and metrics used in the experiments to evaluate the algorithms proposed in chapters Chapter 4 and Chapter 5.

3.1 Datasets

To demonstrate generality, the datasets used in this thesis were specially selected with the purpose of validating the algorithms proposed in this thesis under a range of scenarios (Table 3.1). A range of different sensors were used to acquire the data, including SPECIM, reflective optics system imaging spectrometer (ROSIS-3) and airborne visible/infrared imaging spectrometer (AVIRIS) operating in the VNIR and SWIR spectral ranges. These spectral sensors have a spectral resolution varying from 2 nm (SPECIM) to 10 nm (AVIRIS). Hyperspectral images with very high spatial resolutions captured from field-based sensors (on the ground) are used, as well as images with lower spatial resolutions typically captured from airborne and spaceborne sensors. Coarse spatial resolution increases the likelihood that mixing (linear and non-linear) will occur within an image pixel because the increased pixel size would increase the chances that more than one type of material would be included in it (Bioucas-dias et al., 2012; Mustard and Sunshine, 1999). Hence, spatial resolution also has an impact on algorithms operating in the spectral domain. For generality,

the noisy water absorption bands in the SWIR range were only removed from some of the datasets.

The scenes captured by each sensor consist of vegetation, urban and mining material classes. Spectra therefore have an assortment of different absorption features. The captured scenes also vary in geometric complexity, with some having highly structured geometries due to man-made elements (e.g. captured in an urban environment) and others having highly unstructured geometries because they are made from an assortment of natural materials (e.g. captured in an open pit mine). The datasets were also selected so that they exemplify a variety of illumination conditions. Some scenes have large topographic variations in brightness and others have shadows due to areas that are occluded from direct terrestrial sunlight. There are also timelapse datasets where large changes in illumination occur over time. The range of different data demonstrates the robustness of the algorithms under different illumination conditions.

Some of these images were normalised to apparent reflectance with flat-field calibration using a calibration target of known reflectance placed within the field of view of the sensor. Sometimes the brightness of objects in portions of the scene is greater than the brightness of the calibration panel. This can be caused by spatial variation in illumination or variation in the orientation of objects relative to the calibration panel. Where this occurs, the value of relative reflectance can exceed unity. This has, however, no bearing on the algorithms presented in this thesis.

Table 3.1 – A summary of the datasets used in the experiments. Datasets were captured under different weather conditions, at different times, with different cameras and with different scene contents in order to extensively evaluate the generality of the proposed algorithms. The table indicates the chapter's where the data was were used for experiments.

Dataset Name	Spectrum	Sensor	Chapter 4	Chapter 5	Chapter 6
Simulated USGS	VNIR + SWIR	-	✓	×	×
X-rite	VNIR	A	✓	×	×
Great Hall (VNIR)	VNIR	A	✓	✓	×
Great Hall (SWIR)	SWIR	A	✓	✓	×
Mining timelapse	VNIR	A	✓	✓	✓
Mining	SWIR	A	×	✓	×
Gualtar steps	VIS	D	✓	✓	×
Gualtar timelapse	VIS	D	✓	×	×
Pavia University	VNIR	C	✓	✓	×
KSC	VNIR + SWIR	B	✓	×	×
Indian Pines	VNIR + SWIR	B	×	✓	×
Salinas	VNIR + SWIR	B	×	✓	×

3.1.1 Sensors

The sensors used to acquire the datasets used in this thesis are the:

- **Sensor A: Specim AISA Eagle and Hawk**. A pair of push broom line-scanners manufactured by Specim (Oulu, Finland). The AISA Eagle sensor captures scans in the VNIR range of $400 - 970$ nm with a spectral resolution of 2-4 nm. The AISA Hawk sensor captures scans in the SWIR range of $970 - 2450$ nm with a spectral resolution of 6-8 nm. A second Specim sensor (model unknown) was also used and captures scans in the VNIR range of $378 - 1003$ nm, with 396 channels at a spectral resolution of approximately 2 nm.

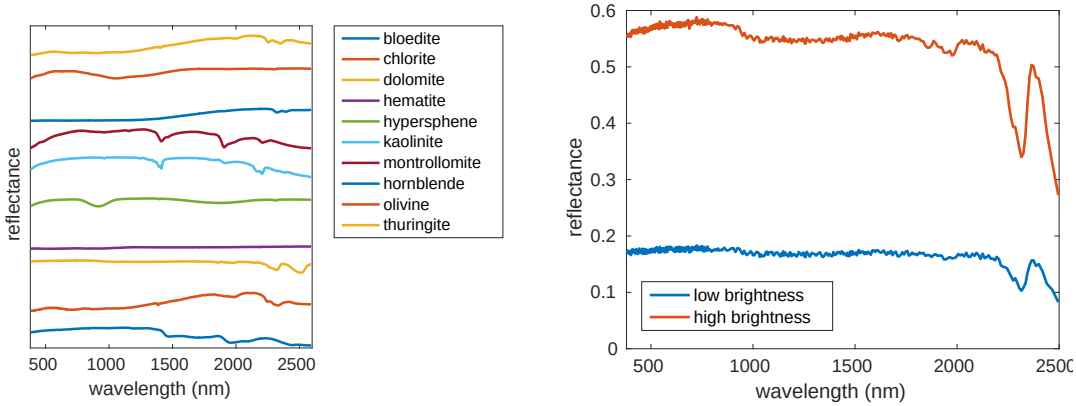
- **Sensor B: AVIRIS** (Eastwood et al., 1987). A push broom line-scanner developed at the Jet Propulsion Laboratory. AVIRIS captures scans with a 614 pixel swathe width in the VNIR and SWIR range of $400 - 2500$ nm with a spectral resolution is 10 nm. AVIRIS is typically operated on-board aircraft platforms, capturing overhead scans.

- **Sensor C: ROSIS-3**. Is a push broom line-scanner developed jointly by MBB Ottobrunn, GKSS Geesthacht and DLR Oberpfaffenhofen. ROSIS-3 captures scans with a 512 pixel swathe width in the VNIR range of $430 - 860$ nm, with a spectral resolution of $4 - 6$ nm. Like AVIRIS, ROSIS-3 was intended for capturing aerial scans.

- **Sensor D: Custom-built hyperspectral imaging system** (Foster et al., 2006). This system is based on a low-noise Peltier-cooled digital camera (Hamamatsu, model C47492-95-12ER, Hamamatsu Photonics K. K., Japan) and uses a fast tunable liquid-crystal filter (VariSpec, model VS-VIS2-10-HC-35-SQ, Cambridge Research and Instrumentation, Inc., Massachusetts). It captures a 1344×1024 pixel image with 33 channels in the visible (VIS) range of $400 - 720$ nm and a spectral resolution of 10 nm.

3.1.2 Dataset 1: Simulated USGS

Mineral spectra from the United States geological survey (USGS) spectral library (Clark et al., 2007) were used to simulate spectra acquired under different illumination conditions - specifically, different brightnesses. Ten minerals were selected. These spectra were used to generate spectral samples comprising 424 channels in the spectral range $383 - 2496$ nm, with a spectral resolution ranging from $2 - 10$ nm resolution, aligned with the characteristics of the spectral measurement device used to obtain the spectra by the USGS (Figure 3.1a). The data was simulated using the principle that the radiance at the sensor $\mathbf{L}$ is the product of the material reflectance $\boldsymbol{\rho}$ and the



(a) Ten different mineral classes. The spectra are offset from each other in the vertical axis for clarity.

(b) Simulated spectra from datasets with different scaling factors and hence different brightnesses. The spectra for dolomite is shown.

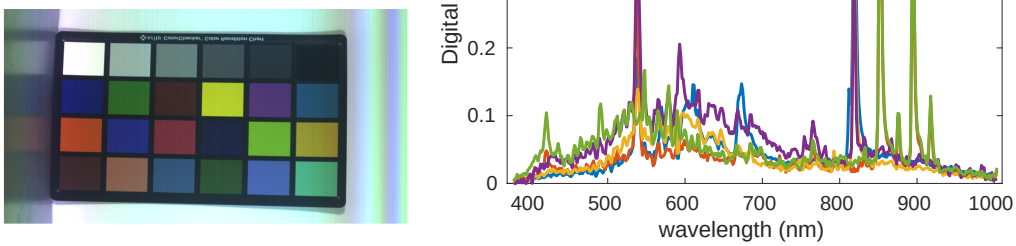
Figure 3.1 – Simulated USGS dataset.

terrestrial sunlight $E_{sun}(\lambda)\tau(\lambda)$ with some additive noise:

$$L(\lambda) = I(\lambda)\rho(\lambda)E_{sun}(\lambda)\tau(\lambda) + noise, \tag{3.1}$$

where **I** is a scale factor controlling the brightness. The terrestrial sunlight was generated using an atmospheric modeller (Gueymard, 2001). Spectra from a calibration panel were also simulated. These data are used to normalise the simulated radiance data to reflectance.

Using this approach, bright and dark datasets were generated. The brightness scaling factors used were 1 and 0.3, producing four datasets in total (datasets with reflectance and DN units for each brightness scaling factor). Each dataset consisted of 10,000 spectral samples (1000 samples for each of the ten classes). The simulated data is equivalent to having two panels with ten materials each and pointing one directly at the sun (scaling factor of 1) and pointing one such that its normal makes an angle of $\cos^{-1} 0.3$ radians (approximately 72 degrees) with the line of sight between the panel and the sun. This simulation does not factor in the variable sky exposure that would occur in the real world. Within one of the simulated datasets (i.e. one panel), there is no variation in illumination. Across the two simulated datasets, there is variability in



(a) A colour composite image of the Xrite dataset under 6500K lighting.

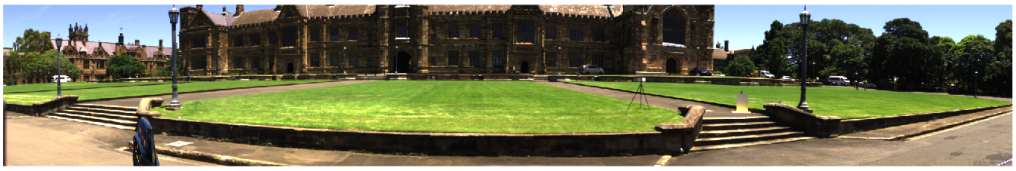
(b) Xrite class spectra.

Figure 3.2 – The X-rite dataset.

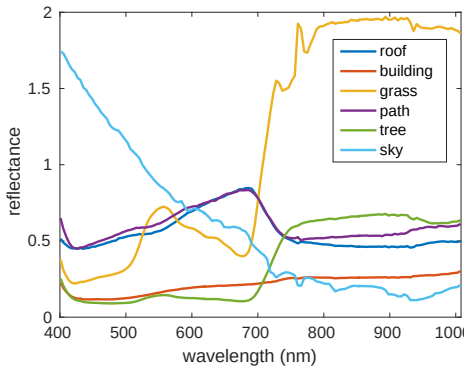
the brightness due to differences in the geometric orientation (Figure 3.1b). In the real world, this scenario could occur by sensing two identical surfaces that have different orientations or by sensing a single surface twice, with the sun changing position and illuminating it from a different angle.

3.1.3 Dataset 2: X-rite panel

The X-rite ColorChecker is an array of 24 different coloured squares. They reflect light in the visible spectrum in the same way as many naturally coloured objects. VNIR images of an X-rite colorChecker were captured indoors (Figure 3.2a) with the SPECIM sensor with the unknown model name (Color-temp). Each image was captured under a different light source, with five different light temperatures used. These were 2500K, 3000K, 3500K, 4300K and 6500K. At different temperatures, both the colour and the brightness of the illumination source varies. For a single image, the geometry is very uniform and there is little cause for variation in the illumination. The image consisted of 697×1312 pixels, with the spectra having DN units. The intensity and shape of the spectrum of each colour changes as the temperature of the illumination source varies (Figure 3.2b). There is an overall shift in the absorption



(a) A colour composite image of the Great Hall VNIR dataset.



(b) Great Hall (VNIR) mean class spectra.

Figure 3.3 – Great Hall (VNIR) dataset.

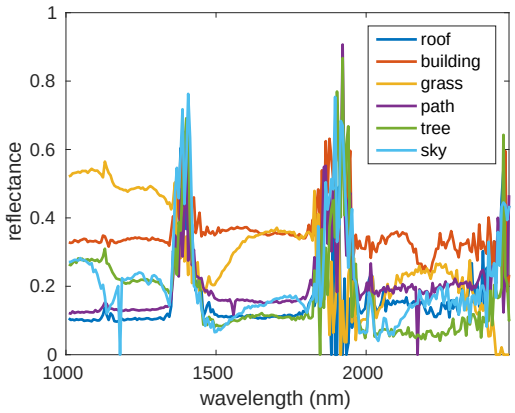
features towards the shorter wavelengths as the temperature rises.

3.1.4 Dataset 3: Great Hall (VNIR)

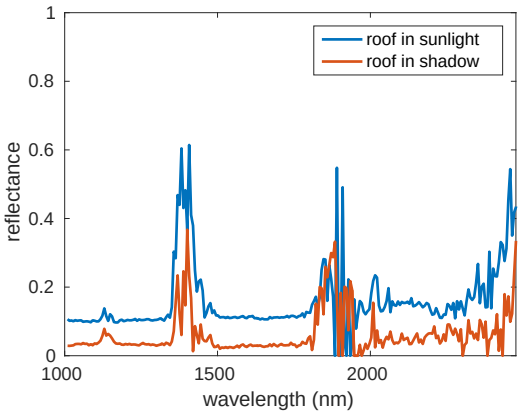
The VNIR scan of the Great Hall of the University of Sydney (Figure 3.3a) was captured with a field-based Specim AISA Eagle sensor in the year 2013 (Ramakrishnan, 2016). The image consists of 320×2010 pixels and 132 spectral channels with a spectral resolution of 4 nm. The outdoor scene consists predominantly of a structured urban environment, with different material classes including roof, sandstone building, grass, path, tree and sky. The roof and path classes are very similar spectrally (Figure 3.3b), however, the other classes are quite discriminable, including the tree and grass. Data were collected under clear-sky conditions, with shadows being evenly distributed across the structure.

A calibration panel of known reflectance is attached to a tripod and placed in the scene so that the incident illumination can be measured. The apparent reflectance

(a) A greyscale image (intensity at wavelength 1633 nm) of the Great Hall SWIR dataset.



(b) Great Hall (SWIR) mean class spectra of pixels in sunlight.



(c) Comparison of mean spectra in sunlight and in shadow for the Great Hall (SWIR) dataset.

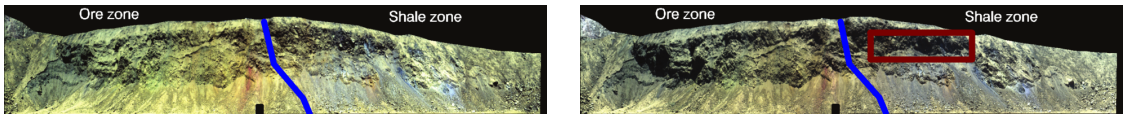
Figure 3.4 – Great Hall (SWIR) dataset.

across the image was calculated using flat-field correction.

3.1.5 Dataset 4: Great Hall (SWIR)

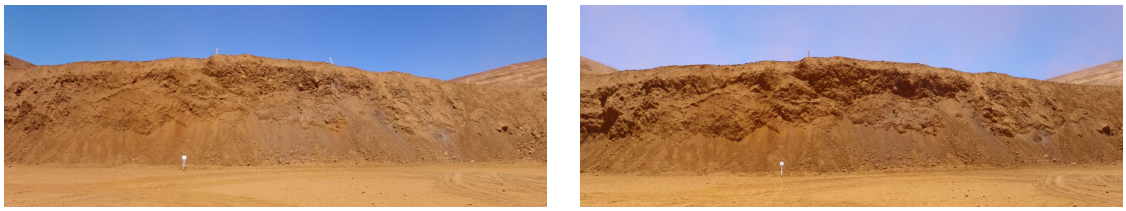
The SWIR hyperspectral scan of the Great Hall of the University of Sydney (Figure 3.4a) was captured with a field-based Specim AISA Hawk sensor in the year 2014 (Ramakrishnan et al., 2015). It is similar to the VNIR scan in that has the same scene constituents (Figure 3.4b), however, the scan was taken on a different day and also from a slightly different viewing angle. The image comprises 293×1306 pixels with 238 spectral channels at a spectral resolution of 6 nm.

Data were captured under partly cloudy skies. Most of the scene is exposed to high intensity illumination, however, large shadows exist, predominately over the building



(a) 11:30 **(b)** 13:30

Figure 3.5 – Colour composite images from the mining timelapse dataset. The blue line indicates the rough geological boundary, identified by an expert, which separates the martite ore from the shale. The red square highlights a region where shadows have become more prominent since the previous image in the timelapse.



(a) 11:30 **(b)** 13:30

Figure 3.6 – Images captured with an RGB camera of the mine at the times of scanning.

and roof regions in the image due to self-occlusion of terrestrial sunlight (the shadows are less evenly distributed than in the VNIR image). There is a distinct change in the shape and intensity of the spectra due to these shadows (Figure 3.4c).

The scene contains a calibration panel which is used to convert the data from DN to apparent reflectance, using flat-field correction. The dataset contains noise due to atmospheric water absorption bands located at 1110-1155, 1338-1514 and 1790-2078 nm, but these bands were not removed. The image is captured from a sensing platform roughly 85 m from the Great Hall.

3.1.6 Dataset 5: Mining timelapse

A timelapse consisting of two VNIR images (Figures 3.5 and 3.6) (Murphy et al., 2012) was acquired of a mine face in Western Australia from a field-based SPECIM AISA Eagle sensor in the year 2009. Each image has 289 × 1443 pixels with 220 channels and a spectral resolution of 2 nm. The sensor was approximately 30 m from the mine face and so the spatial resolution is approximately 6 cm per image pixel.

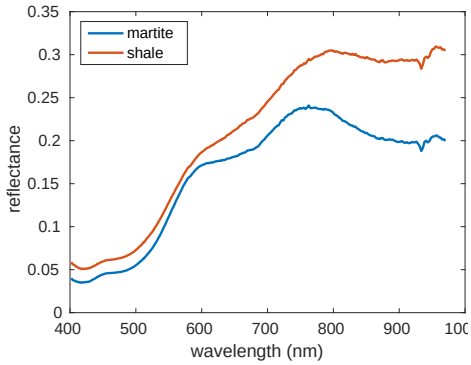


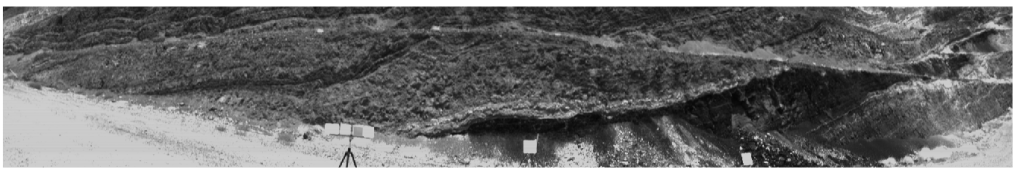
Figure 3.7 – Mining timelapse mean class spectra.

Data were collected under clear-sky conditions. The first image was captured at 11:30 and the second at 13:30. A small shadow in the 11:30 image becomes larger in the 13:30 image as the sun changes position in the sky and more of the mine face becomes occluded by overhanging rocks. The scene exemplifies a natural, unstructured environment. Two types of ore are present; Martite and Shale (Figure 3.7), which are divided by a boundary identified by a geologist. There are actually two types of Shale that can be found in the image, although there is no ground truth information for how these classes are spatially separated. A calibration panel of known reflectance is placed in the scene for normalising the pixel spectra to apparent reflectance. The images are spatially registered to correct any misalignments.

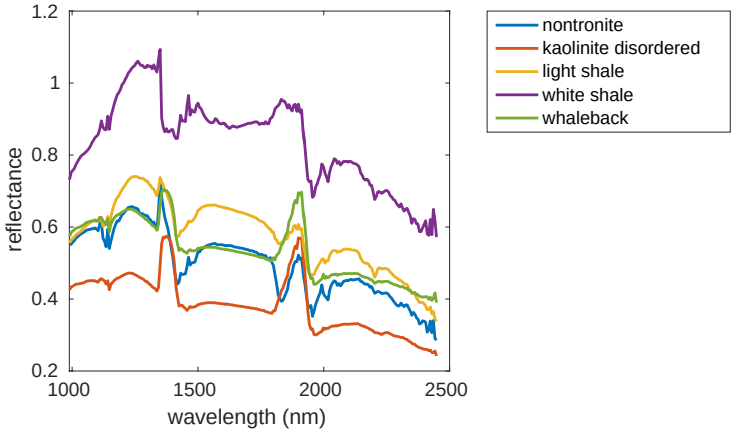
3.1.7 Dataset 6: Mining

The mining image (Figure 3.8a) (Murphy et al., 2014a) consists of a single field-based SWIR scan with a field-based SPECIM AISA Hawk sensor of a mine face (a different to the one in the mining timelapse dataset) captured in 2011. The image has 320×2015 pixels with 235 channels at a spectral resolution of 6 nm. The distance from the sensor to the mine face is about 10 m, hence the spatial resolution is about 0.46 cm per image pixel.

The mine face has an unstructured geometry and large amounts of variation in brightness as well as areas of occlusion from terrestrial sunlight. The five geological classes are nontronite, kaolinite (disordered), light shale, white shale and whaleback shale,



(a) A greyscale image (intensity at wavelength 1633 nm) of the mining dataset.



(b) Mean class spectra.

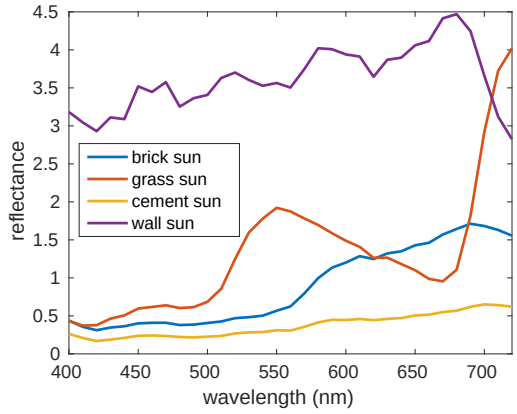
Figure 3.8 – Mining dataset.

with labels provided by Chlingaryan et al. (2016). Because the shale classes are spectrally similar (Figure 3.8b), the optimal distinguishing features in the spectrum are very hard to detect which makes classification a complex task. As with the mining timelapse, a calibration panel of known reflectance is placed in the scene so that the scene was converted to relative reflectance. The noisy water absorption bands were not removed.

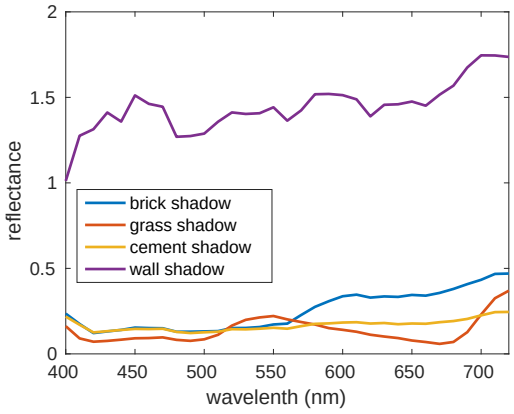
3.1.8 Dataset 7: Gualtar steps

The VIS image of the steps in Gualtar (Figure 3.9a) (Nascimento et al., 2016) has 1024×1344 pixels, captured with the custom-built hyperspectral imaging system with 33 channels and a spectral resolution of 10 nm. The image was captured in the year 2003 from a field-based platform (at a distance of approximately 50 m away), so the

(a) A colour composite image of the gualtar steps dataset.



(b) Gualtar steps mean class spectra of pixels in sunlight.

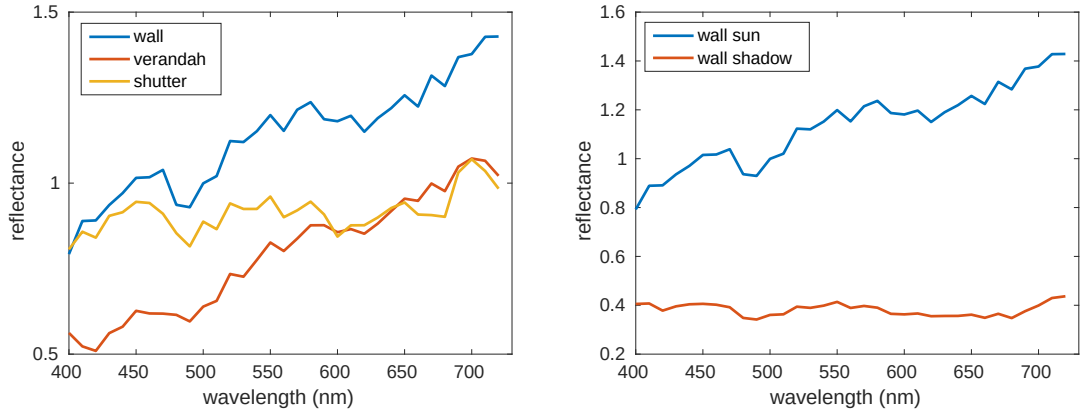


(c) Comparison of mean spectra in sunlight and in shadow for the Gualtar steps.

Figure 3.9 – Gualtar steps dataset.

spatial resolution was high.

The image was captured at midday and consequently there is a large shadow that covers roughly half of the image. The scene constitutes some brick steps with cement in between, grass and a white-painted wall (Figure 3.9b). Part of each class is in the sun and part is in the shadow (Figure 3.9c). In the sunlit region of the image, the brick steps also change orientation slightly with respect to the sun, thus, there is a small difference in the brightness of the light illuminating them. There is a small neutral probe sphere embedded in the scene which is used to convert the image spectra to relative reflectance.



(a) Gualtar timelapse mean class spectra of pixels in sunlight.

(b) Comparison of mean spectra in sunlight and in shadow in the 11:44 image.

Figure 3.10 – Gualtar timelapse dataset.

3.1.9 Dataset 8: Gualtar timelapse

A timelapse consisting of nine VIS images is captured of a building in Gualtar in 2003 (Figure 3.11) (Foster et al., 2015). Each image in the timelapse consists of 1024×1344 pixels with 33 channels having a spectral resolution of 10 nm. Just like the Gualtar steps, the image is captured with the custom-built hyperspectral imaging system mounted on a field-based platform (roughly 50 m away), so the spatial resolution is very high.

The images are captured at approximately one hour intervals starting from 11:45. There are natural changes in the illumination of the scene, with the most prominent change being the movement of a shadow over the building (Figure 3.10b). By 16:45 the building is completely in shadow. As with the Gualtar steps image, a neutral probe sphere in the image is used to convert the data to apparent reflectance.

3.1.10 Dataset 9: Pavia University

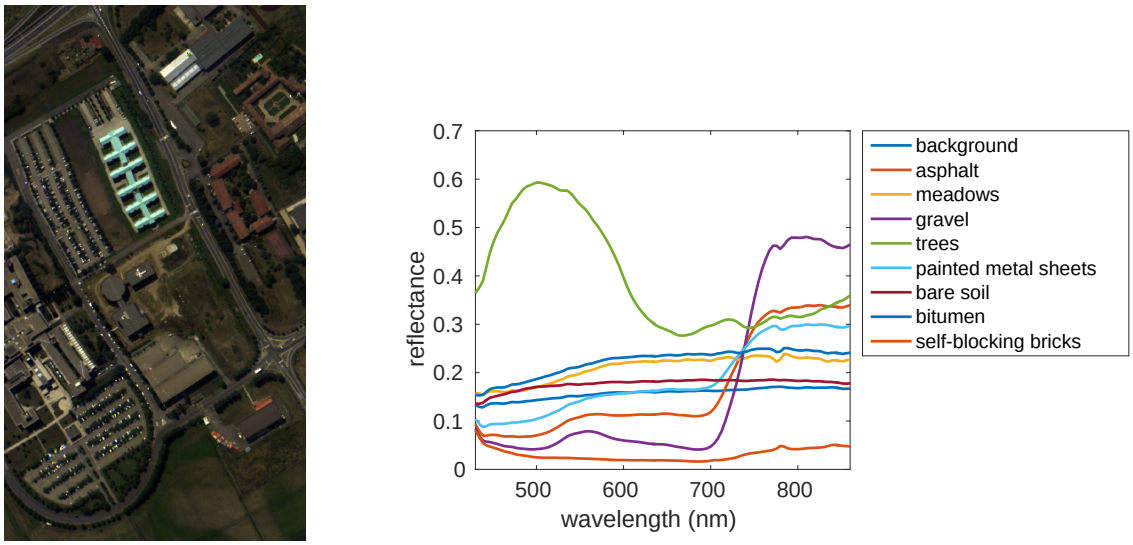
An aerial VNIR image (Figure 3.12a) is acquired over Pavia University using the ROSIS-3 sensor, consisting of 610×340 pixels with 103 spectral channels. The spectral resolution is 6 nm and the spatial resolution is approximately 1.3 m. The spectra in

(a) 11:44 **(b)** 12:45 **(c)** 13:46

(d) 14:47 **(e)** 15:45 **(f)** 16:45

(g) 17:46 **(h)** 18:53 **(i)** 19:44

Figure 3.11 – Colour composite images from the gualtar timelapse dataset.



(a) A colour composite image of the Pavia Uni dataset.

(b) Pavia Uni class spectra.

Figure 3.12 – Pavia Uni dataset.

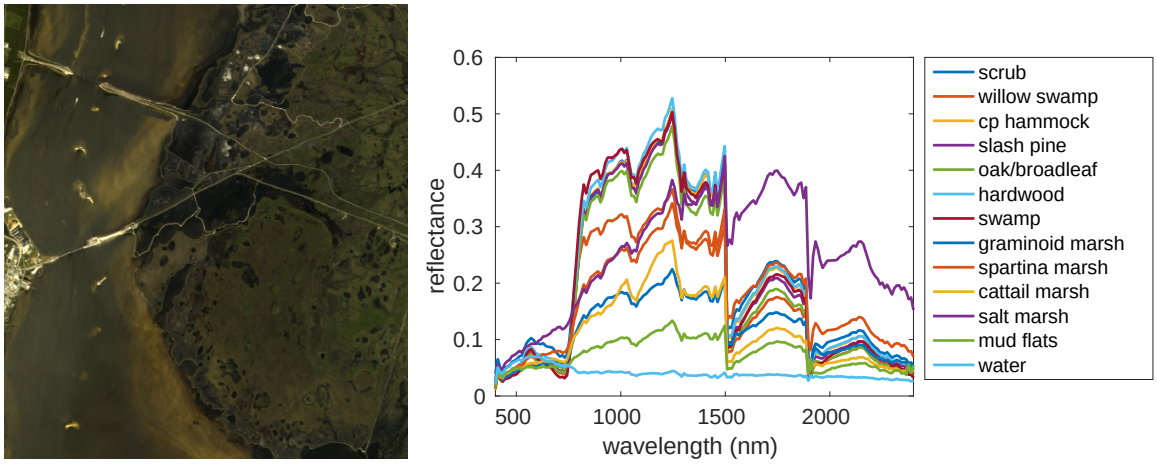
the image have been normalised to reflectance. The scene consists of nine classes, encompassing both urban and natural vegetation classes. Some of the classes are spectrally similar (Figure 3.12b). The Pavia University dataset has been provided by Professor Paolo Gamba, Pavia University.

3.1.11 Dataset 10: Kennedy Space Centre

The Kennedy space station (KSC) image (Figure 3.13a) was acquired by the AVIRIS VNIR/SWIR sensor in 1996. The image has 512×614 pixels and 176 spectral channels (after water absorption bands are removed). The spectral and spatial resolution are 10 nm and 18 m respectively. The 13 classes are a mix of urban and natural (Figure 3.13b). The data are normalised to reflectance.

3.1.12 Dataset 11: Indian Pines

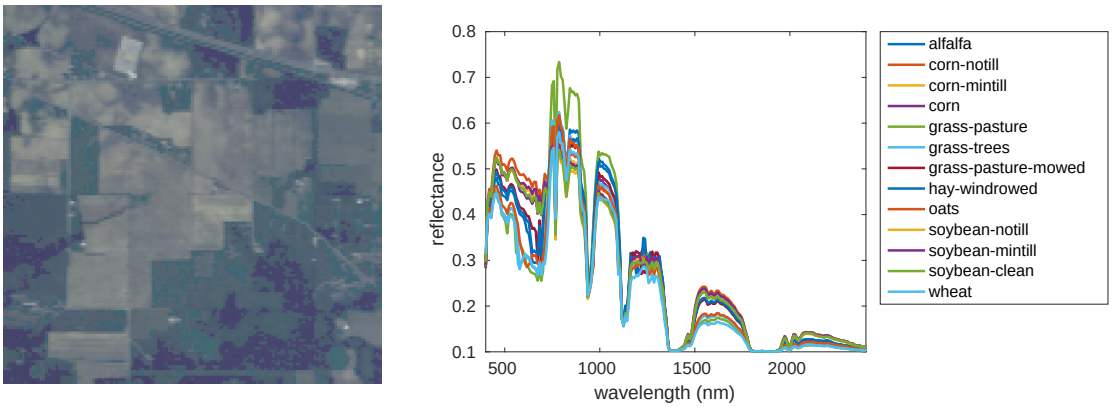
Indian Pines (Landgrebe and Biehl, 1992) is an aerial image (Figure 3.14a) of land that is part agriculture and part forest, captured in 1992. The 145×145 pixel image



(a) A colour composite image of the KSC dataset.

(b) KSC class spectra.

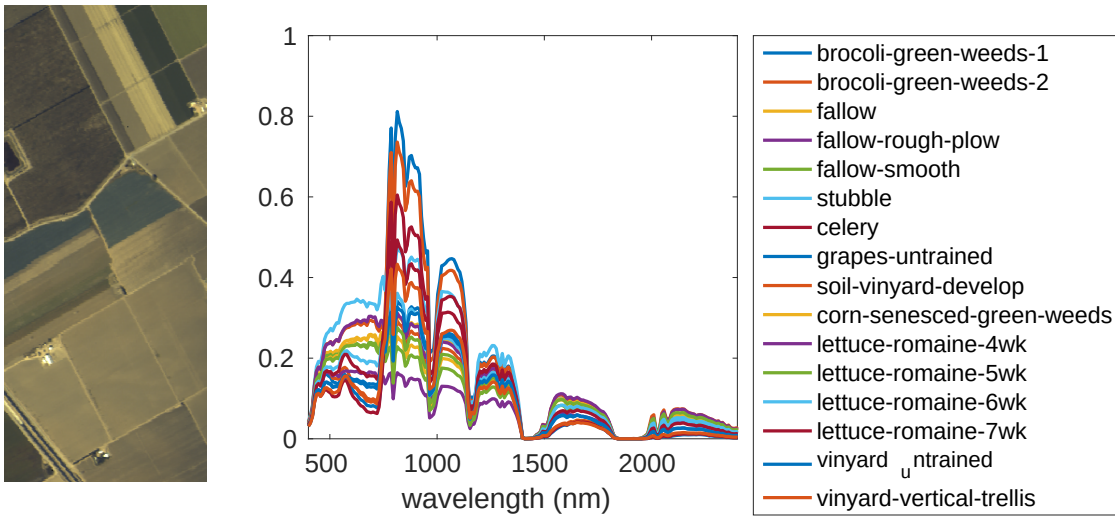
Figure 3.13 – KSC dataset.



(a) A colour composite image of the Indian Pines dataset.

(b) Indian Pines class spectra.

Figure 3.14 – Indian Pines dataset.



(a) A colour composite image of the Salinas dataset.

(b) Salinas class spectra.

Figure 3.15 – Salinas dataset.

is captured with the AVIRIS sensor and has a spatial resolution of 20 m. There are 200 channels once the water absorption bands are removed. The spatial resolution is 20 m and the spectral resolution is 10 nm. The image spectra are normalised to reflectance. Because the classes are all related to vegetation, the spectra all have similar absorption features (Figure 3.12a).

3.1.13 Dataset 12: Salinas

Similar to the Indian Pines dataset, the Salinas dataset is an aerial image of agricultural land captured with the AVIRIS sensor (Figure 3.15a). There are 145×145 pixels with a spatial resolution of 3.7 m and 204 channels once the water absorption bands have been removed. The spectral resolution is 10 nm (Figure 3.15b).

3.2 Metrics

The following describes the evaluation metrics used in this thesis.

3.2.1 Fisher's discriminant ratio

Fisher's discriminant ratio is a measure of how separable classes are in a feature space. When transforming data to a lower dimensional feature space, it is important that classes retain, or if possible improve, their separability in that space. Fisher's discriminant ratio (Theodoridis and Koutroumbas, 1998) is used in feature selection algorithms as a class separability criterion (Lin et al., 2004; Wang et al., 2011)

The measure is calculated for a pair of classes and takes the ratio of the between-class scatter and the within-class scatter. A pair of classes has a high score if their means are far apart and points within each class are close to other points in the same class, indicating good separability. For p dimensional data points from class A and class B, with respective means of $\boldsymbol{\mu_A}$ and $\boldsymbol{\mu_B}$ over all points in each class, the Fisher's discriminant ratio is calculated as:

$$J(A, B) = \frac{\|\boldsymbol{\mu_A} - \boldsymbol{\mu_B}\|_2^2}{S_A^2 + S_B^2}, \tag{3.2}$$

where J is the ratio, $\|\cdot\|_2$ is the L_2 norm, and S_i^2 is the within-class scatter of class i, given by:

$$S_i^2 = \frac{1}{N_i} \sum_{n \in N_i} \|\mathbf{x_n} - \boldsymbol{\mu_i}\|_2^2, \tag{3.3}$$

where $\mathbf{x_n}$ is a point in class i, which has N_i points in total.

An important property of this measure is that it is invariant to the scale of the data points. This allows feature spaces found using different approaches to be compared consistently. It is also important that the measure is invariant to the number of dimensions p so that the data with the original dimensionality can be compared to data with reduced dimensionality.

For multi-class problems, the mean of the Fisher's discriminant ratio for all possible pairs of classes can be found.

3.2.2 Adjusted rand index

The adjusted rand index (ARI) is an evaluation metric for clustering (Hubert and Arabie, 1985; Rand, 1971). It measures the similarity of two data clusterings. If one of those data clustering is the actual class labels, then the ARI can be considered the accuracy of the clustering solution. This measure is useful as an indirect measure of how separable classes are in a feature space (Yeung and Ruzzo, 2001). If the classes are well separated, then they should be easy to cluster and the ARI should be high.

If n_i and n_j are the number of points in class u_i and cluster v_j respectively and n_{ij} is the number of points that are in both class u_i and cluster v_j, then the ARI is calculated as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}\right]/\binom{n}{2}}. \tag{3.4}$$

This adjustment of the rand index makes it more likely that random label assignments will get a value close to zero. The best clustering score has an ARI of 1, and the worst clustering score (equivalent to randomly allocating cluster association) has an adjusted rand index of 0.

3.2.3 Peak signal-to-noise ratio

In an image processing context, the peak signal-to-noise ratio (PSNR) is used to compute how similar two images, given one of the images has been corrupted in some way (eg. it has been compressed, reconstructed or noise has been added). Hence, it is measuring the quality of the corrupted image with respect to the original image (Hore and Ziou, 2010; Huynh-Thu and Ghanbari, 2008). This measure can be used

to evaluate how invariant a spectral feature mapping is to the illumination conditions by computing how similar images captured at different times are when represented by the feature (Xie et al., 2011). For example, if a feature is not illumination invariant, then images represented by the feature captured at times where the illumination conditions are different will not be similar, and the PSNR will be low.

The PSNR, in decibels, is calculated as:

$$PSNR(Im_1, Im_2) = 10 \times log_{10} \frac{peakval^2}{MSE(Im_1, Im_2)}, \tag{3.5}$$

where the mean squared error (MSE) between the two images Im_1 and Im_2 with R rows and C columns is:

$$MSE(Im_1, Im_2) = \frac{1}{RC} \sum_{i \in R} \sum_{j \in C} [Im_1(i,j) - Im_2(i,j)]^2. \tag{3.6}$$

The higher the PSNR is, the more similar the two images are.

3.2.4 Percentage change in classification label

Given two classified images, the percentage of pixels in the first image that had a different classification label in the second image is calculated as:

$$\%change = 100 \times \frac{N_{changed}}{N_{total}}, \tag{3.7}$$

where $N_{changed}$ is the number of pixels in first image that changed classification label and N_{total} is the total number of pixels in one image.

This measure can be used to quantify the impact of variable illumination conditions on classification performance, by looking at the proportion of pixels that change label between images captured at different times (Schneider et al., 2011). It requires a timelapse of spatially registered images.

3.2.5 F1 classification score

Also known as the F-score, the F1 classification score is a measure of binary classification accuracy. It is the harmonic mean of precision and recall (Van Rijsbergen, 1979):

$$F_1 score = 2 \times \frac{precision * recall}{precision + recall}, \tag{3.8}$$

where precision is the fraction of the instances retrieved by the classifier that are relevant, and the recall is the fraction of relevant instances that are retrieved by the classifier. If relevant instances are considered "true" and none-relevant instances considered "false", and if instances retrieved by the classifier are considered "positive" and other instances considered "negative", then the precision and recall can be calculated as:

$$precision = \frac{TP}{TP + FP}, \tag{3.9}$$

$$recall = \frac{TP}{TP + FN}, \tag{3.10}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. An F1 score of 1 indicates the best accuracy, and a score of 0 indicates the worst accuracy.

For a multi-class classification problem, the classification of each class is individually considered as a binary classification task, and the F1 score is determined for that class. Then the mean of the F1 scores for all classes is calculated.

3.2.6 Number of epochs

The number of epochs that the optimisation of a neural network requires to reach within 1% of convergence is a good indication of the training time. Whilst a direct measurement of the training time is more intuitive to interpret, it is highly dependent on the computer processor and software packages used. This makes the result non-generic. If the training time is measured in terms of epochs, then an estimate of the

training time can be calculated for any processor and package if the time taken for one epoch with that processor is known.

Chapter 4

Unsupervised Illumination Invariant Representation of Hyperspectral Data

This chapter proposes a method for unsupervised feature learning/dimensionality reduction of hyperspectral data, such that the low dimensional feature representation is robust to the variability caused by the scene geometry and illumination conditions (e.g. shadows) whilst maintaining the high class discriminability of the full spectrum. Chapter 1 motivated the importance of having robust feature representations of hyperspectral images. Section 2.4.1 described many of the current approaches to compensating for the variability in spectra due to the influence of the atmosphere and geometry on illumination. It remains a largely unsolved problem with most existing solutions either failing to account for the influence of geometry or requiring *a priori* knowledge, labelled data or the use of additional sensor modalities such as LiDAR. Coupled with the fact that additional sensor modalities and *a priori* knowledge are not always available, there exists a need for the development of unsupervised algorithms to process hyperspectral images as there is often only a limited amount of labelled data available due to the large manual effort needed to annotate on a per-pixel scale (see Chapter 1). Many labelled samples would be required to capture the

large amount of variability in the data due to the environmental factors (i.e. scene illumination).

The methods proposed in this chapter integrate domain knowledge into an autoencoder - an unsupervised deep learning framework. The autoencoder has the advantage of reducing the high dimensionality of the hyperspectral data, where correlation results in large amounts of redundancy in the data (Demarchi et al., 2014). In the process of reducing the dimensionality of the data, the autoencoder also extracts features with greater representative power than the raw intensity or reflectance. The features learnt by the final autoencoder proposed in this chapter can be used to discriminate between different classes and are also less sensitive to the influence of the variable geometry of the scene on the incident illumination. Currently, both the illumination variability in the environment and the high dimensionality of hyperspectral images limit the usefulness of these sensors in outdoor robotics, computer vision and remote sensing applications (Ramakrishnan et al., 2015; Schneider et al., 2011). Many high-level algorithms such as those that do mapping/classification, target detection, object detection, segmentation and change detection would benefit from low-dimensional, illumination invariant representations of hyperspectral images.

The advantages of the unsupervised approaches proposed in this chapter over other techniques in the literature (Section 2.4) are that they require no labelled training data, no additional sensors and no *a priori* knowledge of location, season or atmospheric conditions. These approaches also do not make the same assumptions as many of the computer vision approaches, in particular, the Planckian illumination assumption, which is less likely to hold true in the context of hyperspectral imaging (as explained in Section 2.4.1). Instead, they rely on the learning power of the neural networks with incorporated domain knowledge.

In Section 4.1, a set of autoencoders are proposed for finding low dimensional feature representations of hyperspectral data. These autoencoders draw from remote sensing techniques to learn a superior representation. They are also designed to be insensitive to changes in brightness. In Section 4.2, the relit spectral angle-stacked autoencoder (RSA-SAE) is proposed which extends the hyperspectral autoencoder to make it

invariant to shadows. Experimental results validating the proposed algorithms are presented in Section 4.3 and discussed in Section 4.4 before conclusions are made in Section 4.5.

4.1 Hyperspectral Stacked Autoencoders

Autoencoders are a specific type of deep learning algorithm capable of reducing the dimensionality of data with multiple layers of non-linear functions that can learn a mapping from a high dimensional space to a low dimensional feature space. A typical autoencoder uses a reconstruction cost function that is based on the Euclidean distance (Section 2.3.2). In the remote sensing literature, alternative measures to the Euclidean distance are used for comparing spectra to one another. These measures are more suited to spectral data because, unlike the Euclidean distance, they are predominantly dependent on the shape of the spectra rather than their magnitude. The novel autoencoders proposed in this section use reconstruction cost functions based on spectral similarity measures used in the remote sensing literature instead of the standard Euclidean distance. These are the spectral angle (Yuhas et al., 1992), the cosine of the spectral angle and the spectral information divergence (SID) (Chang, 2000). Each of these has been incorporated into the backpropagation learning algorithm for the autoencoder. Because of the dependence of the similarity measures on spectral shape, it is expected that these autoencoders will learn a better low dimensional feature representation than those that use the Euclidean distance. These similarity measures also ostensibly have brightness invariant properties which make the representations learnt by the proposed autoencoders invariant to the brightness. Spectra reflected from surfaces with different orientations with respect to the position of the sun are similar in shape. Thus, it is also expected that the autoencoders that use the remote sensing similarity measures will out-perform those using the Euclidean distance in scenarios with highly variable scene geometry.

4.1.1 Cosine Spectral Angle Stacked Autoencoder

The spectral angle is a similarity measure between two vectors. The angle between vectors is indicative of their shape instead of their magnitude, making it a useful measure for applications utilising spectral data. This measure is also insensitive to differences in brightness, which can be interpreted as changes in the magnitude of the spectral vectors, rather than changes in the direction of the vectors in the multi-dimensional space. Hence, the angle remains the same. By using a reconstruction cost function for an autoencoder based on the spectral angle, the features learnt are insensitive to variations in the brightness of spectra.

The spectral angle θ_{SA} is the angular distance between two spectral vectors, $\mathbf{A}$ and $\mathbf{B}$ of dimensionality T, given by:

$$\theta_{SA} = \cos^{-1} \frac{\sum_{t=1}^{T} A_t B_t}{|\mathbf{A}||\mathbf{B}|}. \tag{4.1}$$

For the first novel autoencoder, the cosine of the spectral angle is incorporated into the reconstruction cost function. That is, the reconstruction cost is calculated as the cosine of the spectral angle between the reconstructed spectrum output by the encoder-decoder network and the spectrum input into the network. As the cosine of the angle gets larger for smaller distances, it must be subtracted from one such that smaller distances, which are desired, correspond to minimising the overall cost. For a single observation, the reconstruction cost is:

$$E_{CSA}(f(\mathbf{z}^{(L)}), \mathbf{y}) = 1 - \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|}, \tag{4.2}$$

where K is the original dimensionality (i.e. number of bands), L is the index of the output layer, y_k is an element of the target data $\mathbf{y}$ which is set to equal the input data $\mathbf{x}$, f is the activation function, $f(z_k^{(L)})$ is an element of the reconstructed input $f(\mathbf{z}^{(L)})$ and:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l-1)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l-1)}, \tag{4.3}$$

$$\mathbf{a}^{(l)} = f(\mathbf{z}^{(l)}), \tag{4.4}$$

$$\mathbf{a}^{(1)} = \mathbf{x} \tag{4.5}$$

for $l = L, L-1, L-2, L-3, ..., 2$, with learnable parameters $\mathbf{W}$ and $\mathbf{b}$.

The reconstruction cost function for all observations, including a regularization term is:

$$E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2, \tag{4.6}$$

where M is the number of observations, λ is the regularization parameter, and I and J are the number of units in layers l and $l+1$ respectively. The regularization term prevents the parameters from getting too large whereby overfitting occurs. The partial derivatives for backpropagation are:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial W_{ji}^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \lambda W_{ji}^{(l)}, \tag{4.7}$$

$$\frac{\partial}{\partial b_{j}^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial b_{j}^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}), \tag{4.8}$$

where for a single observation m, when $l = 1$:

$$\frac{\partial}{\partial W_{ji}^{(1)}} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) = \delta_j^{(2)} x_i. \tag{4.9}$$

and when $l = 2, 3, ..., L-1$:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) = \delta_j^{(l+1)} f(z_i^{(l)}). \tag{4.10}$$

$$\frac{\partial}{\partial b_{j}^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) = \delta_j^{(l+1)}, \tag{4.11}$$

for $l = 1, 2, 3, ..., L-1$. The value of δ is dependent on the layer number l. For $l = L$,

$$\delta_k^{(L)} = \frac{f'(z_k^{(L)})}{|f(\mathbf{z}^{(L)})||\mathbf{y}|} \left[\frac{(f(\mathbf{z}^{(L)}) \cdot \mathbf{y}) f(z_k^{(L)})}{|f(\mathbf{z}^{(L)})|^2} - y_k \right], \tag{4.12}$$

and for $l = L-1, L-2, L-3, ..., 2$,

$$\delta_i^{(l)} = \sum_{j=1}^{J} (\delta_j^{(l+1)} W_{ji}^{(l)}) f'(z_i^{(l)}). \tag{4.13}$$

The parameter update equations for gradient descent optimisation are:

$$W_{ji}^{(l)} := W_{ji}^{(l)} - \alpha \frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{4.14}$$

$$b_j^{(l)} := b_j^{(l)} - \alpha \frac{\partial}{\partial b_j^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{4.15}$$

where α is the learning rate. This approach is called the cosine spectral angle-stacked autoencoder (CSA-SAE), and its derivation is in Appendix A.

4.1.2 Spectral Angle Stacked Autoencoder

The spectral angle itself is also proposed as a novel reconstruction cost function. Like the cosine of the spectral angle, it is a similarity measure between two vectors and is insensitive to differences in brightness. In comparison to its cosine counterpart, the spectral angle requires an additional operation to calculate. It is, however, sometimes preferred. The spectral angle gets smaller when spectra are more similar, so it does not need to be subtracted from one when calculating the cost. The spectral angle reconstruction cost function is calculated as:

$$E_{SA}(f(\mathbf{z}^{(L)}), \mathbf{y}) = \cos^{-1} \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|}, \tag{4.16}$$

and the reconstruction cost function for all observations, including a regularization term is:

$$E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} E_{SA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2. \tag{4.17}$$

The partial derivatives for backpropagation are the same as in 4.7 and 4.8, however the similarity cost is the E_{SA} instead of the E_{CSA}, and the calculation of δ for $l = L$

is now:

$$\delta_k^{(L)} = \frac{1}{\sqrt{1 - \left[\frac{(f(\mathbf{z}^{(L)}) \cdot \mathbf{y})}{|f(\mathbf{z}^{(L)})||\mathbf{y}|}\right]^2}} \frac{f'(z_k^{(L)})}{|f(\mathbf{z}^{(L)})||\mathbf{y}|} \left[\frac{(f(\mathbf{z}^{(L)}) \cdot \mathbf{y}) f(z_k^{(L)})}{|f(\mathbf{z}^{(L)})|^2} - y_k\right], \tag{4.18}$$

with 4.13 still applying for $l = L-1, L-2, L-3, ..., 2,$. The parameter update equations are as in 4.14 and 4.15. This approach is called the spectral angle-stacked autoencoder (SA-SAE), and its derivation is in Appendix B.

4.1.3 Spectral Information Divergence Stacked Autoencoder

The SID is an information-theoretic measure which measures the probabilistic discrepancy between two spectra in order to calculate their similarity. Experiments have shown that it can preserve spectral properties and characterise spectral variability more effectively than the spectral angle (Chang, 2000).

The SID between two spectra $\mathbf{A}$ and $\mathbf{B}$ is given by:

$$SID(\mathbf{A}, \mathbf{B}) = \sum_{n=1}^{N} p_n log \frac{p_n}{q_n} + \sum_{n=1}^{N} q_n log \frac{q_n}{p_n} \tag{4.19}$$

where N is the length of the vectors $\mathbf{p}$ and $\mathbf{q}$, and

$$\mathbf{p} = \frac{\mathbf{A}}{\sum_{t=1}^{T} A_t}, \tag{4.20}$$

$$\mathbf{q} = \frac{\mathbf{B}}{\sum_{t=1}^{T} B_t}. \tag{4.21}$$

To incorporate the SID into the autoencoder cost function, 4.19 is first simplified to:

$$SID(\mathbf{A}, \mathbf{B}) = \sum_{n=1}^{N} (p_n - q_n)(log(p_n) - log(q_n)). \tag{4.22}$$

Then, by making $\mathbf{A}$ the reconstructed spectrum output by the network and $\mathbf{B}$ the

spectrum input into the network, the relevant terms are substituted into 4.22:

$$E_{SID}(f(\mathbf{z}^{(L)}),\mathbf{y}) = \sum_{k=1}^{K}\left[\frac{f(z_k^{(L)})}{\sum_{d=1}^{K} f(z_d^{(L)})} - \frac{y_k}{\sum_{d=1}^{K} y_d}\right][log f(z_k^{(L)}) - log\sum_{d=1}^{K} f(z_d^{(L)}) - log(y_k) + log\sum_{d=1}^{K} y_d], \quad (4.23)$$

and the reconstruction cost function for all observations, including a regularization term is:

$$E(\mathbf{W},\mathbf{b}) = \frac{1}{M}\sum_{m=1}^{M} E_{SID}(\mathbf{W},\mathbf{b};\mathbf{y}^{(m)}) + \frac{\lambda}{2}\sum_{i,j,l=1}^{I,J,L-1}(W_{ji}^{(l)})^2. \quad (4.24)$$

Just like the autoencoder based on the spectral angle, the partial derivatives for backpropagation are the same as in 4.7 and 4.8, but with E_{SID} instead of E_{CSA} and δ for $l = L$ calculated differently:

$$\delta_k^{(L)} = -\frac{f'(z_k^{(L)})}{\sum_{d=1}^{K} f(z_d^{(L)})}\left[\frac{q_k}{p_k} - log\frac{p_k}{q_k} - 1 + \sum_{d=1}^{K}(p_d - q_d + p_d log\frac{p_d}{q_d})\right], \quad (4.25)$$

where

$$\mathbf{p} = \frac{f(\mathbf{z}^{(L)})}{\sum_{c=1}^{K} f(z_c^{(L)})}, \quad (4.26)$$

$$\mathbf{q} = \frac{\mathbf{y}}{\sum_{c=1}^{K} y_c}, \quad (4.27)$$

with 4.13 still applying for $l = L-1, L-2, L-3, ..., 2$. The parameter update equations are once again as in 4.14 and 4.15. This approach is called the spectral information divergence-stacked autoencoder (SID-SAE), and its derivation is in Appendix C.

4.2 Relit Spectral Angle Stacked Autoencoder

Variability in illumination can lead to poor performance by high-level algorithms such as clustering, classification or segmentation. Compensating for this variability is still a largely unsolved problem (see Section 2.4.1). The dimensionality reduction/feature

extraction methods proposed in Section 4.1 naturally have some illumination invariance properties relating to the similarity measures used to derive them. For instance, the spectral angle measure is insensitive to differences in intensity between pixels. When used as the reconstruction cost function of an SAE, the code layer is learnt so that the shape of the reflectance vector is reconstructed instead of its magnitude. Semantically similar pixels that are illuminated with different intensities, will have similarly shaped reflectance vectors, and thus are more likely to appear similar in the learnt code layer. This is similarly the case with the SID-based autoencoder, where differences in brightness are measured as variances in a probability distribution.

The limitation of these methods are their failure to similarly represent pixels belonging to the same material that are within sunlit and shadowed regions, i.e. those that are illuminated predominantly by diffuse skylight and not by sunlight. Under these conditions, the shape of the spectral curve and its magnitude changes because its wavelength-intensity distribution changes. This results in the shadowed pixels taking on different values in the code layer from their sunlit equivalents (e.g. the spectral angle is insensitive to a constant multiplier across all bands, but not changes in the wavelength-intensity distribution). Shadows can occur for reasons related to scene geometry, such as a region being occluded from direct sunlight by another part of the scene, or because the normal vector of a surface is at a 90 degree angle to the line-of-sight with the sun. Shadows can also occur due to variable and patchy cloud cover in the sky occluding the path of light.

By incorporating an outdoor illumination model into the autoencoder framework, the approach proposed in this section aims to overcome the inability of the autoencoders described in (Section 4.1) to encode identical materials in sunlit and shadowed regions similarly. This novel SAE is called the RSA-SAE. The method combines and leverages a physics-based model from a multi-modal technique with a radiative transfer model from the remote-sensing literature, within the unsupervised learning framework of a DAE. Instead of relying on the assumption of Planckian illumination, the proposed approach differs to those based on the Lambertian model for photodetector response (2.24) as it learns the illumination invariant mapping from the data using the DAE's

multiple layers of non-linear functions.

The outcome of the approach is a mapping, learnt from the image spectra, to a low-dimensional, shadow invariant encoding, producing an illumination invariant representation of the pixels in the hyperspectral image. The method does not need any labels for the training data nor, *a priori* knowledge of the scene, such as the scene geometry or atmospheric conditions. Therefore, no data from additional sensors are required. Once trained on spectra from a given scene, the learnt mapping can also be used on similar scenes acquired under different lighting conditions (e.g. captured at different times of the day). The spectra will be consistently mapped to an illumination invariant encoding. The encoded data can be used for high-level pixel-wise applications such as classification and is particularly useful in scenarios where there is limited or no labelled training data available for shadowed regions to train the supervised models.

4.2.1 Overview

Inspired by both the DAE and data augmentation, the RSA-SAE learns a shadow-invariant encoder-decoder network. The encoder phase of the RSA-SAE aims to map identical materials in shadowed and sunlit regions to the same low dimensional representation, as shown in Figure 4.1. This is achieved by first obtaining a hyperspectral image, with each pixel represented either by raw digital numbers or converted to radiance units through radiometric calibration. A sunlit pixel in the image is automatically selected using weighted sampling of the squared image intensity and normalised to reflectance using flat-field correction with a calibration board of known material reflectance properties. Of course, this normalisation only correctly removes the effects of illumination for pixels with the same or similar orientation as the calibration panel and under precisely the same conditions of illumination (i.e. not pixels in shadow). The original data is then relit such that it is illuminated by diffuse skylight only and normalised to reflectance using the same flat-field measurement of the illumination as before. The normalised relit spectra are therefore a representation of what

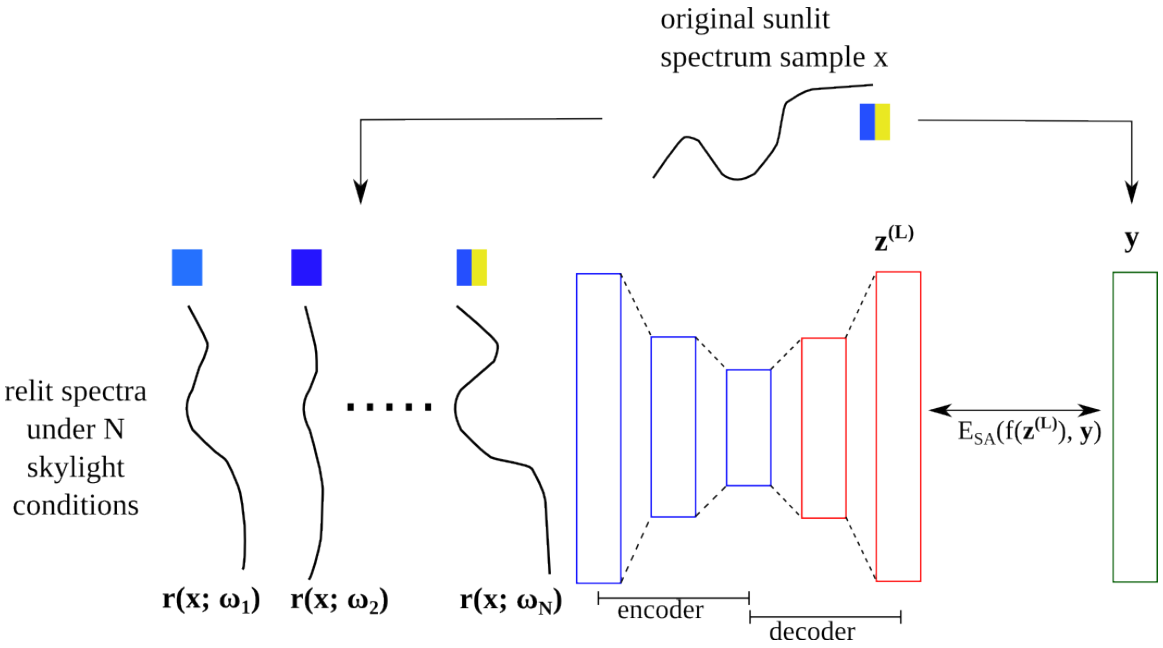


Figure 4.1 – Framework for training the RSA-SAE network. The patches of colour indicate the illumination colour, with blue and yellow representing skylight and sunlight respectively. When the square is completely blue, it indicates that the pixel is illuminated by skylight only and hence is in shadow, with the different shades of blue representing different atmospheric conditions. By minimising the spectral angle cost function E_{SA}, the network is trained to reconstruct the original sunlit spectra (the green bar) from the input spectra after it is relit to be in shadow using one of the candidate atmospheres ω_n and the relighting equation r. After the network is trained, the encoder is expected to map real shadowed and sunlit spectra from the same class to the same values in the code layer, which is the new low dimensional representation of the data.

the reflectance of the pixel would be, if it were in a shadowed region in the original normalised image. The relit reflectance spectra $\tilde{\mathbf{x}}$ and original reflectance spectra $\mathbf{x}$ are used as the corrupted and uncorrupted data, respectively, in the RSA-SAE. The network is trained to reconstruct the uncorrupted spectra in the output layer, from both uncorrupted and corrupted spectral inputs.

Since the RSA-SAE is being trained to reconstruct the same spectra despite the input being in shadow or sunlight, it is expected that the hidden layer will be invariant to the illumination conditions. Hence, the hidden layer is a low dimensional representation of the data, with the encoder providing an illumination invariant mapping to the representation.

4.2.2 Autoencoder Framework

The DAE framework consists of several symmetric, fully connected layers of neurons, with the input data being altered by a relighting corruption process, and the uncorrupted data reconstructed in the output layer. The number of neurons used in each layer is dependent on the dataset. The number of neurons in the first and last layer is always equal to the number of bands in the data and the number of neurons in the code layer is the number of desired dimensions. The number of neurons in the intermittent layers are chosen such that each layer gets smaller from the input to the code layer and then symmetrically larger from the code layer to the output layer. Data are therefore being progressively forced through a bottleneck to learn the most important information required to reconstruct the data. There is a sigmoidal activation function associated with each neuron. This choice of activation function is justified in the next section.

4.2.3 Training

As is the case with standard SAEs, training a network with many layers using back-propagation often results in poor solutions (Hinton and Salakhutdinov, 2006). Hence, an initial solution is found using a pre-training step (Bengio et al., 2007) whereby a set of single-layered autoencoders are first trained on the data in a greedy fashion without any relighting corruption, and by using the squared error reconstruction cost function (2.14). The network parameters learnt in the layerwise pre-training are then used to initialise the full network for training using the RSA-SAE method described above. The RSA-SAE training is effectively an end-to-end fine-tuning step. The RSA-SAE uses the spectral angle reconstruction cost function (4.17) so that that structure in the shape of the reflectance spectra can be learnt rather than the brightness, with:

$$\mathbf{a}^{(1)} = \tilde{\mathbf{x}} = r(\mathbf{x}), \tag{4.28}$$

where r is the relighting equation.

The pre-training of the single-layered autoencoders is done using spectra from the entire hyperspectral image. Since fine-tuning is performed using the relit spectra as well, the dataset increases in size by the number of augmentations used. Hence, training time increases, even though the bulk of the learning has already been done in the pre-training step. Since the RSA-SAE fine-tuning of the network is only training the autoencoder to encode spectra from the same class under different illuminations similarly, a sample of the spectra is used. To do this, spectra from the hyperspectral image are spatially sampled using weighted sampling of the squared intensity of the image pixels. Besides reducing the number of points for training, this has the additional benefit of ensuring most of the points are in the sun (as it is sunlit points which are relit to shadow). Although, preliminary tests indicate it does not hinder the performance if some of the sampled points are in shadow.

A property of the spectral angle reconstruction cost is that it is undefined for $|f(\mathbf{z}^{(L)})| = 0$. Hence, an activation function f must be chosen that does not include zero in its range. This removes the chance of having an undefined reconstruction cost. A sigmoid function is chosen as the activation function because it is bound by a horizontal asymptote at zero such that its range is $0 < f(x) < 1$. It is therefore impossible for $|f(\mathbf{z}^{(L)})|$ to equal zero, which is not the case if functions such as ReLU and the hyperbolic tangent are used.

4.2.4 Spectral Relighting

The spectral relighting augmentation used by the RSA-SAE is based on the outdoor illumination model of equation 2.2. In Ramakrishnan et al. (2015), this model was used to derive a relighting method, where entire images were relit using a common illuminant. In order to relight the radiance from a region (A), to have diffuse skylight as its illuminant, the original spectra $\mathbf{L_A}$ is multiplied by a wavelength dependent scaling factor given by equation 2.3.

The above process ensures all of the spectra in the image appeared as though they were under constant illumination. For the RSA-SAE, radiometrically normalised spectra

are relit using equation 2.3 to have a diffuse skylight illuminant, as this emulates what they would look like under shadow. This is the corruption process which the DAE learns to reverse by reconstructing the spectra to be illuminated by sunlight in the output layer of the autoencoder.

The relighting equation (2.3) requires knowledge of the atmospheric parameters, namely the terrestrial sunlight-diffuse skylight ratio $\frac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}$, and the scenes geometric parameters, the visibility, sun angle and sky factor. This is to accurately emulate the illumination in the scene of interest. In Ramakrishnan et al. (2015), geometric data obtained from the LiDAR could be used to easily determine the visibility, sun angle and sky factor. The terrestrial sunlight-diffuse skylight ratio is also estimated using the LiDAR by selecting two points across a shadow boundary obtained from the same material and with the same geometry. Regarding the RSA-SAE, as no additional sensors such as LiDAR are assumed to be used during data collection, the atmospheric conditions and scene geometry required for relighting are unknown.

To overcome this lack of *a priori* knowledge, a large number of candidate atmospheric models are generated for each spectral sample through the use of the SMARTS radiative transfer model (Gueymard, 2001). The parameters of this model are randomly sampled for each candidate and the network is trained to reconstruct all of the relightings of a given sunlit reflectance spectrum as the same sunlit reflectance (Figure 4.1). The intuition behind this approach is that the network captures commonality in the data and learns to reconstruct sunlit spectra from shadowed spectra regardless of the prevailing atmospheric composition.

The SMARTS radiative transfer model requires several parameters as input, including turbidity, carbon dioxide, oxygen, ozone and water vapour. In this scenario, the only sensor used is the camera, so measurements of the atmospheric composition are unavailable. To generate a candidate atmosphere, each parameter is uniformly sampled between the recommended upper and lower bounds typical for that parameter (Table. 4.1). The model outputs direct normal irradiance (equivalent to $\mathbf{E}_{sun}\boldsymbol{\tau}$) and the diffuse skylight irradiance ($\mathbf{E}_{sky}$). To corrupt the spectra at the input of the autoencoder, the candidate illumination source spectra are input into the relighting

equation (2.3), along with uniformly sampled sun angle and sky factor parameters (Table. 4.2). The sky factor is sampled between 0 where no sky is visible and 1 where the full hemisphere is visible. The sun angle is sampled between 0° where the surface is directly facing the sun and 90° where the surfaces normal is perpendicular to the light of sight to the sun. The visibility is set to 0 to so that the relighting is emulating shadow.

Table 4.1 – The parameters sampled from the SMARTS radiative transfer model to generate the candidate terrestrial sunlight-diffuse skylight ratios needed for relighting. All parameters are sampled uniformly between the given bounds, with the lower bound exemplifying pristine conditions and the upper bound exemplifying severe atmospheric pollution.

Gas	**lower bound concentration (ppm)**	**upper bound concentration (ppm)**
formaldehyde	-0.003	0.007
methane	0	0.4
carbon monoxide	-0.1	9.9
nitrous acid	-9.9E-4	0.01
nitric acid	0	0.012
nitric oxide	0	0.5
nitrogen dioxide	0	0.2
nitrogen trioxide	-4.9E-4	2E-4
ozone	-0.007	0.175
sulphur	0	0.2
carbon dioxide	360	380
visibility (turbidity)	0.77 km	764 km

Table 4.2 – The geometric parameters sampled for the relighting. All parameters are sampled uniformly between the given bounds.

param	**lower bound**	**upper bound**
sky factor Γ	0	1
sun angle θ	0°	90°

Figure 4.2 summarises how the RSA-SAE is trained and used to find low-dimensional, illumination invariant representations of hyperspectral images.

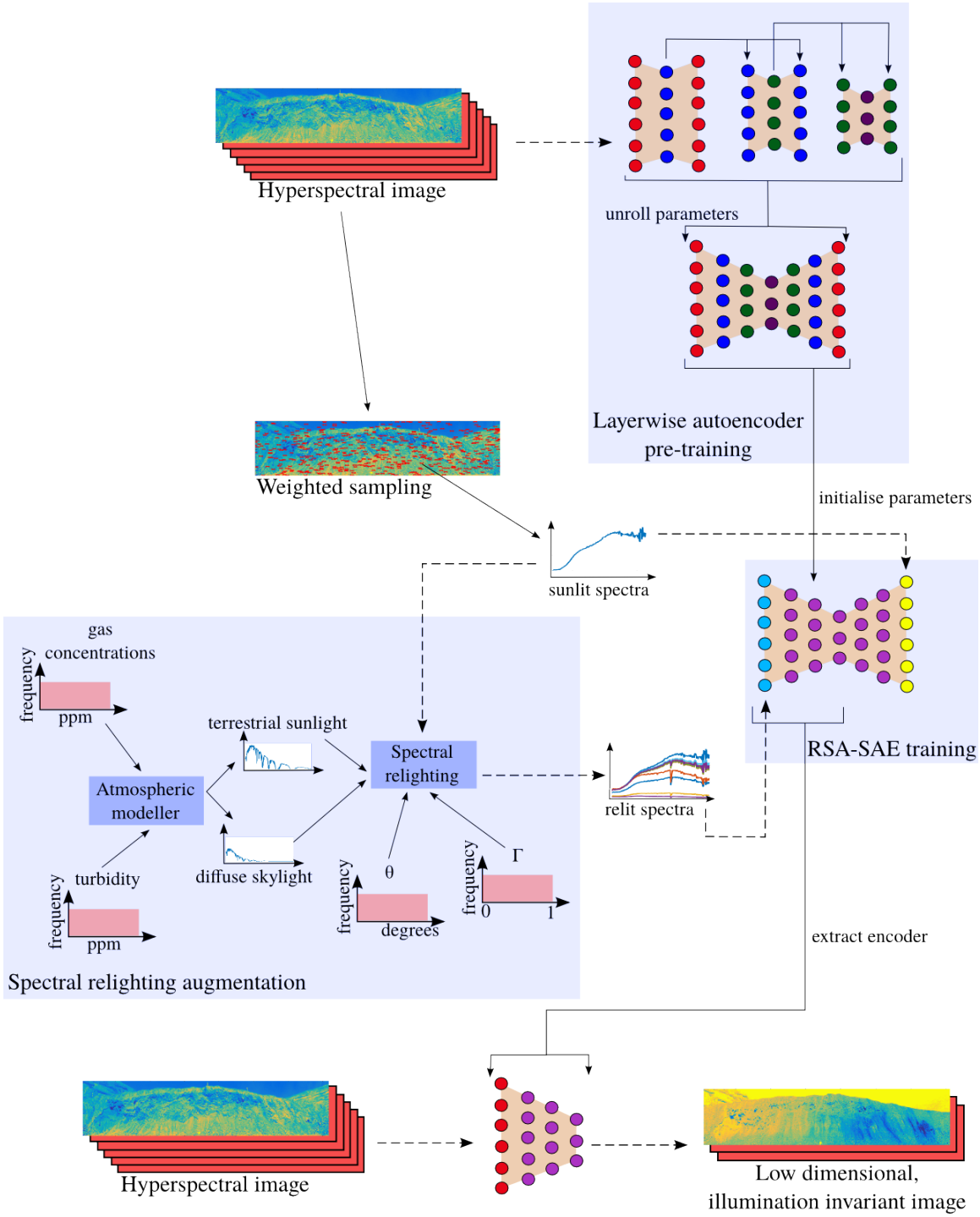


Figure 4.2 – Summary of process for training the RSA-SAE network to find low-dimensional, illumination invariant representations of hyperspectral images. The layerwise autoencoder pre-training and RSA-SAE training are described in more detail in Section 4.2.3 and the spectral relighting augmentation is described in Section 4.2.4.

4.3 Experimental Results

The hyperspectral autoencoders proposed in Section 4.1 and the illumination invariant extension proposed in Section 4.2 were evaluated experimentally with a number of datasets to demonstrate the algorithms ability to generalise to different scenes, illumination conditions and sensors used to capture the images. The datasets covered both simple and complex scene geometries.

4.3.1 Network Architecture and Parameters

The network architecture and tunable parameters of the autoencoder were found in preliminary experiments using a coarse grid search, minimising the reconstruction error. A common architecture was chosen for the experiments to be described in Section 4.3.2 that performed well across all of the datasets. This was not necessarily the best performing architecture for each dataset, however, the purpose of the experiments was to demonstrate improved performance for a given architecture. For the Great Hall VNIR, KSC, Pavia Uni, Simulated and X-rite datasets, an encoder architecture of K-100-50-10 was used with symmetric decoders, where K is the original dimensionality of the data. The Gualtar steps dataset required a different architecture due to its relatively small number of bands. The architecture for the Gualtar Steps and Gualtar timelapse datasets was 33-25-10-5, with symmetric decoders. For the experiments described in Section 4.3.3, the Great Hall SWIR dataset had an architecture of 152-50-40-30 and the mineface timelapse dataset had an architecture of 220-100-50-30. Note, the noisy spectral regions affected by atmospheric water absorption (1110-1155, 1338-1514 and 1790-2078 nm) were removed from the Great Hall SWIR dataset before conducting any experiments, reducing the number of bands from 238 to 152. It was found that the number of layers chosen and width of each layer had a minimal impact on the results as long as the autoencoder layers were wide enough to have the representation power to reconstruct the spectra from the code layer and there was multiple hidden layers (e.g. at least two or three (Tadeusiewicz et al., 2014)). Having a greater width and number of layers in the network than

what was required added redundancy, but did not degrade the results in preliminary experiments (as long as there was no overfitting).

For all datasets, the regularization parameter was set to 10^{-4}, the activation function used was a sigmoid and 1000 epochs of L-BFGS (Lui and Nocedal, 1989) was used to optimise the networks. As this method is unsupervised, the method is trained on the same image it is evaluated on (similar to the evaluation of a dimensionality reduction approach). Hence, for each experiment, the number of data points used to train the autoencoders from Section 4.3.2 and pre-train the RSA-SAE autoencoder from Section 4.3.2 was the number of pixels in the image Section 3.1. For some experiments, a trained encoder was applied to new, unseen images (e.g. experiments using a timelapse dataset). The dimensionality of the new representation was the number of hidden units in the deepest code layer (e.g. five dimensions for the Gualtar Steps dataset).

To fine-tune a network using the RSA-SAE method, weighted sampling of the squared image intensity was used to reduce the training time. From each image, 5000 pixels were sampled prior to the data augmentation step. Ten candidate atmospheres were used to augment the data, creating a training dataset of 50,000 samples. Weighted sampling was necessary such that the majority of pixels used to train the RSA-SAE network were illuminated by sunlight. The geometric parameters for each pixel were individually sampled.

4.3.2 Evaluation of Hyperspectral Stacked Autoencoders

The autoencoders proposed in Section 4.1 (CSA-SAE, SA-SAE and SID-SAE) were compared with other unsupervised dimensionality reduction/feature extraction methods based on their ability to discriminate spectra from different classes and similarly represent spectra from the same class with a reduced number of dimensions. These included PCA, FA, the sum of squared errors-stacked autoencoder (SSE-SAE) using the squared error reconstruction cost function (2.14), and the raw data without any dimensionality reduction. The degree of variability in the scene differed between

datasets, and the algorithms were evaluated on how robust they were to this variability. A number of datasets, described in Section 3.1, were used to evaluate the dimensionality reduction techniques. All of the chosen datasets were used in normalised reflectance form, and the simulated and Great Hall VNIR datasets were also evaluated in DN form. For all experiments, the number of dimensions was reduced to five.

The first set of results compared each autoencoders ability to represent different classes with fewer dimensions. It is desirable that in the new feature space spectra from different classes are separated and spectra from the same class are closer together. This was measured with the Fisher's discriminant ratio (Section 3.2.1). If classes are well represented in the low dimensional space, then it is expected that the data will cluster into semantically meaningful groups (rather than groups that have a similar incident illumination). Hence, as an additional method of evaluation, the low dimensional data was clustered using k-means (Macqueen, 1967), where the number of clusters was set to the number of classes in the dataset. The clustering performance was measured using the ARI (Section 3.2.2).

The Fisher's ratio results showed that for most datasets, the novel hyperspectral autoencoders represented the different classes with fewer dimensions with more between-class discriminability and within-class uniformity than the other approaches (Figure 4.3). Clustering performance was also higher when using the hyperspectral autoencoders to represent the data (Figure 4.4). The hyperspectral autoencoder based on the cosine of the spectral angle (CSA-SAE) had the best overall performance. The results of Figure 4.3 indicated that this method performed well in comparison to the other methods on the simulated reflectance and DN datasets, Great Hall VNIR reflectance and DN datasets and Pavia Uni dataset. The clustering results of Figure 4.4 supported the Fisher's ratio results in terms of relative performance, especially in comparison to the no-dimensionality reduction, PCA and FA results. The CSA-SAE had ARI scores of above 0.7 for the simulated reflectance dataset, Great Hall VNIR reflectance and DN datasets. In terms of distance, the SID-SAE approach performed well on the simulated reflectance and DN datasets and the Pavia

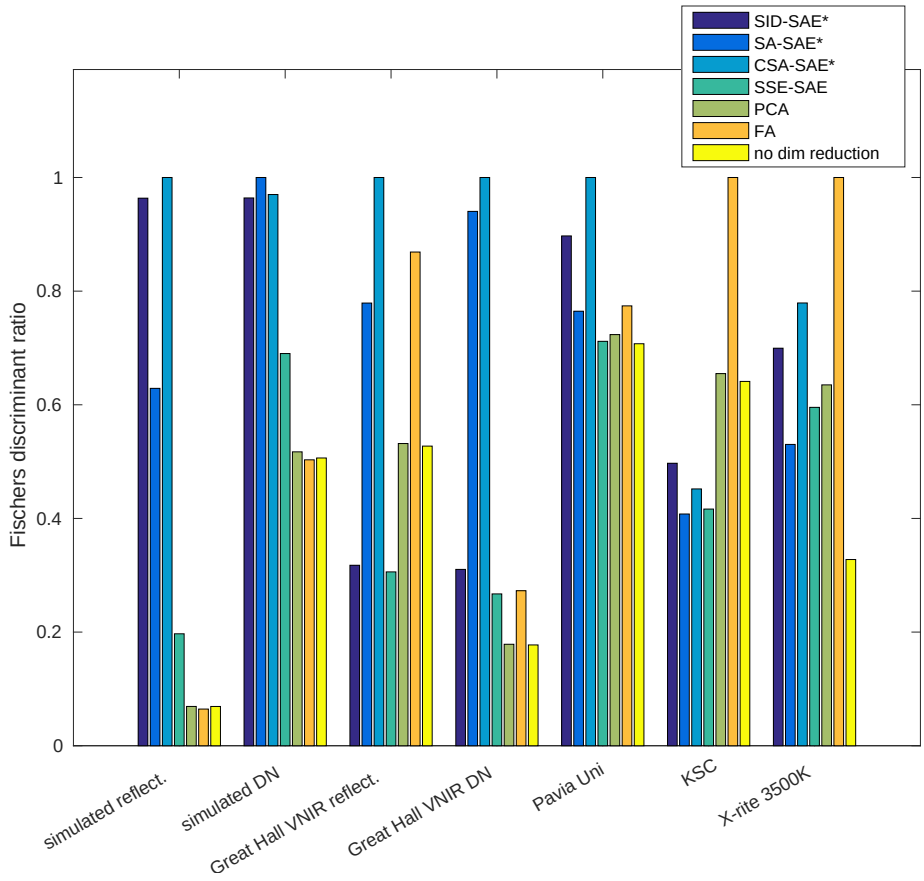


Figure 4.3 – Comparison of representation power of different dimensionality reduction/feature extraction methods, for a range of different datasets (horizontal axis). The methods were evaluated on how well different classes were represented in the low dimensional space using the Fischer's discriminant ratio. The higher the score, the better the representation because spectra from different classes were more separated relative to spectra from the same class. Scores have been standardised for comparison across datasets by dividing each score by the maximum achieved for that dataset. The * in the legend indicates the methods developed and proposed in this thesis.

Uni dataset, but only performed marginally better than the standard autoencoder, the SSE-SAE, on the other datasets. All methods had similar class separation on the Pavia Uni and simulated DN datasets, and the clustering performance of all methods for these datasets was low (ARI below 0.4). FA had the best separation for the KSC and X-rite 3500K datasets.

The next set of results evaluated how invariant the low-dimensional representations were to changes in brightness. Using the data that was simulated with different

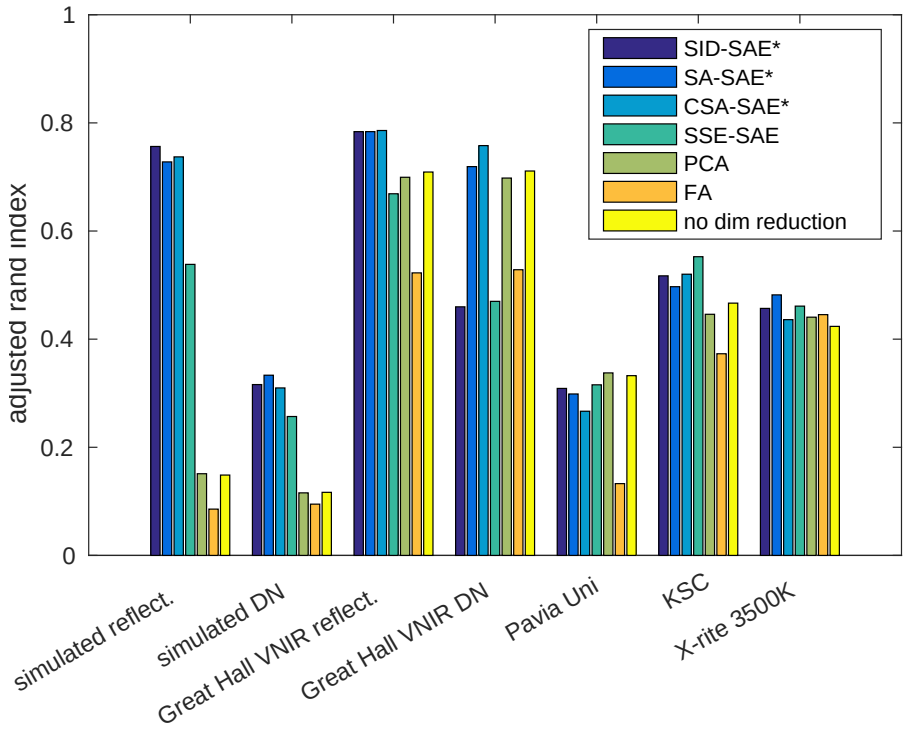


Figure 4.4 – Comparison of clustering results after using different dimensionality reduction/feature extraction methods, for a range of different datasets (horizontal axis). The methods were evaluated using the adjusted rand index of the clustered low dimensional data. The higher the score, the better the clustering performance is. The * in the legend indicates the methods developed and proposed in this thesis.

brightness, dimensionality reduction mappings were learnt by training on data illuminated by a source with one intensity (scaling factor equal to 1), and then applied to the said dataset as well as the data illuminated by a source with a different intensity (scaling factor equal to 0.3). The Fisher's discriminant ratio was then found between spectra of the same class from both datasets to determine how similar the classes of spectra were under different brightnesses (taking the mean over the classes). In a real-world outdoor dataset, differences in brightness can arise due to the variations in geometry with respect to the sun, and also changes in the brightness of the sun (for example, in a dataset of the same scene captured at different times).

The results show that the autoencoder approaches, including the baseline SSE-SAE autoencoder, all had superior brightness invariance to PCA, FA and, at least for the

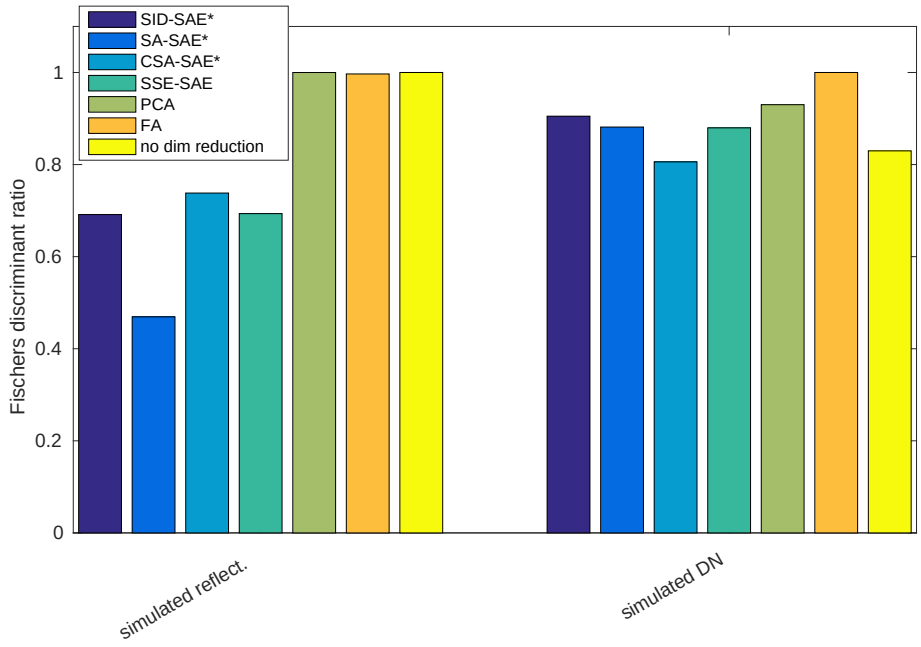
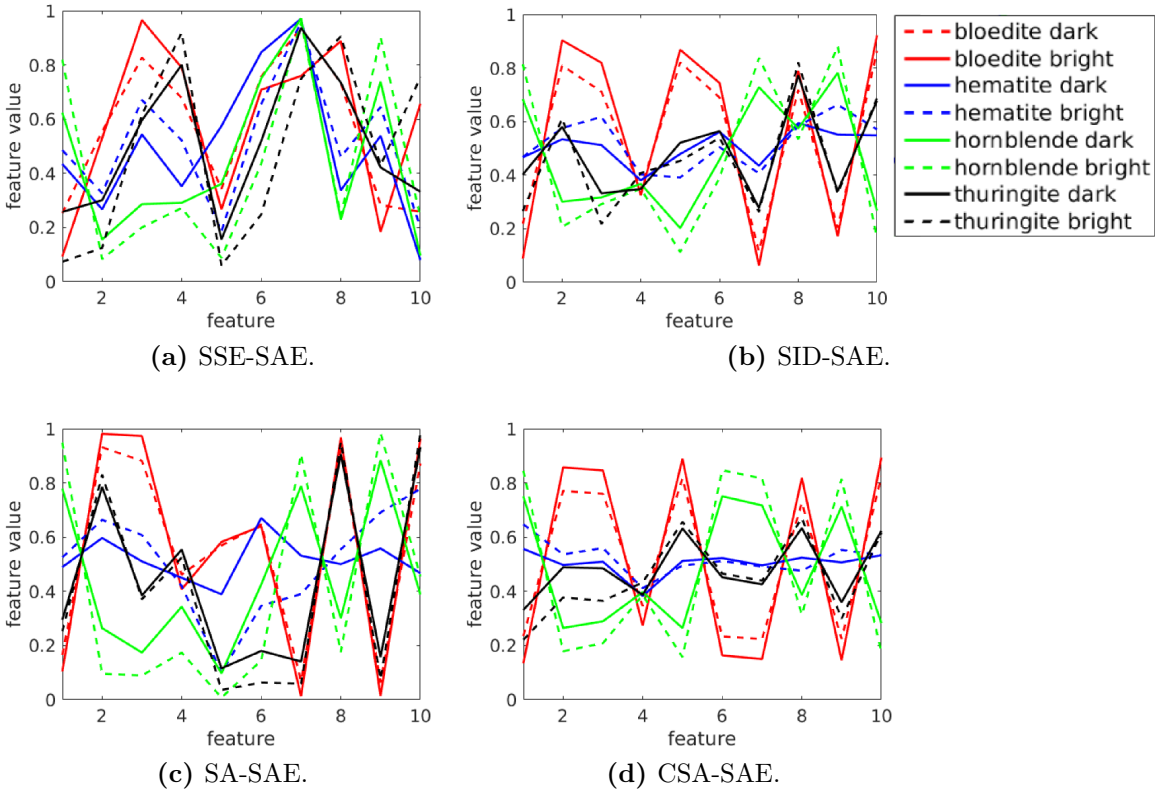


Figure 4.5 – Comparison of brightness invariance of different dimensionality reduction/feature extraction methods, for a simulated reflectance and DN dataset. The methods were evaluated on how well classes with different brightnesses were represented in the low dimensional space. The lower the score, the better the representation because spectra from the same class but with different brightness were more similar relative to spectra from the same class and brightness. Scores have been standardised for comparison across datasets by dividing each score by the maximum achieved for that dataset. The * in the legend indicates the methods developed and proposed in this thesis.

reflectance dataset, the original high dimensional spectra (Figure 4.5). The SA-SAE performed better than the other approaches on the simulated reflectance dataset whilst the CSA-SAE performed marginally better than the other approaches on the simulated DN dataset. The encodings of the material spectra under different brightnesses found using the autoencoder approaches appeared quite similar (Figure 4.6).

Finally, the autoencoders invariance to the wavelength-intensity distribution of the illuminant was evaluated. The first dataset, X-rite, comprised of the same scene imaged under indoor lighting with different temperatures (and thus, different illuminant intensity-wavelength distributions). The Fisher's discriminant ratio is found between spectra of the same class under these different illuminants. The second dataset, Gualtar steps, comprised of a scene half in shadow. The Fisher's discriminant ratio



(a) SSE-SAE. **(b)** SID-SAE.

(c) SA-SAE. **(d)** CSA-SAE.

Figure 4.6 – Comparison of low dimensional autoencoder representation of different classes under different brightnesses, for the simulated reflectance dataset. Results show only four out of the ten classes.

between corresponding classes in shadow and sunlight was found (taking the mean over all classes).

There was no stand-out autoencoder that was the most invariant to the illuminant (Figure 4.7). Regarding invariance to the temperature of the indoor light, the SA-SAE and SSE-SAE were the top performing dimensionality reduction approaches, with the spectra with its original dimensionality representation being the most similar under the different lighting conditions. The other methods all performed quite poorly. For the dataset with the shadow, the CSA-SAE represented the classes in and out of the shadow most similarly.

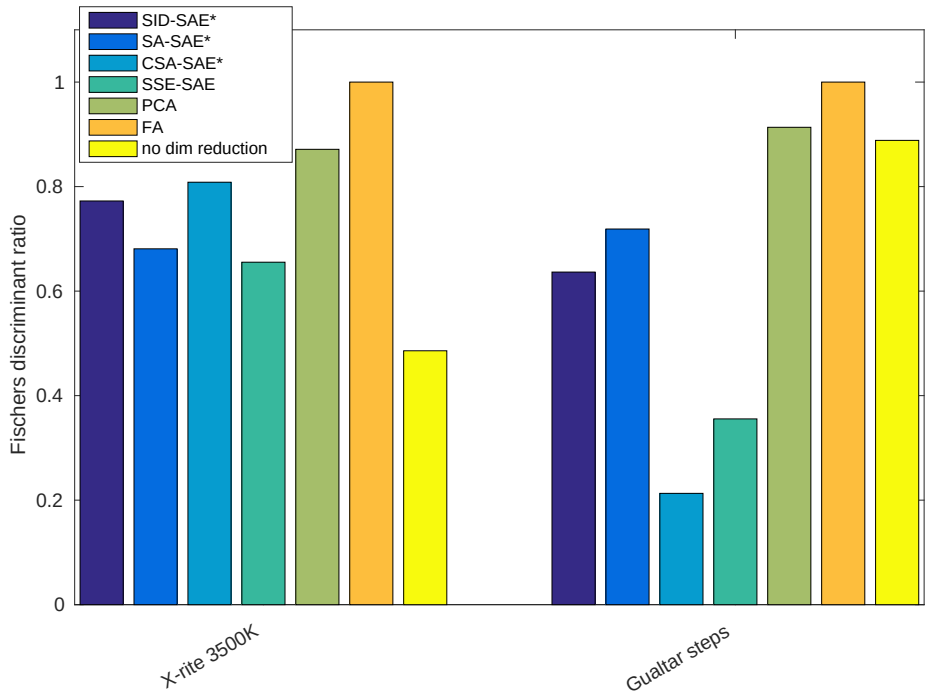
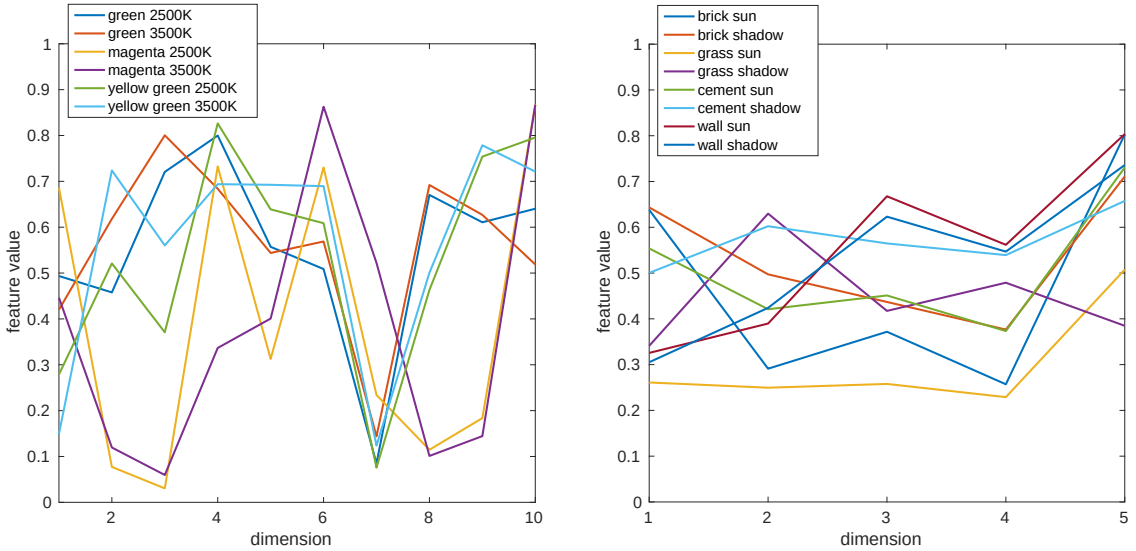


Figure 4.7 – Comparison of illuminant invariance of different dimensionality reduction/feature extraction methods. The X-rite dataset consisted of an image captured indoors under different illuminant temperatures and the Gualtar steps dataset had a large shadow across it, which also represented a change in the illuminant. The methods were evaluated on how well classes with different illuminants were represented in the low dimensional space. The lower the score, the better the representation because spectra from the same class but with different illuminants are more similar. Scores have been standardised for comparison across datasets by dividing each score by the maximum achieved for that dataset. The * in the legend indicates the methods developed and proposed in this thesis.

4.3.3 Evaluation of RSA-SAE

The proposed RSA-SAE was evaluated based on its ability to not only represent spectra with fewer dimensions but, additionally, how illumination invariant its representation of the spectra was. For each of the experiments, features found using the RSA-SAE method were compared against those found with unsupervised dimensionality reduction techniques and state-of-the-art illumination invariant feature extraction techniques. Techniques were picked that were suitable for comparison because they did not require labels, additional sensors or *a priori* knowledge. The dimensionality reduction methods in the comparison included PCA, FA and a standard Autoencoder

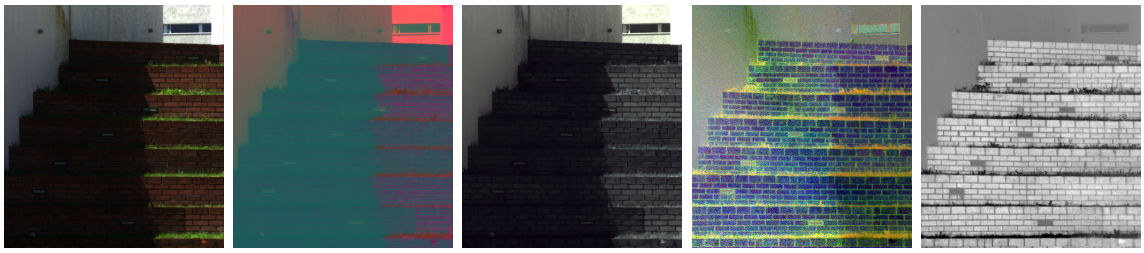


(a) Comparison of X-rite spectra under different illuminants represented using the SA-SAE approach.

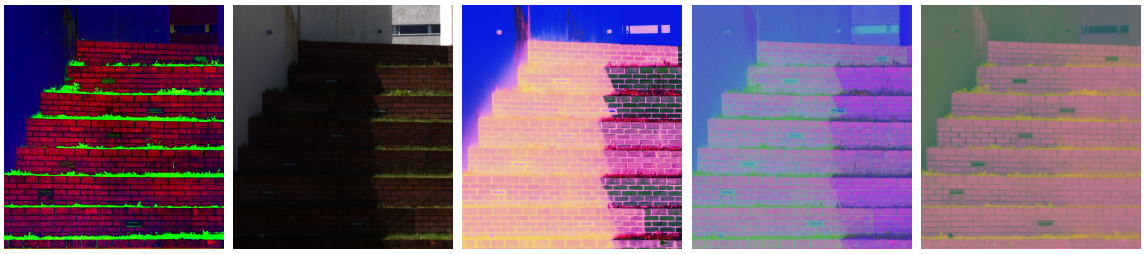
(b) Comparison of Gualtar spectra under different illuminants represented using the CSA-SAE approach.

Figure 4.8 – Comparison of low dimensional autoencoder representation of different classes under different illuminants.

(SSE-SAE) based on the squared error cost function. State-of-the-art illumination invariant feature extraction techniques were the photodector model-based methods including Marchant and Onyango (2002), the popular method for finding 1D shadow invariant images from RGB images using projections in the log-chromaticty space (Finlayson et al., 2004) and the extension of that method to hyperspectral images (Drew and Salekdeh, 2011). Finlayson et al. (2004) was applied to a three-channel image extracted from the hyperspectral image. Other state-of-the-art approaches included were the unsupervised statistical learning method for separating the illuminant from VIS hyperspectral reflection images using low-rank matrix factorization (Zheng et al., 2015) and the CSA-SAE proposed in section Section 4.1.1, developed for learning brightness invariant spectral features that the RSA-SAE extends. A number of datasets, described in Section 3.1, were used to evaluate the proposed autoencoder to demonstrate its generality. Because Zheng et al. (2015) required some data that was only available in the visible spectrum, this method was only evaluated with the Gualtar Steps dataset.



(a) Colour image. **(b)** PCA. **(c)** FA. **(d)** Marchant and Onyango (2002). **(e)** Finlayson et al. (2004).



(f) Drew and Salekdeh (2011). **(g)** Zheng et al. (2015). **(h)** SSE-SAE. **(i)** CSA-SAE* **(j)** RSA-SAE*

Figure 4.9 – Qualitative results from the Gualtar steps. Colour composites of features learnt using various approaches. In the case of dimensionality reduction methods, each dimension is considered a feature. The * indicates the methods developed and proposed in this thesis.

Illumination Invariance of RSA-SAE Representation

The first experiment compared the illumination invariance and discriminability of the image representations/encodings found using each of the methods by analysing the similarity of different materials in the encoding in shadow and when illuminated by sunlight. The Gualtar steps and Great Hall SWIR datasets in DN form were used for evaluation because each image comprises a large shadow. Once all methods had been applied to the images (e.g. dimensions reduced or features extracted), 10 samples were taken from different semantic regions under shadow and under sunlight in the images. To evaluate how good each method was at representing hyperspectral data, the Fisher's discriminant ratio (Section 3.2.1) between the encodings of each of these collections of samples was found, with lower values indicating more similarity and higher values indicating more separability. The mean Fisher's discriminant ratio between corresponding classes in shadow and sunlight was reported as the sun-shadow

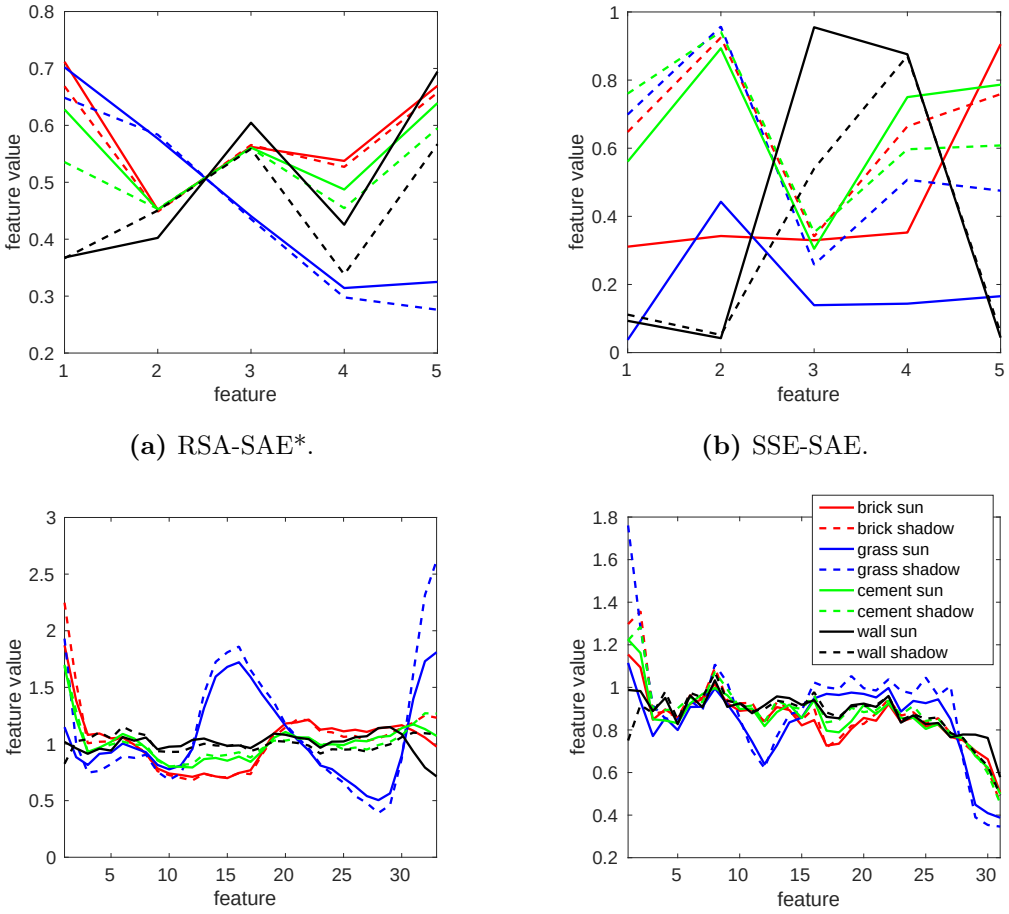
distance (where shadow and sunlit data are considered as different classes). The mean Fisher's discriminant ratio between different classes was reported as the class separability, for the case when classes consisted of only sunlit data and also the case when classes consisted of sunlit and shadowed data. The sun-shadow distance directly measures illumination invariance and the sun-only class separability directly measures class discriminability. The sun and shadow class separability measures both illumination invariance and class discriminability. Qualitatively, for each method, colour composites were created from the three most illumination invariant features from the Gualtar steps dataset. If the features were invariant to illumination then the resulting composite should not show any lighting effects (e.g. shadows). For all methods which reduced the dimensionality of the images, the dimensionality of the Gualtar steps image was reduced from 33 to five, and the dimensionality of the Great Hall SWIR image was reduced from 152 to 30.

Table 4.3 – Quantitative results from the Gualtar steps and Great Hall. The sun-shadow distance is the mean Fisher's discriminant score between classes in shadow and sunlight and the class separability is the mean Fisher's discriminant score between different classes, for the sun-only case and the case with sun and shadow data. Ideally, it is desirable for the sun-shadow distance to be low and the class separability to be high. Note, that whilst the sun-shadow distance is a direct measure of the similarity between pixels in sun and shadow, the class separability (sun+shadow) is also affected by the similarity of sunlit and shadowed pixels, as it contributes to the intra-class scatter. Thus, the class separability (sun+shadow) is the most complete measure of the quality of the feature space. The class separability (sun-only) measures the quality of the feature space when illumination variability due to shadow is not a consideration. The best results from each category are highlighted in bold. The * indicates the methods developed and proposed in this thesis.

Method	**Gualtar steps (VIS)**			**Great Hall (SWIR)**		
	sun-shadow distance	class separability (sun-only)	class separability (sun+shadow)	sun-shadow distance	class separability (sun-only)	class separability (sun+shadow)
reflectance	46.90	79.38	1.74	42.09	73.39	12.47
PCA	48.23	81.15	1.74	46.18	77.90	12.59
FA	52.79	88.80	1.71	49.02	87.21	13.48
Marchant and Onyango (2002)	3.60	5.62	1.32	**0.28**	5.25	0.36
Finlayson et al. (2004)	**1.22**	26.29	**16.94**	13.48	80.38	4.20
Drew and Salekdeh (2011)	12.43	16.29	4.73	0.75	11.74	1.77
Drew and Salekdeh (2011)+PCA	19.71	17.71	5.19	1.53	19.75	3.60
Zheng et al. (2015)	46.74	82.10	1.81	NA	NA	NA
SSE-SAE	18.77	**95.54**	5.79	33.53	37.64	3.41
CSA-SAE*	11.24	21.94	3.25	38.75	156.15	8.61
RSA-SAE*	3.69	52.84	11.88	3.57	**181.46**	**30.91**

The quantitative results (Table 4.3) and qualitative results (Figure 4.9) showed that the basic dimensionality reduction methods PCA, FA and SSE-SAE were good at sep-

arating different classes in the low dimensional space when there were no shadows, but had poor class separability when shadows were present in comparison to other methods. This was evident in the SSE-SAE feature vectors (Figure 4.10) where the feature vectors for all of the classes in sunlight were distinct, but the feature vectors for the brick, grass and cement classes in shadow were very similar. The illumination invariant feature extraction techniques (Drew and Salekdeh, 2011; Finlayson et al., 2004; Marchant and Onyango, 2002) had a low sun-shadow distance (the qualitative results in Figure 4.9 and Figure 4.10 showed that there was almost no difference between the sun and shadow regions for a given class). However, the sun-only class separability was far lower than the basic dimensionality reduction methods, and hence they also had a low class separability when shadows were included (apart from Finlayson et al. (2004)). Even when PCA was used to extract features from the illumination invariant data (Drew and Salekdeh, 2011) that maximised variance, the class separability was still relatively low. The CSA-SAE showed comparable performance with Marchant and Onyango (2002) and Drew and Salekdeh (2011) for the Gualtar steps dataset, and out-performed them on the Great Hall dataset. However, Figure 4.9 reveals that the regions in shadow were represented differently to the regions in sunlight, indicating that the CSA-SAE feature space was not invariant to shadows. Although the regions under shadow and sunlight had a similar colour in the colour composite, suggesting that there were no features dedicated to capturing shadows, there were changes in intensities across the shadow boundary such that all features were effected by the shadow. The source separation method (Zheng et al. (2015)) performed comparably with the reflectance representation, both quantitatively and qualitatively. The RSA-SAE performed well, having both a low sun-shadow distance and a high class separability for the sun-only and sun+shadow cases (the best sun+shadow class separability for the Great Hall and the second best for the Gualtar steps). In Figure 4.9, the shadow was very hard to see suggesting that the low-dimensional representation was invariant to the illumination conditions. In Figure 4.10 the feature vectors for corresponding classes in sun and shadow were quite similar and there was also a distinct difference between the feature vectors of different classes.



(a) RSA-SAE*.

(b) SSE-SAE.

(c) Drew and Salekdeh (2011).

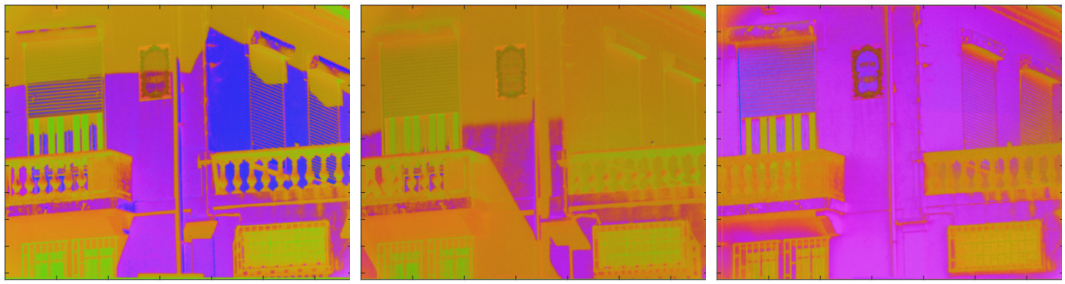
(d) Marchant and Onyango (2002).

Figure 4.10 – Results from the Gualtar steps. Plot of the mean feature vector for each class in sun and shadow for some selected approaches. The SSE-SAE approach was designed for feature extraction, whilst the Drew and Salekdeh (2011) and Marchant and Onyango (2002) approaches were designed for illumination invariance. The proposed RSA-SAE approach was designed for both feature extraction and illumination invariance. The * indicates the methods developed and proposed in this thesis.

A different experiment was also used to compare the illumination invariance of the representations found using the different techniques. The similarity of the encodings of a sequence of images captured at different times of the day was analysed. The Gualtar timelapse dataset in DN form was chosen for the experiment because the position of the sun moved throughout the day that the data was captured on, causing moving shadows and changes in the brightness of surface materials across images. An illumination invariant representation of this set of images should appear similar. The different techniques were used to learn a mapping from the 11:44 image of the Gualtar timelapse dataset (the first of the time sequence). These mappings were then applied to the images captured at all other times of day to obtain the low-dimensional encoded images. To evaluate each method, the PSNR (3.2.3) was found between the encoded 11:44 image and all other encoded images to measure how much the image changed across the day when under the new representation. High PSNR values indicated greater similarity between the two images and thus better performance. For each feature, the mean and standard deviation of the PSNR scores across the day was calculated. For all methods the dimensionality was reduced from 33 to five.

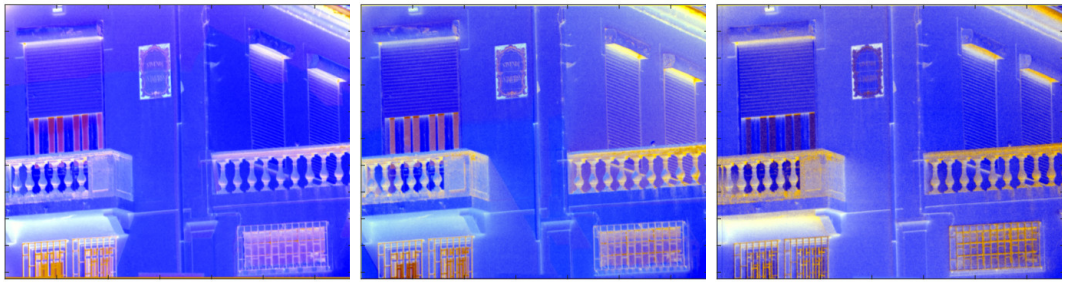
Changes in the colour composites when the shadow in the Gualtar timelapse images moved gave a qualitative indication of the illumination invariance of the features (Figure 4.11). The SSE-SAE encodings had a significant colour change in the composite across the shadow boundary, and because the shadow moved throughout the day the encoded image changed as well. The CSA-SAE changed in intensity rather than colour across the shadow boundary, suggesting that the points in and out of shadow had a more similar representation in the new feature space, but were not totally invariant. The RSA-SAE performed very well, with no colour or intensity change across the shadow boundary. Across the day, the encoded hyperspectral image remained relatively constant despite the significant changes in illumination.

(a) Original.



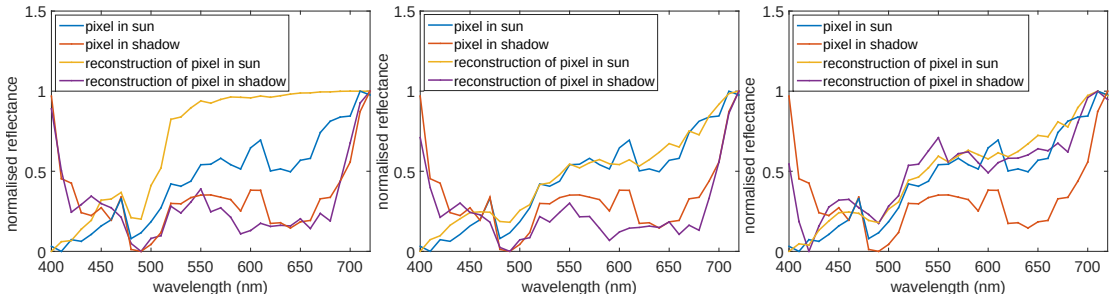
(b) SSE-SAE.

(c) CSA-SAE*.



(d) RSA-SAE*.

Figure 4.11 – Gualtar timelapse qualitative results. Colour composites of three selected timesteps using features/dimensions from each autoencoder method, linearly scaled to increase contrast. The x and o indicate the location of the shadowed and sunlit samples used in Fig. 4.12. The * indicates the methods developed and proposed in this thesis.



(a) SSE-SAE. **(b)** CSA-SAE*. **(c)** RSA-SAE*.

Figure 4.12 – Gualtar timelapse quantitative results. Comparison of the spectrum and reconstruction of a pixels spectrum sampled in shadow and sunlight using each of the three SAEs. Curves were normalised for visualisation by subtracting the minimum and dividing by the maximum. The * indicates the methods developed and proposed in this thesis.

Table 4.4 – Gualtar timelapse quantitative results. Mean±std of PSNR (dB) between the 11:44 image and all other timesteps for each feature/dimension, averaged over all times. The higher the PSNR, the more similar the images were, which was desirable. The top 10 features/dimensions are highlighted in bold, as well as the method with the best overall result. The * indicates the methods developed and proposed in this thesis.

	Feature					
Method	1	2	3	4	5	All
PCA	-10.6±2.5	9.8±5.8	18.1±4.6	21.1±3.2	21.3±3.8	11.9±12.8
Drew and Salekdeh (2011)+PCA	1.1±1.3	10.1±1.0	8.8±2.4	13.5±0.7	18.3±0.3	10.4±5.9
Marchant and Onyango (2002)+PCA	8.4±1.6	15.8±1.4	22.8±3.4	14.4±1.1	19.1±1.4	16.1±5.2
FA	-20.6±2.5	-19.8±2.5	-18.8±2.5	-19.3±2.5	-17.4±2.5	-19.2±2.6
SSE-SAE	14.0±1.8	13.7±2.1	6.8±1.5	11.6±1.9	12.4±3.0	11.7±3.3
CSA-SAE*	24.7±2.6	**26.7 ± 2.1**	25.1±2.7	23.5±2.7	22.3±3.7	24.5±3.0
RSA-SAE*	22.9±1.7	**34.9 ± 1.6**	**28.9 ± 1.4**	**25.5 ± 1.7**	**26.3 ± 3.3**	**27.7 ± 4.6**

The findings of the quantitative results from the timelapse (Table 4.4) were similar to the qualitative results (Figure 4.11). They show the similarity of each feature for all nine images in the timelapse and showed results from the other approaches tested. It can be seen that four of the top five illumination invariant features belonged to

the RSA-SAE. The first of the RSA-SAE features was not as invariant as the other features, however, this was far superior to the other approaches where the illumination effected almost all of the features. With that said, the images made with the features found using the CSA-SAE approach exhibited much less change due to illumination than those found using the SSE-SAE, PCA, FA, Marchant and Onyango (2002) and Drew and Salekdeh (2011) with PCA approaches.

Fig. 4.12 shows the spectral curves of pixels sampled in sunlight and shadow from the 11:44 Gualtar timelapse image, and the spectra that was reconstructed by the decoder of each SAE. The spectral curves had been normalised for visualisation so that it was the shape of each curve being displayed rather than the magnitude (the shadowed pixel would usually have a much lower reflectance than the sunlit pixel). The SSE-SAE reconstructed the shape of the shadowed pixel accurately and the shape of the sunlit pixel poorly. The CSA-SAE accurately reconstructed the shape of both the shadowed and sunlit pixels. The RSA-SAE reconstructed the sunlit spectra from both the sunlit and shadowed spectra.

Classification Application of RSA-SAE Features

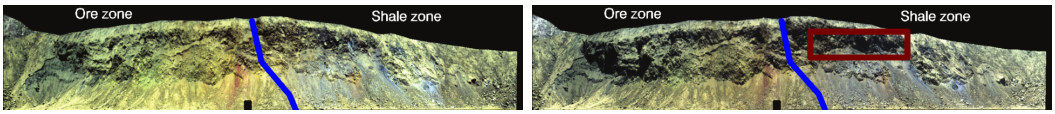
The illumination invariance and discriminability of the RSA-SAE features was evaluated by applying them to a pixel-wise supervised classification application. The purpose of this experiment was to demonstrate the applicability of the RSA-SAE features to high-level algorithms. The mining timelapse dataset (Section 3.1) in DN form was used, whereby mappings were learnt on the 11:30 image only and then these mappings were used as spectral features for KNN and SVM supervised classifiers. Whilst the features were learnt on pixels from the whole image (the methods are unsupervised and hence require no labels), the classifiers were trained on 458 labelled pixels from an image of a small set of rock samples from the scene. No labelled training data were available for the rocks with the same shadows or illumination conditions as the ones in the actual scene. When the learnt classifier was applied to both images, each method was evaluated by calculating the percentage of pixels that changed classification label (Section 3.2.4) between the 11:30 and 13:30 images (smaller numbers indicate greater

robustness to illumination). This measure indicated how robust the classifiers are to illumination. The accuracy of the classification maps was also roughly assessed as a geologist had indicated where the approximate geological boundary between the shale and mineral was (Figure 4.13a). However, there was no semantic knowledge of how the two types of shale were distributed. For all methods, 30 features were learnt from the 220-dimensional image. PCA was used with feature extraction methods that did not reduce the dimensionality of the data.

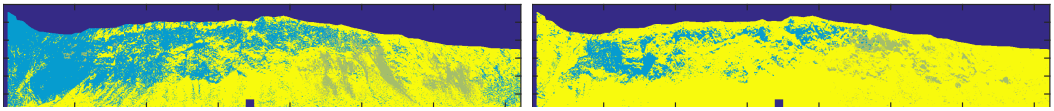
Table 4.5 – Mineface quantitative results. Percentage of pixels that changed classification label from 11:30 to 13:30. The top result from each classifier is highlighted in bold. The * indicates the methods developed and proposed in this thesis.

Features	**KNN**	**SVM**
Reflectance	35.2	33.7
PCA	34.9	34.2
Drew and Salekdeh (2011)+PCA	**14.2**	20.3
Marchant and Onyango (2002)+PCA	18.0	**11.7**
FA	58.2	51.5
SSE-SAE	27.4	26.4
CSA-SAE*	37.7	28.2
RSA-SAE*	26.9	11.8

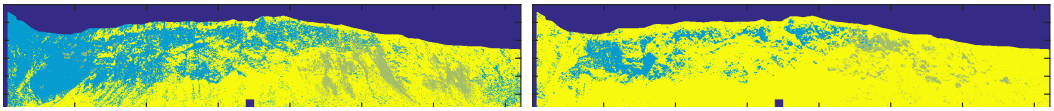
The RSA-SAE features produced some of the most consistent classification results across the timelapse for both classifiers (Table 4.5). The SSE-SAE features produced more consistent results than the CSA-SAE features, but from the visualisation of the KNN results in Fig. 4.13, it can be seen that their labelling of the scene does not align with the ground truth geological regions and it is possible to see where the shadow has resulted in misclassification of the pixels. Fewer pixels changed predicted label across the day when the classifier was trained using the features of Marchant and Onyango (2002) and Drew and Salekdeh (2011), although, the resulting prediction maps show that the classifier struggled to differentiate between the two types of shale



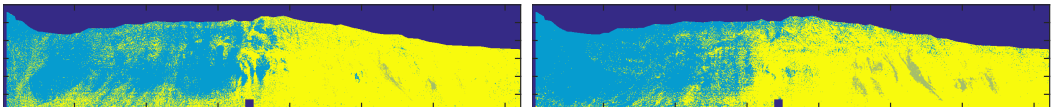
(a) Colour composite.



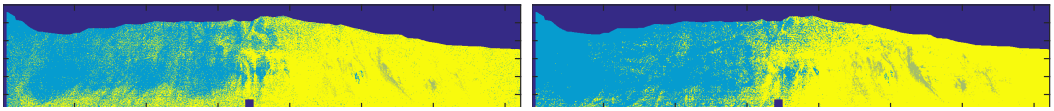
(b) Reflectance.



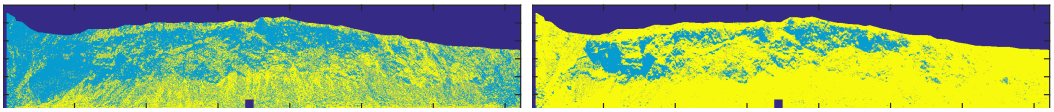
(c) PCA.



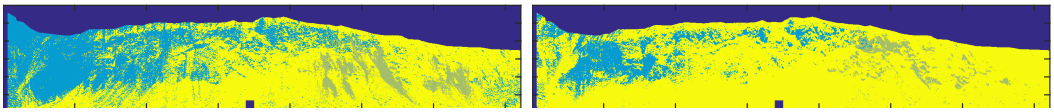
(d) Drew and Salekdeh (2011)+PCA.



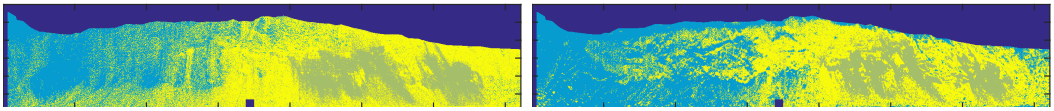
(e) Marchant and Onyango (2002)+PCA.



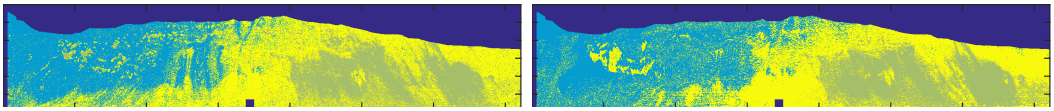
(f) FA.



(g) SSE-SAE.



(h) CSA-SAE*.



(i) RSA-SAE*.

Figure 4.13 – Mineface qualitative results. Classification results over the day using features from the different methods. The first row shows a colour composite of each hyperspectral image with ground truth geological regions (Schneider et al., 2011). Labels assigned by the classifier were Martite (light blue), Shale (yellow) and Magniferous Shale (olive green). The red box indicates one of the regions in shadow. The * indicates the methods developed and proposed in this thesis.

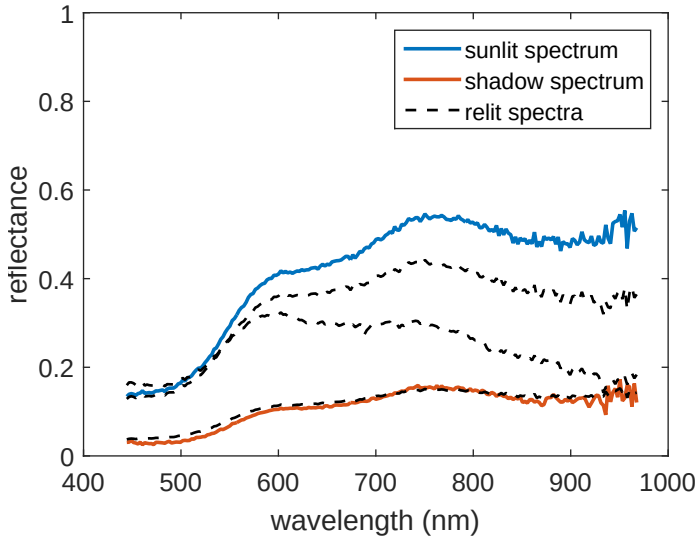


Figure 4.14 – Mineface results. Comparison of the original sunlit spectrum for Martite, original corresponding spectrum in the shadow and several spectra relit to be in shadow (dashed curves) using the original sunlit spectrum. One of the relit spectra closely matched the real spectrum of Martite in shadow, so by training the network to reconstruct the sunlit spectrum from the relit spectra, the proposed approach learnt to encode the real shadowed spectra in the same way as the sunlit spectra.

as almost no pixels had been labelled as magniferous shale. Once again, these methods are achieving illumination invariance at the expense of representation power. The SA-SAE and RSA-SAE features produced reasonably accurate classification maps, with the RSA-SAE result being more consistent across the two times. Both PCA and FA features produced results that were inconsistent across the two times of day and also lacked alignment with the ground truth mapping. Figure 4.14 shows how closely the relit spectra matched actual shadowed spectra, which is critical for the success of the RSA-SAE.

4.4 Discussion

The following section provides an analysis of the illumination invariant autoencoder results of Section 4.3.

4.4.1 Evaluation of Hyperspectral Stacked Autoencoders

The autoencoder approaches were expected to have better overall performance than PCA as they were able to learn non-linear mappings to the low dimensional spaces. This property allowed for more complex transformations to occur. The results (Figures 4.3 and 4.4) supported this for some of the datasets. The proposed hyperspectral autoencoders had a better overall performance than the SSE-SAE, and further, the autoencoders based on the spectral angle performed better than the SID-SAE. By using the spectral angle as the reconstruction cost function, the autoencoders learnt complex mappings that captured features describing the shape of the spectra. This was in comparison to the autoencoders that used the squared error measurement of reconstruction error. These autoencoders learnt features which described the intensity of the spectra. The intensity is highly dependent on the illumination conditions and hence was not the best characteristic on which features should be based on. The autoencoders using the spectral information divergence performed slightly better than the SSE-SAE approaches because the distance function could account for small variations in the probability distribution of each spectra and was also unaffected by wavelength-constant multiplications. Hence, the learnt mapping was invariant to intensity changes. FA performed best on the X-rite 3500K dataset. However, this was likely to be because the X-rite 3500K dataset was the only dataset that had almost no intraclass variability. Thus, features were learnt which maximised the variability between classes and there was very small separation created within each class.

In some cases, the hyperspectral autoencoders did not outperform the other approaches. There was almost no performance improvement in Figure 4.3 from using the angle based methods for the KSC and X-rite 2500K datasets. However, the clustering results (Figure 4.4) were relatively low for these datasets for all methods, suggesting that they were particularly difficult to represent with unsupervised approaches. Clustering results were also low for the simulated DN and Pavia Uni datasets. The KSC, Pavia Uni and X-rite 3500K datasets contain many classes (greater than eight), many of which are very similar. For example, the KSC dataset contains several different 'marsh' classes. With that said, the hyperspectral autoencoders still outperformed

the PCA, FA and no dimensionality reduction approaches for the KSC dataset. Regarding the simulated DN dataset, the angular methods were expected to achieve similar results on both the reflectance and DN datasets, because the shape of the spectra is unique for different classes regardless of normalisation. For the distance results (Figure 4.4) this was true, but for the clustering results, this was only the case for the Great Hall datasets and not the simulated datasets. The clustering results show that all methods performed badly on the simulated DN dataset, suggesting that if not normalised, the differences in the spectra can become very small, making it difficult for any unsupervised methods to separate the classes.

The simulated brightness experiment results (Figure 4.5) show that all of the autoencoders provided an advantage over PCA, FA and no dimensionality reduction when it came to brightness invariant representation of reflectance spectra. The encodings of spectra with different brightnesses appeared similar in the learnt feature space (Figure 4.6), indicating invariance to brightness. Mappings learnt on simulated data, illuminated with a specific brightness, were able to generalise to simulated data illuminated with a brightness that the mappings were not trained on. This was expected from the proposed hyperspectral autoencoders because their reconstruction cost functions are robust to brightness variations. Interestingly, the SSE-SAE also exhibited some brightness invariance despite the reconstruction cost function being dependent on the intensity of the spectra. It was expected that this occurred because there was no brightness variability within the training set. As a result the SSE-SAE learnt to ignore brightness variability because it did not require reconstruction. If, however, there was variability in the brightness of the training data then it is likely that the SSE-SAE would capture the very basic intensity multiplication in the training dataset with its non-linear function approximators and encode these variations in a number of features/dimensions. If the trained encoder was then applied to new data with a different intensity of the source brightness, then some of the features in the encoding would vary depending on the brightness. PCA and FA attempted to decouple the sources of variation, but their representation power was limited by linear functions. Thus, they provided no advantage over using no dimensionality reduction, and for the

simulated DN, they actually had worse brightness invariance than when there was no dimensionality reduction used.

The results for the illuminant experiments (Figure 4.7) differed between the indoor light temperature experiment (X-rite) and the outdoor shadow experiment (Gualtar steps), in terms of which methods performed the best. For the X-rite data, amongst the dimensionality reduction approaches, the SA-SAE appeared to represent the 24 different classes best under the five different illuminants or light bulb temperatures. However, the no dimensionality reduction approach achieved significantly better results than the dimensionality reduction approaches. In the Gualtar steps experiment, the CSA-SAE represented the classes most similarly under shadow and sunlight, but was one of the worst performers for the X-rite dataset. This suggests that none of the dimensionality reduction methods are properly representing the data under different illuminants. To investigate this, Figure 4.8 shows the top performing dimensionality reduction methods from both datasets representation of the spectra under the different illuminants. The SA-SAE's representation of the X-rite spectra was good. There were clear correlations between each corresponding colour across different light temperatures in the low dimensional representation. Dimensions eight and nine in particular seemed to show invariance to the illuminant temperature. However, the other dimensions exhibited a noticeable amount of variation due to the illuminant changing, making it difficult to distinguish between similar colours, such as green and yellow green. Likewise, with the CSA-SAE's representation of the Gualtar steps classes in sun and shadow, the majority of classes exhibited significant variation across all dimensions due to the change in lighting conditions. The only class that seemed to be represented similarly under shadow and sunlight was the wall. The inability of the methods to similarly encode materials under different illuminants was due to the properties of the distance measures from which they were derived. The spectral angle, the cosine of the spectral angle and the spectral information divergence are all brightness invariant measures; however, they are not invariant to the effects of shadows or changes in the illuminant temperature. In the case of the spectral angle, this is because it is invariant to a multiplication of the spectra that is constant across all

wavelengths, but shadows cause a wavelength dependent change in the spectra. For example, there may be a relative increase in reflectance towards the blue wavelengths (Murphy et al., 2012), caused by indirect illumination from skylight. There may also be a biased response to spectral noise where noisy parts of the reflectance spectrum may be greater or smaller than adjacent spectral regions (Murphy et al., 2014a). This change in spectral curve shape makes the spectral angle not invariant to the effects of shadows. For the autoencoders that use the spectral angle, this means that two materials from the same class but under different illuminants will be reconstructed very differently (because the angle between them is large), and hence will appear differently in the low-dimensional code layer. This explains why for the X-rite dataset results, the methods represented the same classes under different illuminants more poorly than the full dimensional, original representation because they were actually increasing the intra-class variation.

4.4.2 Evaluation of RSA-SAE

Illumination Invariance of RSA-SAE Representation

Along with the large shadow in the Gualtar steps image, some of the bricks orientation to the sun also changes with their geometry, which impacts their spectral response. The composites formed from the SSE-SAE, PCA and FA features (Figure 4.9) show that these methods are sensitive to the illumination conditions, not only to the shadow but also to the different amounts of light illuminating the bricks in the sun due to the geometry. This was to be expected as these dimensionality reduction/feature extraction methods capture all of the factors contributing to the data's variability when computing the low-dimensional image, including the illumination factors. This was reflected in the quantitative results (Table 4.3), where the large difference in class separability performance with and without the shadow suggests that these methods are trying to separate spectra within a given class when they are under different illumination. This sensitivity to illumination was less evident in the colour composite of the CSA-SAE features as the different levels of brightness illuminating the bricks

were less visible. The CSA-SAE method was expected to be more robust to the influence of the scene geometry with respect to the position of the sun on the incident illumination. This is because of the dependency of its reconstruction cost function on the shape of spectra rather than the magnitude, and the similar spectral shape of materials with different geometries. It can be seen, however, that it was not robust to the shadow which is still visible in the encoded image (the reasoning for this was discussed in Section 4.4.1).

The method of Finlayson et al. (2004) was effective at finding an image that was invariant to the illumination effects in the scene. However, the resulting 1D image lost much of its discriminability as the target of the method was not to preserve information, which explains its poor performance on the Great Hall dataset. The features found using Marchant and Onyango (2002) and Drew and Salekdeh (2011) displayed good illumination invariance. Despite this, Drew and Salekdeh (2011) had a relatively high sun-shadow distance. This is because it represented the wall class under the different illuminations poorly, evident from the different colours in the colour composite (Figure 4.9) and also in the feature vector for the wall in sun and shadow in Figure 4.10. The method for separating illumination from reflectance spectra using low-rank matrix factorisation (Zheng et al., 2015) was ineffective at removing the different effects of illumination from the image. This could be because some of the pixels in the scene are illuminated differently to others due to the occlusions (i.e. in shadow), whereas this method was designed for separating out a constant illuminant.

The proposed RSA-SAE method encoding of the image pixels displayed almost no evidence of the shadow in the scene or the different lighting conditions on the brick due to geometry, whilst maintaining good discrimination amongst the different classes. The results suggest that the method is operating effectively as a dimensionality reduction algorithm whilst also ensuring illumination invariance in the new low-dimensional feature space, as these two targets have both been incorporated into the algorithm. By utilising the autoencoder architecture, it is learning a highly discriminable hidden layer, and the incorporation of the physics-based model for illumination makes this hidden layer invariant to the illumination. The other approaches failed to meet

both objectives as effectively. Also, since the algorithm is representing the sunlit and shadowed materials similarly, the relighting process must be accurately emulating how the material spectra appear in shadow. It should be noted that the RSA-SAE had a higher class separability for the sunlit-only classes than the SA-SAE on both datasets, suggesting that the illumination invariant extension is not degrading the autoencoders dimensionality reduction/feature extraction performance.

The results from the Gualtar timelapse dataset (Figure 4.11) further demonstrate the robustness of the RSA-SAE to the illumination. The SSE-SAE network learns features dedicated to capturing shadows so that it can reconstruct them. This was evident by the different colours of the composite image across the shadow boundary. The composite image changed significantly over the day as the shadow moved. Given the shadowed regions are the same colour as the sunlit regions in the colour composite, the CSA-SAE features are less dedicated to capturing shadows, however, the shadows still affected the output of the features, evident by the change in intensity across the shadow boundary of the colour composite (similar to the results in Figure 4.9). As a result there was a small change in the colour composites over the day and the PSNR was relatively high (Table 4.4). The RSA-SAE features showed no obvious change in colour or intensity across the shadow boundary, suggesting that the feature representation is shadow invariant. There was, however, a distinct difference in the pixels below the balcony. This is suspected to be due to indirect illumination, which introduces effects to which the network is not trained to be robust. From Table 4.4, the first of the RSA-SAE features was not as invariant as the other features, however, this is far superior to the other approaches where the illumination effects almost all of the features.

The SSE-SAE (Figure 4.12a) and CSA-SAE (Figure 4.12b) reconstructed the shape of the shadowed pixel best, however, this was actually a disadvantage as it resulted in the shadow and sunlit spectra having different encodings because of the reconstructions having different shapes, despite belonging to the same class. The RSA-SAE (Figure 4.12c) was trained on corrupted relit data and hence reconstructed the sunlit spectra from the shadowed input, which is why the encoding of sunlit and shadowed

pixels were more similar. The CSA-SAE and RSA-SAE also reconstructed the sunlit spectra from the sunlit input more accurately than the SSE-SAE, because they used the spectral angle reconstruction cost function (4.2) which promotes reconstruction of the spectral shape rather than magnitude.

Classification Application of RSA-SAE Features

From the mining results of Figure 4.13 and Table 4.5, given a small training set was used, it was difficult for the classification model to capture the variability in the mineral classes due to illumination. Hence, the methods that used the spectral angle reconstruction cost function, which were more robust to changes in illumination due to the highly variable geometry of the surface and the position of the sun, produced a more accurate mineral classification. As the RSA-SAE is invariant to the changes in shadow, it also produced a more temporally consistent classification. However, illumination invariance did not guarantee good performance. The approaches of Marchant and Onyango (2002) and Drew and Salekdeh (2011) represented the spectra so that shadowed material appeared similar to sunlit material, evident by the similar maps at different times of the day. However, just as in the previous experiments (Figure 4.9 and Table 4.3), they have obtained illumination invariance at the expense of some representation power. The two types of shale were represented so similarly that the classifier could not distinguish between them. This is because these methods were not designed for feature extraction. Using PCA afterwards did not help as the specific spectral absorption features that differentiate the two shale classes were already lost once the illumination invariant representation was found. The RSA-SAE was able to represent the spectra in an illumination invariant space without compromising any of the features that make classes unique, making it the only approach with an accurate classification map that was also temporally consistent.

Figure 4.14 shows plots of the sunlit and shadowed spectra of Martite, as well as some of the relit spectra used by the RSA-SAE. It can be seen that one of the relit spectra matched the shadowed spectrum very closely, allowing the network to learn to be invariant to the real shadows in the image.

4.5 Summary

The complex ways that illumination interacts with the outdoor environment can result in unwanted variability in image data captured from these environments. Hence, illumination invariance is an important characteristic for platforms in computer vision, remote sensing and robotics that utilise hyperspectral data. It is important that illumination invariance does not come at the cost of the class discriminability of the data. Different classes should either remain or be more separable in an illumination invariant feature space. For platforms to use hyperspectral data efficiently, it is also important that the dimensionality of the data is reduced.

In this chapter, a method for finding low dimensional illumination invariant feature representations of hyperspectral data was proposed. Brightness invariant autoencoders developed specifically for hyperspectral data were presented, followed by an extension which made the autoencoder invariant to shadows. This means that in the new low dimensional image, pixels from the same material that are in shadow and sunlight, or have different brightnesses due to geometry or change in the suns position, will have a very similar representation. Experiments showed that the class discriminability was also preserved in the new representation. This representation is useful for higher-level algorithms that use hyperspectral data, such as classification, clustering or segmentation.

The key advantages of the proposed approach are that it does not require additional sensor modalities such as LiDAR, labelled training data or *a priori* knowledge of the atmospheric conditions or scene geometry. It also does not make the assumption of Planckian illumination, which is limiting for hyperspectral data because the incident illumination is greatly effected by the atmosphere. The proposed method is able to outperform similar methods as it is designed to simultaneously achieve both feature extraction and illumination invariance.

The dimensionality reduction/feature extraction techniques proposed in this chapter are illumination invariant. One of the applications of these representations is supervised spectral classifiers. However, this requires the training of a separate classifier

with labelled data. In Chapter 5, neural networks for end-to-end classification are proposed under the constraint of limited available training data.

Chapter 5

Supervised Classification of Hyperspectral Data with Limited Training Samples

This chapter proposes a method for training a supervised hyperspectral CNN classifier when there are only a limited number of labelled training samples available. Pixel-wise classification is one of the most useful applications for hyperspectral imaging systems, as this can be used to autonomously map the materials within a scene (Camps-Valls et al., 2015). Because they can learn features for classification rather than relying on hand-crafted ones, CNNs have been shown to be a powerful tool for supervised image classification (He et al., 2016; Krizhevsky and Hinton, 2012). CNNs also provide a good platform for transfer learning (Ahmed et al., 2008; Oquab et al., 2014), whereby data from other domains can be leveraged for tasks where training data is limited. Recently, CNNs have been adapted for material classification in hyperspectral images (Chen et al., 2016; Hu et al., 2015a).

The problem with CNNs is that they require significant amounts of training data because there are many learnable parameters, otherwise overfitting can occur (Larochelle et al., 2009). Parallel to this issue is the difficulty in acquiring labelled hyperspectral training samples (Yu et al., 2016), and the enormous amounts of variability in

hyperspectral data due to the illumination (see Chapter 1). As was discussed in Section 2.4.3, very few of the CNN approaches proposed for hyperspectral data are tackling the problem of limited labelled training samples.

To address the above problem, approaches for training a hyperspectral CNN under the constraint of a limited amount of labelled data are proposed and developed. In Chapter 4, an unsupervised learning method for finding low dimensional feature representations of hyperspectral data, using autoencoders, was proposed. The approach did not require any annotated training data. Classifiers for the autonomous identification of materials require annotated training data. It is possible to use the unsupervised method proposed in Chapter 4 to first learn features from the data that suppress the illumination variability, before adding a supervised classification layer on top of the encoder, similar to Chen et al. (2014). However, this would require training the classifier separately in order to preserve the illumination invariance in the features, and there remains a high chance of overfitting of the parameters of the classifier if there are not enough labelled training samples available. Also, if some annotated data is available then better features can be learnt using a supervised training scheme, instead of an unsupervised training scheme as in Chapter 4. In this chapter, an approach is taken where the labelled training data is expanded with augmentation to capture more variability and knowledge is leveraged from other labelled datasets with transfer learning. With this process, an end-to-end CNN classifier, where the parameters of the features and classifier are jointly optimised, can be trained to perform with high accuracy when there are limited available training samples.

The architecture of the hyperspectral CNN is presented in Section 5.1. A method to make the CNN more robust when there are limited labelled training data available is proposed by utilising transfer learning, which is developed in Section 5.2, and spectral relighting augmentation, which is developed in Section 5.3. Experiments are conducted in Section 5.4 to evaluate the methods for making the CNN more robust to the limited training samples, and analyse the filters learnt by the CNN, drawing links to the material spectra. Finally, the results are analysed in Section 5.5 before conclusions are drawn.

5.1 CNN Architecture for Hyperspectral Classification

Figure 5.1 shows a simplified version of an example of the architecture that was used for the hyperspectral CNN. Each input into the CNN is a single pixel spectra with dimensions $1 \times 1 \times K \times 1$ from a datacube (where K is the number of channels). The first stage of the architecture of the hyperspectral CNN consists of the convolutional layers. The first convolutional layer filters the input spectra with k_1 filters of size $1 \times 1 \times M \times 1$ (such that each filter spans M channels of the pixels spectrum). The best filter size (M) for the first convolutional layer of the pre-trained network is determined experimentally in Section 5.4.2. The second convolutional layer filters the output of the first layer with 10 filters of size $1 \times 1 \times 10 \times k_1$. All subsequent convolutional layers filter the previous layers output with 10 filters of size $1 \times 1 \times 10 \times 10$. No pooling or downsampling layers are included in this network so to preserve the spectral resolution. The second stage of the architecture consists of the fully connected layers, which each have 20 neurons. A ReLU non-linearity is applied to the output of each convolutional and fully connected layer. A normalisation layer follows the ReLU and is applied to the convolutional layers only. A softmax classifier is appended to the output of the last fully connected layer. This architecture was loosely based on successful image CNN architectures such as AlexNet (Krizhevsky and Hinton, 2012) as well as other hyperspectral CNN architectures (e.g. Hu et al. (2015b)), with preliminary experiments used to determine some of the required parameters.

The hyperspectral CNN proposed in this work is designed to classify individual pixel spectra. This is done using spectral information only, hence the spatial information in the image is not being utilised for training. This is due to the unreliable nature of texture for classification purposes, as discussed in Chapter 1. For example, a sandstone brick and crushed sandstone are the same material, but are visually different, and so the spectral information is more reliable for classification purposes. Hence, these networks are only trained by convolving filters over the spectral domain, and each pixel is classified independently of its neighbouring pixels. This approach is similar

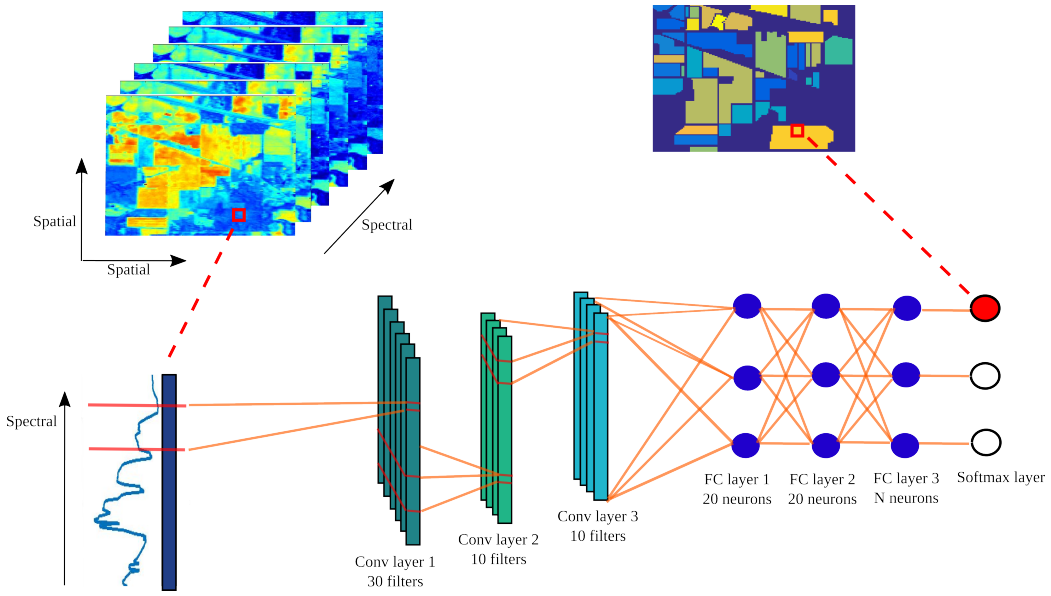


Figure 5.1 – Simplified diagram of example CNN architecture used. N stands for the number of classes, the red circle indicates the predicted class. Each convolutional layer has a normalization layer. All layers have an associated ReLU activation layer. This particular network has three convolutional and three fully connected layers.

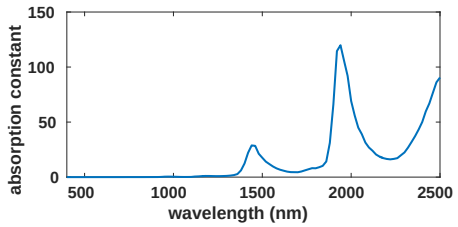
to the 1-D spectral CNN classifiers of Hu et al. (2015a) and Chen et al. (2016) and reduces the number of learnable parameters and any potential overfitting on small contextual regions that are not representative of the entire scene. In doing this, each pixel spectra in the hyperspectral image is assigned a classification label. Of course, this work provides a baseline upon which future efforts can be built to incorporate neighbouring pixels into the learning (i.e. spatial information processing). The experiments will touch on this by showing one way that the spatial information can be incorporated into the classification process when there is limited data available.

CNNs are much more efficient than their fully-connected counterparts (used in Chapter 4). Unlike fully-connected networks, which make no assumptions about the local statistics of the data and hence have many connections between neurons, CNNs exploit statistics generated from a local neighbourhood (LeCun and Bengio, 1995). That is, they assume that features in the data are localised to smaller regions of the data, and these localised features can occur in the same way anywhere in the data. In terms of image data and spatially convolving CNNs, this means that low-level features such

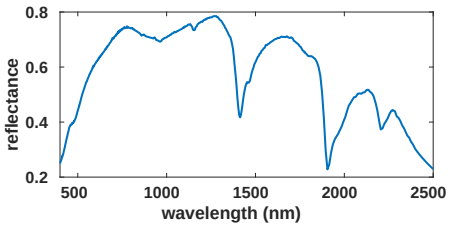
as edges and corners are localised to small regions of the image rather than the whole image, and they can also occur anywhere in the image. By exploiting local statistics with convolving filters, the weights of the network are effectively shared between neurons and there are far fewer parameters to train (Ba and Caruana, 2014). This is a form of regularisation that helps to prevent the network from overfitting. Given the constraint of limited labelled hyperspectral data, it is beneficial to use a CNN rather than a fully-connected network to classify pixels in a hyperspectral image because there are fewer parameters. Hence, less data is required to train generalisable models.

By using a CNN, the assumption about local statistics is made. It assumes that there are spectral features that occur at localised regions of the spectrum. This assumption is quite reasonable because although there are often large-scale trends in the spectra such as positive or negative gradients, many of the characteristic absorptions only occur over a small number of wavelengths. An example of this is the absorption features that occur due to the water held by materials. Water absorbs light over specific regions of wavelengths (Figure 5.2a), thus, a material such as montmorillonite has localised absorption features in its reflectance spectrum due to the water in its structure (Figure 5.2b)(Murphy, 2015). These specific absorptions are also superimposed on a generalised reduction in reflectance towards longer wavelengths caused by the exponential increase in absorption by water (Figure 5.2a) (Milliken and Mustard, 2005). Although this generalised trend is not localised to a subset of wavelengths, it could be learnt by the deeper layers of the CNN as the overall trend is superimposed onto other effects. The filters in the deeper layers have a larger receptive field and thus span a larger number of wavelengths, allowing them to learn broader features.

The CNN is highly advantageous to use if the localised features can also occur anywhere along the spectrum. Although this is not a strict requirement, it is one of the main reasons CNNs are so effective. However, this characteristic of hyperspectral data is less intuitive. In natural images, edges are invariant to their spatial location in the image. However, absorption features in spectra occur at specific wavelengths due to a material's unique chemical and structural properties. Whilst a specific material



(a) Absorption of water at wavelengths between 400 and 2500 nm.



(b) Reflectance spectrum of montmorillonite between 400 and 2500 nm (Clark et al., 2007).

Figure 5.2 – Absorption by water at specific wavelengths superimposed on a generalised increase towards larger wavelengths and its impact on the reflectance spectrum of a material which holds water (montmorillonite).

absorption will only ever occur at the same wavelength region, the shapes of these absorptions are not necessarily unique. That is, even though the underlying chemical and physical structures of materials are different, the statistics in the data could be similar at different wavelengths. An example of this is the sharp rise in reflectance that occurs at around the 680 nm wavelength for vegetation spectra (Figure 5.3). This absorption feature is unique to vegetation spectra and is caused by a sharp increase in the absorption of chlorophyll A (at about 680 nm) and increased scattering of light at longer wavelengths caused by the internal structure of the leaf (Clark, 1999). Similarly, the mineral geothite has an intense absorption at 850-900 nm followed by a sharp increase in reflectance towards longer wavelengths (Haest et al., 2012; Murphy and Monteiro, 2013; Murphy et al., 2014b). Although the absorption feature in geothite is caused by a different physical process to that observed in vegetation, its generalised shape is similar - an intense absorption followed by a relatively sharp increase in reflectance. Hence, it is possible for a CNN to learn a filter for capturing this general shape of the feature at both points in the spectrum.

5.2 Transfer Learning

CNNs have become increasingly popular in the field of computer vision and image processing, with their usage spanning diverse subject areas including image recognition

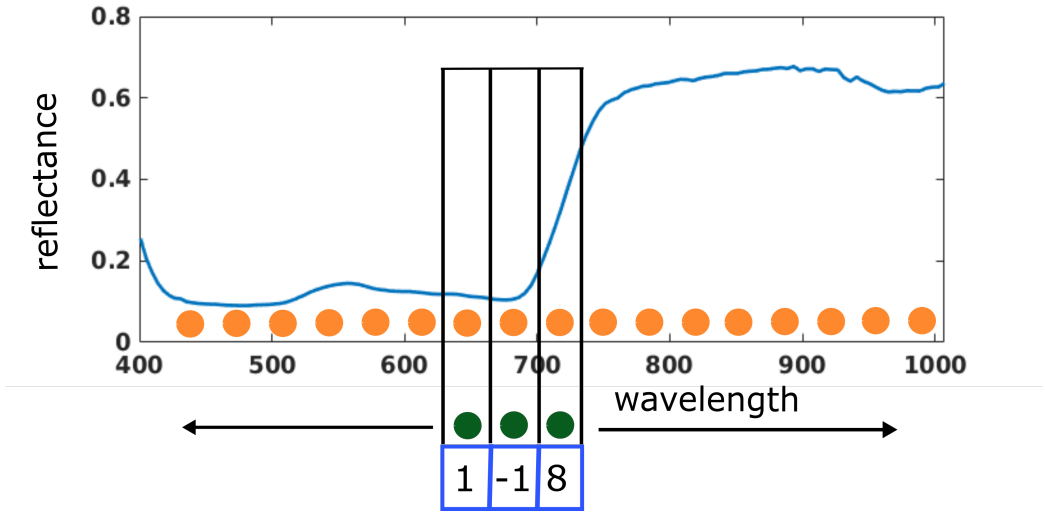


Figure 5.3 – Simplified diagram showing an example CNN filter convolving over a vegetation spectra. The orange circles represent the neurons in the input layer and the green circles represent a small learnable filter. There is a localised characteristic absorption feature that occurs at around 680 nm. Whilst this feature is unique to vegetation and only occurs at this wavelength (relating to the physiology of vegetation), features with a similar shape can occur in different parts of the spectrum for different materials composed of completely different elements. These statistics are exploited by the CNN, which learns a filter for capturing features with this shape anywhere in the spectrum.

(Krizhevsky and Hinton, 2012), object detection (Ren et al., 2015), semantic segmentation (Long et al., 2015a) and image captioning (Xu et al., 2015). The widespread use of CNNs in these communities has been facilitated by pre-trained networks which have been previously trained on large amounts of data and tailored slightly for new tasks. This transfer of knowledge makes it less time consuming to train a new classifier and reduces the need for a large labelled dataset. It is now common practice to use an off-the-shelf pre-trained CNN such as AlexNet (Krizhevsky and Hinton, 2012) as a starting point for image classification related tasks (Marmanis et al., 2016; Romero et al., 2016). The knowledge learnt by a pre-trained CNN can be transferred to a new classification task by fine-tuning it on new data (Bengio, 2012; Glorot et al., 2011; Yosinski et al., 2014) (Figure 5.4). In doing this, the large amount of data used to pre-train the CNN can be leveraged, producing better results and requiring less training for the new classification task. This is particularly useful for scenarios where

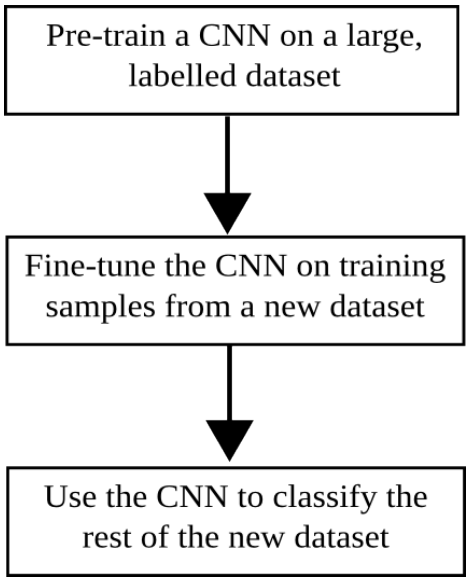


Figure 5.4 – The basic transfer learning process.

there are limited amounts of training data available for the new task.

The reason that this is possible is because the most popular pre-trained CNNs are typically quite deep and have been trained on very large, manually annotated datasets (in the order of 1000 object categories and 1.2 million images). As a result, the earlier layers in the pre-trained CNN architectures tend to learn very generic, basic image features such as edges and corners at different orientations (similar to Gabor filters). These types of features are common in most images, not just the types of images in the training sets. In the deeper layers, more abstract features are learnt such as combinations of basic features that take on the shapes of the object classes in the training dataset. These features are less generic, and are most useful for classifying the types of images specifically in the training set (Long et al., 2015b; Yosinski et al., 2014; Zeiler and Fergus, 2014). When one of these pre-trained CNNs is fine-tuned on data from new classes, then during training the network will alter the deeper weights such that they resemble the more relevant classes, and will most likely preserve the lower level features as they remain useful. The benefit of this is that the network does not have to learn all of the parameters from scratch. Training from scratch is time consuming and produces poor results when the quantity of training data for the new task is low.

Although CNNs have been used successfully for pixel-wise classification of hyperspectral images, their performance and widespread use is still very far behind the CNNs adoption in the computer vision community. A large contributing factor for this is the lack of off-the-shelf networks available for training hyperspectral CNNs. The large majority of hyperspectral CNNs in the literature (e.g. Cao et al. (2016); Hu et al. (2015a); Yang et al. (2016)) have been trained and tested on publicly available, thoroughly annotated datasets (e.g. Pavia University, Indian Pines, Salinas, Kennedy Space Center from Chapter 3). The sufficient labelling of these datasets has allowed these networks to be trained from scratch. But for many applications, datasets will not have been annotated to the same degree, making it much harder to train a CNN classifier as they require a sufficient number of training samples. This is particularly true for CNNs with many layers, where training with small amounts of data will often lead to overfitting of the parameters (Larochelle et al., 2009). For these scenarios, it would be of great benefit to have some pre-trained networks available that could be fine-tuned for new classification problems. This would reduce the training time and increase the performance of hyperspectral CNNs, especially for applications where the amount of annotated training data is limited. Once these networks are trained they can be used multiple times to make training new networks much faster. This would help facilitate the wide-spread usage of CNNs in hyperspectral applications, and the remote sensing community working with hyperspectral data could see the significant classification performance gains already seen in the computer vision community.

There are many questions that must be considered when training these off-the-shelf pre-trained CNNs for hyperspectral classification. For instance, regarding the datasets used for pre-training, the popular pre-trained CNN AlexNet (Krizhevsky and Hinton, 2012) was trained on a very large set of natural images with a variety of classes - ImageNet (Deng et al., 2009). In the hyperspectral community, there are not many public datasets that are thoroughly labelled, so if they do not cover similar material classes to the new task, it must be determined if it is still worthwhile doing pre-training. The datasets used for pre-training will have been captured from a different sensor to the dataset for the new task, posing compatibility issues. Hence, a suitable wavelength

interval for the pre-training dataset must be investigated. This will determine the resolution that data for new tasks that use the pre-trained network must have. A suitable filter size for the first convolutional layer is also an important consideration. Each of these factors will affect the overall performance of how well hyperspectral CNNs can be fine-tuned for new applications.

Another major concern which must be addressed, is that many of the publicly available, annotated hyperspectral datasets have been captured from either airborne or satellite platforms. Very few annotated public datasets have been captured from a field-based platform. Hence, it must be determined if it is possible to use the pre-trained networks, that have been trained on data from an overhead perspective and at large sensor to target distances, to initialise networks for field-based applications. Field-based hyperspectral sensing is important for many robotics and autonomous applications in mining (Murphy and Monteiro, 2013; Murphy et al., 2012; Schneider et al., 2012), agriculture (Wendel and Underwood, 2016), urban sensing (Ramakrishnan et al., 2015) and disaster response (Trierscheid et al., 2008). Since the potential of transferring knowledge from airborne platforms to field-based platforms has not been formally determined by previous works, this is the focus of the transfer learning experiments in this thesis.

The overall objective of this transfer learning analysis is to consider all of these questions and to facilitate the use of powerful CNN classifiers for hyperspectral data through transfer learning. The findings of this work make training CNNs (a state-of-the-art classification tool) for hyperspectral applications much easier and faster and, in some cases, with improved results.

5.2.1 Datasets to use for pre-training

Table 5.1 – A selection of the most common, annotated hyperspectral datasets which are publicly available.

dataset	**spatial size (pixels)**	**spatial res (m/pixel)**	**# channels**	**spectral range (nm)**	**spectral res (nm)**	**# classes labelled**	**sensed by**
Indian Pines	145 x 145	20	200	400-2500	10	16	AVIRIS
Salinas	512 x 217	3.7	204	400-2500	10	16	AVIRIS
Kennedy Space Centre	512 x 614	18	176	400-2500	10	13	AVIRIS
Pavia Univeristy	610 x 340	1.3	103	430-860	4	9	ROSIS-3
Pavia Centre	1096 x 715	1.3	102	430-860	4	9	ROSIS-3

To train the proposed CNN which will be used to initialise new CNNs, a dataset must be chosen. Several publicly available hyperspectral datasets that have been thoroughly annotated, exist (Table 3.1). These datasets are all captured from airborne platforms and have different spatial and spectral resolutions. It is also worth noting that all datasets, because captured from air, have quite a large spatial resolution (in the order of meters) and are all overhead in perspective.

The datasets used to train the proposed pre-trained CNN should have balanced class sizes (Kubat and Matwin, 1997). Hence, the public datasets with the bigger class sizes are chosen to be used for the pre-training. The number of samples per class are higher on average in the Salinas and Pavia University datasets, which is desirable, whereas the classes in the Indian Pines, Kennedy Space Centre and Pavia Centre datasets have far fewer samples, making it more difficult to train a deep CNN. In order to increase the number of samples per class, some of the more similar vegetation classes can be grouped together. This has the added advantage of helping the pre-trained

Table 5.2 – Classes and number of samples for Indian Pines.

Class	Number of samples
Alfalfa	46
Corn-notill	1428
Corn-mintill	830
Corn	237
Grass-pasture	483
Grass-trees	730
Grass-pasture-mowed	28
Hay-windrowed	478
Oats	20
Soybean-notill	972
Soybean-mintill	2455
Soybean-clean	593
Wheat	205
Woods	1265
Buildings-Grass-Trees-Drives	386
Stone-Steel-Towers	93

network to avoid learning filters to do fine-grained classification (e.g. classifying lettuce at different stages of growth). Since these pre-trained networks are to be used to initialise other networks, there is no need for such a fine-grained classification, and it is more beneficial to have bigger class sizes (Huh et al., 2016). The datasets chosen for pre-training with their original class names and sizes are shown in Tables 5.2,5.3 and 5.4.

5.2.2 Forming a composite dataset for pre-training

Rather than use a single dataset, it is more advantageous to use a composite of several of the datasets. This has the benefit of increasing the number of classes in the dataset so that more diverse features can be learnt and more knowledge can be transferred to new learning tasks (Huh et al., 2016).

These datasets have different properties and come from multiple sensors, so they must be combined together in a reasonable way. This involves interpolating all datasets so that they all have a common spectral resolution and range. If the common resolution

Table 5.3 – Classes and number of samples for Salinas.

Class	Number of samples
Brocolo green weeds 1	2009
Brocolo green weeds 2	3726
Fallow	1976
Fallow rough plow	1394
Fallow smooth	2678
Stubble	3959
Celery	3579
Grapes untrained	11271
Soil vineyard develop	6203
Corn senesced green weeds	3278
Lettuce romaine 4wk	1068
Lettuce romaine 5wk	1927
Lettuce romaine 6wk	916
Lettuce romaine 7wk	1070
Vineyard untrained	7268
Vineyard vertical trellis	1807

Table 5.4 – Classes and number of samples for Pavia University.

Class	Number of samples
Asphalt	6631
Meadows	18649
Gravel	2099
Trees	3064
Painted metal sheets	1345
Bare soil	5029
Bitumen	1330
Self-blocking bricks	3682
Shadows	947

is too fine then interpolation errors will arise for datasets with a coarse resolution, but if it is too coarse then much of the important information in the spectrum will be lost. The best spectrum for this composite dataset is determined experimentally in Section 5.4.2.

Another way the different datasets must be made more similar, is through pre-processing of the spectra prior to input into the CNN. This also helps the CNN to learn better features (Pal and Sudeep, 2016). To do this, first the spectra must

be normalised to reflectance via a flat-field calibration panel or other method. By measuring each spectrum relative to the incident illumination, most of the difference arising from capturing images with different cameras, is removed. At this point the spectrum should be between zero and one. Then, the spectrum is offset so that the reflectance at an arbitrarily selected wavelength (e.g. half way through the spectrum) is set to zero. This makes some of the input units negative and some positive without compromising the reflectance spectrum's natural structure, and vertically centres the spectra about the zero axis, further removing intensity differences due to capturing images with different cameras. Similar pre-processing is used to train spatial CNNs on image data.

5.2.3 Pre-training and fine-tuning a network

Algorithm 5.1: Procedure for pre-training a CNN and then transferring the knowledge learnt - by fine-tuning for a new task.

Input : composite dataset D_{comp}, dataset for new task D_{new} with N classes
Output: CNN for new task CNN_{new}

1 pre-process D_{comp} with radiometric normalisation and wavelength zeroing;
2 train a CNN, CNN_{comp}, on D_{comp};
3 **for** *k=1:number of new tasks* **do**
4 pre-process D_{new} with radiometric normalisation and wavelength zeroing;
5 interpolate D_{new} so that the spectral resolution matches those of D_{comp};
6 replace the last fully connected layer of CNN_{comp} with a fully connected layer that has N neurons;
7 randomly initialise the parameters of the last fully connected layer of CNN_{comp} using the Xavier method (He et al., 2015);
8 fine-tune this network using D_{new} ;

Algorithm 5.1 details how to pre-train a CNN with the architecture outlined above on the composite airborne dataset and then fine-tune that network on data for a new task (e.g. data captured from a ground-based sensor). As is shown in the algorithm, the CNN only needs to be trained on the composite dataset once and after that can be re-used many times by fine-tuning it on new datasets. In this sense, the knowledge

acquired from the composite dataset can be transferred to many new tasks (Dong et al., 2014).

5.3 Spectral Relighting Augmentation

When there are a limited number of training samples, it is difficult for a supervised classifier to capture the variability in the data, even if transfer learning is used. Incorporating illumination variability into the training of supervised hyperspectral classifiers is an ongoing research question, with two main approaches being utilised. The first is by using large amounts of training data that sufficiently represent the variability within the scene. However, a significant amount of data is required in order to capture such variability and this becomes increasingly difficult in complex scenes. The geometric structure of surfaces occludes regions from being illuminated evenly by terrestrial sunlight and diffuse skylight, with incident illumination and intensity varying on a per-pixel basis. Further, because of the high-dimensionality of hyperspectral data, more data is required to adequately fill the space (i.e. the curse of dimensionality). The second method is to use pre-processing to convert the data to a form that is less dependent on illumination. This is typically done by converting the raw DN values to reflectance using a process by which the hyperspectral image is normalised against a material of known reflectance within the scene such as a calibration board (e.g. flat-field correction, see Section 2.4.1). Flat-field correction is only correct for the region in which the calibration board was placed. In areas with significantly different incident illumination, the incident illumination varies based on the illumination sources and geometry of the surface. Also, placing additional hardware within the scene is impractical in hazardous environments, or for robotic platforms operating in dynamic environments.

A common tactic for making a classifier more robust to undesirable variances, is to use data augmentation to artificially simulate those variances in the training data so that the classifier can learn to account for them (Gupta et al., 2014; Jaderberg et al., 2014; Zhang et al., 2014). The strategy developed in this section involves

augmenting the input spectra using a relighting technique (Ramakrishnan et al., 2015), similar to Section 4.2, in order to make the classifier robust to illumination variability. Critical to the relighting is the calculation of the ratio between the two primary outdoor illumination sources (terrestrial sunlight and diffuse skylight), for which a novel, image based method is proposed.

The advantages of the proposed augmentation strategy are that it does not require multiple sensor modalities (e.g. hyperspectral camera with LiDAR) or computational atmospheric models. It allows training data to be collected from within sunlit regions, which are commonly easier to label, while classification can be performed on both sunlit and shadowed regions.

5.3.1 Augmenting Spectra

Training supervised classification algorithms using small labelled regions fails to account for illumination variability induced by the complex geometry of a scene. This thesis proposes the use of relighting (Ramakrishnan et al., 2015) as a data augmentation strategy, in order to encompass the illumination variations typically found in the outdoor environment. Such variations include the surface orientation with respect to the sun, the amount of sky exposure and shadows. This allows training data to be obtained from regions in the image that are either easy to access or easy to label, and inference can then be performed on the remaining data. The following relighting equation focuses on obtaining labels from sunlit regions, and classifying on shadowed and sunlit data, though the approach is easily transferable to the reverse scenario (labelling shadowed regions and inferring on sunlit and shadowed regions).

Relighting is the process of simulating the spectral appearance of a region under different illumination and geometrical conditions that are not encompassed by the training set. The relighting equations 2.3 and 2.4 were derived for the individual cases of relighting sunlit spectra to shadow and full exposure respectively. However, for augmentation purposes, in this work it is desirable to have one equation to sample, which will cover both the sunlit and shadowed scenarios. Hence, a generic equation is

formulated by including a binary visibility term V. The visibility determines whether the simulated region has a line-of-sight with the sun and hence whether the spectra are relit to shadow or sunlight. Also, rather than relighting spectra to full exposure for the sunlit case, the equation is modified to include the sun angle θ and sky factor Γ of the relit region so that the geometry of the relit spectrum's surface can be augmented. By doing this, the variability of the geometry of sunlit pixels in the scene captured by the training data can also be increased via augmentation.

For a single sunlit datapoint at region i, the augmented version L_j is calculated by multiplying the training spectra L_i (prior to pre-processing) by a wavelength dependent scaling factor:

$$L_j(\lambda) = L_i(\lambda)\frac{V_j\dfrac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_j + \Gamma_j}{\dfrac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_i + \Gamma_i}, \tag{5.1}$$

where θ_i and Γ_i are the geometric parameters describing the sun angle and sky factor of the original training datapoint, while V_j, θ_j and Γ_j are the parameters of the augmented datapoint. When V_j is 0, relighting has the effect of simulating the appearance of the original datapoint within a shadowed region, while setting it to 1 simulates the same datapoint with a different orientation. Relighting alters both the brightness and spectral curve shape of the datapoint. The derivation for this relighting equation is in Section D.3.

In order to relight the data during training, several illumination and geometric parameters are required. The first is the terrestrial sunlight-diffuse skylight ratio $\frac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}$, which describes the relationship, in terms of both spectral distribution and intensity, between the primary illumination sources in the outdoor environment. The geometric parameters, such as the sun angle and sky factors are typically known when utilising multi-modal systems, where geo-registered point cloud data can be used explicitly to estimate these values (Ramakrishnan et al., 2015). However, for image based methods, these parameters remain unknown, therefore a sampling procedure is used during training to augment the data with the relighting equation. During each batch

Algorithm 5.2: Augmenting a batch of spectra for training the CNN. $\mathcal{U}$ and B represent uniform and Bernoulli distributions respectively.

Input : batch of spectra **data**, number of irradiance ratio estimates M
Output: augmented batch of spectra **dataAug**

dataAug $\leftarrow$ **data**
for *k = 1 to M* **do**
  sample $\theta_A \sim \mathcal{U}[0, \frac{\pi}{2})$, $\Gamma_A \sim \mathcal{U}[0, 1]$
  irradianceRatio$\leftarrow$ eq.(5.6)(θ_A, Γ_A)
  for *l = 1 to size of batch* **do**
    $V_j \sim B(1, \frac{1}{2})$, $\theta_i \sim \mathcal{U}[0, \frac{\pi}{2})$, $\theta_j \sim \mathcal{U}[0, \frac{\pi}{2}]$, $\Gamma_i \sim \mathcal{U}[0, 1]$, $\Gamma_j \sim \mathcal{U}[0, 1]$
    relitSpectra$\leftarrow$ eq.(5.1)(**data**; V_j,θ_i,θ_j, Γ_i,Γ_j,**irradianceRatio**)
    dataAug $\leftarrow$ **dataAug** $\cup$ **relitSpectra**

of gradient descent optimisation of the network, the geometric parameters V_j, θ_j, Γ_j, θ_i and Γ_i are sampled as shown in Algorithm 5.2.

5.3.2 Image Based Estimation of the Terrestrial Sunlight-Diffuse Skylight Ratio

The terrestrial sunlight-diffuse skylight ratio is integral to the relighting process. Manual methods involving the user selecting two adjacent points obtained from the same material (Ramakrishnan et al., 2015), or the use of computational atmospheric models may be used, but these require the knowledge of parameters such as turbidity, gas concentration, water vapour and humidity (Berk et al., 1987). Automatic methods also exist (Ramakrishnan, 2016), but these require extra sensor modalities (e.g. LiDAR). Within the colour-constancy literature (see Section 2.4.1), there are numerous techniques for estimating the illuminant from an image (Barron, 2015). However, these techniques often assume that there is only one illuminant in the scene. Methods that assume multiple illuminants (Cheng et al., 2016) usually extract the combination of illuminants incident at each region, rather than the individual terrestrial sunlight and diffuse skylight illuminants. Whilst the combination of illuminants is useful for correcting the colour of a region, it cannot be used for computing the ratio of illumi-

nants necessary for the relighting process.

In Chapter 4, the terrestrial sunlight-diffuse skylight ratio did not need to be accurately estimated. Instead, many candidate ratios were generated so that commonality in the data could be learnt through an unsupervised process. The resulting autoencoder mapped spectra to a representation which was independent of the prevailing atmospheric conditions. For the supervised CNN classifier to learn to map spectra to their correct label, the distribution of the training data must accurately portray the distribution of the test data. Hence, the training data must be augmented using a terrestrial sunlight-diffuse skylight ratio that closely reflects the scenes prevailing atmospheric conditions.

Thus, in this work, a novel image based method for estimating the terrestrial sunlight-diffuse skylight ratio from the scene itself is proposed. Because the ratio is derived from the inherent properties of the data collected from the scene, it is a better estimate of the prevailing atmospheric conditions. The method consists of a three stage process of candidate generation, refinement and smoothing.

If two spectra from regions (A, A') of the same material are selected from a sunlit and shadowed region respectively, both of which have the same orientation, the terrestrial sunlight-diffuse skylight ratio can be approximated by equating the reflectance of two regions having the same material, but one being in shadow. From the model (2.2):

$$L_A(\lambda) = \frac{\rho_A}{\pi} \left[E_{sun}(\lambda)\tau(\lambda)\cos\theta_A + \Gamma_A E_{sky}(\lambda) \right], \tag{5.2}$$

$$L_{A'}(\lambda) = \frac{\rho_{A'}}{\pi} \left[\Gamma_{A'} E_{sky}(\lambda) \right], \tag{5.3}$$

where L_A is the radiance of the material in sunlight and L'_A is the radiance of the material in shadow. Equating the reflectances ρ_A and $\rho_{A'}$ and simplifying gives:

$$\frac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)} = \frac{\Gamma_{A'} L_A(\lambda) - \Gamma_A L_{A'}(\lambda)}{L_{A'}(\lambda) cos\theta_A}, \tag{5.4}$$

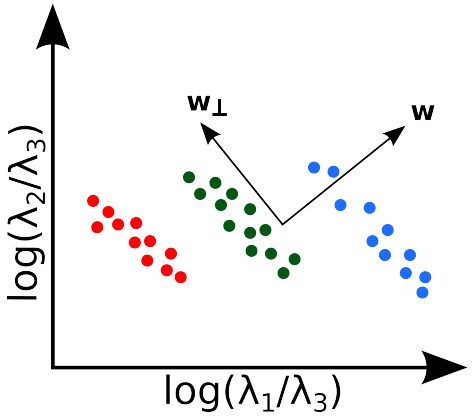


Figure 5.5 – Example of how points of different material, given by the different colours, are projected into a 2D log-chromaticity space by taking the log of the ratio of two wavelengths. The spread intra-class variation within each colour is due to variations in the illuminant, such as when the material is in shadow. In this space, there exists an illumination invariant direction w which captures changes in material and an orthogonal direction $w_\perp$ which captures changes in the illuminant.

and if regions A and A' are at the same orientation, then:

$$\Gamma_A = \Gamma_{A'} \tag{5.5}$$

$$\begin{aligned} \frac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)} &= \frac{\Gamma_A}{\cos\theta_A}\left[\frac{L_A(\lambda)}{L_{A'}(\lambda)} - 1\right], \\ &\propto \frac{L_A(\lambda)}{L_{A'}(\lambda)} - 1. \end{aligned} \tag{5.6}$$

Since the scene geometry is considered to be unknown due to the use of only image data, it is assumed that the scene is spatially consistent. Hence, the selection of pairs of adjacent points in sunlit and shadowed regions from the same material are assumed to be at the same orientation and can be used to obtain candidate terrestrial sunlight-diffuse skylight ratios that are a scalar multiple of the underlying ratio using 5.6. The scalar multiple is dependent on the orientation of the pair of regions.

To automatically select adjacent pairs of points in and out of the shadowed regions, three bands are generated to form a pseudo RGB image. Hypothetically these can be any bands, however they should be chosen to maximise the discriminability of

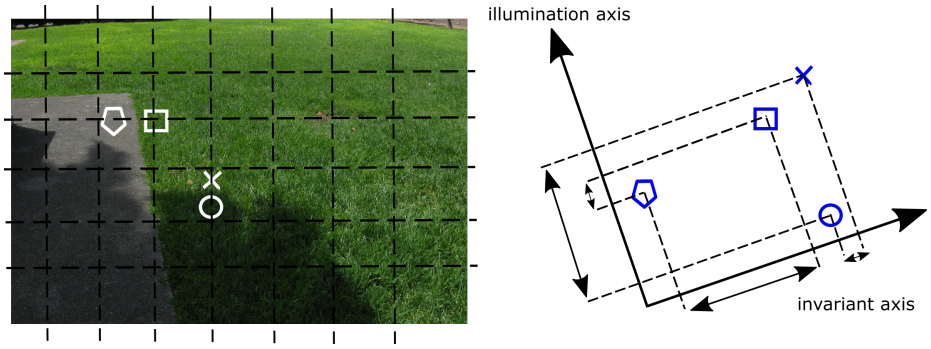


Figure 5.6 – Depiction of candidate pairs along horizontal and vertical transects projected onto illumination and invariant axes for an example image. Pairs on a class boundary have a large separation on the invariant axis, whilst pairs on a shadow boundary have a small separation on the invariant axis and a large separation on the illumination axis.

the classes. If in the visible domain, these bands can be 450nm, 550nm and 600nm (peak wavelengths of an RGB camera), and if in the Short-Wave Infrared (SWIR) domain the bands can be 1060nm, 1250nm and 1630nm (the middle of sections of the spectrum outside the destructible water bands). Next, the three channel image is converted to 2D log-chromaticity space (Figure 5.5) where the illumination invariant direction is found through entropy minimization (Corke et al., 2013; Finlayson et al., 2004). The $1D$ projection of the image onto this axis should be invariant to changes in the illumination. The orthogonal axis is found (deemed the illumination axis) which captures large changes in illumination resulting from either shadow or spectrally discrete class boundaries.

Candidate pairs of adjacent points, which are assumed to have similar orientation, are taken from horizontal and vertical transects (Figure 5.6) of the pseudo RGB image and projected onto the invariant and illumination axes:

$$I_{inv_i} = e^{\mathbf{X}_i \mathbf{w}}, \tag{5.7}$$

$$I_{ill_i} = e^{\mathbf{X}_i \boldsymbol{w}^{\perp}}, \tag{5.8}$$

where $\mathbf{X}_i$ is the log-chromaticity of point i, the vector $\mathbf{w}$ is the direction of the invariant axis and I_i is the exponential of the points location on either the invariant

or illumination axis. If there is a large difference between a pair along the illumination axis (corresponding to either a shadow or material class boundary) but a small difference between the pair along the invariant axis (ruling out the class boundary), then the pair is considered to be valid (Figure 5.6) and the ratio between the spectra is calculated. For a candidate pair of points, the validity of them constituting a sun-shadow pair can be determined using:

$$\frac{|I_{inv_1} - I_{inv_2}|}{I_{inv_2}} < \mu, \tag{5.9}$$

$$\frac{|I_{ill_1} - I_{ill_2}|}{\min(I_{ill_1}, I_{ill_2})} > \xi, \tag{5.10}$$

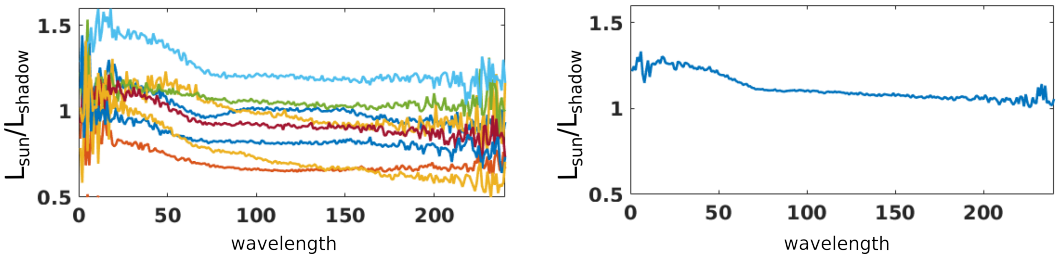
where reasonable values for μ and ξ are 0.3 and 1.2 respectively. The average ratio is subsequently taken over all valid candidate pairs (Figure 5.7) before smoothing with an Savitzky-Golay filter (Savitzky and Golay, 1964). The result is used to estimate $\frac{L_A(\lambda)}{L_{A'}(\lambda)}$ in equation 5.6. The geometric parameters θ_A, Γ_A in equation 5.6 are unknown, although, they can be sampled along with the geometric parameters needed for relighting as in Algorithm 5.2.

5.4 Experimental Results

The improvement in hyperspectral CNNs which use transfer learning and spectral relighting augmentation when there is limited labelled training data are evaluated experimentally in Section 5.4.2 and Section 5.4.3 respectively. To obtain a more intuitive understanding of what the CNN is learning from the spectral data, an attempt is made to draw links between the filters, the spectra and which features in the spectra are activating the filters Section 5.4.4.

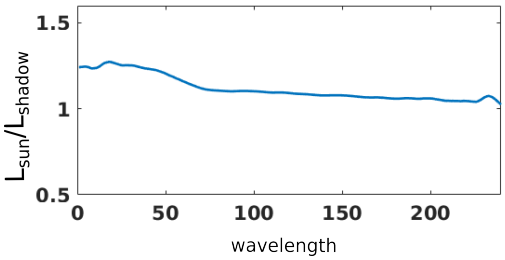
5.4.1 Network Architecture and Parameters

For the experiments where networks were trained with an unspecified filter size, the filter size in the first layer (M) was 30 for the Great Hall SWIR and mining timelapse



(a) Ratios from all valid sun-shadow pairs of spectra.

(b) Mean of ratios from valid pairs.



(c) Smoothed ratio.

Figure 5.7 – Synthetic example of the process of extracting L_{sun}/L_{shadow} once valid sun-shadow pairs have been identified from the candidate pairs.

datasets, and was 10 for the Gualtar steps dataset. The smaller filter size for the Gualtar steps dataset was due to the reduced number of channels. These filter sizes were chosen based on the filter size experiments of Section 5.4.2. Unless otherwise stated, two convolutional and two fully connected layers were used for the networks for all datasets. These values were chosen based on preliminary experimentation. For some of the experiments, it is shown that the performance improvements are not highly dependent on the architecture used.

The networks were optimised using 200 epochs of stochastic gradient descent with a learning rate of 10^{-5}, momentum of 0.9 and batch size of 50. This is with the exception of the networks in Section 5.4.2 which were pre-trained using 1000 epochs, and then the fine-tuned networks were left to optimise until convergence.

5.4.2 Evaluation of Transfer Learning

The experiments for the transfer learning work encompassed two key aspects. The first was to propose a suitable way to train a CNN that could be used as a pre-trained network and fine-tuned for new spectral classification tasks. The second aspect was comparing the networks which were pre-trained and fine-tuned against networks that were trained from scratch, which is the current method for training many hyperspectral CNNs (Chen et al., 2016; Hu et al., 2015b). The focus of this analysis was field-based applications, so networks were pre-trained from data captured from airborne sensor platforms, and fine-tuned on data captured from field-based sensor platforms.

In the first experiments, the properties of the pre-trained network, including the wavelength interval length of the input data and the size of the filters in the first convolutional layer, were determined experimentally. Given that knowledge is being transferred between datasets captured from different sensors, with different spectral and spatial resolutions, it is important to determine the impact of these properties on the learning. Separate pre-trained CNNs were made for VNIR and SWIR datasets. For these experiments, only the Pavia University dataset and the Salinas dataset were used for pre-training the VNIR and SWIR networks respectively (no composite datasets were used). Eight classes were used for each pre-training dataset, with 1000 samples per class used for training. For the Pavia University VNIR dataset the eight classes used were asphalt, meadows, gravel, trees, painted metal, bare soil, bitumen and self-blocking bricks (Table 5.4). Some of the classes from Salinas were grouped together to avoid fine-grained classification: 'broccoli green weeds 1' and 'broccoli green weeds 2' were be grouped as 'broccoli'; 'fallow','fallow rough plow' and 'fallow smooth' are grouped as 'fallow' and 'lettuce romaine' weeks 4, 5, 6 and 7 are grouped as 'lettuce' (Table 5.3). For the Salinas SWIR dataset the eight classes used were broccoli, fallow, stubble, celery, grapes, soil-vine, corn weeds and lettuce. Only the SWIR component of the Salinas dataset was used ($1020 - 2400$ nm).

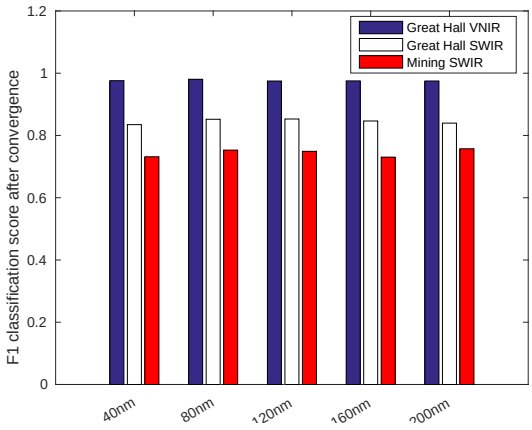
The first two target datasets that the pre-trained networks were fine-tuned on, where 'target' defines a new dataset requiring classification that differs to the ones that the

CNN was pre-trained on, were the field-based VNIR and SWIR hyperspectral images of the University of Sydney's Great Hall. These datasets were chosen because the scene contains structured (e.g. a building) and unstructured (e.g. a tree) elements and within-class variations in illumination, requiring the CNN classifier to be robust. The third target dataset is the field-based SWIR hyperspectral mining image. The scene is completely unstructured, with large variations in surface geometry. This dataset was used in the evaluation because the light shale, white shale and whaleback shale classes of spectra are similar, with the distinguishing features being very hard to detect making classification a complex task. It is worth noting that whilst the Great Hall dataset contains some similar classes to the pre-training dataset, the classes of the mining dataset are completely different to those in the pre-training dataset. Section 3.1 provides a summary of the datasets used for pre-training and also the target datasets that the pre-trained CNNs were fine-tuned on. All data was normalised to reflectance using flat-field correction.

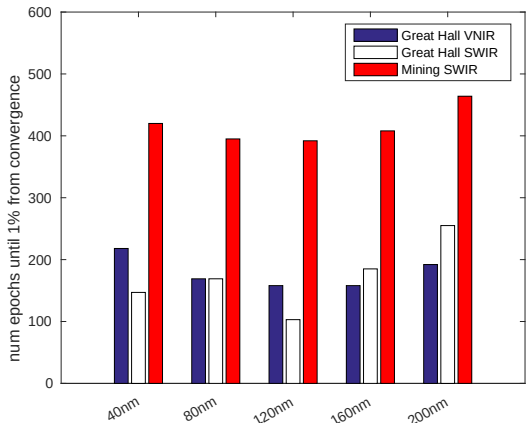
The pre-trained networks were evaluated using the classification performance from fine-tuning them on the target datasets. The dataset for the target classification task consisted of 200 training samples per class, 50 validation samples per class and 800 test samples per class. The metrics used for comparison were the F1 classification score (Section 3.2.5) after convergence and the number of epochs of optimisation that were required to reach to within one percent of convergence (Section 3.2.6). With these metrics, the added benefit of pre-training was evaluated in terms of classification error and the speed of convergence. All experiments were repeated five times, with the mean result reported.

Firstly, different spectral widths of the filters in the first convolutional layer of the pre-trained network were compared. The wavelength interval was fixed at 4 nm. The filter size was varied to cover 40 nm, 80 nm, 120 nm, 160 nm and 200 nm which corresponded to 10, 20, 30, 40 and 50 input neurons respectively (4 nm per neuron).

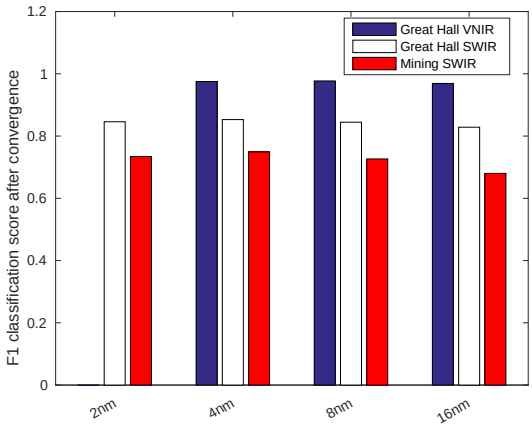
The results (Figure 5.8) suggest that the filter width doesn't have a significant impact on the classification score (at least, within the bounds experimented on). Filter width did have an impact on the convergence time. Across all datasets, the 120 nm filter (30



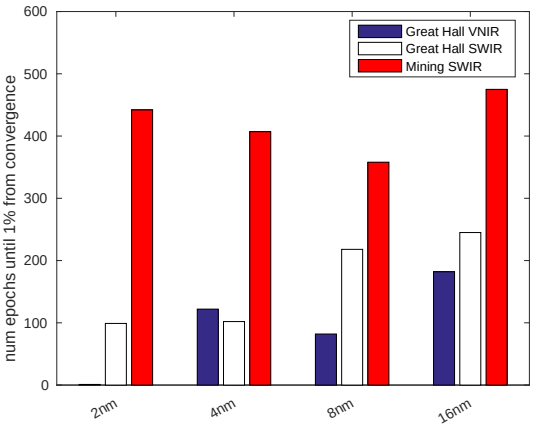
(a) Classification score after convergence for different filter sizes



(b) Number of epochs to reach 1% of classification score for different filter sizes



(c) Classification score after convergence for different wavelength intervals



(d) Number of epochs to reach 1% of classification score for different wavelength intervals

Figure 5.8 – Transfer learning results: Comparison of the results of using different filter sizes for the first convolutional layer of the pre-trained network and different wavelength intervals.

neurons for a 4 nm wavelength interval) required the shortest time for the optimisation to converge, and hence was concluded to be the best size.

Secondly, different wavelength intervals for the data to do the pre-training were compared. This is important because the pre-training and target datasets must both have the same wavelength interval, so either one will have to be down-sampled or the other will have to be up-sampled if they are different. For the experiment, the wavelength interval was set to 2 nm, 4 nm, 8 nm and 16 nm. The spectral width

of the first convolutional layer filters was fixed at 120 nm (which required adjusting the number of units due to the changing wavelength interval). For both experiments, a CNN architecture with three convolutional layers and three fully connected layers was used.

For the classification score, the results (Figure 5.8) were very similar across all datasets apart from the mining SWIR dataset where the 4 nm spectral interval marginally had the best classification score. The convergence times showed a bit more variation, with the Great Hall SWIR dataset optimising much faster with shorter resolutions ($2-4$ nm) but the Great Hall VNIR and Mining SWIR datasets optimising slightly faster with the mid-length resolutions, particularly 8 nm, than with the shorter resolutions. For all datasets, the larger 16 nm optimised the slowest. No result could be obtained for the Great Hall VNIR dataset for 2 nm. It was concluded that 4 nm be the most appropriate resolution due to its sound performance on all datasets.

For the second set of transfer learning experiments, the performance of a CNN pre-trained with a composite dataset was compared to a CNN trained from scratch, that is, without pre-training, where the learn-able parameters were all randomly initialised. For these experiments, VNIR and SWIR composite datasets were formed for the pre-training (coming from multiple datasets). The VNIR dataset was made up of eight classes from Pavia University and eight classes from Salinas, with 1200 samples per class used for training. The similar classes from Salinas that were grouped together in the first set of experiments were once again grouped together to avoid fine-grained classification. The spectrum of the VNIR dataset was $430-860$ nm. The SWIR dataset was made up of eight classes from Salinas and five classes from Indian Pines, with 1200 samples per class used for training. Some of the classes from Indian Pines were also grouped together to avoid fine-grained classification: 'corn-notill', 'corn-mintill' and 'corn' are grouped as 'corn', 'grass-pasture', 'grass-trees' and 'grass-pasture-mowed' were grouped as 'grass' and 'soybean-notill', 'soybean-mintill' were grouped as 'soybean-till' (Table 5.2). The eight classes from Pavia University and Salinas were the same as in the first experiments, and the classes for Indian Pines were corn, grass-trees, soybean-till, soybean-clean and woods. The classes selected

Table 5.5 – The classes chosen from each dataset to pre-train the composite CNNs.

VNIR	SWIR
Pavia University	**Indian Pines**
asphalt	corn
meadows	grass-trees
trees	soybean-till
painted metal	soybean-clean
bare soil	woods
bitumen	
self-blocking bricks	
Salinas	**Salinas**
broccoli	broccoli
fallow	fallow
stubble	stubble
celery	celery
grapes	grapes
soil vineyard	soil vineyard
corn-senesced	corn-senesced
lettuce	lettuce

from each of the datasets to make the composite datasets used for pre-training are summarised in Table 5.5. The spectrum of the SWIR dataset was $1020 - 2400$ nm. Prior to combining datasets, the wavelength interval was made to be 4 nm (based on the previous wavelength interval results).

The target datasets on which the pre-trained networks were fine-tuned, were the same hyperspectral images of the University of Sydney's Great Hall and mining dataset from the first experiments. In this experiment, the number of samples used for training and testing on the new datasets depended on the experiment. All data was normalised to reflectance using flat-field correction.

The performance of the network as it optimised when pre-trained was compared against a network trained from scratch. The parameters of the network trained from scratch were randomly initialised using the state-of-the-art improved Xavier method (He et al., 2015), which has been shown to produce good results. The metric used for evaluation was the F1 classification score after convergence and also the number of epochs taken for the classification score to get to within one percent of that score.

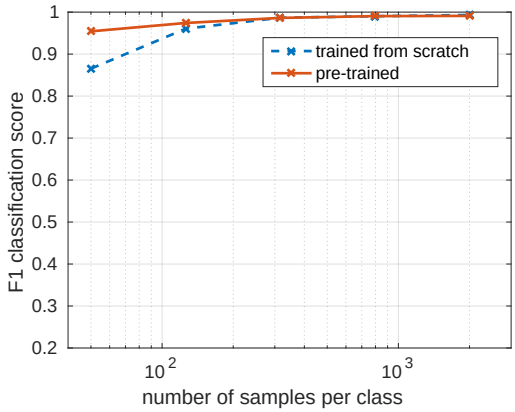
All experiments were repeated five times, with the mean result reported.

Firstly, the impact of pre-training with a varying number of training samples for fine-tuning in the target dataset was analysed. A CNN architecture with three convolutional layers and three fully connected layers was used. The number of samples in the target training set was varied logarithmically between 50 and 2000 and the testing set was fixed at 800 samples per class, with 50 samples used for validation.

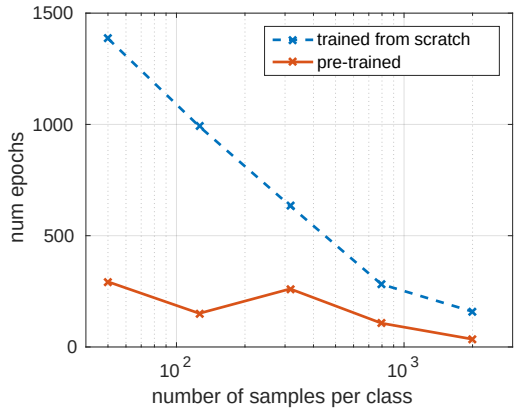
The results (Figure 5.9) compare training from scratch with pre-training (i.e. utilising transfer learning) and suggest that transfer learning (i.e. pre-training) improves the classification error when there is a small amount of training data (less than approximately 100 samples per class). This was consistent for all datasets. However, when there were a larger number of training samples, the classification error was the same regardless of whether knowledge was transferred via pre-training or not.

The most significant impact of the transfer learning was seen in the time taken for convergence. For the Great Hall VNIR results, training from scratch took a long time to optimise when there was a small amount of training data, and this time decreased quite rapidly as the amount of training data increased. When pre-training was done to initialise the network, the training time was relatively consistent for all numbers of training samples tested. For small training set sizes, the pre-trained networks optimised much faster than the networks trained from scratch (roughly four times fewer epochs were required to reach one percent of convergence). For larger dataset sizes, there was still an improvement in the optimisation time, but it was smaller because the time taken when training from scratch improved significantly. For the Great Hall SWIR and Mining SWIR datasets, when pre-training, the optimisation convergence time was more dependent on the number of samples than it was for the Great Hall VNIR dataset. Just as with training from scratch, it took more time to converge for small amounts of training data and became faster as the size of the training set increased. However, the convergence time when pre-training was better than the convergence time when training from scratch, particularly for the smaller training sets. For the larger datasets, convergence time became more similar, as did the classification error.

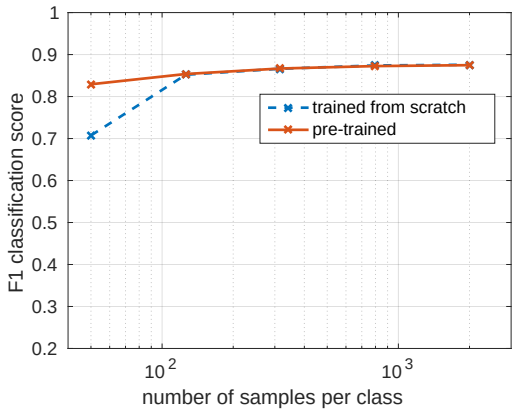
Secondly, the generality of the results for different architectures was analysed. The performance when training from scratch and pre-training for several different CNN architectures was compared. The number of training samples was fixed at 200, the number of validation samples was fixed at 50 and the number of testing samples was fixed at 800. The number of convolutional layers and fully connected layers in the CNN was varied between two and five. From the filter width experimental results, the spectral width of the first layer convolving filter was fixed at 120 nm for both experiments.



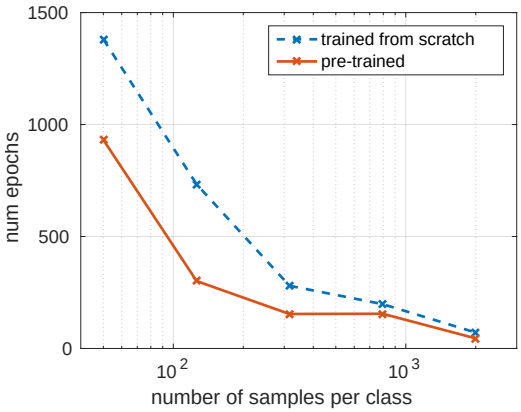
(a) Great Hall VNIR F1 classification score after convergence



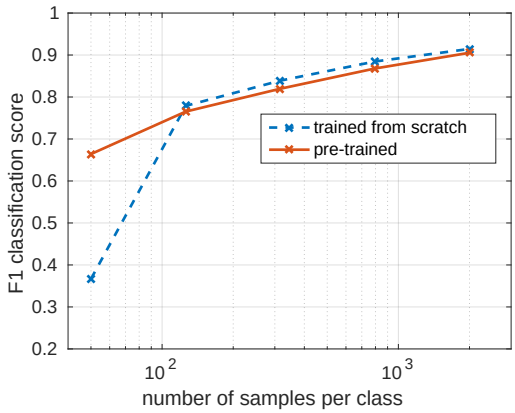
(b) Great Hall VNIR number of epochs to reach 1% of convergence of classification score



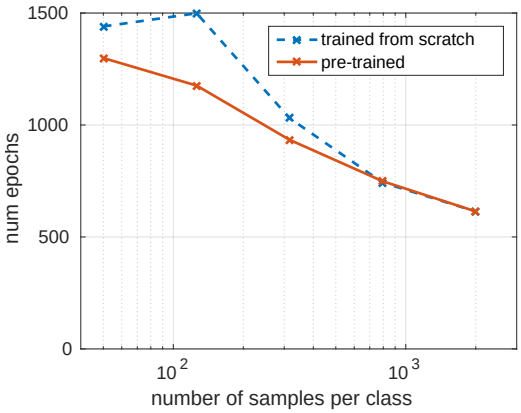
(c) Great Hall SWIR F1 classification score after convergence



(d) Great Hall SWIR number of epochs to reach 1% of convergence of classification score

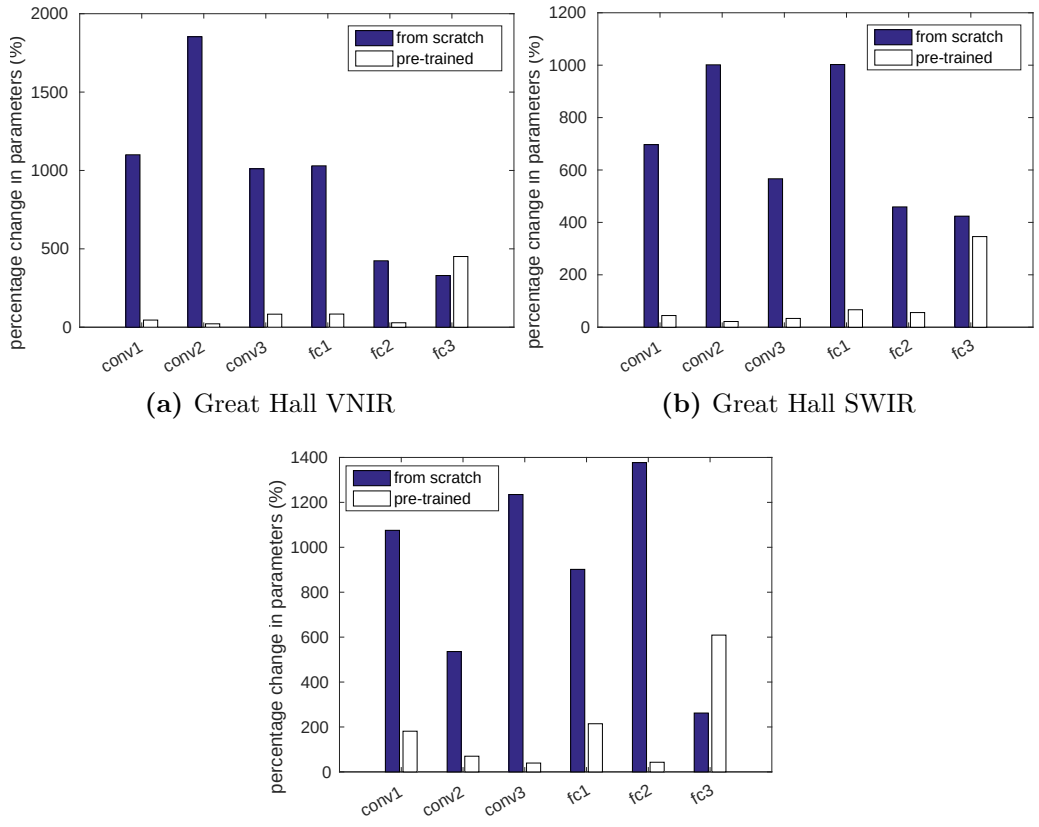


(e) Mining SWIR F1 classification score after convergence



(f) Mining SWIR number of epochs to reach 1% of convergence of classification score

Figure 5.9 – Transfer learning results: Comparison of pre-training and training from scratch for different numbers of training samples available for the target dataset. For the Great Hall VNIR dataset, the CNN was pre-trained on the VNIR composite dataset (Table 5.5). For the Great Hall SWIR and Mining SWIR datasets, the CNN was pre-trained on the SWIR composite dataset (Table 5.5).

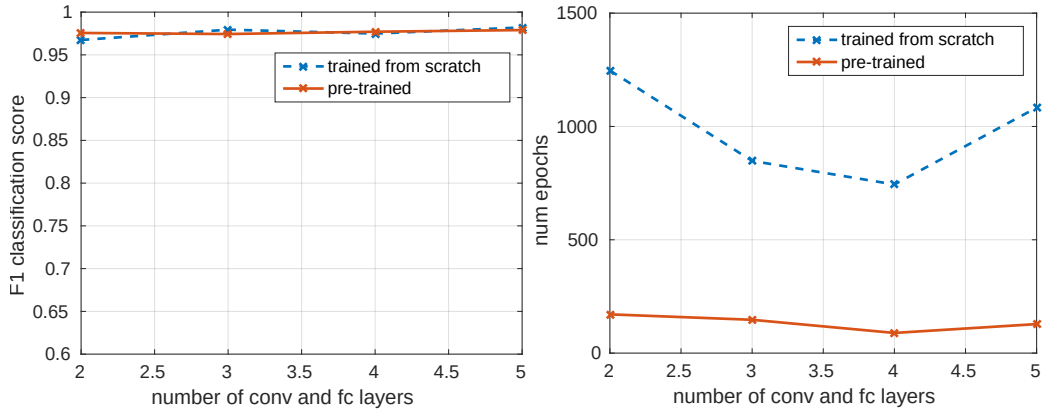


(a) Great Hall VNIR

(b) Great Hall SWIR

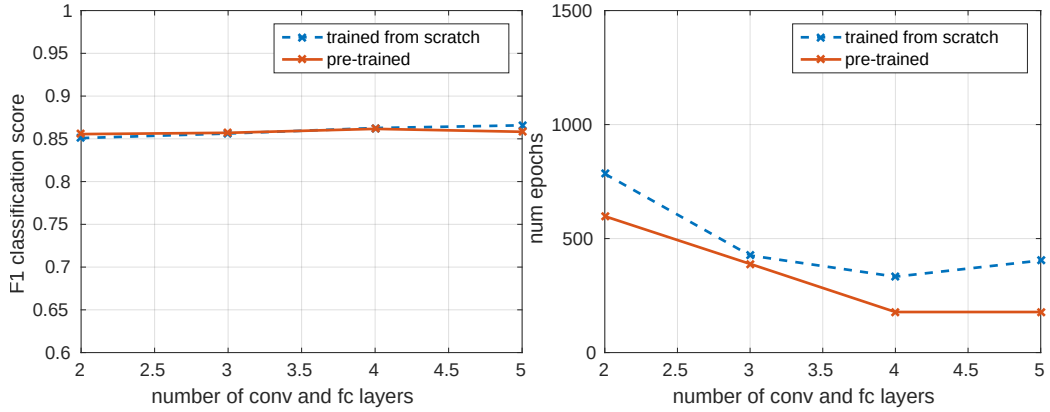
(c) Mining SWIR

Figure 5.10 – Transfer learning results: Mean percentage change in parameters from initialisation to convergence for each layer of the hyperspectral CNN, for the networks trained on 316 samples per class. For each parameter, the difference between the converged parameter value and the initialised parameter value was taken as a percentage of the initialised parameter value.



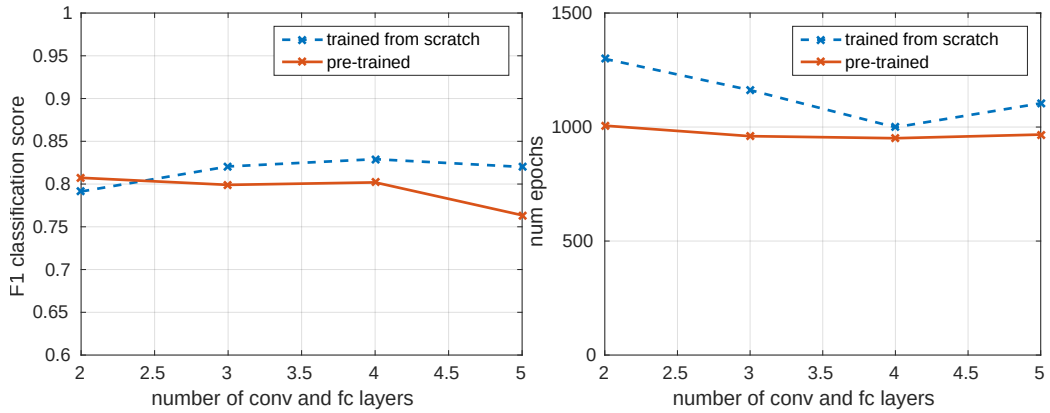
(a) Great Hall VNIR F1 classification score after convergence

(b) Great Hall VNIR number of epochs to reach 1% of convergence of classification score



(c) Great Hall SWIR F1 classification score after convergence

(d) Great Hall SWIR number of epochs to reach 1% of convergence of classification score



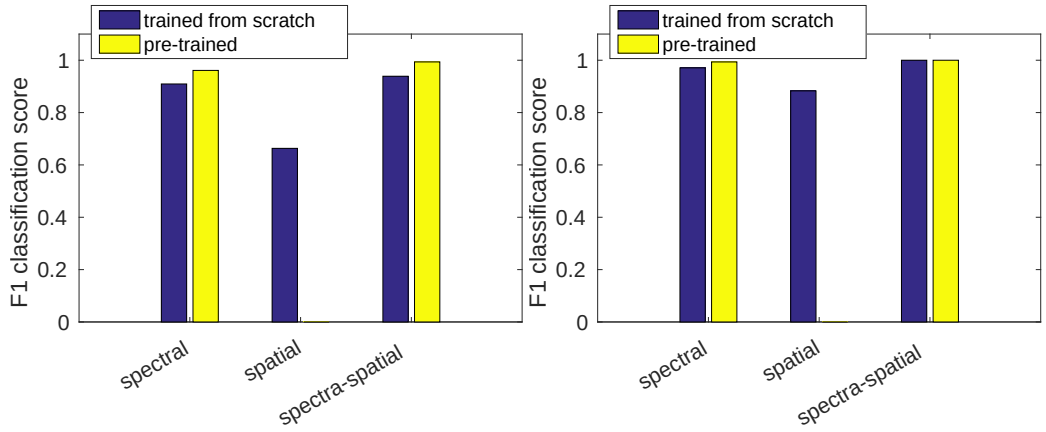
(e) Mining SWIR F1 classification score after convergence

(f) Mining SWIR number of epochs to reach 1% of convergence of classification score

Figure 5.11 – Transfer learning results: Comparison of pre-training and training from scratch for different architectures. For the Great Hall VNIR dataset, the CNN was pre-trained on the VNIR composite dataset (Table 5.5). For the Great Hall SWIR and Mining SWIR datasets, the CNN was pre-trained on the SWIR composite dataset (Table 5.5). The horizontal axis gives the number of convolutional (conv) layers and the number of fully connected (fc) layers used in the network.

The architecture results (Figure 5.11) show that the training set size used was sufficient enough for the classification error to be the same for pre-training and training from scratch, however, the time until convergence differed between the two initialisation approaches. The key observation was that the convergence time for the networks that were pre-trained was less sensitive to the architecture than the networks trained from scratch, where there was an optimal architecture for fastest convergence time. For all datasets, the results show that the fastest convergence time comes from training a network with four convolutional and four fully connected layers from scratch, and then increasing the network to five layers results in a decrease in convergence time. However, when using pre-training, five layers can be trained without noticeably increasing the convergence time.

To further show the generality of the performance gains from transfer learning regarding CNN architecture, a spectral-spatial network for hyperspectral classification was evaluated. The architecture was based on Yang et al. (2017), where there were separate spectral and spatial CNN branches that merged together with fully-connected layers before the softmax layer. This particular architecture works on the principle of learning the low-level spectral and spatial features individually and then learning the higher level dependencies between them, rather than simply learning filters which convolve in the spatial and spectral dimensions in the hypercube, where low-level dependencies are learnt. The pre-trained networks learnt in this chapter can be used to initialise the parameters of the spectral CNN branch of the spatial-spectral network with very little adjustment. The parameters of the spatial branch and the fully connected layers that merge the spectral and spatial CNNs were initialised randomly. Minor adjustments made to the architecture from Yang et al. (2017) were the substitution of the spectral branch with the pre-trained VNIR composite network architecture, and the increase in spatial filter size from 3×3 to 5×5 due to the higher spatial resolution of the field-based images. A spatial-spectral network was trained on a small number of samples collected from the Great Hall VNIR dataset (100 and 500 samples per class), and only allowed to optimise for 50 epochs. The results of pre-training the spectral CNN branch and training it from scratch were compared,



(a) 100 training samples per class. **(b)** 500 training samples per class.

Figure 5.12 – Transfer learning results: Comparison of pre-training and training from scratch for a spectral-spatial network. For the pre-trained case, the spectral CNN branch was pre-trained on the VNIR composite dataset (Table 5.5) and the spatial CNN branch and fully connected layers were randomly initialised. All networks were optimised for 50 epochs. Spatial-only and spectral-only results were included for comparison.

using the improved Xavier method (He et al., 2015) as before.

The results (Figure 5.12) show that with 100 training samples, there was a performance increase of about 5% in the spectral-spatial network when the spectral CNN branch was pre-trained. In the 500 training sample case, the classification score of learning all of the parameters from scratch was already too high, and so the pre-training barely improved it any more. Although the results of only using the spectral networks were very good, there was a small improvement in the results when the spatial texture information was added.

5.4.3 Evaluation of Spectral Relighting Augmentation

The spectral relighting augmentation strategy proposed to improve CNN performance when the number of training samples is limited was evaluated using three outdoor field-based datasets: Great Hall SWIR, mining timelapse and Gualtar steps. These datasets comprise a variety of classes and illumination conditions. For the evaluation, separate CNNs were trained on two sets of labelled training data. One set (referred

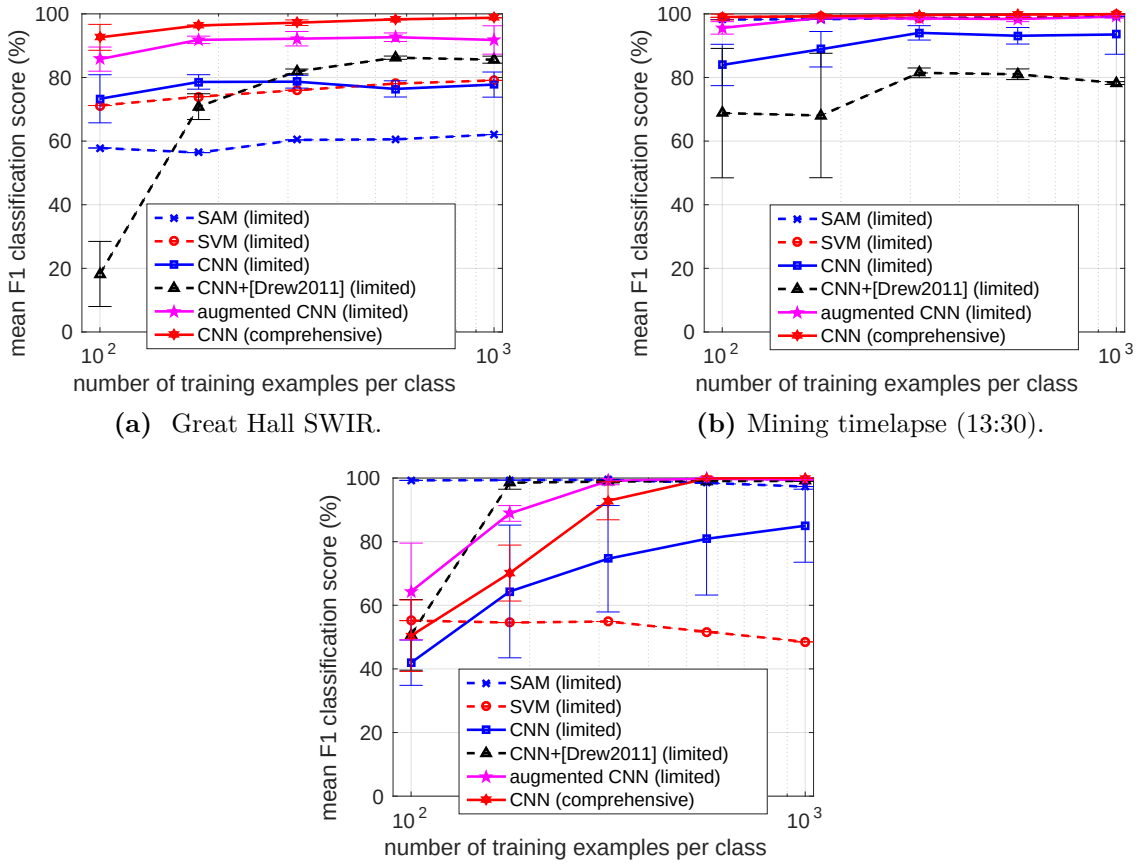
to as 'comprehensive') was collected from both sunlit and shaded regions of the image, and also had a large spatial coverage such that it best represented the variation in scene geometry and incident illumination. The second labelled training dataset (referred to as 'limited') was only collected in sunlit regions and the spatial coverage was small - limited to a patch (red squares in Figures 5.14a and 5.15b), such that there was a very poor representation of the scene geometry and incident illumination. Networks were trained on the limited dataset using the proposed data augmentation. Their performance was compared to networks that were trained on the limited and comprehensive datasets without any augmentation. The results of the networks trained on the comprehensive training data can be seen as roughly an upper bound for the results of training on the limited datasets. The performance of the CNN trained on the limited data, both augmented and not augmented, was compared against other classification approaches trained on the limited data including SAM (Yuhas et al., 1992), an SVM (Melgani and Bruzzone, 2004), and a CNN trained on spectra projected into an illumination invariant space using a log-chromaticity method (Drew and Salekdeh, 2011). SAM is insensitive to differences in brightness between the training data and the target data. For the mining timelapse dataset, networks were trained on the 13:30 image only.

To evaluate the classifiers, test set sizes of roughly 225000, 90000 and 230000 pixels spectra were chosen for the Great Hall, mining timelapse and Gualtar steps datasets respectively. The number of training examples was varied logarithmically between 100 and 1000 examples per class (5 increments: 100, 178, 316, 562, 1000). The lower bound was selected in order to demonstrate the performance of CNNs with spectral relighting augmentation under the condition of limited labelled training samples. The upper bound was selected to demonstrate its performance in scenarios where more labelled training data is available. Only the wall, brick and grass classes were used from the Gualtar steps dataset as not enough pixels from the cement class could be labelled. For the mining timelapse dataset, the two shale classes were combined into one class due to a lack of ground truth data for the distributions of the individual classes. Ten candidate terrestrial sunlight-diffuse skylight ratios were generated for

the augmentation such that each pixel spectra was relit ten times, with roughly half of those relightings being to shadow and half remaining sunlit but with different orientations. This expanded the training dataset to 11 times its original size. The validation dataset consisted of 50 examples per class, sampled with the same restrictions as the training data (i.e. limited or comprehensively). The mean and standard deviation for five randomly initialised networks was recorded.

Unless otherwise stated, all data passed to the CNN were in DN form and pre-processed using the zero-wavelength technique (described previously). So that one of the cases maintained convention with the remote sensing community, the Gualtar steps dataset was pre-processed with flat-field correction, such that it was normalised to apparent reflectance. Data used by the SVM and SAM classifiers for all datasets were normalised to reflectance using flat-field correction as this was found to be the best pre-processing method in preliminary experiments.

The results (Figure 5.13) show the impact of the spectral augmentation with different amounts of training data. A visualisation of some of the results in image form is also shown (Figures 5.14, 5.15, 5.16). The results from all three datasets showed that there was a clear advantage to augmenting the labelled data used to train the CNN. The CNN achieved better classification scores with augmentation in comparison to not using augmentation for all training set sizes and all datasets for the limited case, with most of the gap between training the CNN with limited and comprehensive training sets being bridged. Regions in shadow that were not included in the training set were misclassified by the CNN when trained on the non-augmented data, but mostly classified correctly by the CNN when trained on the augmented data. The results of training with the augmented data were also superior to the CNN trained on data that was projected into an illumination invariant space (Drew and Salekdeh, 2011). The SAM and SVM classifiers performed comparably to the augmented CNN on the mine face and the SAM classifier also performed comparably on the Gualtar steps datasets. Both of these datasets had quite discriminative classes. However, the SAM and SVM classifiers performed significantly worse on the Great Hall SWIR dataset which had more spectrally similar classes. The performance of the augmented and



(a) Great Hall SWIR.

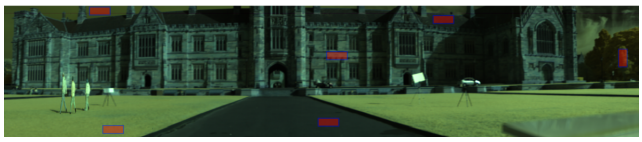
(b) Mining timelapse (13:30).

(c) Gualtar steps.

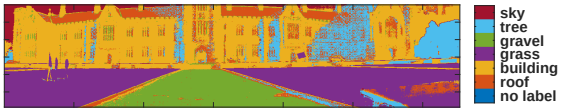
Figure 5.13 – Spectral relighting augmentation results: Results showing the benefit of training a CNN with the proposed data augmentation approach. The performance of the augmented CNN was compared with other approaches. 'Limited' indicates that a classifier was trained on localised, well-lit regions of the image, and 'comprehensive' indicates that a classifier was trained on the whole image (both well-lit and shadowed regions, see Figures 5.14a and 5.15b). The number of training samples was varied logarithmically between 100 and 1000, and the mean and standard deviation of the F1 classification score was reported for three different hyperspectral datasets.

non-augmented CNN as the amount of training samples was increased, was relatively consistent for the Great Hall SWIR and Gualtar steps datasets.

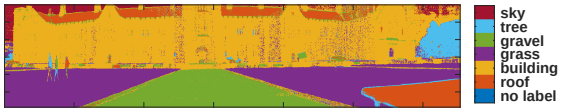
Figure 5.17 shows the improvement that augmented training data provided specifically in the shadowed and sunlit areas for two of the Great Hall classes. In both sun and shadow, most of the gap between training the CNN with limited and comprehensive data was bridged by the augmentation. Also, in contrast to the SVM,



(a) False colour image of Great Hall building with example of limited sunlit regions where labels are extracted.

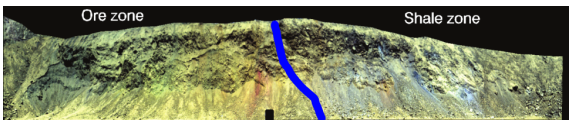


(b) No augmentation.

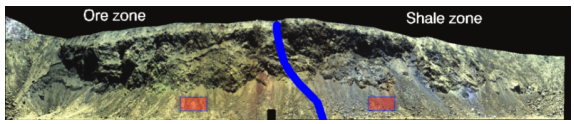


(c) With augmentation.

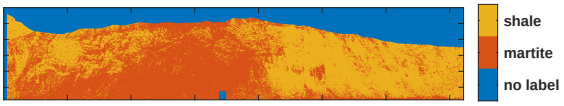
Figure 5.14 – Spectral relighting augmentation results: CNN classification of the Great Hall using 1000 training examples per class, drawn from the limited regions (red squares).



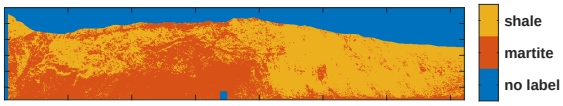
(a) Mine face image captured at 11:30.



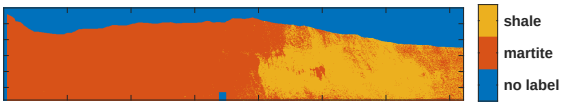
(b) Mine face image captured at 13:30.



(c) No augmentation 11:30.



(d) No augmentation 13:30.

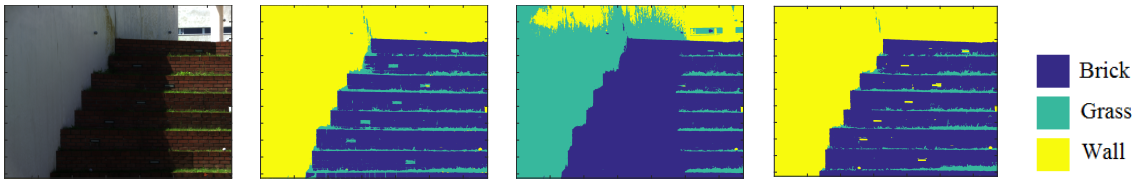


(e) With augmentation 11:30.

(f) With augmentation 13:30.

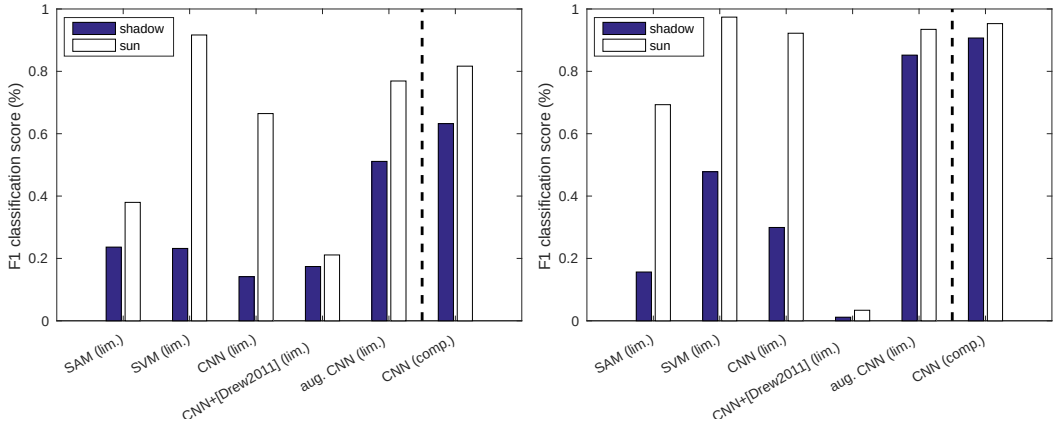
Figure 5.15 – Spectral relighting augmentation results: CNN classification of the mining timelapse using 100 training examples per class drawn from the limited regions (red squares). The same network was used to classify the image at both times of the day. The blue line is a rough class boundary identified by geologists.

which performed well on the sunlit regions for these two classes but poorly on the shadowed regions, the augmented CNNs performance in the shadow approached its performance in the sun. The CNN trained on the illumination invariant projection of the data (Drew and Salekdeh, 2011) had similar performance in sunlit and shadowed regions but was significantly worse than the augmented CNN, whose performance in the shadow surpassed all other methods trained on limited data.



(a) Gualtar steps colour composite. **(b)** Trained on sun and shadow data (comprehensive). **(c)** Trained on sun only data (limited). **(d)** Trained on sun only data (limited) with augmentation.

Figure 5.16 – Spectral relighting augmentation results: Improvement in CNN per-pixel classification of a hyperspectral image with a shadow when using spectral relighting augmentation. Training labels were extracted from either the entire image or only sunny areas. Results shown for the Gualtar steps dataset using 316 training examples per class.



(a) Roof class. **(b)** Building class.

Figure 5.17 – Spectral relighting augmentation results: F1 classification score of Great Hall classes in shadow and sunlight, for several different methods, trained on limited regions and the whole image (comprehensive), for training set size 100 samples per class.

Temporally, the CNN trained with augmentation exhibited greater invariance over the day in comparison to all other approaches, given by the reduced percentage of pixels that changed classification label between 11:30 and 13:30 (Figure 5.18).

Figure 5.19 shows an example of the valid candidate pairs of sun-shadow pixels from the same material that the image based algorithm automatically selects for the mining timelapse image. Many of the pairs fall along the shadow boundaries. Figure 5.20 shows a comparison between using the automatic approach of Section 5.3.2 and man-

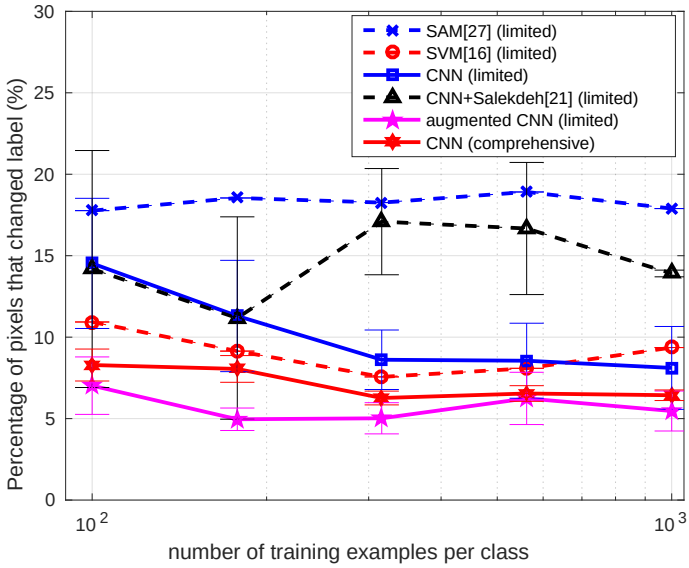


Figure 5.18 – Spectral relighting augmentation results: Percentage of pixels that changed classification label from the 11:30 to 13:30 mine face images for several different methods trained on localised regions and the whole image for different sized training sets.

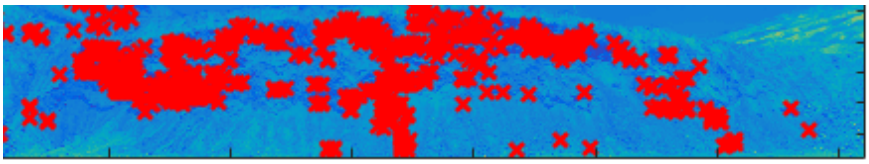


Figure 5.19 – Spectral relighting augmentation results: Example of the valid candidate pairs of sun-shadow pixels found on the mine face using the automated image based approach for determining the diffuse skylight-terrestrial sunlight ratio. Result shown for the 13:30 mining timelapse image.

ual selection of points for extracting the ratio between sunlit spectra and shadowed spectra used in 5.6 to estimate the terrestrial sunlight-diffuse skylight ratio for the Great Hall SWIR dataset. The manually selected points were carefully chosen to be on shadow boundaries of the same material and with the same geometric orientation. Hence, the ratio calculated using these points can be considered a gold-standard. The ratio extracted with the automatic approach had a similar shaped curve, but its mean value was lower than the gold-standard. Once the terrestrial sunlight-diffuse skylight ratio is found, it can be used for relighting. Figure 5.21 shows an example from the Gulatar steps image of how close a sunlit spectrum augmented to be in shadow using

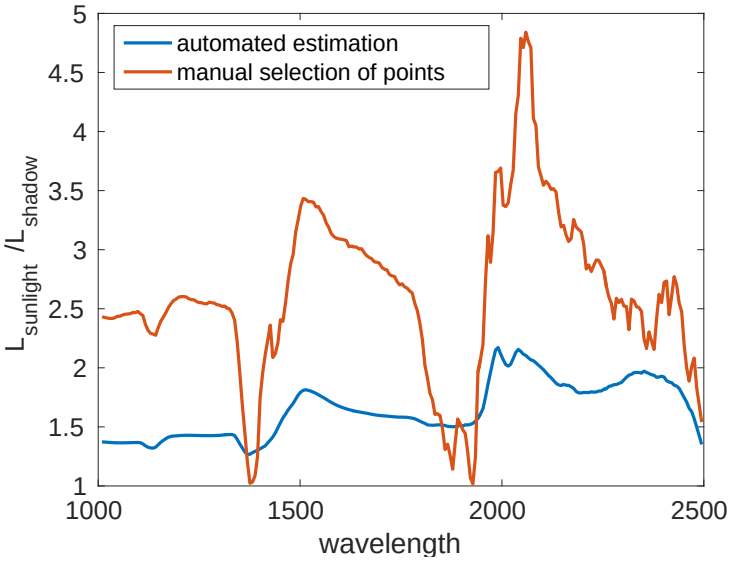


Figure 5.20 – Spectral relighting augmentation results: Comparison of automatic extraction of $L_A/L_{A'}$ in 5.6 using the approach in Section 5.3.2 with extraction using manually selected points from the image, for the Great Hall SWIR dataset.

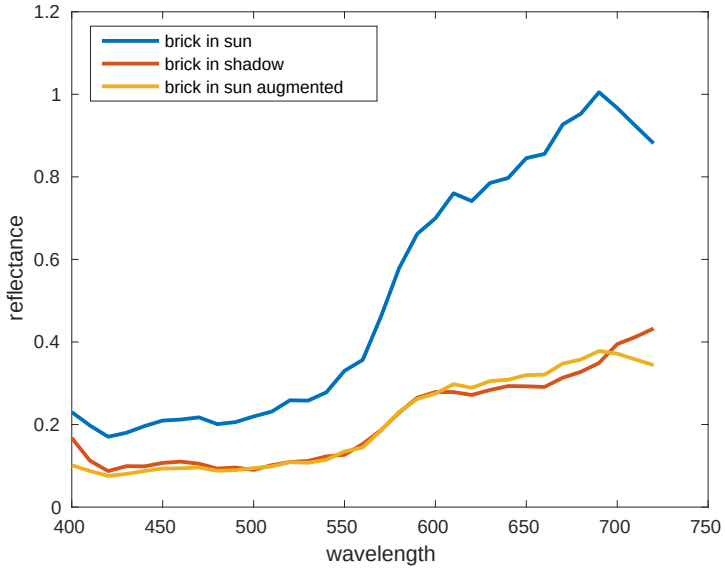


Figure 5.21 – Spectral relighting augmentation results: Example showing the accuracy of the spectral relighting augmentation for the Gualtar steps dataset.

relighting is to the spectrum in shadow.

Figure 5.22 shows that, for all three datasets, the spectral relighting augmentation of the training data improved the results regardless of whether the data was in the raw DN form or had been normalised to reflectance. It also shows that the spectral re-

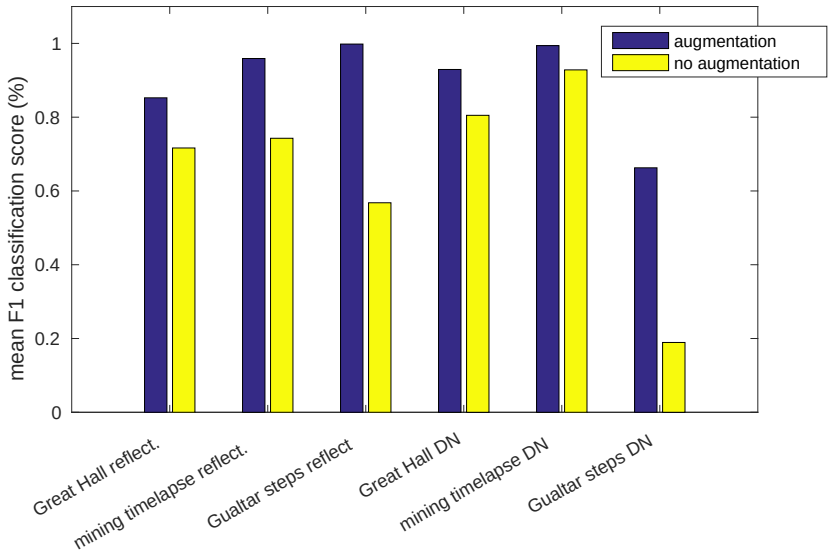


Figure 5.22 – Spectral relighting augmentation results: Comparison of training CNN with augmented and non-augmented spectra, in reflectance and DN form, for training set size 500 samples per class. The classifiers were trained on localised, well-lit (limited) regions of the image (see Figures 5.14a and 5.15b).

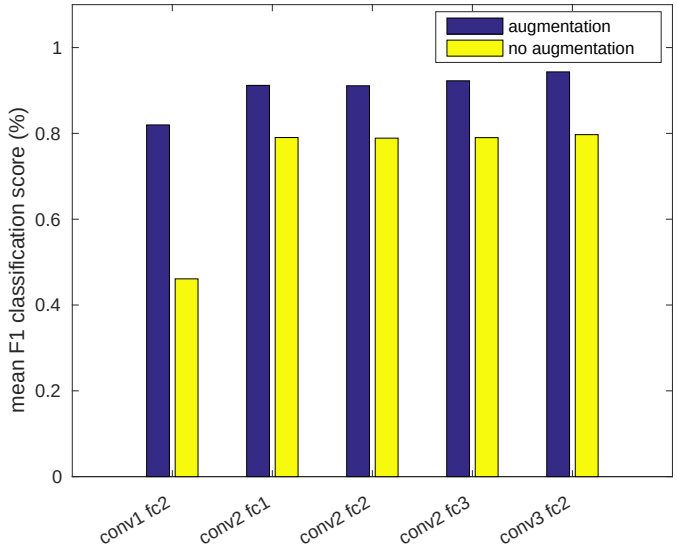


Figure 5.23 – Spectral relighting augmentation results: Comparison of training CNN with augmented and non-augmented spectra, with different architectures, for training set size 500 samples per class. The classifiers were trained on localised, well-lit (limited) regions of the image (see Figures 5.14a and 5.15b). *Conv* indicates the number of convolution layers and *fc* indicates the number of fully connected layers. The Great Hall SWIR dataset was used. Zero-wavelength pre-processing was used.

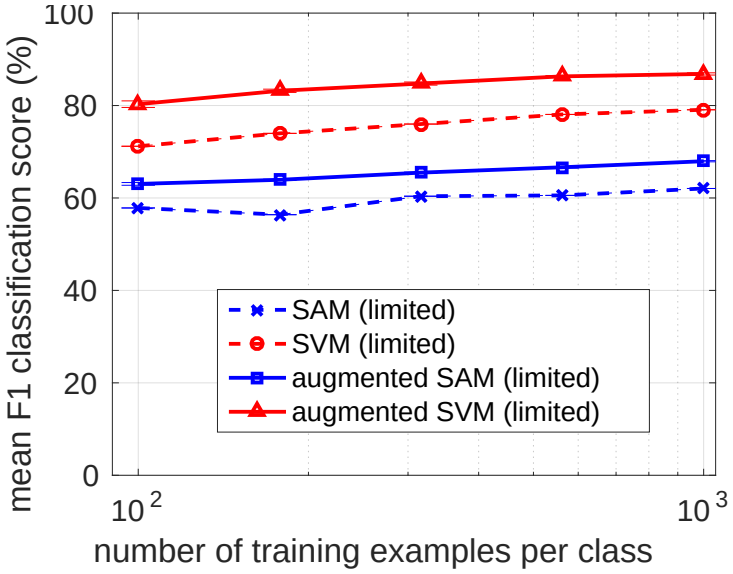


Figure 5.24 – Spectral relighting augmentation results: Results showing the benefit of training other classifiers (SVM and SAM) with the proposed data augmentation. The classifiers were trained on localised, well-lit (limited) regions of the image (see Figures 5.14a and 5.15b). The number of training samples was varied logarithmically between 100 and 1000, taken from the limited regions, and the mean and standard deviation of the F1 classification score was reported for the Great Hall dataset. The spectra was normalised to reflectance.

lighting augmentation improved the CNNs performance for several different numbers of convolutional and fully connected layers used (Figure 5.23). That is, the success of the relighting augmentation was not dependent on the method of pre-processing or the architecture. Figure 5.24 shows that the improvement in classification performance was not limited to CNNs. The classification accuracy of SVM and SAM classifiers also improved when their limited training datasets were augmented, demonstrating that the relighting approach is agnostic to the classifier.

The improvement in a spectral-spatial CNN classifier when spectral relighting augmentation is used, was evaluated. The same spatial-spectral network architecture used to evaluate transfer learning (Figure 5.12) was used, with the spectral branch trained via spectral relighting augmentation. The spatial branch was not trained with any augmented data. The evaluation was done with the Great Hall SWIR dataset, chosen because the different classes have a distinguishable texture, making it a suitable dataset to incorporate spatial information. A limited training set size of 100

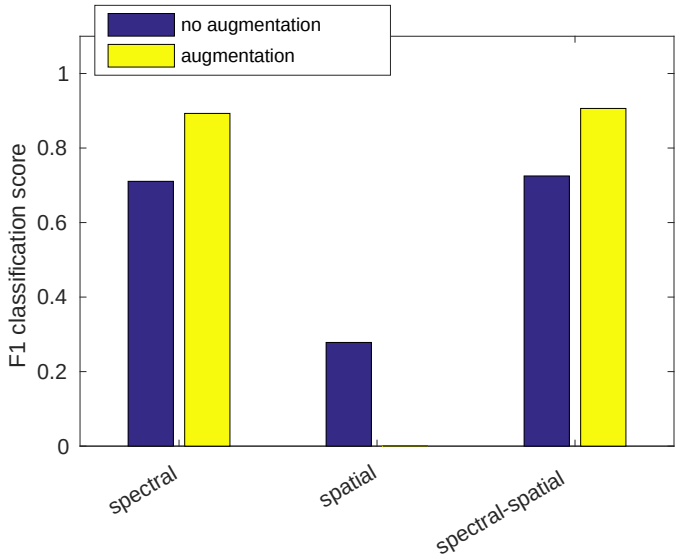
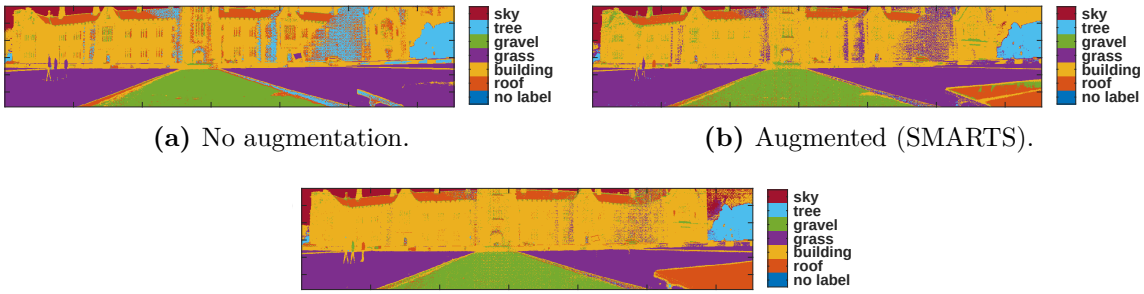


Figure 5.25 – Spectral relighting augmentation results: Comparison of relighting augmentation versus no augmentation for a spectral-spatial network, on the Great Hall SWIR dataset. For the augmented case, the spectral CNN branch was trained with spectral relighting augmentation and the spatial CNN branch and fully connected layers were trained normally. All networks were optimised for 200 epochs. Spatial-only and spectral-only results were included for comparison.

samples collected in a fully illuminated, localised region was used for training. The CNN was allowed to optimise for 200 epochs. The results compare using spectral relighting augmentation on the spectral branch of the network and not using any augmentation. All learnable parameters were initialised from scratch.

The results (Figure 5.25) show that augmentation improved the performance of a spectral-spatial architecture, when the architecture was set up as similar to Yang et al. (2017). There was only a minor increase in performance between the spectral and spectral-spatial networks, for both the augmentation case and the non-augmentation case. The increase in performance when using augmentation with the spatial-spectral network was very similar to the increase in performance when using augmentation with the spectral-only network.

The final set of spectral relighting augmentation results compared the CNN performance when the spectra was augmented using the terrestrial sunlight-diffuse skylight ratio found by randomly sampling the parameters of the SMARTS atmospheric mod-



(a) No augmentation.

(b) Augmented (SMARTS).

(c) Augmented (image-based).

Figure 5.26 – Spectral relighting augmentation results: Comparison between augmenting the spectra with different methods for extracting the terrestrial sunlight-diffuse skylight ratio. The augmented (SMARTS) approach used a ratio generated by randomly sampling the parameters of the SMARTS atmospheric modeller (as in Section 4.2) and the augmented (image-based) approach used throughout Section 5.3 estimates the ratio from the image. The no augmentation case is also displayed. CNN classification of the Great Hall using 500 training examples per class, drawn from the limited regions (red squares). Zero-wavelength pre-processing was used.

Table 5.6 – Spectral relighting augmentation results: Quantitative form of the results in Figure 5.26, with the mean and standard deviation reported for five repetitions.

Method	Roof	Building	Grass	Path	Tree	Sky	Mean
No augmentation	54.76±8.90	87.12±4.10	97.70±3.19	92.54±6.16	63.37±18.54	82.07±16.00	79.59±4.38
Augmented (SMARTS)	62.14±13.67	88.71±4.80	89.19±9.12	81.16±11.67	85.97±19.90	90.26±3.73	82.91±5.72
Augmented (image-based)	**82.48±3.01**	**94.76±1.78**	**94.73±3.83**	**91.98±1.90**	**97.78±0.54**	**93.54±4.57**	**92.55±1.55**

eller, against estimating the ratio from the image using the approach proposed in Section 5.3.2. The results (Figure 5.26 and Table 5.6) show that the CNN using the atmospheric modeller approach (used in Section 4.2) incorrectly classified the shaded regions in the image, and the overall classification accuracy was lower. The CNN which used the proposed image based approach to obtain the terrestrial sunlight-diffuse skylight ratio performed much better in the shaded regions of the image.

5.4.4 Analysis of the Learnt Filters

To attempt to get a better intuition for what the filters in these networks are learning from the spectral data, ten of the first layer filters were plotted (Figure 5.27) from the

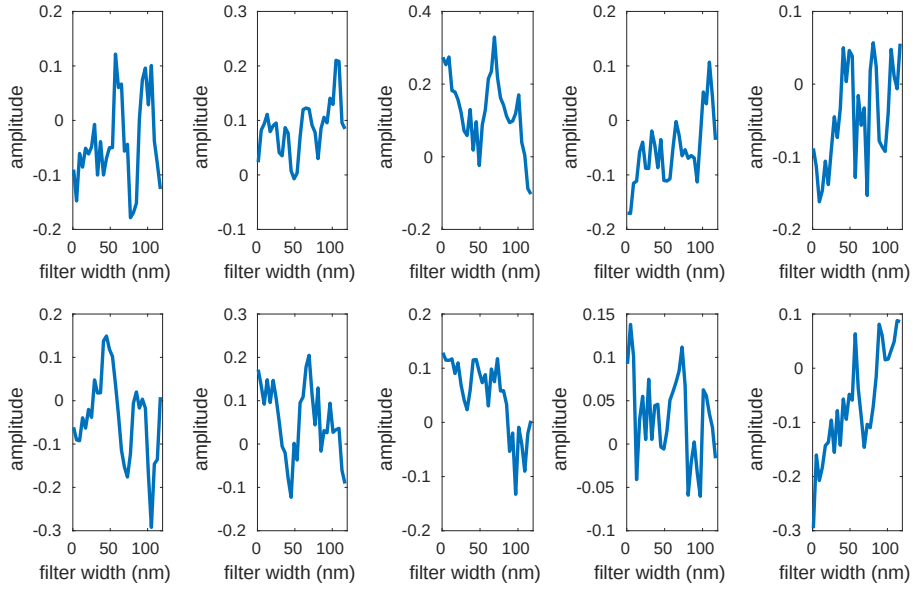


Figure 5.27 – Filter analysis results: Visualisation of the first 10 filters of the first layer of the network fine-tuned on the Great Hall VNIR dataset. The top row is the first five and the bottom row is the last five.

VNIR composite CNN after it was fine-tuned on the Great Hall VNIR dataset for the transfer learning experiment Section 5.4.2. Each filter attempts to capture a unique characteristic in the spectrum. It is difficult to gauge from the filters alone what sort of features they are capturing. Thus, an example spectra from each of the six classes in the Great Hall VNIR dataset was plotted (Figure 5.28) along with the activation of the ReLU units associated with the filters in Figure 5.27 when convolved over each spectrum. In a similar set of plots, the third convolutional layer activations for the Great Hall VNIR were also shown (Figure 5.29).

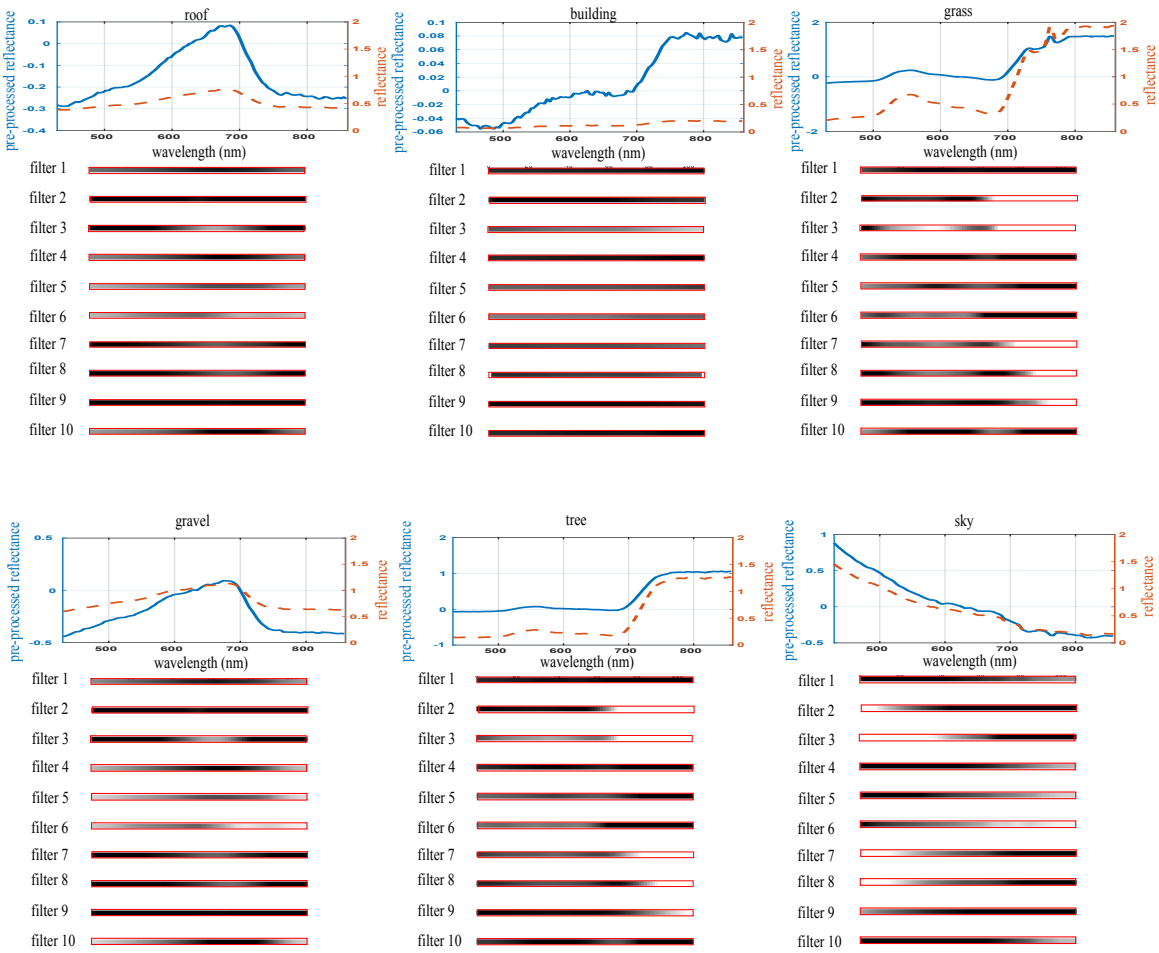


Figure 5.28 – Filter analysis results: Visualisation of where in the spectrum is activated by the first 10 filters of the first convolutional layer of the network fine-tuned on the Great Hall VNIR dataset. Each greyscale bar corresponds to a filter, and is aligned with the spectrum above it to show how the reflectance at each wavelength activates each filter. The whiter the region of the greyscale bar, the more that part of the spectrum is activating that particular filter. The original reflectance spectrum is dashed and the pre-processed spectrum is solid. Note that there is no padding on the edges of the spectrum, so the filtering does not extend to the end of the spectrum and does start at the beginning.

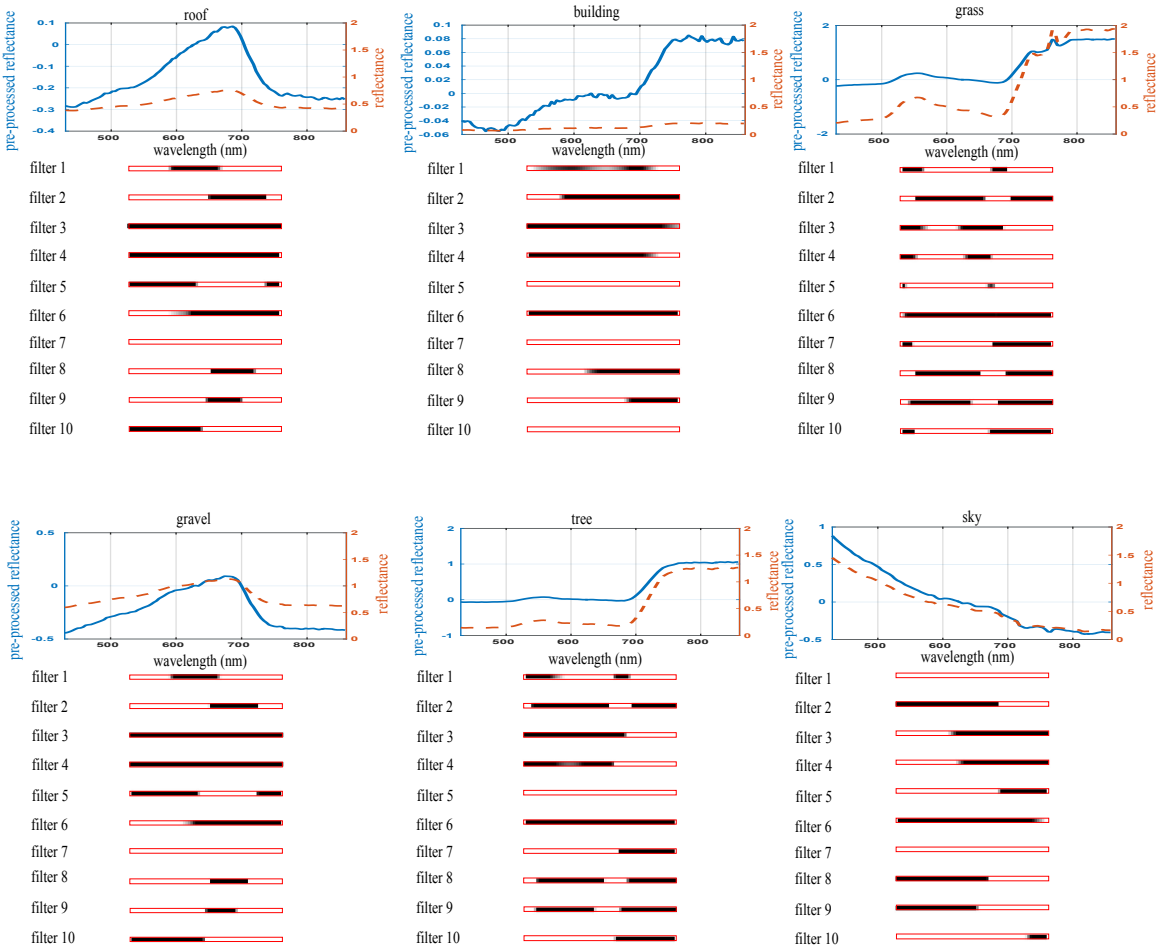


Figure 5.29 – Filter analysis results: Visualisation of where in the spectrum is activated by the first 10 filters of the third convolutional layer of the network fine-tuned on the Great Hall VNIR dataset. Each greyscale bar corresponds to a filter, and is aligned with the spectrum above it to show how the reflectance at each wavelength activates each filter. The whiter the region of the greyscale bar, the more that part of the spectrum is activating that particular filter. The original reflectance spectrum is dashed and the pre-processed spectrum is solid. Note that there is no padding on the edges of the spectrum, so the filtering does not extend to the end of the spectrum and does start at the beginning.

5.5 Discussion

The following section provides an analysis of the results of Section 5.4.

5.5.1 Evaluation of Transfer Learning

The kernels in the first convolutional layers filter the spectrum. Hence, the choice of width of the filters in the first layer is important as a bigger filter provides more context, but if it is too big it can drown out important absorption features. Ideally, it must be of a sufficient size to capture the most significant features. The results (Figure 5.8) suggest that the best filter size is 120 nm as it had the shortest convergence time and similar classification error to the others. Vibrational absorption features, which tend to be the most diagnostic features in the spectrum, typically occur over a shorter width relative to electronic absorption features (Clark, 1999). A 120 nm filter is a good size for capturing such features. For example, in the mining dataset kaolinite spectra has an aluminium hydroxide absorption feature that occurs at about 2.2 µm, with a width that roughly matches 120 nm. If the filters were smaller than this, then the broader electronic features would be captured poorly, and if they were wider then perhaps the critical vibrational features would not be captured as well.

Regarding the wavelength interval results (Figure 5.8), it is important to use a resolution which balances the issues associated with up-scaling and down-scaling. For example, the Salinas dataset, which was used to train the SWIR pre-trained network, has a relatively coarse resolution (10 nm). If the selected wavelength interval is too small (e.g. 2 nm), then errors could be introduced when the coarser data used for pre-training is up-sampled to much finer resolutions. This was seen in the convergence time results for the mining dataset. Conversely, if too big, as is the case with 16 nm, then the performance drops, as much of the fine spectral detail in the data is lost once it is interpolated, as seen in the convergence time results for the Great Hall SWIR and mining datasets. These considerations are also applicable to the datasets being used to do the fine-tuning. For this reason, the mid-length resolutions of $4 - 8$

nm are determined to be the safest resolutions to use for the data to pre-train and fine-tune the network.

The improvement in convergence time that was observed when pre-training rather than initialising randomly (Figure 5.9) is related to the change (as a percentage) in the parameters in each of the layers from their initialisation until convergence (Figure 5.10). As the networks being trained from scratch were initialised randomly, there was a large change in the parameters in all of the layers before they converge. The networks that were pre-trained and then fine-tuned experienced a large amount of change in the final layer. This was expected as the final layer was randomly initialised due to the fine-tuning dataset having a different number of classes to the pre-training dataset, and hence a different number of output neurons. However, all of the other layers, which were initialised using the parameters of the pre-trained network, barely changed at all over the optimisation. Hence, the optimisation required less time to reach convergence.

The fact that the pre-trained filters did not change much when being used for a completely new dataset, suggests that they are quite generic and pick up absorption characteristics in a spectrum that are not specific to any one class. This is useful for transfer learning because it means that these features can be learnt once, elsewhere, and then transferred to new scenarios as required. For example, the mining dataset contains completely different classes to the classes of spectra used to pre-train the CNN, yet the filters were still transferable because the pre-training produced an improvement in the results. Interestingly, filters learnt on the SWIR pre-training dataset, which has no whaleback shale samples, have transferred to the mining problem and are detecting the subtle diagnostic absorption features that discriminate the different classes of shale. With that said, there was a larger change in the filter parameters in each layer for the mining dataset than the Great Hall datasets (Figure 5.10). This was because of the very different composition of the materials in the mining dataset compared to the materials in the pre-training dataset. However, the change was still small with respect to the change in parameters when the filters were learnt from scratch. It is also very interesting to see that the filters learnt from

the airborne data were transferable to classification of data acquired from field-based platforms. This was unexpected because the airborne spectra were integrated across a much larger area of ground and therefore are likely to be spectral mixtures. Spectra collected from ground-based platforms are more pure as the sensors scan the scene with a finer spatial resolution. It is suspected that the filters learnt by the CNN transfer well between low spatial resolution airborne data and high spatial resolution ground-based data because they are learnt on localised parts of the spectrum rather than the entire spectrum. If a fully-connected network was used where filters were learnt from the entire spectrum and there were no convolutions, then it is presumed that the filters would be less useful when fine-tuning on ground-based data.

The architecture experimental results (Figure 5.11) showed that it was possible to train deeper networks (e.g. five layers instead of four) without the associated increase in convergence time when using transfer learning. This is often an advantage, as was the case for the Great Hall VNIR dataset as the five layer architecture had a slightly higher classification score than the four layer architecture. Nevertheless, a disadvantage of transfer learning is that the architecture of the network being trained on the target data is constrained to be the same as the one that was pre-trained for all of the layers to be transferred. However, this was not a limiting constraint, as seen in the results, which show that the architecture choice had minimal impact on the classification score. It was also shown that transfer learning improved the results when using a spectral-spatial architecture (Figure 5.12). The results were consistent with those in Figure 5.9, as the biggest benefit to using transfer learning came from performance increases when there was a small number of training samples as well as not having to train for as long because the network was being fine-tuned rather than trained from scratch. Thus, good results (99%+) were achieved with only 100 training samples per class and 50 epochs. Interestingly, for the 100 training sample case, the performance improvements from transfer learning for the spectral network were bigger than the performance improvements of adding the spatial component to the network architecture. This suggests that the spectral knowledge transferred from the airborne composite dataset was of greater value for the ground-based classification

task than the additional ground-based spatial information.

5.5.2 Evaluation of Spectral Relighting Augmentation

All of the results indicated that the spectral relighting augmentation was effective at improving the training of CNNs with limited amounts of data. In Figure 5.13 and Figure 5.17, most of the gap between training the CNN with limited and comprehensive data was bridged by augmenting the limited training data. These results suggest that the augmentation allowed the limited training data to capture the variability as if pixels from all over the image were labelled - covering areas with different geometry (there is improved performance for the classes in sunlight in Figure 5.17) and occlusions (e.g. Figure 5.21 and the improved performance for the classes in shadow in Figure 5.17). Despite this, in Figure 5.14, there was still some sporadic misclassification of pixels in the shadow. It is suspected that this is potentially because of increased prevalence of indirect illumination from nearby parts of the wall. For simplicity, indirect illumination was not incorporated into the outdoor model (E_{ind} in equation 2.1 was set to zero) and hence the spectra has not been augmented to account for this type of variability. However, the impact of the indirect illumination on the image was minimal in comparison to the impact of the shadows. Also, despite the diffuse assumption made in the model (2.2), the results show that this is not a particularly strong constraint on the data, with each dataset possessing non-Lambertian materials.

The consistent performance of the CNNs as the number of training samples varied was expected (Figure 5.13). This is because the training examples were being sampled from such a small region of the image such that increasing them allowed the CNN to capture very little extra variability. Since the augmented CNN simulated the missing variability there was little dependence of the classifier's performance on the number of training examples.

The SAM classifiers poor performance on the Great Hall dataset (Figure 5.13a) suggests that its use may be limited to scenarios with very spectrally distinguishable

classes (such as those in the Gualtar steps dataset). Although SAM is insensitive to brightness, it failed to correctly match shadowed spectra to their sunlit equivalents in the reference library (comprised of the limited training set). This is due to the change in shape that a shadowed spectra undergoes, resulting in large angles between shadowed and sunlit spectra from the same class. This may not have been the case for the Gualtar steps and Mining timelapse datasets, because even though there were large angles between the shadowed and sunlit spectra, they were spectrally distinguishable enough that corresponding classes maintained the smallest angle of separation. The proposed CNN was trained to be invariant to shadows and is able to learn more complex mappings due to its highly non-linear architecture, and hence could be trained to encompass the very subtle differences in spectra that are not easily distinguishable.

The method for extracting points in order to obtain the ratio between sunlit spectra and shaded spectra required for estimating the terrestrial sunlight-diffuse skylight ratio (Section 5.3.2), was compared against a gold-standard ratio in Figure 5.20 acquired by manually selecting points along shadow boundaries. Whilst the curve shape of the automatically extracted ratio matched the manually selected one, the overall mean was slightly lower. This could be because of the outlier pairs of points that contain materials under the same lighting conditions (rather than one under shadow and one in sunlight). The mean ratio of these pairs across all wavelengths is closer to unity and hence the overall ratio was pulled down. However, this did not have a significant impact on the performance of the augmented CNN because the actual terrestrial sunlight-diffuse skylight ratio is a scalar multiple of the extracted ratio, and so the extracted ratio gets multiplied by a wavelength-independent constant in the relighting process anyway, reducing the importance of the overall mean. Hence, the most important thing is that the shape of the curve is accurate, which in Figure 5.20 it is.

The image based method ratio extraction method relies on pixels from a pseudo RGB image projected into an illumination invariant space in order to determine which pairs contain pixels from the same material. It is possible that certain materials that are distinguishable in the hyperspectral space may look similar in the pseudo RGB image

and hence will be mistaken as belonging to the same class. Thus, there would be some invalid pairs left over after the refinement process. Hence, it is important that the majority of pairs are valid so that the impact of these invalid pairs is smoothed when the average is taken. If there are too many invalid pairs, it is possible to project points from the entire spectrum instead of the pseudo RGB image onto the invariant axis. This would ensure better class discrimination and fewer invalid pairs.

The image based approach for estimating the terrestrial sunlight-diffuse skylight ratio was compared against an atmospheric modeller approach (Figure 5.26 and Table 5.6). The better performance of the CNN with the image based approach indicates that it was able to model the appearance of the spectra in shadow better than if the parameters of the atmospheric modeller were randomly sampled. This result was expected given that the modeller requires many parameters to be sampled, generating many possible ratios with the hope that the correct one appears, whereas the image based approach only requires the sampling of a geometric-based scaling parameter. Whilst the atmospheric modeller approach was not effective for use with the spectral relighting augmentation for the CNN classifier, it was effective in Section 4.2 for augmenting the input layer of the RSA-SAE. This is because, unlike the CNN approach, it did not rely on the correct ratio being generated. It is possible to also use the image based approach to do the relighting in Section 4.2 as well, however, this would limit the invariance of the RSA-SAE to scenarios with the terrestrial sunlight-diffuse skylight ratio that was used for training. It is more beneficial to use the atmospheric modeller for the simulating many possible ratios for relighting, as this makes the RSA-SAE work for many different scenarios, including a temporally variable dataset where the atmospheric conditions change.

Figure 5.18 shows that the CNNs learnt are not only invariant to illumination variability across the image, but can also be invariant across different capture times. The images in figure 5.15a and 5.15b undergo a change in the illumination conditions due to shadows moving in the image as the day progressed. The relighting augmentation simulated how the pixels would appear in shadow with many different possible geometries, thus maintaining invariance even when the shadows moved in the image.

However, if the temporal change in the illumination was related to a change in the atmospheric conditions, then the CNN would not be invariant and would have to be re-trained. This is because the relighting augmentation is done using the terrestrial sunlight-diffuse skylight ratio extracted from the training image - the consequence being that the classifier does not work for fundamentally different ratios. If the atmospheric conditions change, then so do the illuminants themselves, and hence there is a new ratio. It is highly likely that across multiple days the atmospheric conditions will change and hence a new terrestrial sunlight-diffuse skylight ratio will have to be estimated in order to do the relighting and train the CNN. As mentioned before, this is not a problem for the RSA-SAE.

The results show the generality of the approach in terms of reflectance normalisation (Figure 5.22), architecture (Figures 5.23, 5.25) and classifier (Figure 5.24). This is quite intuitive, as it is expected that whatever pre-processing, classifier or architecture is being used, expanding the variability of the training data will always improve the results.

5.5.3 Analysis of the Learnt Filters

Figure 5.28 reveals what features of each class spectrum are activating the filters in Figure 5.27. In the first layer, very simple characteristics like positive and negative gradients are activating the filters. The convolving filters are not wavelength dependent, which can be seen by the way that some of them are active for gradients at the beginning of the spectrum of some classes and those same filters are active for gradients occurring in the latter of the spectrum of other classes. An example of this is filter 7, which is activated by features occurring at the shorter wavelengths of the tree spectrum and the longer wavelengths of the sky spectrum. This shows that the weight sharing that occurs in convolutional neural networks is justified for hyperspectral data, as many of the absorption features that the network attempts to capture are not wavelength dependent. The filters in the third layer (Figure 5.29) are activated by more complex shapes in the spectrum and are more class specific than

the first layer filters. This is expected as they are actually non-linear combinations of filters from the first two convolutional layers. These filters are also detecting much finer features in the spectrum. For example, despite the spectrum of grass and tree being very similar, the activation response of applying the third filter in the third layer to both spectra is different. The small peak that occurs at around 550 nm is marginally more prominent in the grass spectrum, and this causes filter 3 to activate at this region. However, filter 3 does not activate at the peak in the tree spectrum. Hence, this filter has learnt to differentiate these two very similar classes.

5.6 Summary

CNNs are achieving state-of-the-art classification results on many benchmark computer vision datasets. They require a significant amount of labelled training data, which is not as readily available when they are being used for hyperspectral applications due to the difficulties in annotating the data.

In this chapter, a method was proposed to train CNNs for hyperspectral applications where there is a limited amount of training data available. The method utilised data augmentation based on spectral relighting to expand the variability that the labelled data captured, and transfer learning to leverage knowledge from a composite of other well-labelled datasets. These approaches made it possible to significantly improve the performance of CNN classifiers and also reduce the time required to train them. Links were also made between the physical spectra and the filters learnt in the layers of the CNN to get a more intuitive understanding of what the CNN is learning.

In Chapter 6, the methods proposed in Chapter 4 and Chapter 5 are utilised in a single pipeline to show how a scene can be mapped with reasonable accuracy given a hyperspectral image with no initial annotated data.

Chapter 6

Case Study: A Unified Pipeline

This chapter presents an example of how to practically use several of the elements proposed in this thesis in a unified pipeline. The problem scenario addressed is the clustering of a hyperspectral image without any prior annotations. Usually labelled data is available to train a classifier, and Chapter 5 proposed techniques for dealing with scenarios where the amount of labelled data is limited. However, in some cases, it is desirable to map a scene when there is no training data available. This could be an initial undertaking that preludes the collection of data and a supervised classification process.

This pipeline will demonstrate how some of the contributions proposed in this thesis can be used together to solve a difficult task. It will utilise techniques from Chapter 4 and Chapter 5. The problem scenario is described in more detail in Section 6.1, followed by the pipeline in Section 6.2 and some experimental results in Section 6.3.

6.1 Problem Definition

A pipeline is constructed in order to cluster each pixel of a hyperspectral image of an open-cut mine when there is no labelled training data or knowledge of classes available. In this problem, the only prior knowledge available is the number of classes

in the scene. Everything else, including the names and whereabouts of the classes in the image, is unknown. Hence, if using a spectral library-based approach, many classes will have to be searched through in order to estimate which ones are in the scene, and even then, it is highly unlikely that the reference data in the library exhibits the variability of the spectra in the image.

The mining timelapse data (Section 3.1) is used as the label-less hyperspectral image. Labelled data for this image does exist, however, it will only be used to evaluate the results. The VNIR image has 220 channels in the range $401 - 970$ nm, with the geological classes of Martite and Shale and some sky in the background (Figure 6.1). Hence, there are three predominant classes in the image. As evidenced in the spectra of Figure 6.1, separating the sky from the geological classes is a trivial task, although, separating the two geological classes is difficult due to their similarity coupled with illumination variability across the scene. Usually these classes are separated using their reflectance in the SWIR (Murphy et al., 2012), but it is shown that this pipeline is capable of separating them in the VNIR.

In real-world applications for autonomous mining, hyperspectral images of open-cut mines can reveal the material distribution on the surface. Due to the rocks uniformity of colour and texture, it can be difficult to label data from these images via visual inspection for training learning algorithms. Experts would be required to analyse spectrometer data samples in order to confidently assign class labels. In autonomous mineral exploration, it is highly desirable to bypass this entire process. For example, a robot exploring Mars might be looking for minerals, and because there is no one on-site to analyse the samples it is difficult to annotate the data. Similarly, when sensing in inaccessible or dangerous locations such as unpredictable volcanic terrain, it is hazardous to annotate data or collect samples to be later studied. Thus, it is beneficial to be able to estimate the spatial distributions of materials in a hyperspectral image without requiring labelled data.

The spectral data in the image of the mining scene has variability due to the interaction of incident light with its complex topology. There are many shaded regions throughout the image. Classifying or clustering spectra with significant variability is

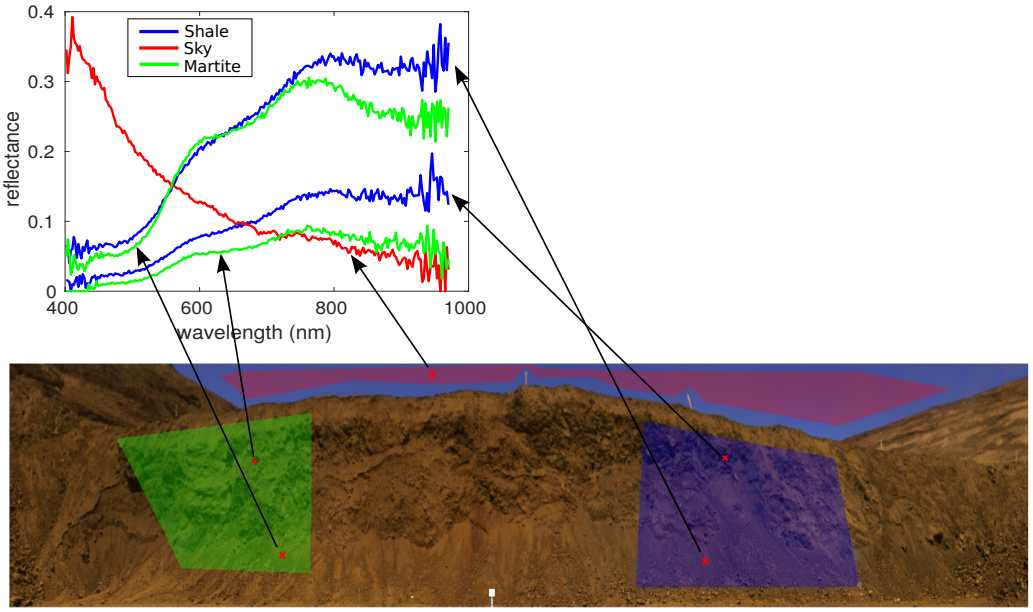


Figure 6.1 – The mining 11:30 timelapse image with ground truth areas highlighted. An example spectra from each class is shown. The similarity of the shale and martite classes coupled with the within-class variability of the spectra across the scene makes this is a difficult clustering problem.

difficult. Even if there was prior knowledge of which classes were in the scene, the variability would still make it difficult to accurately classify the hyperspectral image (e.g. using SAM with a reference library).

Note that in the mining timelapse dataset there are two similar variants of shale. However, for the experiments in this chapter, these two classes are merged together to make one shale class, due to the lack of ground truth data for validation.

6.2 Pipeline

There are two main parts to this pipeline. Firstly, an unsupervised process is used to categorise the data and extract some class representative points. This process only extracts points of high confidence, and hence does not capture the variability of the class. But it removes the need for any spectral libraries or reference data,

and doesn't require knowledge of the labels of the classes in the scene. Once high confidence class representatives have been extracted from the scene, a self-supervised process is used to predict the class association of all pixels in the image. Within this process, the variability in the data, due to illumination and the scene's complex topology, is accounted for. This is essentially clustering the data a second time, but this time, more accurately. Note that whilst supervised algorithms are used, the entire process requires no labelled data. The data to train the supervised algorithm comes from the unsupervised process (hence self-supervised).

The result of the pipeline is a function which can produce a per-pixel thematic map of the distribution of the classes in a hyperspectral image. The pipeline is summarised in Figure 6.2.

6.2.1 Unsupervised Process

There are two DN images in the hyperspectral mining timelapse dataset, one captured at 11:30 and one captured at 13:30. An RSA-SAE network (Section 4.2) is trained on the hyperspectral data from the 11:30 image. Note that it could be trained on the 13:30 image, or data from both images, but the 11:30 image is selected to show that the pipeline generalises to new images captured with different illumination conditions. In doing this, a mapping is learnt from the high dimensional hyperspectral reflectance image to a low dimensional, illumination invariant image. For the mining dataset, the same process was used as in Section 4.3 to train the RSA-SAE, with the same network architecture and parameters used. Hence, a 30 dimensional illumination invariant representation is learnt.

The data in the illumination invariant feature space is clustered using k-means, searching for three clusters (it is assumed that prior knowledge is available that there are three classes in the image). Then, the dataset of high confidence class representatives for training the classifier is built by automatically extracting the 200 points closest to each of the three cluster centroids. The distance is measured using the Euclidean distance. These 200 points have the highest confidence of belonging to each of the

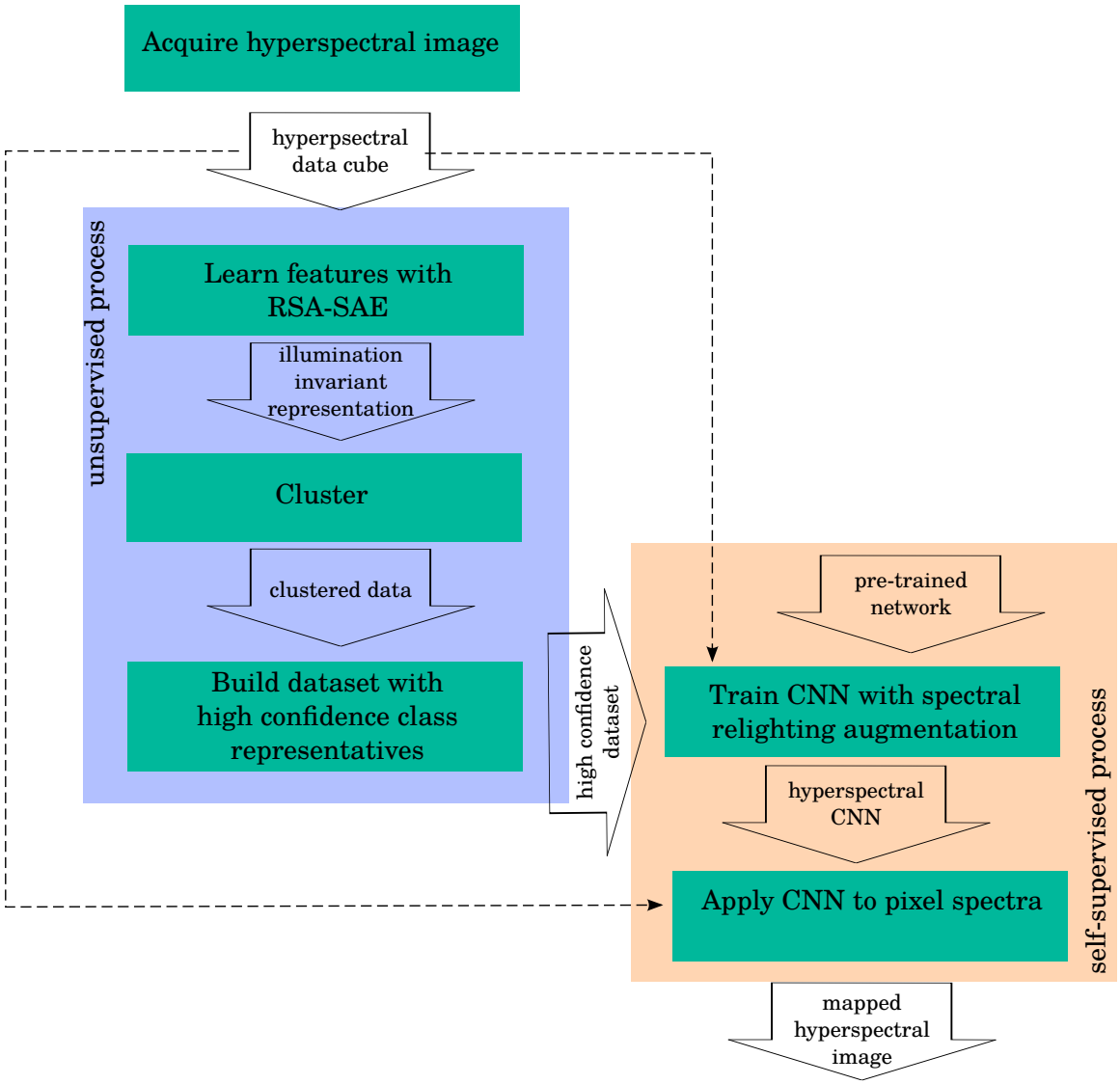


Figure 6.2 – A flowchart summarising the pipeline for clustering the pixels of a hyperspectral image without any prior labelled data.

classes in the scene. Hence, the dataset has 600 points in total.

6.2.2 Self-Supervised Process

The pre-trained composite VNIR network used in Section 5.4.2 is used to initialise the parameters of a hyperspectral CNN. The hyperspectral CNN has an architecture with the same general structure as those used in previous experiments (Section 5.1).

This particular network has three convolutional and three fully connected layers.

The dataset automatically extracted from the high confidence points via the unsupervised process is divided into training and validation data, with 180 training points per class and 20 validation points per class. To train the CNN, data in DN format is extracted in batches. Each batch is then augmented using spectral relighting as described in Section 5.3, which expands the size of the batch by many times. The augmented batch is then normalised to reflectance using flat-field correction, interpolated to the spectral wavelengths used by the pre-trained network, and pre-processed using the zero-wavelength approach (Section 5.2.2), before being input into the CNN. Note that all of the parameters used in the spectral relighting augmentation are the same as those used in the experiments in Section 5.4.3. Once the CNN is trained, it can be applied to the remaining pixels in the image and to all of the other images in the mining timelapse to produce thematic classification maps.

6.3 Results and Discussion

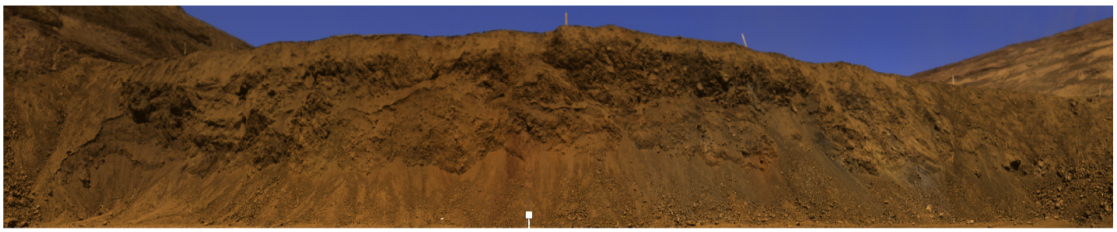
This section presents the results of using the pipeline in Section 6.2 to cluster the mining timelapse images. An analysis of the results is also provided.

6.3.1 Implementation Specifications

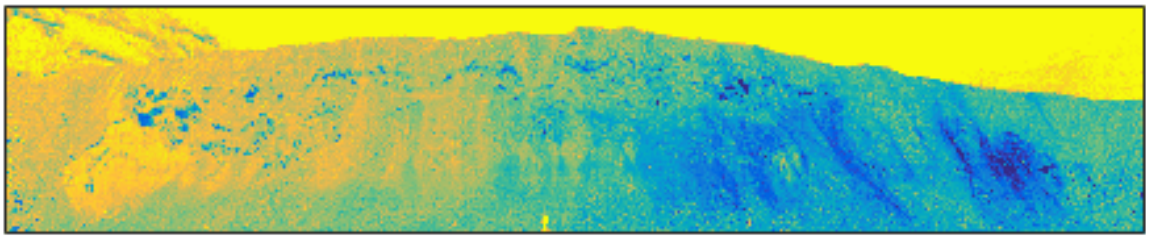
The pipeline was implemented in MATLAB 2015b on a 64-bit computer with an Intel Core i7-4770 CPU @ 3.40GHz × 8 processor and GeForce GTX 760/PCIe/SSE2 graphics card. All CNNs were trained and implemented with the matconvnet-1.0-beta20 software package (Vedaldi and Lenc, 2015).

6.3.2 Results

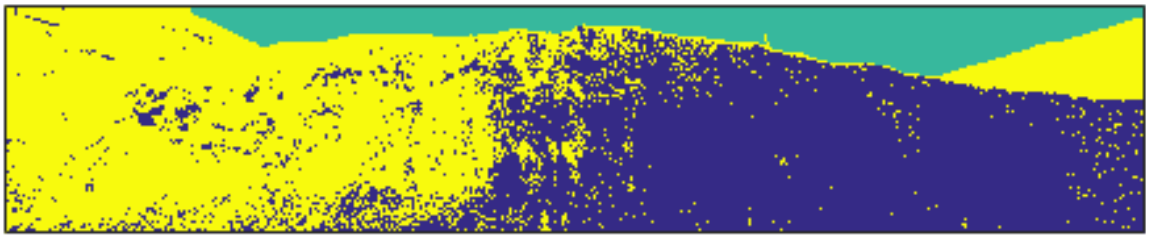
The results at different stages of the pipeline for the unsupervised process are shown in Figure 6.3. The invariant image (Figure 6.3b) visualises the representation of the



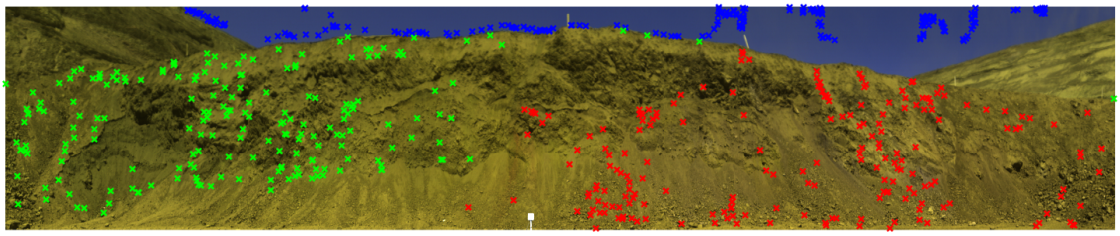
(a) Pseudo RGB image.



(b) The image represented with one of the illumination invariant RSA-SAE features.



(c) The image after clustering in the RSA-SAE space.



(d) The points of highest confidence automatically extracted and used to train the CNN, chosen due to their proximity from cluster centroids in the RSA-SAE space.

Figure 6.3 – Results from the different steps of the unsupervised process, in the extraction of the annotated dataset for training the CNN. The results are from the 11:30 mining timelapse dataset.

hyperspectral image with one of the RSA-SAE features. It was easier to see the how the materials are distributed across the mine face in the invariant image in comparison

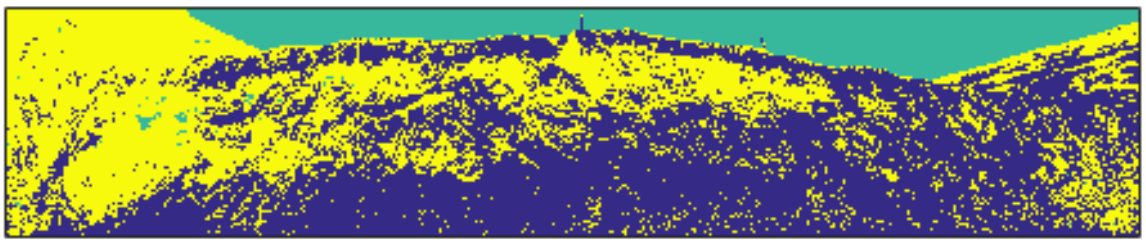
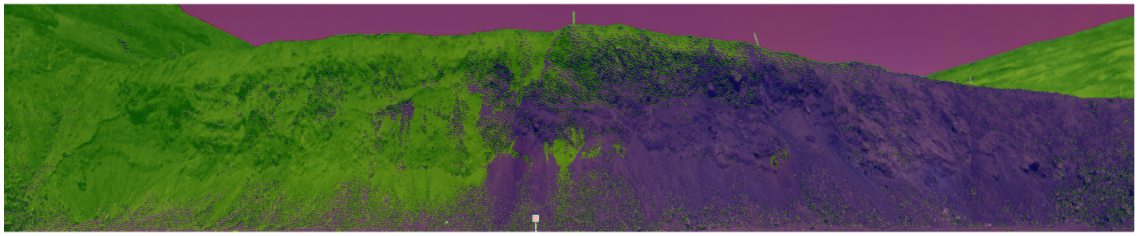


Figure 6.4 – The result of clustering in the original reflectance space.

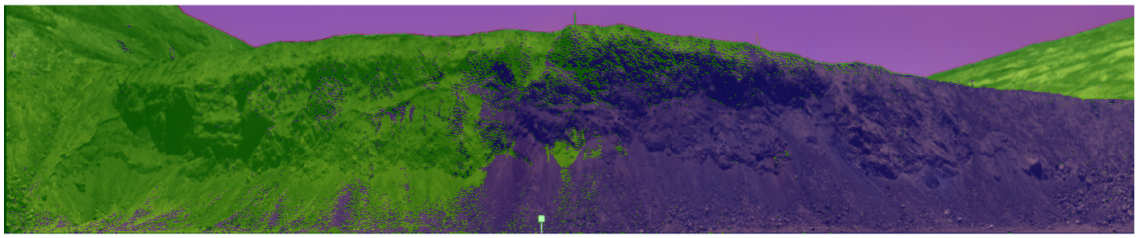
to the pseudo RGB image, where everything is uniform in colour and texture. The shadows in the invariant image were also far less prominent. The clustered image (Figure 6.3c) aligned well with the ground truth information (Figure 6.1). The result was far less affected by the illumination than if the data was clustered in the original reflectance space (Figure 6.4). The high-confidence points that were automatically identified by being closest to the cluster centroids (Figure 6.3d) appeared to belong to the correct class (according to the boundary identified by geologists). These points, which were used to train the self-supervised CNN, did not represent the different classes under shadow.

The training dataset that was extracted using the unsupervised process was then used to train a CNN in the self-supervised component of the pipeline. A visualisation of applying the trained classifiers to the 11:30 and 13:30 images is shown in Figure 6.5 and the progression in classification score during optimisation is shown in Figure 6.6 for four different training strategies, using a test set of about 120,000 labelled samples from the 11:30 image. The four training strategies included a transfer learning approach where the network was pre-trained from the VNIR composite network from Section 5.4.2 (with no augmentation), an augmentation strategy where the parameters were trained from scratch with spectral relighting augmentation, and a combined approach of spectral relighting augmentation and pre-training the network from the VNIR composite. A baseline approach with no pre-training or data augmentation was also compared against for benchmarking.

The visualisation of the mining timelapse images with the CNN clustering result overlaid showed good correspondence to the ground truth information and the geological boundary. There was also insignificant change in the clustered pixels across the two



(a) Clustered 11:30 image.



(b) Clustered 13:30 image.

Figure 6.5 – Images of the mineface captured at different times of the day but assigned categories with the same CNN, thus showing the generality of the pipeline. Green represents martite prediction, purple represents shale prediction, and pink represents sky prediction.

times. Most of the regions which had a significant change in the illumination conditions between 11:30 and 13:30 remained in the same predicted cluster (Figure 6.5).

The result in Figure 6.6 compares the individual significance of each of the key elements of the self-supervised process, that is, the transfer learning and spectral relighting augmentation. When used on its own, in comparison to the baseline CNN, the transfer learning via initialisation with a pre-trained network reduced the number of epochs required to reach convergence, and the F1 score at convergence was also better. When the spectral relighting augmentation was used on its own, it took more epochs for the optimisation to converge than the baseline, but the F1 score it converged at was nearly perfect, significantly higher than the baseline. When using both transfer learning and spectral relighting augmentation, the CNN converged to a very good result with relatively few epochs. The F1 score was higher than both the baseline and transfer learning only approaches, and was just under the spectral

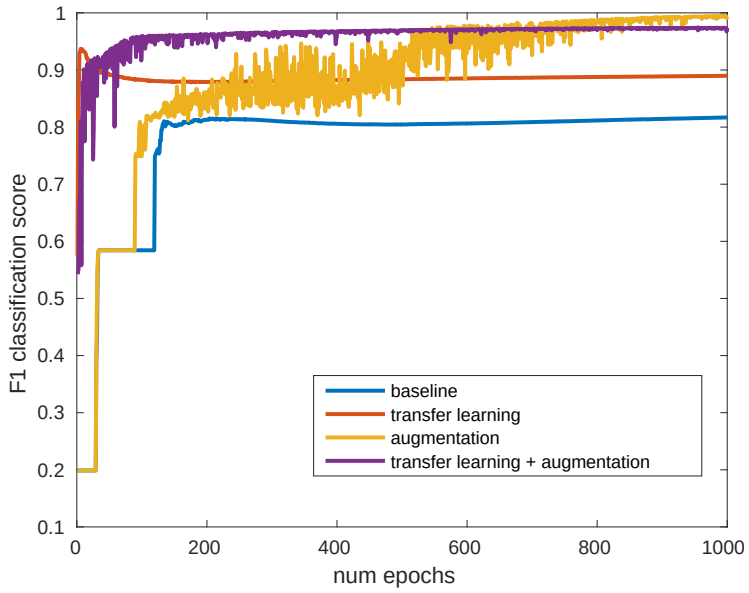


Figure 6.6 – A comparison of the change in the F1 score as the optimisation progresses for the different elements of the self-supervised process.

relighting augmentation-only result. The number of epochs required for convergence was comparable to the transfer learning-only approach, and far fewer than the number of epochs required for the spectral relighting augmentation-only result.

Table 6.1 – Runtime of the different stages of the pipeline.

Stage	Processor	Time
pre-train RSA-SAE	CPU	60m 31s
train RSA-SAE	CPU	14m 51s
apply RSA-SAE	CPU	<1s
clustering	CPU	1s
train CNN	GPU	23m 43s
apply CNN	GPU	10s

6.3.3 Discussion

Converting the hyperspectral image into the illumination invariant RSA-SAE space (Figure 6.3b) was critical for the success of the pipeline. When the data was clustered in the original reflectance space (Figure 6.4), the shadows had a negative effect on

the clusters, making them inaccurate representations of the class distributions on the mine face. This is demonstrated by the correlation between the shape of the clusters and the distribution of shadow on the image. It is possible to search for extra clusters, such that the shadows effects become confined to dedicated clusters. The problem with this is that it is essentially creating extra classes. If the self-supervised learning were to be conducted using these extra clusters, it would be very hard to automatically determine which clusters belonged to the same class or which clusters were from shadows. However, by clustering in the illumination invariant space, the number of clusters can be set to the number of classes in the scene, and the shadows have almost no impact on the clustering accuracy, with shadowed and sunlit data from corresponding classes being correctly clustered together (Figure 6.3c). The lack of correlation between the image clustered in the illumination invariant space and the distribution of shadows demonstrates the effectiveness of the RSA-SAE for finding illumination invariant feature representations.

The points of high confidence that were close to the cluster centroids were all in sunlit regions (Figure 6.3d), which is to be expected, as they were most similar to the mean spectra of that cluster. This means that they do not capture the variability needed to train a robust classifier. At this stage of the pipeline, the scenario was very similar to those explored in Chapter 5, because there was a small amount of labelled training data available that captured a limited amount of the variability in the scene. This justified the importance of using the methods developed in Chapter 5 to classify the entire scene correctly.

The results (Figure 6.5) show that when spectral relighting augmentation and transfer learning were used with the CNN, the pixels were correctly grouped despite the training data not capturing the variability. By transferring knowledge from the airborne VNIR datasets, the classification accuracy on the mining data was improved and the convergence time was reduced (Figure 6.6). The F1 classification score at convergence for the network using purely transfer learning was better than the baseline approach. This is impressive, because in order to do the transfer learning, the data had to be reduced in spectral range from $401-970$ nm to $430-860$ nm, which is

the spectral range of the VNIR pre-trained network. This means that despite having less information from the mining dataset, the network was able to produce better classification accuracies by leveraging information from other datasets, highlighting the advantage of the transfer learning. Although convergence was significantly faster, the purely transfer learning CNNs F1 score at convergence was not as high as the purely augmented CNNs score. This was expected, because it was only trained on the high confidence points extracted from the image via the unsupervised process. The dataset did not capture the variability, and the spectral relighting augmentation added the missing variability needed to train the classifier. The best result was achieved by combining the two techniques (transfer learning and spectral relighting augmentation) together. By doing this, the merits of both approaches were achieved (fast convergence and high accuracy). The F1 classification score at convergence was slightly lower for the combined approach than when the augmentation was used on its own. This is expected to be because the combined approach also required that the wavelength range be reduced to $430 - 860$ nm, as it used the pre-trained VNIR network for initialisation. This could be changed by training a new pre-trained network which uses the full VNIR spectrum, and transferring knowledge from that dataset instead.

The F1 score fluctuated more for the methods that used spectral relighting augmentation (Figure 6.5). This is because in each batch that was used to train the CNN, the parameters for the relighting were being randomly sampled, meaning the same batch of data from two different epochs could be very different. But as the CNN came closer to converging, the fluctuations reduced in size.

The runtimes in Table 6.1 were measured using 1000 epochs for training the autoencoder and CNN. The time to train the CNN with transfer learning and spectral relighting augmentation was 1.42 seconds per epoch (about 24 minutes to train it for 1000 epochs), but, due to the transfer learning, at about 250 epochs, the convergence reached within one percent of its final value. Thus, a good classifier could be trained in about six minutes. Once trained, to forward propagate every pixel in the image through the CNN only took 10 seconds, which was significantly shorter than

the training time. Additionally, running new images of this size through the network would take a similarly shorter amount of time. The longest stage of the pipeline was the pre-training of the RSA-SAE. In this process, each layer was trained individually for 1000 epochs. It could be possible to skip this step all together if a generic network was trained on lots of data and used to pre-train RSA-SAE networks for any new image (similar to how the CNNs are pre-trained in Section 5.2). It is also possible to speed up the training of the RSA-SAE by using a GPU instead of the CPU.

6.4 Summary

New environments are often encountered which have little prior knowledge. It is valuable to be able to make predictions in these environment at the early stages of an investigation. Hence, this chapter has presented a method, comprising of some of the algorithms proposed in this thesis, to cluster the pixels in an image given no labelled data, for a mining application. This process comprised both an unsupervised and supervised component (requiring no human annotations).

The predictions this method makes about the semantic association of the pixels could be part of a higher-level autonomous process for a mining operation. However, the methods value is not constrained to mining applications as it could be applied to any scenario where a hyperspectral sensor is used to collect data from the environment (e.g. on-board a robot) and a human is not on-hand to annotate the data.

Chapter 7

Conclusions

The purpose of this thesis was to develop approaches to learning illumination invariant representations and classification models for hyperspectral data under natural illumination, using limited or no labelled training samples. This chapter provides a summary of the content in the thesis, a list of contributions to the field, and a discussion of potential future work.

7.1 Summary

In **Chapter 2**, the relevant background theory was presented and current work in the field was reviewed. The datasets and evaluation metrics were described in **Chapter 3**. In **Chapter 4**, unsupervised approaches to learning illumination invariant representations were proposed and evaluated. In **Chapter 5**, classification models trained with limited data that were robust to the illumination of the scene, were proposed and evaluated. Finally, **Chapter 6** developed a unified pipeline for implementing the algorithms proposed in Chapter 4 and Chapter 5 in a special case study.

7.2 Contributions

There is a significant amount of variability in hyperspectral data attributed to illumination. This problem has been approached by different communities of researchers. Computer vision techniques derive illumination invariant feature spaces but rely on assumptions that are not as justified for hyperspectral images because they were developed for RGB imagery. Remote sensing techniques often use atmospheric modellers which require *a priori* knowledge of the atmospheric composition. They also will often disregard the geometry of the scene. Multi-modal approaches can account for the scene geometry but require additional sensors which are often not available. Finally, learning-based approaches require enough labelled data to capture all of the variability in the scene. Due to the difficulty in labelling hyperspectral data and the small amount of publicly available annotated datasets (in comparison to more widely used sensors like RGB cameras), there are often limited amounts of labelled samples available for learning. This thesis proposes approaches which integrate domain knowledge into learning algorithms in order to learn illumination invariance without needing large amounts of labelled data.

The proposed approach for learning low-dimensional illumination invariant feature representations of hyperspectral images is unsupervised, so it is trained entirely on unlabelled data. It requires no additional sensors, no *a priori* knowledge about the atmospheric conditions and does not make all of the assumptions that computer vision approaches make. In such a feature space, a material retains the highly discriminative information relating to its underlying properties from the hyperspectral representation, and is less influenced by the effects of illumination relating to geometric orientation, sun position and occlusions which cause shadows in the scene. The approaches developed were shown to robustly represent the materials in a number of images with a high-degree of discrimination compared to other techniques. The shadow invariant approach was shown to represent a static scene very similarly when captured at different times of the day, despite shadows moving across the image. The illumination invariant representation was also shown to produce robust results when used as a feature space for a classifier.

An approach to learning a classification model with a limited amount of labelled data was also proposed. Transfer learning was investigated as a means of using knowledge from other labelled datasets to improve classification performance for problems with limited knowledge. When knowledge was transferred from data acquired from airborne platforms to data acquired from field-based platforms, the training time in the latter decreased and the classification accuracy improved when field-based training sets had a limited number of labelled samples. An image-based method was also proposed for extracting the terrestrial sunlight-diffuse skylight ratio from a hyperspectral image. This could be used in the proposed spectral relighting-based classifier for predicting the labels of pixels in hyperspectral images using limited amounts of labelled data. The performance of a classifier trained with spectral relighting augmentation on just a limited, localised set of sunlit samples closely rivalled the performance of a classifier trained on samples collected from the entire image. In the experiments evaluating both transfer learning and spectral relighting, the results showed that the methods worked well even when the training datasets had fewer than 200 labelled samples per class.

The specific contributions of this thesis are:

- An unsupervised approach to learning low-dimensional feature representations designed for hyperspectral data (Chapter 4). By incorporating spectral similarity measures into the learning process, these approaches learn a more discriminative feature representation of spectra. The representations are also insensitive to brightness, and less effected by geometric orientation with respect to sun position than other approaches. They can be used for dimensionality reduction or as an unsupervised feature extraction technique, where the features are useful for classification clustering, or other high-level tasks.

- An extension to the above unsupervised low-dimensional feature learning approach which makes the representation invariant to shadows (Chapter 4). This approach integrates physics-based relighting into the learning algorithm. The method produced superior results to a number of similar techniques on multiple

datasets.

- A transfer learning scheme to reduce training time and improve classification performance on field-based datasets by leveraging information from data captured from airborne platforms (Chapter 5).
- An image-based approach for determining the terrestrial sunlight-diffuse skylight ratio (Chapter 5).
- An approach to training classification models to be robust to illumination effects using spectral relighting augmentation (Chapter 5). This allowed for accurate pixel-wise classification of hyperspectral imagery using limited amounts of training data, and had superior performance to several other approaches.
- A pipeline for clustering hyperspectral data in the presence of illumination variability that uses no annotated samples (Chapter 6).

These methods exploit domain knowledge as well as powerful learning algorithms to achieve illumination invariance, which allows them to work with no labelled data in the unsupervised cases and limited labelled data in the supervised cases. They also do not exhibit the limitations of many of the techniques from the literature. The model accounts for the scene geometry and how it interacts with the multiple sources of illumination arising from direct sunlight and light reflecting off the sky dome. When the model is coupled with state-of-the-art learning algorithms, the unknown variables required to obtain robust feature representations and classification models can be solved for using considerably less annotated data. As the use of hyperspectral sensors increases in many fields of research and application, including robotics and ground-based sensing, the work presented in this thesis provides a framework for feature learning for the scenarios where there are limited amounts of labelled data available.

7.3 Future Work

As has been outlined in Chapter 2, there is a need to further improve and propose new illumination invariant algorithms for the limited scenario, which was done in Chapter 4 and Chapter 5 within the scope of this thesis. Some specific directions of work which have been highlighted by the analyses in this thesis as areas for further improvement are:

Investigating the temporal invariance of the RSA-SAE. The extent of the temporal invariance of the low-dimensional, illumination invariant feature representation has not yet been determined. Because the feature space is learnt from many different randomly sampled atmospheric compositions, there is nothing constraining it to only working with the conditions of the current atmosphere, at least in the shadow. Whilst this was tested to some degree with temporal datasets, this should be rigorously investigated with more temporal hyperspectral datasets, particularly those captured over different days and even different seasons.

Indirect illumination extension. Light reflecting off different regions in the scene can also act as a source of illumination. Although far weaker than the dominant sunlight and skylight sources, indirect sources of illumination can still have an influence on the spectra returned to the camera. Indirect illumination can be incorporated into the model, and hence it is possible to derive relighting equations that can simulate its effect. Thus, it can be integrated into the learning algorithm.

Adding spatial information. In some applications the texture in the image provides useful semantic information. Although this was touched upon, the results for these scenarios would improve if the spatial dimension was to be incorporated into the algorithms presented in this thesis. It would also be useful to know which kinds of scenarios would see a significant gain in performance when the spatial information is included.

Investigations into overcast conditions. It is unknown whether the outdoor model can accurately depict an overcast scenario. Experiments should be conducted

to determine how similar images of the same scene represented with the proposed approaches are when captured under sunny conditions and overcast conditions.

List of References

Elhadi Adam, Onisimo Mutanga, and Denis Rugege. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. *Wetlands Ecology and Management*, 18(3):281–296, 2010.

Steven M Adler-golden, Robert Y Levine, Michael W Matthew, Steven C Richtsmeier, Lawrence S Bernstein, John Gruninger, Gerald Felde, Michael Hoke, Gail Anderson, and Anthony Ratkowski. Shadow-insensitive material detection/classification with atmospherically corrected hyperspectral imagery. *Proc. SPIE 4381, Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VII*, 4381:461, 2001. ISSN 0277786X.

Steven M Adler-golden, Michael W Matthew, Gail P Anderson, Gerald W Felde, and James A Gardner. An algorithm for de-shadowing spectral imagery. *Imaging Spectrometry VIII Proceedings of SPIE*, 4816:203–210, 2002.

Vivek Agarwal, Andreas Koschan, and Mongi A Abidi. An Overview of Color Constancy Algorithms. *Journal of Pattern Recognition Research*, 1(1):42–54, 2006.

Amr Ahmed, Kai Yu, Wei Xu, Yihong Gong, and Eric Xing. Training Hierarchical Feed-Forward Visual Recognition Models Using Transfer Learning from Pseudo-Tasks. In *Computer Vision–ECCV 2008*, pages 69–82, 2008.

Fahim Irfan Alam, Jun Zhou, Alan Wee-Chung Liew, and Xiuping Jia. CRF Learning with CNN Features for Hyperspectral Image Segmentation. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 6890–6893, 2016. ISBN 9781509033324.

José M Álvarez and M Antonio. Road Detection Based on Illuminant Invariance. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):184–193, 2011.

Lei Jimmy Ba and Rich Caruana. Do Deep Nets Really Need to be Deep? In *Advances in neural information processing systems*, pages 2654–266, 2014. ISBN 3135786504.

Steve De Backer, Pieter Kempeneers, Walter Debruyn, and Paul Scheunders. A Band Selection Technique for Spectral Classification. *IEEE Geoscience and Remote Sensing Letters*, 2(3):319–323, 2005.

Haris Baltzakis, Antonis Argyros, and Panos Trahanias. Fusion of laser and visual data for robot motion planning and collision avoidance. *Machine Vision and Applications*, 15(2):92–100, 2003.

Suchet Bargoti and James Underwood. Deep Fruit Detection in Orchards. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on.*, pages 3626–3633, 2017.

Jonathan T Barron. Convolutional Color Constancy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 379–387, 2015.

Etienne Beauchesne and R Sbastien. Automatic Relighting of Overlapping Textures of a 3D Model. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, pages II—-166, 2003.

Yoshua Bengio. Deep Learning of Representations for Unsupervised and Transfer Learning. In *ICML Unsupervised and Transfer Learning*, pages 17–36, 2012.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy Layer-Wise Training of Deep Networks. *Advances in neural information processing systems*, 19:153, 2007.

Alexander Berk, Lawrence S. Bernstein, and David C. Robertson. MODTRAN: A moderate resolution model for LOWTRAN. Technical report, Spectral Sciences Inc Burlington MA, 1987.

José M Bioucas-dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn 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.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Marcus Borengasser, William S Hungate, and Russell Watkins. *Hyperspectral remote sensing: principles and applications*. Crc Press, 2007.

Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

Joshua Broadwater and Amit Banerjee. Improved atmospheric compensation of hyperspectral imagery using LIDAR. In *Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International*, pages 2200—-2203, 2013.

Gershon Buchsbaum. A Spatial Processor Model for Object Colour Perception. *Journal of the Franklin institute*, 310(1):1—-26, 1980.

Gustavo Camps-Valls, Devis Tuia, Lorenzo Bruzzone, and Jón Atli Benediktsson. Advances in Hyperspectral Image Classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, 31(1):45—-54, October 2014.

Gustavo Camps-Valls, Devis Tuia, and Lorenzo Bruzzone. Advances in hyperspectral image classification. *IEEE Signal Processing Magazine*, 31(1): 45–54, 2015.

Jiayan Cao, Zhao Chen, and Bin Wang. Deep Convolutional Networks with Superpixel Segmentation for Hyperspectral Image Classification. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 3310–3313, 2016. ISBN 9781509033324.

Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *arXiv preprint arXiv:1508.00092*, 2015.

Chein-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.

Guangyi Chen and Shen-en Qian. Dimensionality reduction of hyperspectral imagery using improved locally linear embedding. *Journal of Applied Remote Sensing*, 1(1):013509—-013509, 2007.

Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep Learning-Based Classification of Hyperspectral Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2094–2107, 2014.

Yushi Chen, Xing Zhao, and Xiuping Jia. Spectral – Spatial Classification of Hyperspectral Data Based on Deep Belief Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2381—-2392, 2015.

Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251, 2016.

Dongliang Cheng, Abdelrahman Kamel, Brian Price, Scott Cohen, and Michael S Brown. Two Illuminant Estimation and User Correction Preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 469–477, 2016.

Anil Cheriyadat and LM Bruce. Why principal component analysis is not an appropriate feature extraction method for hyperspectral data. *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International*, 6:3420–3422, 2003.

Shao-Shan Chiang, Chein-I Chang, and Irvin W Ginsberg. Unsupervised hyperspectral image analysis using independent component analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 7:3136–3138, 2000.

Anna Chlingaryan, Arman Melkumyan, Richard J Murphy, and Sven Schneider. Automated Multi-class Classification of Remotely Sensed Hyperspectral Imagery Via Gaussian Processes with a Non-stationary Covariance Function. *Mathematical Geosciences*, 48(5):537–558, 2016. ISSN 1874-8953.

Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella, and Jurgen Schmidhuber. Deep, Big, Simple Neural Nets for Handwritten. *Neural computation*, 22(12):3207–3220, 2010.

R.N Clark, G.A Swayze, R.A Wise, K.E Live, T.M Hoefen, R.F Kokaly, and S.J Sutley. USGS Digital Spectral Library splib06a: U.S. Geological Survey Data Series 231. 2007.

Roger N Clark. Spectroscopy of rocks and minerals, principles of spectroscopy. In Andrew N. Rencz, editor, *Remote Sensing for the Earth Sciences. Volume 3*, chapter 1, pages 3–58. 3 edition, 1999.

Roger N Clark and Ted L Roush. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 89(B7):6329–6340, 1984. ISSN 2156-2202.

Color-temp. Hyperspectral and Colour Imaging. `https://sites.google.com/site/hyperspectralcolorimaging/dataset`. Accessed: 2017-02-02.

Peter Corke, Rohan Paul, Winston Churchill, and Paul Newman. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2085–2092. Ieee, November 2013. ISBN 978-1-4673-6358-7.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2012.

Luca Demarchi, Frank Canters, Claude Cariou, Giorgio Licciardi, and Jonathan Cheung-Wai Chan. Assessing the performance of two unsupervised dimensionality reduction techniques on hyperspectral APEX data for high resolution urban

land-cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87 (August 2015):166–179, January 2014. ISSN 09242716.

Jia Deng, Wei Dong, Richard Socher, Li-jia Li, Kai Li, and Li Fei-fei. ImageNet : A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

L Deng, M Seltzer, D Yu, A Acero, A Mohamed, and G Hinton. Binary Coding of Speech Spectrograms Using a Deep Auto - encoder. In *Eleventh Annual Conference of the International Speech Communication Association*, pages 1692–1695, 2010.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a Deep Convolutional Network for Image Super-Resolution. In *European Conference on Computer Vision*, pages 184–199, 2014.

David L Donoho, Iain Johnstone, Bob Stine, and Gregory Piatetsky-shapiro. High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality. *AMS Math Challenges Lecture*, 1:1–33, 2000.

Mark S Drew and Amin Yazdani Salekdeh. Multispectral Image Invariant to Illumination Colour , Strength , and Shading. *IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics*, (January):78760A–78760A, 2011.

Qian Du. Modified Fisher ' s Linear Discriminant Analysis for Hyperspectral Imagery. *IEEE geoscience and remote sensing letters*, 4(4):503–507, 2007.

Michael L Eastwood, Charles M Sarture, Thomas G Chrien, Mikeal Aronsson, Bruce J Chippendale, Jessica A Faust, Betina E Pavri, Christopher J Chovit, Manuel Solis, and Martin R Olah. Imaging Spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote sensing of environment*, 65(3):227—-248, 1987.

Michael T Eismann, Craig R Schwartz, Jack N Cederquist, John A Hackwell, and Ronald J Huppi. hyperspectral sensors for military target detection applications. *SPIE's 1996 International Symposium on Optical Science, Engineering, and Instrumentation*, pages 91—-101, 1996.

Gamal Elmasry, Mohammed Kamruzzaman, Da-wen Sun, and Paul Allen. Principles and Applications of Hyperspectral Imaging in Quality Evaluation of Agro-Food Products : A Review. *Critical reviews in food science and nutrition*, 52(11):999—-1023, 2012.

M Fauvel, J Chanussot, and J A Benediktsson. A spatial–spectral kernel-based approach for the classification of remote-sensing images. *Pattern Recognition*, 45 (1):381–392, 2012. ISSN 0031-3203.

Mathieu Fauvel, Jon Atli Benediktsson, Jocelyn Chanussot, and R Johannes. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46 (11):3804—-3814, 2008.

Yao-ze Feng and Da-wen Sun. Application of Hyperspectral Imaging in Food Safety Inspection and Control : A Review. *Critical reviews in food science and nutrition*, 52(11):1039—-1058, 2012.

G D Finlayson, S D Hordley, C Lu, and M S Drew. On the Removal of Shadows From Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006.

Graham D Finlayson, Steven D Hordley, and Mark S Drew. Removing Shadows from Images. In *European Conference on Computer Vision*, pages 823–836, 2002.

Graham D Finlayson, Mark S Drew, and Cheng Lu. Intrinsic Images by Entropy Minimization. In *European Conference on Computer Vision*, pages 582–595, 2004.

Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy Minimization for Shadow Removal. *International Journal of Computer Vision*, 85(1):35—-57, 2009.

David H Foster, Kinjiro Amano, Sérgio M C Nascimento, and Michael J Foster. Frequency of metamerism in natural scenes. *Josa a*, 23(10):2359–2372, 2006.

David H Foster, Kinjiro Amano, and Sérgio M C Nascimento. Time-lapse ratios of cone excitations in natural scenes. *Vision Research*, 120:45–60, 2015. ISSN 0042-6989.

Ola Friman, Jorgen Ahlberg, and Gustav Tolt. Illumination and shadow compensation of hyperspectral images using a digital surface model and non-linear least squares estimation. In *Image and Signal Processing for Remote Sensing XVII*, 2011.

Pedram Ghamisi, Yushi Chen, and Xiao Xiang Zhu. A Self-Improving Convolution Neural Network for the Classification of Hyperspectral Data. *IEEE Geoscience and Remote Sensing Letters*, 13(10):1537–1541, 2016.

Arjan Gijsenij, Rui Lu, and Theo Gevers. Color Constancy for Multiple Light Sources. *IEEE Transactions on Image Processing*, 21(2):697–707, 2012.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain Adaptation for Large-Scale Sentiment Classification : A Deep Learning Approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

Pavel Golik, Patrick Doetsch, and Hermann Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Interspeech*, pages 1756—-1760, 2013.

M Govender, K Chetty, and H Bulcock. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa*, 33(2): 145–151, 2007.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645—-6649, 2013.

Zhang Guangjun, Dong Yongsheng and Ji Song. Dimensionality reduction of hyperspectral data based on ISOMAP algorithm. In *Electronic Measurement and Instruments, 2007. ICEMI'07. 8th International Conference on*, pages 3—-935, 2007.

Christian A Gueymard. Parameterized Transmittance Model for Direct Beam and Circumsolar Spectral Irradiance. *Solar Energy*, 71(5):325–346, 2001.

Ruiqi Guo, Dai Qieyun, and Derek Hoiem. Single-Image Shadow Detection and Removal using Paired Regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2033–2040, 2011.

Saurabh Gupta, Ross Girshick, Pablo Arbelaez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer International Publishing, 2014.

Driss Haboudane, John R Miller, Nicolas Tremblay, Pablo J Zarco-tejada, and Louise Dextraze. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote sensing of environment*, 81(2):416–426, 2002.

Maarten Haest, Thomas Cudahy, Carsten Laukamp, and Sean Gregory. Quantitative Mineralogy from Infrared Spectroscopic Data . I . Validation of Mineral Abundance and Composition Scripts at the Rocklea Channel Iron Deposit in Western Australia. *Economic Geology*, 107(2):209–228, 2012.

Jisoo Ham, Yangchi Chen, Melba M Crawford, and Joydeep Ghosh. Investigation of the Random Forest Framework for Classification of Hyperspectral Data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):492–501, 2005.

Tian Han and David G Goodenough. Nonlinear Feature Extraction of Hyperspectral Data Based on Locally Linear Embedding (LLE). In *Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. Proceedings. 2005 IEEE International*, pages 1237–1240, 2005. ISBN 0780390504.

Bruce Hapke. Bidirectional reflectance spectroscopy: 1. Theory. *Journal of Geophysical Research: Solid Earth*, 86(B4):3039–3054, 1981.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Glenn Healey and David Slater. Models and Methods for Automated Material Identification in Hyperspectral Imagery Acquired Under Unknown Illumination and Atmospheric Conditions. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(6):2706–2717, 1999.

Christoph Hecker, Mark Van Der Meijde, Harald Van Der Werff, and Freek D Van Der Meer. Assessing the Influence of Reference Spectra on Synthetic SAM Classification Results. *IEEE Transactions on geoscience and remote sensing*, 46 (12):4162–4172, 2008.

G E Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313(5786):504–507, July 2006. ISSN 1095-9203.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep neural networks for accoustic modelling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Alain Hore and Djemel Ziou. Image quality metrics : PSNR vs . SSIM. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2366–2369, 2010.

Lukàš Hošek and Alexander Wilkie. Adding a Solar-Radiance Function to the Hošek-Wilkie. *IEEE computer graphics and applications*, 33(3):44—-52, 2013.

Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *Journal of Sensors*, 2015, 2015a.

Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *Journal of Sensors*, 2015b.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.

Minyoung Huh, Pulkit Agrawa, and Alexei A. Efros. What makes ImageNet good for transfer learning? *arXiv preprint*, arXiv:1608, 2016.

Quan Huynh-Thu and M Ghanbari. Scope of validity of PSNR in image / video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

Emmett J Ientilucci. Leveraging Lidar Data to Aid in Hyperspectral Image Target Detection in the Radiance Domain. *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*, 8390:839007—-1, 2012.

Emma Izquierdo-Verdiguier, Valero Laparra, Luis Gomez-Chova, and Gustavo Camps-Valls. Encoding invariances in remote sensing image classification with SVM. *IEEE Geoscience and Remote Sensing Letters*, 10(5):981–985, 2013. ISSN 1545598X.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep Features for Text Spotting. In *European Conference on Computer Vision*, pages 512–528. Springer International Publishing, 2014.

Ronald Kemker and Christopher Kanan. Self-Taught Feature Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2693–2705, 2017.

David H Kim and Leif H Finkel. Hyperspectral Image Processing Using Locally Linear Embedding. In *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*, number March, pages 316—-319, 2003.

Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. ISSN 1547-5905.

Alex Krizhevsky and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

F A Kruse, A B Lefkoff, J W Boardman, K B Heidebrecht, A T Shapiro, P J Barloon, and A F H Goetz. The Spectral Image Processing System (SIPS) Interactive Visualization and Analysis of Imaging Spectrometer Data. In *AIP Conference Proceedings*, volume 283, pages 192—-201, 1993.

Fred A Kruse. Use of airborne imaging spectrometer data to map minerals associated with hydrothermally altered rocks in the northern Grapevine Mountains, Nevada, and California. *Remote Sensing of Environment*, 24(1): 31–51, 1988.

Miroslav Kubat and Stan Matwin. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *ICML*, volume 97, 1997.

Bor-chen Kuo, Cheng-hsuan Li, and Jinn-min Yang. Kernel Nonparametric Weighted Feature Extraction for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(4):1139–1155, April 2009. ISSN 0196-2892.

Jean-francois Lalonde and Alexei Efros. Detecting Ground Shadows in Outdoor Consumer Photographs. In *European Conference on Computer Vision*, pages 322–335, 2010.

D Landgrebe and L Biehl. Multispec and datasets flightline c1 Tippecanoe County and aviris NW Indiana's Indian Pines. 1992.

Hugo Larochelle, Aaron Courville, and James Bergstra. An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480, 2007.

Hugo Larochelle, Yoshua Bengio, J{\'e}r{\^o}me Louradour, and Pascal Lamblin. Exploring Strategies for Training Deep Neural Networks. *Journal of Machine Learning Research*, 10(Jan):1–40, 2009. ISSN 15324435.

Yann LeCun and Yoshua Bengio. Convolutional Networks for Images , Speech , and Time-Series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten Digit Recognition with a Back-Propagation Network. In *Advances in neural information processing systems*, pages 396–404, 1990.

Chulhee Lee and David A Landgrebe. Analyzing High Dimensional Multispectral Data. *IEEE Transactions on Geoscience and Remote Sensing*, 31(4):792—-800, 1993.

Hyungtae Lee and Heesung Kwon. Contextual Deep CNN Based Hyperspectral Classification. *arXiv*, 1604.03519:2–4, 2016.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and

large-scale data collection. *The International Journal of Robotics Research*, pages 1–16, 2017.

Tong Li, Junping Zhang, and Ye Zhang. Classification of hyperspectral image based on deep belief networks. *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5132–5136, October 2014.

Wei Li, Guodong Wu, and Fan Zhang. Hyperspectral Image Classification Using Deep Pixel-Pair Features. *IEEE Transactions on Geoscience and Remote Sensing*, 2016.

G. Licciardi, F. Del Frate, and R. Duca. Feature reduction of hyperspectral data using autoassociative neural networks algorithms. In *Geoscience and Remote Sensing Symposium (IGARSS), 2009 IEEE International*, number 1, pages 176–179, 2009. ISBN 9781424433957.

G. A. Licciardi, X. Ceamanos, S. Doute´ E, and J. Chanussot. Unsupervised nonlinear spectral unmixing by means of NLPCA applied to hyperspectral imagery. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 1369–1372, 2012a. ISBN 9781467311595.

G.A. Licciardi and J. Chanussot. Nonlinear PCA for visible and thermal hyperspectral images quality enhancement. *IEEE Geoscience and Remote Sensing Letters*, 12(6):1228—-1231, 2015.

Giorgio Licciardi, Prashanth Reddy Marpu, Jocelyn Chanussot, and Jon Atli Benediktsson. Linear Versus Nonlinear PCA for the Classification of Hyperspectral Data Based on the Extended Morphological Profiles. *Geoscience and Remote Sensing Letters, IEEE*, 9(3):447–451, 2012b.

Giorgio Licciardi, Jocelyn Chanussot, Gabriel Vasile, and A Piscini. Enhancing Hyperspectral Image Quality using Nonlinear PCA. In *IEEE Internation Conference on Image Processing*, pages 5087 – 5091, 2014.

Giorgio A Licciardi and F. Del Frate. Pixel Unmixing in Hyperspectral Data by Means of Neural Networks. *IEEE transactions on Geoscience and remote sensing*, 49(11):4163–4172, 2011.

Thy-hou Lin, Huang-te Li, and Keng-chang Tsai. Implementing the Fisher ' s Discriminant Ratio in a k -Means Clustering Algorithm for Feature Selection and Data Set Trimming. *Journal of chemical information and computer sciences*, 44 (1):76–87, 2004.

Zhouhan Lin, Yushi Chen, Xing Zhao, and Gang Wang. Spectral-Spatial Classification of Hyperspectral Image Using Autoencoders. In *Information, Communications and Signal Processing (ICICS) 2013 9th International Conference on*, pages 1—-5, 2013.

Yazhou Liu, Guo Cao, Quansen Sun, and Mel Siegel. Hyperspectral classification via deep networks and superpixel segmentation. *International Journal of Remote Sensing*, 36(13):3459–3482, July 2015. ISSN 0143-1161.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015a.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning Transferable Features with Deep Adaptation Networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, volume 37, 2015b.

Guolan Lu and Baowei Fei. Medical hyperspectral imaging : a review. *Journal of biomedical optics*, 19(1):010901—-010901, 2014.

D C Lui and J Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Stuart E Lynch, Mark S Drew, and Graham D Finlayson. Colour Constancy from Both Sides of the Shadow Edge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 899—-906, 2013. ISBN 9781479930227.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the 30th International Conference on Machine Learning*, volume 30, page 3, 2013.

J Macqueen. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14):281–297, 1967.

Konstantinos Makantasis, Konstantinos Karantzalos, Anastasios Doulamis, and Nikolaos Doulamis. Deep Supervised Learning for Hyperspectral Data Classification through Convolutional Neural Networks. In *IGARSS 2015. 2015 IEEE International Geoscience and Remote Sensing Symposium. Proceedings*, pages 4959–4962, 2015. ISBN 9781479979295.

John Marchant and Christine Onyango. Shadow-invariant classification for scenes illuminated by daylight. *JOSA A*, 17(11):1952–1961, 2000.

John Marchant and Christine Onyango. Spectral invariance under daylight illumination changes. *JOSA A*, 19(5):840–848, 2002.

Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2016.

Stephen R Marschner and Donald P Greenberg. Inverse Lighting for Photography. In *Color and Imaging Conference*, pages 262—-265, 1997.

S. E. Marsh and J. B. McKeon. Integrated analysis of high-resolution field and airborne spectroradiometer data for alteration mapping. *Economic Geology*, 78 (4):618–632, 1983. ISSN 03610128.

Colin Mcmanus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. Shady Dealings : Robust , Long-Term Visual Localisation using Illumination Invariance. In *IEEE International Conference on Robotics and Automation*, pages 901–906, 2014.

Shaohui Mei, Jingyu Ji, Qianqian Bi, Junhui Hou, Qian Du, and Wei Li. Integrating Spectral and Spatial Information into Deep Convolutional Neural Networks for Hyperspectral Classification. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, number 61201324, pages 5067–5070, 2016. ISBN 9781509033324.

Shaohui Mei, Jingyu Ji, Junhui Hou, Xu Li, and Qian Du. Learning Sensor-Specific Spatial-Spectral Features of Hyperspectral Images via Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8): 4520–4533, 2017.

Farid Melgani and Lorenzo Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8):1778–1790, 2004.

Ralph E Milliken and John F Mustard. Quantifying absolute water content of minerals using near-infrared reflectance spectroscopy. *Journal of Geophysical Research: Planets*, 110(E12), 2005.

Richard J. Murphy. Evaluating simple proxy measures for estimating depth of the ˜ 1900 nm water absorption feature from hyperspectral data acquired under natural illumination. *Remote Sensing of Environment*, 166:22–33, 2015. ISSN 0034-4257.

Richard J. Murphy and Sildomar T. Monteiro. Mapping the distribution of ferric iron minerals on a vertical mine face using derivative analysis of hyperspectral imagery (430–970nm). *ISPRS Journal of Photogrammetry and Remote Sensing*, 75:29–39, January 2013. ISSN 09242716.

Richard J. Murphy, AJ Underwood, TJ Tolhurst, and M G Chapman. Field-based remote-sensing for experimental intertidal ecology : Case studies using hyperspatial and hyperspectral data ... *Remote Sensing of Environment*, 112(8): 3353–3365, 2008.

Richard J. Murphy, Sildomar T. Monteiro, and Sven Schneider. Evaluating Classification Techniques for Mapping Vertical Geology Using Field-Based Hyperspectral Sensors. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8):3066–3080, August 2012. ISSN 0196-2892.

Richard J. Murphy, Sven Schneider, and Sildomar Monteiro. Mapping Layers of Clay in a Vertical Geological Surface Using Hyperspectral Imagery: Variability in Parameters of SWIR Absorption Features under Different Conditions of Illumination. *Remote Sensing*, 6(9):9104–9129, September 2014a. ISSN 2072-4292.

Richard J. Murphy, Sven Schneider, and Sildomar T Monteiro. Consistency of Measurements of Wavelength Position From Hyperspectral Imagery : Use of the Ferric Iron Crystal Field Absorption at ˜ 900 nm as an Indicator of Mineralogy. *IEEE Transactions on Geoscience and Remote Sensing*, 52(5):2843–2857, 2014b.

John F Mustard and Carle M Pieters. Quantitative abundance estimates from bidirectional reflectance measurements. *Journal of Geophysical Research: Solid Earth*, 92(B4):617–626, 1987.

John F Mustard and Jessica M Sunshine. Spectral analysis for earth science: investigations using remote sensing data. In *Remote sensing for the earth sciences: Manual of remote sensing*, volume 3, pages 251–307. John Wiley and Sons, Inc., New York, New York, 1999.

Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807—-814, 2010.

Sérgio M C Nascimento, Kinjiro Amano, and David H Foster. Spatial distributions of local illumination color in natural scenes. *Vision Research*, 120:39–44, 2016.

Andrew Ng. CS294A Lecture notes Sparse autoencoder. *CS294A Lecture notes*, 72: 1–19, 2011.

David Nistér, Oleg Naroditsky, and James Bergen. Visual Odometry. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages 1–8, 2004.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014.

Kuntal Kumar Pal and K. S. Sudeep. Preprocessing for Image Classification by Convolutional Neural Networks. In *Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference on*, pages 1778–1781, 2016. ISBN 9781509007745.

Mahesh Pal and Giles M Foody. Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5): 2297–2307, 2010.

Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade Residual Learning : A Two-stage Convolutional Neural for Stereo Matching. *arXiv preprint arXiv:1708.09204*, 2017.

R Perez, R Seals, and J Michalsky. All-weather model for sky luminance distribution—preliminary configuration and validation. *Solar Energy*, 50(3): 235–245, 1993.

A J Preetham, Peter Shirley, and Brian Smits. A Practical Analytic Model for Daylight. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 91–100, 1999. ISBN 0201485605.

Shen-en Qian and Guangyi Chen. A New Nonlinear Dimensionality Reduction Method with Application to Hyperspectral Image Analysis. In *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, pages 270–273, 2007. ISBN 1424412129.

Rishi Ramakrishnan. *Illumination Invariant Outdoor Perception*. PhD thesis, 2016.

Rishi Ramakrishnan, Juan Nieto, and Steve Scheding. Shadow Compensation for Outdoor Perception. In *International Conference on Robotics and Automation*, pages 4835–4842, 2015. ISBN 9781479969227.

William M Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Michael Rast, Simon J Hook, Ronald E Alley, and Christopher D Elvidge. An evaluation of techniques for the extraction of mineral absorption features from high spectral resolution remote sensing data. Technical report, 1991.

Frédéric Ratle, Gustavo Camps-valls, and Jason Weston. Semi-Supervised Neural Networks for Efficient Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2271—-2282, 2009.

Sivalogeswaran Ratnasingam and T. Martin McGinnity. Chromaticity Space for Illuminant Invariant Recognition. *IEEE Transactions on Image Processing*, 21(8): 3612–3623, 2012.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in neural information processing systems*, pages 1–9, 2015.

R Richter and A Müller. De-shadowing of satellite/airborne imagery. *International Journal of Remote Sensing*, 26(15):3137–3148, 2005.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive Auto-Encoders : Explicit Invariance During Feature Extraction. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 833—-840, 2011.

Craig Rodarmel and Jie Shan. Principal Component Analysis for Hyperspectral Image Classification. *Surveying and Land Information Science*, 62(2), 2002.

Adriana Romero, Carlo Gatta, and G Camps-Valls. Unsupervised deep feature extraction of hyperspectral images. In *Proc. of WHISPERS*, pages 2–5, 2014.

Adriana Romero, Carlo Gatta, and Gustau Camps-valls. Unsupervised Deep Feature Extraction for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1349–1362, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and Alexander C Berg. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. ISSN 0920-5691.

Amin Yazdani Salekdeh. *Multispectral and Hyperspectral Images Invarinat to Illumination*. PhD thesis, 2011.

Yoichi Sato and Katsushi Ikeuchi. Reflectance analysis under solar illumination. In *Physics-Based Modeling in Computer Vision, 1995., Proceedings of the Workshop on. IEEE*, pages 180–187, 1995.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

Juergen Schmidhuber. Deep Learning in Neural Networks: An Overview. page 75, April 2014.

S Schneider, R J Murphy, A Melkumyan, and E Nettleton. Autonomous Mapping of Mine Face Geology Using Hyperspectral Data. In *35th APCOM Symposium*, number September, pages 24–30, 2011.

Sven Schneider, Arman Melkumyan, Richard J. Murphy, and Eric Nettleton. Gaussian Processes with OAD Covariance Function for Hyperspectral Data Classification. In *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, pages 393–400. Ieee, October 2010. ISBN 978-1-4244-8817-9.

Sven Schneider, Arman Melkumyan, Richard J. Murphy, and Eric Nettleton. A geological perception system for autonomous mining. *2012 IEEE International Conference on Robotics and Automation*, pages 2986–2991, May 2012.

Robert A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Elsevier Inc., San Diego, third edition, 2007. ISBN 978-0-12-369407-2.

Stephen Se, David Lowe, and Jim Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *The international Journal of robotics Research*, 21(8):735–758, 2002.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Geoffrey M Smith and Edward J Milton. The use of the empirical line method to calibrate remotely sensed data to reflectance. *International Journal of remote sensing*, 20(13):2653—-2662, 1999.

David Stein, Jon Schoonmaker, and Eric Coolbaugh. Hyperspectral Imaging for Intelligence , Surveillance , and Reconnaissance. Technical report, SPACE AND NAVAL WARFARE SYSTEMS CENTER SAN DIEGO CA, 2001.

Weiwei Sun, Avner Halevy, John J Benedetto, Wojciech Czaja, Chun Liu, Hangbin Wu, Beiqi Shi, and Weiyue Li. UL-Isomap based nonlinear dimensionality reduction for hyperspectral imagery classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 89:25–36, 2014. ISSN 0924-2716.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

Ryszard Tadeusiewicz, Rituparna Chaki, and Nabendu Chaki. *Exploring neural networks with C*. CRC Press, 2014.

Chao Tao, Hongbo Pan, Yansheng Li, and Zhengrou Zou. Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geoscience and remote sensing letters*, 12(12):2438—-2442, 2015.

Y Tarabalka, J Chanussot, and J A Benediktsson. Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, 43 (7):2367–2379, 2010. ISSN 0031-3203.

Yuliya Tarabalka, Jón Atli Benediktsson, and Jocelyn Chanussot. Spectral - Spatial Classification of Hyperspectral Imagery Based on Partitional Clustering

Techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8): 2973–2987, 2009.

S Theodoridis and K Koutroumbas. *Recognition, Pattern*. San Diego: CA: Academic Press, 1998.

Marina Trierscheid, Johannes Pellenz, Dietrich Paulus, and Dirk Balthasar. Hyperspectral Imaging for Victim Detection with Rescue Robots. In *IEEE International Workshop on Safety, Security and Rescue Robotics*, number October, pages 7–12, 2008. ISBN 9781424420322.

Alejandro Troccoli and Peter K Allen. Relighting Acquired Models of Outdoor Scenes. In *3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on*, pages 245—-252, 2005.

Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586—-591, 1991.

Ben Upcroft, Colin Mcmanus, Winston Churchill, Will Maddern, and Paul Newman. Lighting Invariant Urban Street Classification. In *IEEE International Conference on Robotics and Automation*, pages 1712–1718, 2014.

Freek D Van der Meer, Harald MA Van der Werff, Frank Ja Van Ruitenbeek, Chris A Hecker, Wim H Bakker, Marleen F Noomen, Mark Van Der Meijde, E John M Carranza, J Boudewijn De Smeth, and Tsehaie Woldai. Multi-and Hyperspectral geologic remote sensing : A review. *International Journal of Applied Earth Observations and Geoinformation*, 14(1):112–128, 2017. ISSN 0303-2434.

CJ Van Rijsbergen. Information retrieval. dept. of computer science, university of glasgow. 14, 1979.

A. Vedaldi and K. Lenc. MatConvNet – Convolutional Neural Networks for MATLAB. In *Proceeding of the {ACM} Int. Conf. on Multimedia*, 2015.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-antoine Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked Denoising Autoencoders : Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

Suge Wang, Deyu Li, Xiaolei Song, Yingjie Wei, and Hongxia Li. A feature selection method based on improved fisher ' s discriminant ratio for text sentiment classification. *Expert Systems With Applications*, 38(7):8696–8702, 2011. ISSN 0957-4174.

Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304—-3308, 2012.

Alexander Wendel and James Underwood. Self-Supervised Weed Detection in Vegetable Crops Using Ground Based Hyperspectral Imaging. In *IEEE International Conference on Robotics and Automation*, pages 5128–5135, 2016. ISBN 9781467380263.

John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust Principal Component Analysis : Exact Recovery of Corrupted Low-Rank Matrices by Convex Optimization. In *Advances in Neural Information Processing Systems 22*, pages 2080–2088, 2009.

Xiaohua Xie, Wei-shi Zheng, and Jianhuang Lai. Normalization of Face Illumination Based on Large- and Small- Scale Features. *IEEE Transactions on Image Processing*, 20(7):1807–1821, 2011.

Chen Xing, Li Ma, and Xiaoquan Yang. Stacked Denoise Autoencoder Based Feature Extraction and Classification for Hyperspectral Images. *Journal of Sensors*, 2015.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, and Aaron Courville. Show, Attend and Tell : Neural Image Caption Generation with Visual Attention. *arXiv preprint arXiv:1502.03044 2.3*, 2015.

Jingxiang Yang, Yongqiang Zhao, Jonathan Cheung, Wai Chan, Chen Yi, and Vrije Universiteit Brussel. Hyperspectral Image Classification using Two Channel Deep Convolutional Neural Network. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 5079–5082, 2016. ISBN 9781509033324.

Jingxiang Yang, Yong-qiang Zhao, and Jonathan Cheung-wai Chan. Learning and Transferring Deep Joint Spectral – Spatial Features for Hyperspectral Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

Jinn-Min Yang, Pao-Ta Yu, and Bor-Chen Kuo. A Nonparametric Feature Extraction and Its Application to Nearest Neighbor Classification for Hyperspectral Image Data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(3):1279–1293, March 2010. ISSN 0196-2892.

Qingxiong Yang, Kar-han Tan, and Narendra Ahuja. Shadow Removal Using Bilateral Filtering. *IEEE Transactions on Image Processing*, 21(10):4361–4368, 2012.

K Y Yeung and W L Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks ? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

Lu Yu, Jun Xie, Songcan Chen, and Lei Zhu. Generating labeled samples for hyperspectral image classification using correlation of spectral bands. *Frontiers of Computer Science*, 10(2):292–301, 2016.

Shiqi Yu, Sen Jia, and Chunyan Xu. Convolutional neural networks for hyperspectral image classification. *Neurocomputing*, 219:88–98, 2017. ISSN 0925-2312.

P W T Yuen and M Richardson. An introduction to hyperspectral imaging and its application for security , surveillance and target acquisition. *The Imaging Science Journal*, 58(5):241—-253, 2010.

RH Yuhas, Alexander F H Goetz, and Joe W Boardman. Descrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm. *Summaries of the Third Annual JPL Airborne Geoscience Workshop, JPL Publ. 92–14*, 1:147–149, 1992.

Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision*, pages 818–833, 2014.

Liangpei Zhang, Lefei Zhang, Dacheng Tao, and Xin Huang. Tensor Discriminative Locality Alignment for Hyperspectral Image Spectral – Spatial Feature Extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1): 242—-256, 2013.

Liangpei Zhang, Lefei Zhang, and Bo Du. Deep Learning for Remote Sensing Data. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22—-40, 2016.

Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow Remover : Image Shadow Removal Based on Illumination Recovering Optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial Landmark Detection by Deep Multi-task Learning. In *European Conference on Computer Vision*, pages 94–108. Springer International Publishing, 2014.

Wenzhi Zhao and Shihong Du. Spectral-Spatial Feature Extraction for Hyperspectral Image Classification : A Dimension Reduction and Deep Learning Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8): 4544–4554, 2016.

Yinqiang Zheng, Imari Sato, and Yoichi Sato. Illumination and Reflectance Spectra Separation of a Hyperspectral Image Meets Low-Rank Matrix Factorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1779–1787, 2015.

Zilong Zhong, Jonathan Li, Lingfei Ma, Han Jiang, and He Zhao. Deep Residual Networks for Hyperspectral Image Classification. In *IEEE 2017 International Geoscience & Remote Sensing Symposium (IGARSS 2017)*, 2017.

Jiejie Zhu, Kegan G G Samuel, Syed Z Masood, and Marshall F Tappen. Learning to Recognize Shadows in Monochromatic Natural Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 223–230, 2010.

Appendix A

Derivation of the CSA-SAE

This section includes a derivation for the content in Section 4.1.1.

For a single observation is, the reconstruction cost that uses the cosine of the spectral angle is:

$$E_{CSA}(f(\mathbf{z}^{(L)}), \mathbf{y}) = 1 - \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|}, \tag{A.1}$$

where:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l-1)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l-1)}, \tag{A.2}$$

$$\mathbf{a}^{(l)} = f(\mathbf{z}^{(l)}), \tag{A.3}$$

$$\mathbf{a}^{(1)} = \mathbf{x} \tag{A.4}$$

for $l = L, L-1, L-2, L-3, ..., 2$, with learnable parameters $\mathbf{W}$ and $\mathbf{b}$.

The reconstruction cost function for all observations, including a regularization term is:

$$E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2, \tag{A.5}$$

and give that:

$$\frac{\partial}{\partial W_{ji}^{(l)}} \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2 = \lambda W_{ji}^{(l)}, \tag{A.6}$$

$$\frac{\partial}{\partial b_j^{(l)}} \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2 = 0, \tag{A.7}$$

the partial derivatives for backpropagation are:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial W_{ji}^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \lambda W_{ji}^{(l)}, \tag{A.8}$$

$$\frac{\partial}{\partial b_j^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial b_j^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}). \tag{A.9}$$

For a single observation m, for $l = 1, 2, ..., L-1$:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(l+1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(l+1)}}{\partial W_{ji}^{(l)}}, \tag{A.10}$$

$$\frac{\partial}{\partial b_j^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(l+1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(l+1)}}{\partial b_j^{(l)}}. \tag{A.11}$$

For $l = L-1$, equations A.10 and A.11 become:

$$\frac{\partial}{\partial W_{ki}^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_k^{(L)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_k^{(L)}}{\partial W_{ki}^{(L-1)}}, \tag{A.12}$$

$$\frac{\partial}{\partial b_k^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_k^{(L)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_k^{(L)}}{\partial b_k^{(L-1)}}, \tag{A.13}$$

where, from A.1:

$$\begin{aligned}
\frac{\partial}{\partial z_k^{(L)}} E_{CSA}(\mathbf{W}, \mathbf{b}) &= \frac{\partial}{\partial z_k^{(L)}} \left[1 - \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|} \right] \\
&= -\frac{1}{|\mathbf{y}|} \frac{\partial}{\partial z_k^{(L)}} \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})|} \\
&= \frac{f'(z_k^{(L)})}{|f(\mathbf{z}^{(L)})||\mathbf{y}|} \left[\frac{(f(\mathbf{z}^{(L)}) \cdot \mathbf{y}) f(z_k^{(L)})}{|f(\mathbf{z}^{(L)})|^2} - y_k \right] \\
&= \delta_k^{(L)},
\end{aligned} \tag{A.14}$$

and:

$$\begin{aligned}\frac{\partial z_k^{(L)}}{\partial W_{ki}^{(L-1)}} &= \frac{\partial}{\partial W_{ki}^{(L-1)}} \sum_{i=1}^{I} W_{ki}^{(L-1)} f(z_i^{(L-1)}) + b_k^{(L-1)} \\ &= f(z_i^{(L-1)}). \end{aligned} \tag{A.15}$$

$$\begin{aligned}\frac{\partial z_k^{(L)}}{\partial b_k^{(L-1)}} &= \frac{\partial}{\partial b_k^{(L-1)}} \sum_{i=1}^{I} W_{ki}^{(L-1)} f(z_i^{(L-1)}) + b_k^{(L-1)} \\ &= 1. \end{aligned} \tag{A.16}$$

Substitute A.14 and A.15 into A.12:

$$\frac{\partial}{\partial W_{ki}^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_k^{(L)} f(z_i^{(L-1)}), \tag{A.17}$$

and substitute A.14 and A.16 into A.13:

$$\frac{\partial}{\partial b_k^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_k^{(L)}. \tag{A.18}$$

For $l = L - 2$, equations A.10 and A.11 become:

$$\frac{\partial}{\partial W_{ji}^{(L-2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(L-1)}}{\partial W_{ji}^{(L-2)}}, \tag{A.19}$$

$$\frac{\partial}{\partial b_j^{(L-2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(L-1)}}{\partial b_j^{(L-2)}}, \tag{A.20}$$

where:

$$\frac{\partial}{\partial z_j^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_k^{(L)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_k^{(L)}}{\partial z_j^{(L-1)}}. \tag{A.21}$$

Substituting A.14 into A.21:

$$\frac{\partial}{\partial z_j^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_k^{(L)} \frac{\partial z_k^{(L)}}{\partial z_j^{(L-1)}}, \tag{A.22}$$

where:

$$\begin{aligned} \frac{\partial z_k^{(L)}}{\partial z_j^{(L-1)}} &= \frac{\partial}{\partial z_j^{(L-1)}} \sum_{j=1}^{J} W_{kj}^{(L-1)} f(z_j^{(L-1)}) + b_k^{(L-1)} \\ &= W_{kj}^{(L-1)} \frac{\partial}{\partial z_j^{(L-1)}} f(z_j^{(L-1)}) \\ &= W_{kj}^{(L-1)} f'(z_j^{(L-1)}). \end{aligned} \tag{A.23}$$

Substituting into A.23 gives A.22:

$$\begin{aligned} \frac{\partial}{\partial z_j^{(L-1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) &= \sum_{k=1}^{K} \delta_k^{(L)} W_{kj}^{(L-1)} f'(z_j^{(L-1)}) \\ &= \delta_j^{(L-1)}. \end{aligned} \tag{A.24}$$

Given:

$$\begin{aligned} \frac{\partial z_j^{(L-1)}}{\partial W_{ji}^{(L-2)}} &= \frac{\partial}{\partial W_{ji}^{(L-2)}} \sum_{i=1}^{I} W_{ji}^{(L-2)} f(z_i^{(L-2)}) + b_j^{(L-2)} \\ &= f(z_i^{(L-2)}), \end{aligned} \tag{A.25}$$

and:

$$\frac{\partial z_j^{(L-1)}}{\partial b_j^{(L-2)}} = \frac{\partial}{\partial b_j^{(L-2)}} \sum_{i=1}^{I} W_{ji}^{(L-2)} f(z_i^{(L-2)}) + b_j^{(L-2)}$$

$$= 1, \tag{A.26}$$

substituting A.24 and A.25 into A.19:

$$\frac{\partial}{\partial W_{ji}^{(L-2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(L-1)} f(z_i^{(L-2)}), \tag{A.27}$$

and substituting A.24 and A.26 into A.20:

$$\frac{\partial}{\partial b_j^{(L-2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(L-1)}. \tag{A.28}$$

For $l = L-3, L-4, ..., 2$:

$$\frac{\partial}{\partial z_i^{(l+1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(l+2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(l+2)}}{\partial z_i^{(l+1)}}, \tag{A.29}$$

where:

$$\frac{\partial}{\partial z_j^{(l+2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(l+2)}, \tag{A.30}$$

and:

$$\begin{aligned} \frac{\partial z_j^{(l+2)}}{\partial z_i^{(l+1)}} &= \frac{\partial}{\partial z_i^{(l+1)}} \sum_{i=1}^{I} W_{ji}^{(l+1)} f(z_i^{(l+1)}) + b_j^{(l+1)} \\ &= W_{ji}^{(l+1)} \frac{\partial}{\partial z_i^{(l+1)}} f(z_i^{(l+1)}) \\ &= W_{ji}^{(l+1)} f'(z_i^{(l+1)}). \end{aligned} \tag{A.31}$$

Substituting A.30 and A.31 into A.29:

$$\frac{\partial}{\partial z_i^{(l+1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \sum_{j=1}^{J} \delta_j^{(l+2)} W_{ji}^{(l+1)} f'(z_i^{(l+1)}) \\ = \delta_i^{(l+1)}. \tag{A.32}$$

Given:

$$\frac{\partial z_j^{(l+1)}}{\partial W_{ji}^{(l)}} = \frac{\partial}{\partial W_{ji}^{(l)}} \sum_{i=1}^{I} W_{ji}^{(l)} f(z_i^{(l)}) + b_j^{(l)} \\ = f(z_i^{(l)}), \tag{A.33}$$

and:

$$\frac{\partial z_j^{(l+1)}}{\partial b_j^{(l)}} = \frac{\partial}{\partial b_j^{(l)}} \sum_{i=1}^{I} W_{ji}^{(l)} f(z_i^{(l)}) + b_j^{(l)} \\ = 1, \tag{A.34}$$

substituting A.33 and A.32 into A.10:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(l+1)} f(z_i^{(l)}), \tag{A.35}$$

and substituting A.34 and A.32 into A.11:

$$\frac{\partial}{\partial b_j^{(l)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(l+1)}. \tag{A.36}$$

For $l = 1$, equations A.10 and A.11 become:

$$\frac{\partial}{\partial W_{ji}^{(1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(2)}}{\partial W_{ji}^{(1)}}, \tag{A.37}$$

$$\frac{\partial}{\partial b_j^{(1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(2)}}{\partial b_j^{(1)}}, \tag{A.38}$$

where:

$$\frac{\partial}{\partial z_i^{(2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(3)}} E_{CSA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(3)}}{\partial z_i^{(2)}}. \tag{A.39}$$

Given:

$$\frac{\partial}{\partial z_j^{(3)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(3)}, \tag{A.40}$$

and:

$$\begin{aligned} \frac{\partial z_j^{(3)}}{\partial z_i^{(2)}} &= \frac{\partial}{\partial z_i^{(2)}} \sum_{i=1}^{I} W_{ji}^{(2)} f(z_i^{(2)}) + b_j^{(2)} \\ &= W_{ji}^{(2)} \frac{\partial}{\partial z_i^{(2)}} f(z_i^{(2)}) \\ &= W_{ji}^{(2)} f'(z_i^{(2)}), \end{aligned} \tag{A.41}$$

substituting A.40 and A.41 into A.39:

$$\begin{aligned} \frac{\partial}{\partial z_i^{(2)}} E_{CSA}(\mathbf{W}, \mathbf{b}) &= \sum_{j=1}^{J} \delta_j^{(3)} W_{ji}^{(2)} f'(z_i^{(2)}) \\ &= \delta_i^{(2)}. \end{aligned} \tag{A.42}$$

Given:

$$\frac{\partial z_j^{(2)}}{\partial W_{ji}^{(1)}} = \frac{\partial}{\partial W_{ji}^{(1)}} \sum_{i=1}^{I} W_{ji}^{(1)} x_i + b_j^{(1)}$$
$$= x_i, \tag{A.43}$$

and:

$$\frac{\partial z_j^{(2)}}{\partial b_j^{(1)}} = \frac{\partial}{\partial b_j^{(1)}} \sum_{i=1}^{I} W_{ji}^{(1)} f(z_i^{(1)}) + b_j^{(1)}$$
$$= 1, \tag{A.44}$$

substituting A.42 and A.43 into A.37:

$$\frac{\partial}{\partial W_{ji}^{(1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(2)} x_i, \tag{A.45}$$

and substituting A.42 and A.44 into A.38:

$$\frac{\partial}{\partial b_j^{(1)}} E_{CSA}(\mathbf{W}, \mathbf{b}) = \delta_j^{(2)}. \tag{A.46}$$

The parameter update equations for gradient descent optimisation are:

$$W_{ji}^{(l)} := W_{ji}^{(l)} - \alpha \frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{A.47}$$

$$b_j^{(l)} := b_j^{(l)} - \alpha \frac{\partial}{\partial b_j^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{A.48}$$

These derivatives were validated numerically using the technique described in Ng (2011).

A.1 Derivative of the Sigmoid Activation Function

The above derivation for the parameter update equations is independent of the activation function $f(\cdot)$. Thus, different activation functions can be used as long as $\frac{\partial}{\partial z} f(z)$, referred to as $f'(z)$ in this thesis, can be computed.

The implementation of the CSA-SAE in this thesis used the sigmoid activation function, given by:

$$f(z) = \frac{1}{1 + \exp -z}. \tag{A.49}$$

The derivative of this function is given by:

$$\begin{aligned}
\frac{\partial}{\partial z} f(z) &= \frac{\partial}{\partial z} \frac{1}{1 + \exp -z} \\
&= -\frac{1}{(1 + \exp -z)^2} \frac{\partial}{\partial z} (1 + \exp -z) \\
&= \frac{\exp -z}{(1 + \exp -z)^2} \\
&= \frac{1 + \exp -z - 1}{(1 + \exp -z)^2} \\
&= \frac{1 + \exp -z}{(1 + \exp -z)^2} - \frac{1}{(1 + \exp -z)^2} \\
&= \frac{1}{1 + \exp -z} - \left(\frac{1}{1 + \exp -z} \right)^2
\end{aligned} \tag{A.50}$$

Substitute A.49 into A.50:

$$\begin{aligned}
\frac{\partial}{\partial z} f(z) &= f(z) - f(z)^2 \\
&= f(z)(1 - f(z))
\end{aligned} \tag{A.51}$$

Appendix B

Derivation of the SA-SAE

This section includes a derivation for the content in Section 4.1.2.

For a single observation is, the reconstruction cost that uses the spectral angle is:

$$E_{SA}(f(\mathbf{z}^{(L)}), \mathbf{y}) = \cos^{-1} \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|}, \tag{B.1}$$

where:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l-1)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l-1)}, \tag{B.2}$$

$$\mathbf{a}^{(l)} = f(\mathbf{z}^{(l)}), \tag{B.3}$$

$$\mathbf{a}^{(1)} = \mathbf{x} \tag{B.4}$$

for $l = L, L-1, L-2, L-3, ..., 2$, with learnable parameters $\mathbf{W}$ and $\mathbf{b}$.

The reconstruction cost function for all observations, including a regularization term is:

$$E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} E_{SA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2, \tag{B.5}$$

and give that:

$$\frac{\partial}{\partial W_{ji}^{(l)}} \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2 = \lambda W_{ji}^{(l)}, \tag{B.6}$$

$$\frac{\partial}{\partial b_j^{(l)}} \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2 = 0, \tag{B.7}$$

the partial derivatives for backpropagation are:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial W_{ji}^{(l)}} E_{SA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \lambda W_{ji}^{(l)}, \tag{B.8}$$

$$\frac{\partial}{\partial b_j^{(l)}} E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial b_j^{(l)}} E_{SA}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}). \tag{B.9}$$

For a single observation m, for $l = 1, 2, ..., L-1$:

$$\frac{\partial}{\partial W_{ji}^{(l)}} E_{SA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(l+1)}} E_{SA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(l+1)}}{\partial W_{ji}^{(l)}}, \tag{B.10}$$

$$\frac{\partial}{\partial b_j^{(l)}} E_{SA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_j^{(l+1)}} E_{SA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_j^{(l+1)}}{\partial b_j^{(l)}}. \tag{B.11}$$

For $l = L-1$, equations B.10 and B.11 become:

$$\frac{\partial}{\partial W_{ki}^{(L-1)}} E_{SA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_k^{(L)}} E_{SA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_k^{(L)}}{\partial W_{ki}^{(L-1)}}, \tag{B.12}$$

$$\frac{\partial}{\partial b_k^{(L-1)}} E_{SA}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial z_k^{(L)}} E_{SA}(\mathbf{W}, \mathbf{b}) \frac{\partial z_k^{(L)}}{\partial b_k^{(L-1)}}, \tag{B.13}$$

where, from B.1:

$$\begin{aligned}
\frac{\partial}{\partial z_k^{(L)}} E_{SA}(\mathbf{W}, \mathbf{b}) &= \frac{\partial}{\partial z_k^{(L)}} \left[\cos^{-1} \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|} \right] \\
&= -\frac{1}{\sqrt{1 - \left[\frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|} \right]^2}} \frac{1}{|\mathbf{y}|} \frac{\partial}{\partial z_k^{(L)}} \frac{\sum_{k=1}^{K} f(z_k^{(L)}) y_k}{|f(\mathbf{z}^{(L)})|}
\end{aligned}$$

$$= \frac{f'(z_k^{(L)})}{|f(\mathbf{z}^{(L)})||\mathbf{y}|\sqrt{1 - \left[\frac{\sum_{k=1}^{K} f(z_k^{(L)})y_k}{|f(\mathbf{z}^{(L)})||\mathbf{y}|}\right]^2}} \left[\frac{(f(\mathbf{z}^{(L)}) \cdot \mathbf{y})f(z_k^{(L)})}{|f(\mathbf{z}^{(L)})|^2} - y_k\right]$$
$$= \delta_k^{(L)}, \tag{B.14}$$

and:

$$\frac{\partial z_k^{(L)}}{\partial W_{ki}^{(L-1)}} = \frac{\partial}{\partial W_{ki}^{(L-1)}} \sum_{i=1}^{I} W_{ki}^{(L-1)} f(z_i^{(L-1)}) + b_k^{(L-1)}$$
$$= f(z_i^{(L-1)}). \tag{B.15}$$

$$\frac{\partial z_k^{(L)}}{\partial b_k^{(L-1)}} = \frac{\partial}{\partial b_k^{(L-1)}} \sum_{i=1}^{I} W_{ki}^{(L-1)} f(z_i^{(L-1)}) + b_k^{(L-1)}$$
$$= 1. \tag{B.16}$$

Substitute B.14 and B.15 into B.12:

$$\frac{\partial}{\partial W_{ki}^{(L-1)}} E_{SA}(\mathbf{W}, \mathbf{b}) = \delta_k^{(L)} f(z_i^{(L-1)}), \tag{B.17}$$

and substitute B.14 and B.16 into B.13:

$$\frac{\partial}{\partial b_k^{(L-1)}} E_{SA}(\mathbf{W}, \mathbf{b}) = \delta_k^{(L)}. \tag{B.18}$$

The partial derivatives with respect to the parameters in layers $l = L-2, ..., 2, 1$ are the same as in Appendix A. The parameter update equations for gradient descent optimisation are:

$$W_{ji}^{(l)} := W_{ji}^{(l)} - \alpha \frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{B.19}$$

$$b_j^{(l)} := b_j^{(l)} - \alpha \frac{\partial}{\partial b_j^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{B.20}$$

These derivatives were validated numerically using the technique described in Ng (2011).

Appendix C

Derivation of the SID-SAE

This section includes a derivation for the content in Section 4.1.3.

For a single observation is, the reconstruction cost that uses the spectral information divergence is:

$$E_{SID}(f(\mathbf{z}^{(L)}), \mathbf{y}) = \sum_{k=1}^{K} \left[\frac{f(z_k^{(L)})}{\sum_{d=1}^{K} f(z_d^{(L)})} - \frac{y_k}{\sum_{d=1}^{K} y_d} \right] [log f(z_k^{(L)}) - log \sum_{d=1}^{K} f(z_d^{(L)}) - log(y_k) + log \sum_{d=1}^{K} y_d]. \quad \text{(C.1)}$$

where:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l-1)} \mathbf{a}^{(l-1)} + \mathbf{b}^{(l-1)}, \quad \text{(C.2)}$$

$$\mathbf{a}^{(l)} = f(\mathbf{z}^{(l)}), \quad \text{(C.3)}$$

$$\mathbf{a}^{(1)} = \mathbf{x} \quad \text{(C.4)}$$

for $l = L, L-1, L-2, L-3, ..., 2$, with learnable parameters $\mathbf{W}$ and $\mathbf{b}$.

The reconstruction cost function for all observations, including a regularization term is:

$$E(\mathbf{W}, \mathbf{b}) = \frac{1}{M} \sum_{m=1}^{M} E_{SID}(\mathbf{W}, \mathbf{b}; \mathbf{y}^{(m)}) + \frac{\lambda}{2} \sum_{i,j,l=1}^{I,J,L-1} (W_{ji}^{(l)})^2, \quad \text{(C.5)}$$

and give that:

$$\frac{\partial}{\partial W_{ji}^{(l)}}\frac{\lambda}{2}\sum_{i,j,l=1}^{I,J,L-1}(W_{ji}^{(l)})^2 = \lambda W_{ji}^{(l)}, \tag{C.6}$$

$$\frac{\partial}{\partial b_{j}^{(l)}}\frac{\lambda}{2}\sum_{i,j,l=1}^{I,J,L-1}(W_{ji}^{(l)})^2 = 0, \tag{C.7}$$

the partial derivatives for backpropagation are:

$$\frac{\partial}{\partial W_{ji}^{(l)}}E(\mathbf{W},\mathbf{b}) = \frac{1}{M}\sum_{m=1}^{M}\frac{\partial}{\partial W_{ji}^{(l)}}E_{SID}(\mathbf{W},\mathbf{b};\mathbf{y}^{(m)}) + \lambda W_{ji}^{(l)}, \tag{C.8}$$

$$\frac{\partial}{\partial b_{j}^{(l)}}E(\mathbf{W},\mathbf{b}) = \frac{1}{M}\sum_{m=1}^{M}\frac{\partial}{\partial b_{j}^{(l)}}E_{SID}(\mathbf{W},\mathbf{b};\mathbf{y}^{(m)}). \tag{C.9}$$

For a single observation m, for $l = 1, 2, ..., L-1$:

$$\frac{\partial}{\partial W_{ji}^{(l)}}E_{SID}(\mathbf{W},\mathbf{b}) = \frac{\partial}{\partial z_{j}^{(l+1)}}E_{SID}(\mathbf{W},\mathbf{b})\frac{\partial z_{j}^{(l+1)}}{\partial W_{ji}^{(l)}}, \tag{C.10}$$

$$\frac{\partial}{\partial b_{j}^{(l)}}E_{SID}(\mathbf{W},\mathbf{b}) = \frac{\partial}{\partial z_{j}^{(l+1)}}E_{SID}(\mathbf{W},\mathbf{b})\frac{\partial z_{j}^{(l+1)}}{\partial b_{j}^{(l)}}. \tag{C.11}$$

For $l = L-1$, equations C.10 and C.11 become:

$$\frac{\partial}{\partial W_{ki}^{(L-1)}}E_{SID}(\mathbf{W},\mathbf{b}) = \frac{\partial}{\partial z_{k}^{(L)}}E_{SID}(\mathbf{W},\mathbf{b})\frac{\partial z_{k}^{(L)}}{\partial W_{ki}^{(L-1)}}, \tag{C.12}$$

$$\frac{\partial}{\partial b_{k}^{(L-1)}}E_{SID}(\mathbf{W},\mathbf{b}) = \frac{\partial}{\partial z_{k}^{(L)}}E_{SID}(\mathbf{W},\mathbf{b})\frac{\partial z_{k}^{(L)}}{\partial b_{k}^{(L-1)}}, \tag{C.13}$$

where, from C.1:

$$\frac{\partial}{\partial z_{k}^{(L)}}E_{SID}(\mathbf{W},\mathbf{b}) = \frac{\partial}{\partial z_{k}^{(L)}}\Bigg\{\sum_{k=1}^{K}\left[\frac{f(z_{k}^{(L)})}{\sum_{d=1}^{K}f(z_{d}^{(L)})} - \frac{y_k}{\sum_{d=1}^{K}y_d}\right]\Bigg[log f(z_{k}^{(L)}) - log\sum_{d=1}^{K}f(z_{d}^{(L)}) - log(y_k) + log\sum_{d=1}^{K}y_d\Bigg]\Bigg\} \tag{C.14}$$

$$= -\frac{f'(z_k^{(L)})}{\sum_{d=1}^{K} f(z_d^{(L)})}\left[\frac{q_k}{p_k} - log\frac{p_k}{q_k} - 1 + \sum_{d=1}^{K}(p_d - q_d + p_d log\frac{p_d}{q_d})\right]$$
$$= \delta_k^{(L)}, \tag{C.15}$$

where

$$\mathbf{p} = \frac{f(\mathbf{z}^{(L)})}{\sum_{c=1}^{K} f(z_c^{(L)})}, \tag{C.16}$$

$$\mathbf{q} = \frac{\mathbf{y}}{\sum_{c=1}^{K} y_c}, \tag{C.17}$$

and:

$$\frac{\partial z_k^{(L)}}{\partial W_{ki}^{(L-1)}} = \frac{\partial}{\partial W_{ki}^{(L-1)}} \sum_{i=1}^{I} W_{ki}^{(L-1)} f(z_i^{(L-1)}) + b_k^{(L-1)}$$
$$= f(z_i^{(L-1)}). \tag{C.18}$$

$$\frac{\partial z_k^{(L)}}{\partial b_k^{(L-1)}} = \frac{\partial}{\partial b_k^{(L-1)}} \sum_{i=1}^{I} W_{ki}^{(L-1)} f(z_i^{(L-1)}) + b_k^{(L-1)}$$
$$= 1. \tag{C.19}$$

Substitute C.15 and C.18 into C.12:

$$\frac{\partial}{\partial W_{ki}^{(L-1)}} E_{SID}(\mathbf{W}, \mathbf{b}) = \delta_k^{(L)} f(z_i^{(L-1)}), \tag{C.20}$$

and substitute C.15 and C.19 into C.13:

$$\frac{\partial}{\partial b_k^{(L-1)}} E_{SID}(\mathbf{W}, \mathbf{b}) = \delta_k^{(L)}. \tag{C.21}$$

The partial derivatives with respect to the parameters in layers $l = L-2, ..., 2, 1$ are the same as in Appendix A. The parameter update equations for gradient descent optimisation are:

$$W_{ji}^{(l)} := W_{ji}^{(l)} - \alpha \frac{\partial}{\partial W_{ji}^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{C.22}$$

$$b_{j}^{(l)} := b_{j}^{(l)} - \alpha \frac{\partial}{\partial b_{j}^{(l)}} E(\mathbf{W}, \mathbf{b}), \tag{C.23}$$

These derivatives were validated numerically using the technique described in Ng (2011).

Appendix D

Derivation of the relighting equations

This appendix includes derivations for the relighting equations used throughout this thesis.

D.1 Relighting with respect to diffuse skylight

This section derives the scaling factor for relighting the radiance $\mathbf{L}$ of a region i in sunlight with respect to diffuse skylight (occluded from the sun). This relighting equation is used in Section 4.2. The derivation comes from Ramakrishnan (2016).

From the model (2.2), a region j, which is the same as the region i, but is only illuminated by diffuse skylight, is given by:

$$\begin{aligned} L_j(\lambda) &= \frac{\rho(\lambda)}{\pi}\Big[E_{sky}(\lambda)\Big] \\ &= \frac{\rho(\lambda)}{\pi}\Big[E_{sky}(\lambda)\Big]\left[\frac{E_{sun}(\lambda)\tau(\lambda)cos\theta_i + \Gamma_i E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \end{aligned}$$

$$= \frac{\rho(\lambda)}{\pi}\Big[E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)\Big]\left[\frac{E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right]. \tag{D.1}$$

From the model (2.2), the same region in sunlight is given by:

$$L_i(\lambda) = \frac{\rho(\lambda)}{\pi}\Big[E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)\Big]. \tag{D.2}$$

Substituting D.2 into D.1:

$$\begin{aligned} L_j(\lambda) &= L_i(\lambda)\left[\frac{E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \\ &= L_i(\lambda)\frac{1}{\dfrac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_i + \Gamma_i} \end{aligned} \tag{D.3}$$

D.2 Relighting with respect to full terrestrial sunlight and diffuse skylight

This section derives the scaling factor for relighting the radiance $\mathbf{L}$ of a region i in sunlight with respect to full terrestrial sunlight and diffuse skylight exposure. The derivation comes from Ramakrishnan (2016).

From the model (2.2), a region j, which is the same as the region i, but illuminated by full terrestrial sunlight and diffuse skylight exposure ($\cos\theta = 1$, $\Gamma = 1$), is given by:

$$\begin{aligned} L_j(\lambda) &= \frac{\rho(\lambda)}{\pi}\Big[E_{sun}(\lambda)\tau(\lambda) + E_{sky}(\lambda)\Big] \\ &= \frac{\rho(\lambda)}{\pi}\Big[E_{sun}(\lambda)\tau(\lambda) + E_{sky}(\lambda)\Big]\left[\frac{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \end{aligned}$$

$$= \frac{\rho(\lambda)}{\pi}\Big[E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)\Big]\left[\frac{E_{sun}(\lambda)\tau(\lambda) + E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \tag{D.4}$$

Substituting D.2 into D.4:

$$\begin{aligned} L_j(\lambda) &= L_i(\lambda)\left[\frac{E_{sun}(\lambda)\tau(\lambda) + E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \\ &= L_i(\lambda)\frac{\dfrac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)} + 1}{\dfrac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_i + \Gamma_i} \end{aligned} \tag{D.5}$$

D.3 Relighting with respect to a generalised illuminant

This section derives the scaling factor for relighting the radiance $\mathbf{L}$ of a region i in sunlight with respect to either diffuse skylight only (occluded from the sun) or a combination of terrestrial sunlight and diffuse skylight with arbitrary exposure. This relighting equation is used in Section 5.3.

From the model (2.2), a region j, which is the same as the region i, but illuminated by either diffuse skylight only ($V_i = 0$) or a combination of terrestrial sunlight and diffuse skylight ($V_i = 1$), is given by:

$$\begin{aligned} L_j(\lambda) &= \frac{\rho(\lambda)}{\pi}\Big[V_i E_{sun}(\lambda)\tau(\lambda)\cos\theta_j + \Gamma_j E_{sky}(\lambda)\Big] \\ &= \frac{\rho(\lambda)}{\pi}\Big[V_i E_{sun}(\lambda)\tau(\lambda)\cos\theta_j + \Gamma_j E_{sky}(\lambda)\Big]\left[\frac{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \\ &= \frac{\rho(\lambda)}{\pi}\Big[E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)\Big]\left[\frac{V_i E_{sun}(\lambda)\tau(\lambda)\cos\theta_j + \Gamma_j E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \end{aligned} \tag{D.6}$$

Substituting D.2 into D.6:

$$\begin{aligned} L_j(\lambda) &= L_i(\lambda)\left[\frac{V_i E_{sun}(\lambda)\tau(\lambda)\cos\theta_j + \Gamma_j E_{sky}(\lambda)}{E_{sun}(\lambda)\tau(\lambda)\cos\theta_i + \Gamma_i E_{sky}(\lambda)}\right] \\ &= L_i(\lambda)\frac{V_i\dfrac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_j + \Gamma_j}{\dfrac{E_{sun}(\lambda)\tau(\lambda)}{E_{sky}(\lambda)}\cos\theta_i + \Gamma_i} \end{aligned} \tag{D.7}$$